Word Meaning Representation and Interpretation in Vector Space

Marianna Apidianaki

MARIANNA APIDIANAKI

HABILITATION À DIRIGER DES RECHERCHES EN INFORMATIQUE

Word Meaning Representation and Interpretation in Vector Space

Jury

Marco Baroni, ICREA, Universitat Pompeu Fabra, Rapporteur
Marie-Catherine de Marneffe, FNRS, UCLouvain, The Ohio State University, Rapporteure
Emmanuel Morin, Nantes University, Rapporteur
Philippe Blache, CNRS, Aix-Marseille Université, Examinateur
Katrin Erk, University of Texas at Austin, Examinatrice
Aline Villavicencio, University of Sheffield, Examinatrice

Nicholas Asher, CNRS, Université Toulouse III - Paul Sabatier, Tuteur

November, 9 2022

Abstract

The analysis and representation of lexical meaning is a central topic in computational linguistics research, with both theoretical and application-oriented interest. It allows to study phenomena related to language acquisition and semantic modelling, and to improve the inference and common-sense reasoning capabilities of computational systems. The work presented in this thesis covers a period of time where the field of computational linguistics has been marked by an important paradigm shift due to the wide adoption of neural language models, which have progressively replaced older methods for semantic modelling.

This paradigm shift has greatly impacted my research and constitutes the central axis of this thesis. The content is organised in a way that highlights important aspects of this transition, and which puts forward the methodological changes that have been imposed. The introduction of deep learning models in the field of computational semantic analysis has brought about new research questions, but older questions also remain relevant and are now being studied from a different angle. Moreover, the recent area of interpretability, which attempts to explain the success of neural models, brings a novel perspective to open questions related to semantics.

After explaining the transition to neural model representations in an introductory chapter, I then propose a systematic comparison of older and more recent studies that I conducted individually, and in collaboration with students and colleagues. These have been selected with the aim to put side by side the older and novel methodology that has been used for addressing semantics-related questions, including lexical polysemy and clusterability, in-context paraphrasing, and the modelling of abstract semantic notions such as scalar adjective intensity. The last chapter of this thesis includes a discussion of the limitations of current methodology, and a presentation of future perspectives aimed at addressing these shortcomings. These include the grounding of semantic knowledge into different modalities, the development of adversarial methods for improving the robustness and generalisation capability of neural models, and the development of interpretation methods for capturing causality and explaining model behaviour.

Résumé

L'analyse et la représentation des sens lexicaux sont des sujets de recherche centraux en linguistique computationnelle, qui présentent de l'intérêt tant théorique que pratique. Elles permettent d'étudier des phénomènes liés à l'acquisition des langues et à la modélisation sémantique, et d'améliorer les capacités d'inférence et de raisonnement des systèmes computationnels. Le travail présenté dans cette thèse couvre une période de temps où le domaine de la sémantique computationnelle a été marqué par un changement de paradigme important dû à l'adoption des modèles de langue neuronaux, qui ont progressivement remplacé les méthodes traditionnelles de modélisation sémantique.

Ce changement de paradigme a eu un grand impact sur ma recherche et constitue l'axe central de cette thèse, dont le contenu est organisé de manière à mettre en valeur les aspects importants de cette transition, et les changements méthodologiques qui ont été imposés. L'introduction de modèles d'apprentissage profond dans le champ de l'analyse sémantique computationnelle a suscité de nouvelles questions de recherche, néanmoins celles étudiées auparavant restent pertinentes et sont actuellement explorées sous un angle différent. En outre, le nouveau domaine de l'interprétabilité qui vise à expliquer les succès des modèles neuronaux, amène une nouvelle perspective aux questions liées à la sémantique.

Suite à un chapitre d'introduction qui explique la transition vers les représentations neuronales, je propose une comparaison systématique d'études que j'ai menées pendant cette période à titre individuel, et en collaboration avec mes étudiant(e)s et collègues. Ces travaux ont été sélectionnés avec l'objectif de juxtaposer la méthodologie utilisée auparavant et la méthodologie actuelle pour l'analyse de la sémantique, comme la polysémie lexicale et la clusterabilité des sens, la substitution lexicale en contexte, et la modélisation de notions sémantiques abstraites comme l'intensité des adjectifs scalaires. Le dernier chapitre de cette thèse comporte une discussion des limites de la méthodologie actuelle, et une présentation de perspectives de recherche futures destinées à pallier ces faiblesses. Celles-ci incluent l'ancrage des connaissances sémantiques dans des modalités variées, le développement d'attaques contradictoires dans le but d'améliorer la robustesse et la capacité de généralisation des modèles neuronaux, et le développement de méthodes d'interprétation visant à capter la notion de causalité et à expliquer le comportement des modèles.

Acknowledgments

I would like to thank my students and collaborators for their engagement, and for their insightful feedback and contribution to our joint projects. Our collaborations have been, and still are, a valuable source of inspiration for my research.

I also wish to thank the members of this committee for finding the time in their busy schedules to provide feedback on my work. I am grateful to Marco Baroni, Marie-Catherine de Marneffe and Emmanuel Morin for their thorough reports; to Philippe Blache, Katrin Erk and Aline Villavicencio for accepting to act as examiners for my HDR; and to my tutor, Nicholas Asher.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Computational Approaches to Meaning Representation

Meaning representation and inference are core mechanisms that allow us humans to reason about our experiences and the world. Children acquire these skills early in life through exposure to language in a situated perceptual context. They know, for example, that *cats* are animals, and that *a fun game* is worth playing. Devising computational systems endowed with human-like language understanding capabilities has been a long-term goal in linguistics, cognitive science and artificial intelligence. From a theoretical perspective, computational models of meaning are of enormous interest: Leveraging knowledge from large amounts of data, they provide insights into how humans themselves perceive meaning. Meaning representation is also at the center of applied research, since any computational system that needs to be "smart" and understand language in order to perform specific tasks (e.g., question answering, translation or summarization) should possess semantic reasoning and inference skills.

My research addresses the meaning representation and reasoning capabilities of computational systems, with specific focus on lexical meaning. This domain has seen a big paradigm shift in recent years. We have witnessed an important change from traditional distributional models which encoded meaning by keeping track of words' co-occurrences in texts, to neural language models which encode this type of knowledge using self-supervision, through exposure to huge volumes of text. These deep artificial neural networks (ANNs) are currently the dominant paradigm in artificial intelligence and natural language processing. Their architecture is inspired by the neural structure of the brain with artificial neurons (processing units) organised in multiple layers, allowing to form representations at different levels of abstraction (LeCun et al., 2015). These models learn to identify and encode rich knowledge through complex calculations during exposure to raw text data during training, which they can subsequently refine when fine-tuned for specific tasks. The traditional Natural Language Processing (NLP) pipeline which included grammatical, morphological, syntactic and semantic processing tasks (such as part of speech tagging, parsing and word sense disambiguation) has become obsolete, since neural models can learn to perform the same tasks in a black-box manner, without explicit (pre-)processing and feature engineering steps.

My research has followed this path from dedicated lexical semantic analysis modules for specific NLP applications, to interpretation studies aimed at deciphering the semantic knowledge encoded in neural networks and enhancing its quality. My PhD and early research work focused on word sense induction and disambiguation (WSI and WSD) for cross-lingual applications (Apidianaki, 2006, 2007, 2008b, 2009b,a, 2011, 2012; Apidianaki and Gong, 2015), specifically Machine Translation (MT) (Apidianaki

et al., 2012b) and the integration of semantic knowledge in MT evaluation metrics (Apidianaki and He, 2010; Marie and Apidianaki, 2015), but also the use of cross-lingual WSD techniques for enhancing tasks such as Bilingual Lexicon Induction (BLI) (Apidianaki et al., 2012a, 2013), lexicon development (Apidianaki and Sagot, 2012), and Semantic Role Labelling (van der Plas and Apidianaki, 2014; van der Plas et al., 2014). Cross-lingual applications (for example, Machine Translation) have been the targeted downstream task in these studies, but I have also used translations as a proxy for meaning in my work on lexical semantic analysis, often combined with distributional information for word sense identification and cross-lingual clustering (Apidianaki, 2009a, 2011; Apidianaki et al., 2014). Alignments and translation annotations have served to identify senses at the level of word types (Apidianaki and Gong, 2015; Marie and Apidianaki, 2015), but also for estimating the meaning and similarity of contextualised word instances (McCarthy et al., 2016; Garí Soler et al., 2019a).

The shift to neural language models in the field of NLP couldn't leave the field of lexical semantics untouched. Low-dimensional word embedding representations generated by neural models have been shown to represent lexical semantic similarity better than traditional count-based distributional approaches, and to offer greater generalization potential (Baroni et al., 2014b). The questions posed until their introduction in the lexical semantics field could thus be revisited under this new representation paradigm. My research has followed this evolution. Notably, this change has occurred in the beginning of my research project MULTISEM "Advanced Models for Multilingual Semantic Processing" which was granted funding by the ANR Young Researchers (JCJC) program in 2016.[1] Consequently, the work carried out during the five-year duration of the project has been mainly focused on exploring and leveraging the knowledge encoded in neural language models for semantic analysis.

In spite of this paradigm and methodological shift, a common thread underlying these works is the data-driven approach to meaning, where knowledge about words is inferred from their distribution in texts without access to external semantic resources. This synthetic document first explains the changes that have occurred in the domain due to this paradigm shift and the current state of affairs in terms of meaning representation. Additionally, I will explain and analyse the divergences witnessed in the field in terms of the unit of representation addressed in lexical semantics. I will explain how meaning is represented at the level of word types and that of word tokens or individual instances, and the advantages of the two types of representation. I will then present a selection of my articles which address lexical semantic questions using distributional models possibly enriched with other knowledge sources (for example, annotations or external lexicons), and others that demonstrate how these same questions can be addressed using neural language model representations. The presented works have been carried out in collaboration with my colleagues Diana McCarthy (University of Cambridge), Katrin Erk (University of Texas in Austin) and Chris Callison-Burch (University of Pennsylvania); Anne Cocos and other PhD and Master's students at the University of Pennsylvania; and my PhD student Aina Garí Soler at the University Paris-Saclay. Aina did her thesis in the frame of the ANR JCJC project MULTISEM for which I have been the Principal Investigator from 2016 to 2021. The results and findings of these studies highlight the potential of neural representation methods for representing meaning, demonstrate the richness of the encoded information, and open up multiple promising perspectives for future research.

[1] Information about MULTISEM can be found on the project website: https://sites.google.com/view/multisem/.

1.2 Lexical Semantic Analysis Across Paradigms

The paradigm shift that has occurred with the wide adoption of neural language models in the field of NLP and computational linguistics has raised novel research questions. It has also allowed to approach old research topics from a different perspective than previous computational analysis methods. I propose to illustrate this paradigm shift by presenting studies that address specific research questions from different standpoints and using different methodology. The research questions studied are: (A) Lexical Polysemy and Word Sense Clusterability, (B) In-Context Lexical Substitution, and (C) Adjective Intensity Identification.

(A) **Lexical Polysemy and Word Sense Clusterability**

In this chapter, I present two studies which approach the questions of lexical polysemy and word sense clusterability using manual meaning annotations (substitutes and translations), and representations generated by neural language models. The first paper was published in the Computational Linguistics journal (MIT Press) in 2016. The second paper has been published in the Transactions of the ACL (TACL) journal (MIT Press) in 2021.

(i) McCarthy et al. (2016): Diana McCarthy, Marianna Apidianaki and Katrin Erk, "*Word Sense Clustering and Clusterability*", Computational Linguistics Journal, Vol. 42(2), p. 245-275.

(ii) Garí Soler and Apidianaki (2021a): Aina Garí Soler and Marianna Apidianaki, "*Let's Play Mono-Poly: BERT Can Reveal Words' Polysemy Level and Partitionability into Senses*", TACL Journal, Vol. 9, p. 825–844.

In our Computational Linguistics paper (i), we propose to study the question of word sense partitionability using clusterability metrics from the Machine Learning literature (Ackerman and Ben-David, 2009). We adopt a graph-based method to sense identification which discovers word senses by relying on substitute and translation annotations of English words (Erk et al., 2013). These annotations serve as proxy for words' meaning in context. We show that it is possible to identify words with clear sense distinctions by analysing the semantic space occupied by their instances, and to distinguish them from words with fuzzy sense boundaries where different meanings overlap with each other (as in the case of regular polysemy).

In our paper published in the TACL journal (ii), we extend and scale up the McCarthy et al. (2016) clusterability approach using representations generated by contextual language models. This overcomes the limitation of the previous study which focused on words in the annotated datasets, and allows us to apply the method to an open vocabulary and to different languages. We specifically conduct experiments using monolingual language models in English, French, Spanish and Greek (Devlin et al., 2019; Le et al., 2020; Cañete et al., 2020; Koutsikakis et al., 2020), as well as multilingual models covering these languages. We show that state-of-the-art language models encode rich knowledge about lexical polysemy, and that their representations can be effectively used to define words' ease of partitionability into senses.

(B) **In-context Lexical Substitution**

In this chapter, I present two studies which address the question of in-context lexical substitution using distributional and neural language models (Apidianaki, 2016; Garí Soler et al., 2019b). In-context lexical substitution is the process where a word (or phrase) is substituted by a synonym (or paraphrase) with similar meaning, which is also a good fit in the context. Hence, a lexical

substitution method needs to consider both the semantics of the substitutes and the context (e.g., sentence) where the substitution will take place, in order to select the best candidate.

Lexical substitution methods can be useful for language learners and in applications that require rewriting, such as text summarization and simplification systems. They can also aid systems address unknown words (words not seen during system training) by substituting them with plausible alternatives. Furthermore, word and phrase substitution is highly useful for matching alternative wordings in evaluation metrics, for example in order to award systems that generate text containing words not found in the reference. Finally, substitution models are very useful for generating adversarial examples that are fluent and preserve the meaning of the original text, which can serve to test the robustness of computational systems (Alzantot et al., 2018; Jin et al., 2020).

I present two papers addressing in-context lexical substitution which appeared at the Empirical Methods for Natural Language Processing (EMNLP) conference in 2016, and at the International Conference on Computational Semantics (IWCS) in 2019.

(i) Apidianaki (2016): Marianna Apidianaki "*Vector-space models for PPDB paraphrase ranking in context*", Proceedings of the Empirical Methods for Natural Language Processing (EMNLP) Conference - Short Papers, p. 2028–2034.

(ii) Garí Soler et al. (2019b): Garí Soler, Aina and Cocos, Anne and Apidianaki, Marianna and Callison-Burch, Chris (2019) "*A Comparison of Context-sensitive Models for Lexical Substitution*", Proceedings of the 13th International Conference on Computational Semantics (IWCS) - Long Papers, p. 271–282.

In the EMNLP 2016 paper (Apidianaki, 2016), I show how paraphrases of words in the Paraphrase Database (Ganitkevitch et al., 2013; Pavlick et al., 2015) can be used for in-context lexical substitution. Specifically, I demonstrate how a syntax-based distributional model (Thater et al., 2011) can be used to filter and rank the unigram paraphrases of words (i.e. their synonyms) according to their context of use in order to select the best substitutes. In our Garí Soler et al. (2019b) paper, we present more recent neural lexical substitution methods, which explicitly model the context of substitution and the semantics of individual lexical items. In the end of the chapter, I briefly discuss substitution approaches which rely on the capability of the BERT model (Devlin et al., 2019) to perform cloze-style slot filling.

(C) **Scalar Adjective Intensity Identification**

This chapter presents two studies addressing scalar adjective intensity. Scalar adjectives describe a property of an entity (e.g., BEAUTY, TEMPERATURE, SIZE) at different degrees of intensity (e.g., *beautiful/gorgeous* beach; *hot/scalding* drink; *compact/big/huge* car). Adjectives that express intensity can serve to assess the emotional tone of a given text (Hatzivassiloglou and McKeown, 1993; Pang et al., 2008), as opposed to relational adjectives (e.g., *wooden*, *chemical*) which contribute to its descriptive content (McNally and Boleda, 2004). Intensity estimation is also useful for detecting the directional textual entailment relationship (*wonderful* $\models$ *good* but *good* $\not\models$ *wonderful*) (Van Tiel et al., 2016; McNally, 2016), for product review analysis and recommendation systems, as well as for emotional chatbots, conversational agents and question answering applications (de Marneffe et al., 2010). Lastly, this type of knowledge can assist language learners in distinguishing and learning to use semantically similar words (Sheinman and Tokunaga, 2009). I will discuss two papers addressing this question which have been published at the Empirical Methods for Natural Language Processing (EMNLP) Conference in 2018 and in 2020.

(i) Cocos et al. (2018): Anne Cocos, Skyler Wharton, Ellie Pavlick, Marianna Apidianaki, Chris Callison-Burch, "*Learning Scalar Adjective Intensity from Paraphrases*", Proceedings of the Empirical Methods for Natural Language Processing (EMNLP) Conference, p. 1752-1762.

(ii) Garí Soler and Apidianaki (2020): Aina Garí Soler and Marianna Apidianaki, "*BERT Knows Punta Cana is not just beautiful, it's gorgeous: Ranking Scalar Adjectives with Contextualised Representations*", Proceedings of the Empirical Methods for Natural Language Processing (EMNLP) Conference, p. 7371–7385.

In our EMNLP 2018 paper, we proposed to learn the intensity relationship of scalar adjectives using paraphrases from the automatically created Paraphrase Database (PPDB) (Ganitkevitch et al., 2013; Pavlick et al., 2015). We test this method on a scalar adjective ranking task, and combine it with information extracted from corpora using patterns and with knowledge from polarity lexicons. The paraphrase-based approach guarantees the large coverage of the method, while the use of patterns and lexicons increases its precision in the ranking task.

In our EMNLP 2020 paper, we again address the question of adjective intensity identification, this time using representations generated by contextual language models. We demonstrate that information about intensity is encoded in scalar adjectives' contextualised representations. We conceive intensity as a continuum going from less intense to more intense words (e.g., *pretty* $>$ *beautiful* $>$ *gorgeous*). We construct a vector in the space built by BERT which represents the semantic notion of intensity, and which can serve to rank adjectives across this axis. Our methodology for detecting the intensity dimension is inspired from gender bias detection works, and involves simple vector calculations in the vector space constructed by the neural language model. One of its strong advantages is that it can be easily applied to other languages for which such models are available. We demonstrate the cross-lingual applicability of the method in a sequel study presented at NAACL 2021 (Garí Soler and Apidianaki, 2021b).

Chapter 2

Representations of Word Meaning

2.1 Introduction

Word representation in vector space lies in the core of distributional approaches to language processing. The idea that words' collocations describe their meaning (Harris, 1954; Firth, 1957) underlies traditional distributional vector space models (DSMs) and the structure of the semantic space built by neural language models (NLMs). Different approaches, however, address different units of meaning representation.

Traditional DSMs represent words by aggregating over their usages in a corpus of documents (Landauer and Dumais, 1997; Lund and Burgess, 1996). Similarly, classical word embedding approaches (such as word2vec, GloVe and FastText) generate a static vector per word type which groups its different senses (Mikolov et al., 2013a; Pennington et al., 2014; Bojanowski et al., 2017). Current NLMs, on the contrary, generate "dynamic" representations that differ for every new occurrence of a word in texts and directly encode the contextualised meaning of individual tokens (Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019c). For example, although a static word embedding model would create one vector for the word *bug*, a contextual model would generate different representations for instances of the word in context (e.g., "*There is a bug in my food*", "*There is a bug in my code*").

Contextualised representations constitute a powerful feature of state-of-the-art NLMs, and contribute to their impressive performance in downstream tasks. The flexibility of contextualised representations confers them an undeniable advantage over static embeddings which, by aggregating information from different contexts in the same word vector, often lead to the "meaning conflation" problem (Pilehvar and Camacho-Collados, 2020). Additionally, the dynamic nature of contextualised embeddings provides a more straightforward way for capturing meaning variation than previous sense representation methodologies which relied on semantic resources or a clustering algorithm (Reisinger and Mooney, 2010; Iacobacci et al., 2015; Camacho-Collados and Pilevar, 2018).

Interestingly, in spite of the success of this new representation paradigm, in recent works we witness a critical stance and a rise of scepticism regarding the quality of these representations and of the similarity estimates that can be drawn from them, accompanied with a resurgence of interest towards word type level vectors. This is due to several reasons. Although modelling word usage is one of contextualised representations' recognized merits and a highly useful methodological tool for studying linguistic structure (Linzen et al., 2016; Hewitt and Manning, 2019), the observed context variation makes the study of the encoded semantic knowledge challenging. Specifically, it has been shown that context specificities and the token position within a sentence have a negative impact on the quality of the semantic

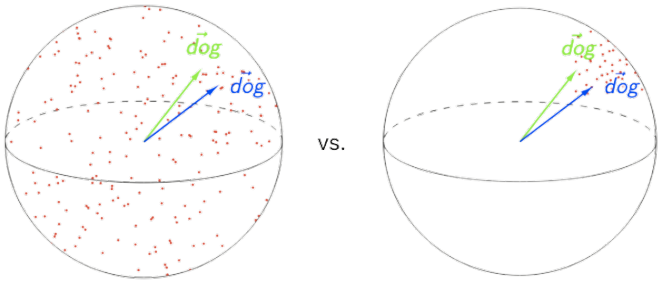


Figure 2.2: Comparison of an isotropic vector space where embeddings are uniformly distributed in all directions (left) with a highly anisotropic space (right).

similarity estimates that can be drawn from them (Mickus et al., 2020). Hence, two occurrences of the same word expressing the same meaning in slightly different contexts, or at different positions in the sentence, might get substantially different representations. Figure 2.1 shows the pairwise cosine similarity of contextualised vectors obtained for the highlighted words from the last layer of the BERT (`bert-base-uncased`) model. We observe that vectors assigned to different words in the same context are more similar than the ones assigned to semantically similar instances of the same word in different contexts. The vectors for *interesting* and *great* in examples (i) and (iii) have a cosine similarity of 0.753, while the vectors for the instances of the noun *field* in (b) and (c) get a score of 0.411. Surprisingly, the vectors for antonymous words (*interesting* ↮ *boring* and *great* ↮ *horrible*) are among the most similar. The instances of these adjectives in sentences (i) and (ii), and (iii) and (iv), have a cosine similarity score of 0.691 and 0.670, respectively. We would, however, expect a high quality semantic vector space to reflect the dissimilarity between these semantically opposite words in any context.

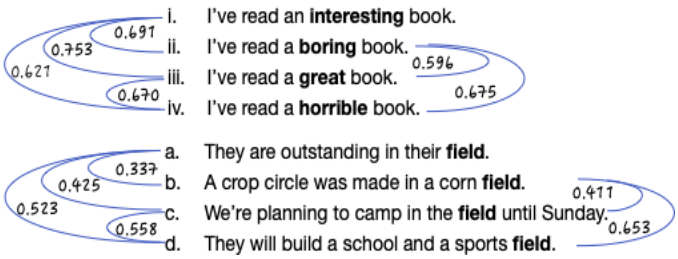


Figure 2.1: Similarity of vectors for word instance pairs.

The issue of the problematic, or distorted, similarity estimates is accentuated by the geometry of the highly anisotropic semantic space that is constructed by contextual models (Ethayarajh, 2019a). In this space, word vectors are concentrated in a narrow (cone-shaped) area, instead of being uniformly distributed in the constructed space, as shown in Figure 2.2. Since similarity is a relative notion (Arora et al., 2017), this particularly close proximity of all word instance vectors gives rise to problematic similarity estimations. Ethayarajh (2019a) specifically highlights the high similarity of random words in the anisotropic space constructed by the models. Hence, although reflecting the meaning of word usages is one of the recognized merits of these dynamic vectors, the observed variation makes the exploration of semantic knowledge challenging. This explains to some extent the resurgence of interest towards word-level representations, deemed to provide a more solid basis for meaning exploration. Naturally, this trend is mainly observed in the lexical semantics field where the notion of lexical concept is central (Lauscher et al., 2020b; Liu et al., 2020; Bommasani et al., 2020; Vulić et al., 2020b; Garí Soler and Apidianaki, 2021a).

The prevalence of contextual models has also brought about a shift from methodologies traditionally used to evaluate the quality of distributed representations (e.g., out-of-context word similarity and analogy tasks) (Mikolov et al., 2013b) to interpretation tools common in human language learning studies (e.g., cloze tasks and probes) (Linzen et al., 2016; Kovaleva et al., 2019; Tenney et al., 2019b; Ettinger,

2020). These serve to assess the linguistic and world knowledge encoded in contextualised vectors, and are often complemented with methods that explore the models' inner workings, such as their self-attention mechanism or the information flow between layers (Voita et al., 2019a; Hewitt and Manning, 2019; Clark et al., 2019; Voita et al., 2019b; Tenney et al., 2019a). In lexical semantics, the probing methodology is used to explore the knowledge models encode about semantic properties and relationships (Petroni et al., 2019; Bouraoui et al., 2020; Ravichander et al., 2020; Apidianaki and Garí Soler, 2021). Nevertheless, evaluations that rely on cloze-task prompts are not really indicative of the knowledge the models encode about language, since results strongly depend on prompt quality (Ettinger, 2020; Apidianaki and Garí Soler, 2021). Additionally, semantic information might be implicitly encoded and not explicitly stated in texts due to the reporting bias phenomenon, in which case it is hard to recover this knowledge from language model representations using probing (Gordon and Van Durme, 2013; Shwartz and Choi, 2020; Apidianaki and Garí Soler, 2021). These two factors undermine the reliability of the cloze-task probing methodology for semantics. The variability of the results obtained with different prompts has brought attention back to word similarity and analogy tasks, considered as more established and mature for exploring the concept-related knowledge encoded in language model representations (Vulić et al., 2020b; Bommasani et al., 2020).

This section provides an overview of word and meaning representation methodologies that rely on distributional approaches and language models. We present methods that generate distributed representations (embeddings) at the level of word types, senses, and individual instances. We explain the evolution from distributional to distributed embedding representations, as well as the advantages of the latter over traditional distributional models. For a full account of distributional approaches and their origins, we point the reader to the surveys by Turney and Pantel (2010), Erk (2012) and Clark (2015). For a discussion of their relationship with theoretical and formal linguistics, we refer the reader to Boleda and Herbelot (2016) and Boleda (2020). Finally, a thorough account into embeddings generated by different types of language models is given in the essay of Pilehvar and Camacho-Collados (2020).

2.2 Distributional Representations

Semantic spaces are a popular framework for meaning representation, encoding words as high-dimensional vectors (Schütze, 1998). Distributional Semantic Models (DSMs) construct such spaces by collecting vectors that keep track of lexical co-occurrence patterns in large text corpora, following the distributional hypothesis of meaning (Harris, 1954). The central idea behind the distributional approach is that "the meaning of a word is its use in the language" (Wittgenstein, 1953) or, put in other words, "you shall know a word by the company it keeps" (Firth, 1957).

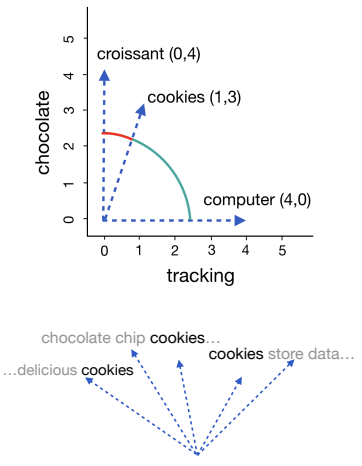


Figure 2.3: Toy examples of word similarity in vector space.

Since related words occur in similar contexts, distributional vectors also provide a robust model of semantic similarity which is reflected in the proximity of words in the constructed high-dimensional vector space (Miller and Charles, 1991). In this space, a word is represented as a point where the dimensions stand for context items (co-occurring words), and the word's coordinates represent its context counts (Erk, 2012). Geometric distance (or proximity) between vectors (as measured by the cosine of their angle, or their Euclidean distance) tends to correlate well with human semantic similarity judgements. This is illustrated by the artificially small vector space on top of Figure 2.3, where *croissant* is close to *cookies*

because they often co-occur with *chocolate*, but *cookies* also has a computer-related meaning as tracking devices. As we have shown in previous work, vector similarity can serve to identify word senses (Apidianaki, 2008a,b, 2009a; Van de Cruys and Apidianaki, 2011; Apidianaki et al., 2014; McCarthy et al., 2016), detect semantic relationships (Rajana et al., 2017; Cocos et al., 2018), and determine word substitutability in context (Apidianaki, 2016; Cocos et al., 2017; Garí Soler et al., 2019b).

Distributional models of meaning can be learned from a corpus in an unsupervised fashion (Erk, 2012). As a consequence, the provided model of word meaning is independent of dictionary senses and underlying linguistic theories (Kilgarriff, 1998; Ide and Wilks, 2007). One of the downsides of traditional distributional methods is that the use of raw co-occurrence counts leads to sparse representations. A common approach to alleviate this sparseness and improve the performance of the representations is to apply some type of transformation to the raw vectors. This involves reweighting the counts for context informativeness and smoothing them with dimensionality reduction techniques (e.g., Singular Value Decomposition (SVD)) (Turney and Pantel, 2010). The applied optimization process might be unsupervised and based on independent (for example, information-theoretic) considerations. I can also rely on some sort of indirect supervision where the best parameter settings are chosen based on performance on some semantic task selected for tuning (Baroni et al., 2014b). This is an important point of difference from word embedding techniques which replace the stacking of vector transforms with a single supervised learning step, generating low-dimensional vector representations at no manual annotation cost. We explain this procedure in more detail in Sections 2.4 and 2.5.

An important aspect that deserves our attention is how contextualization occurs in distributional models. As explained above, a word (lemma) in a distributional model is represented as a point in space, the position of which is defined by its co-occurrences over a large corpus. This representation is the result of an aggregation over multiple contexts (Landauer and Dumais, 1997; Lund and Burgess, 1996), it will hence encode information for different meanings of the word. In the toy example given in Figure 2.3, the position of *cookies* shows its semantic proximity to *croissant* as well as its relation to *computer*. However, every new occurrence of a word in a text might instantiate a different meaning (e.g., *Chocolate chip cookies are my favourite* vs. *You can disable unnecessary cookies*). In order to characterise the meaning of individual instances, different methods have proposed to compute meaning in context from lemma vectors. The problem is then addressed as a matter of vector composition (Schütze, 1998; Mitchell and Lapata, 2008): the meaning of a target occurrence a in context b is a single new vector c that is a function of the two vectors: $c = a \odot b$. Mitchell and Lapata apply this method to calculating phrase meaning, and test different composition methods, such as vector addition and component-wise multiplication. The good results of the latter has made of it a widely used approach for calculating phrase meaning as in verb phrases (e.g., *catch a ball* vs. *attend a ball*) (Erk and Padó, 2008) and noun phrases involving a modifier (e.g., *green chair* vs. *green initiative*) (Baroni and Zamparelli, 2010; Zanzotto et al., 2010).

When considering entire sentences, it is important to integrate syntax into the computation of word meaning in context, because the position and syntactic role of a word in the sentence greatly impacts its meaning (Erk and Padó, 2008; Thater et al., 2011). A model which does not account for syntax would, for example, generate the same representation for *school* in "*law school*" and in "*school law*". A good quality semantic model must also be able to distinguish "*Mary peeled the avocado*" from "*The avocado peeled Mary*", similar to human hearers, and to draw different conclusions from them, specifically that the event described by the second sentence is not as probable as that described by the first one. The dependency-based vector space of Padó and Lapata (2007) models coocurrences of words linked with dependency relations in a parsed corpus, viewing syntactic and argument structure

as a reflection of lexical meaning (Levin, 1993). They precisely model meaning by quantifying the degree to which words are attested in similar syntactic environments. The structure vector space model of Erk and Padó (2008) also accounts for selectional preferences; the meaning of a word is represented as a combination of a vector that models lexical meaning, and a set of vectors which represent the semantic expectations for each relation that the word supports (e.g., *catch a cold* vs. *catch a ball*). In the model of Thater et al. (2011), the vector of a target word is modified according to the words in its syntactic context. Contextualization of a vector is performed by reweighting its components based on distributional information about the context words.

Still, it is difficult to generate a vector which properly describes the meaning of an entire sentence. As noted by Erk and Padó (2008), classical composition methods which result in a single vector are not informative enough to provide interesting meaning representations for entire sentences. It is hard to conceive how a single vector can encode deeper semantic properties like predicate- argument structure (cf. the "*avocado*" example above) which are crucial for sentence-level semantic tasks such as the recognition of textual entailment (Dagan et al., 2006). There has, however, been important evolution on this topic in recent years with neural models which generate sentence representations that achieve impressive performance in downstream natural language understanding tasks. It, however, remains difficult to find out how this is done, and what is really encoded in these sentence representations (Conneau et al., 2018). Furthermore, these vectors seem to not capture semantic similarity well (Reimers and Gurevych, 2019).

2.3 Lexical Meaning in Formal Semantics

Formal Semantics provides semantic representations of linguistic expressions using logic and other symbolic mathematical tools (Montague, 1973). It is mainly centred around the inferential properties of language and compositionality, the mechanism by which the meaning of sentences is incrementally built by combining the meanings of their constituents. Formal Semantics research has been mainly preoccupied with logical form and the mapping from a sentence-level syntactic representation to a logical representation, with a few exceptions of approaches that recognise the central role of lexical meaning in semantic inference and explicitly account for it (Asher, 2011). The Generative Lexicon (Pustejovsky, 1995), for example, is a system that involves different levels of semantic representation, and a set of generative devices which connect these levels and account for the compositional interpretation of words in context. It provides an expressive and flexible formal statement of language which can capture the generative nature of lexical creativity and sense extension, contrary to approaches aimed at the exhaustive listing of word senses.

Formal Semantics approaches to meaning interpretation are highly elegant but face the issue of lexicon coverage. It is not feasible to manually specify how the meaning of content words in the lexicon is affected by context, be it for regular polysemy or for non-systematic meaning shifts (Baroni et al., 2014a). DSMs, instead, provide extremely rich data for lexical semantic analysis, and can contribute to building large coverage semantic systems. They specifically provide good representations of the descriptive content of linguistic expressions by encoding abstractions over contexts of use observed in large amounts of data (Schütze, 1998; Pantel and Lin, 2002), as explained in Section 2.2. They also account for polysemy, as reflected in word usage, and capture lexical relations through geometric distance in the constructed vector space. DSMs can also capture fuzzy aspects of meaning and subtle meaning shifts by accounting for the graded similarity of word usages (Erk et al., 2013; Jurgens and Klapaftis, 2013), while their representations can be combined into more complex meanings (Baroni et al., 2014a). Recent works attempt to combine the strengths of Formal Semantics relative to the

rigorous account of linguistic structure, with the ability of DSMs to account for the descriptive content of linguistic expressions. Asher et al. (2016), for example, effectively combine Type Composition Logic (TCL) (Asher, 2011), a detailed formal model of the interaction between composition and lexical meaning, with DSMs that provide the information needed for constructing the functors within the TCL construction process. This trend is becoming increasingly important and a new framework is being established under the umbrella term of "Formal Distributional Semantics" (Boleda and Herbelot, 2016).

Contextualised representations generated by state-of-the-art models (Devlin et al., 2019; Liu et al., 2019c) can open up new avenues for Formal Semantics research since they encode rich information about word usage. They can provide representations of the descriptive content of words and also reflect minor meaning shifts, thanks to their high sensitivity to syntactic variation which can be beneficial in this setting. This can also help in handling types of polysemy that are central in Formal Semantics but difficult to handle with traditional DSM models, which involve related (or complementary) senses that are less sensible to contextual priming than contrastive senses and, thus, more difficult to disambiguate (Haber and Poesio, 2020, 2021).

2.4 Language Modelling with Deep Neural Networks

Deep artificial neural networks (ANNs) belong to the "representation learning" paradigm that seeks to automatically induce useful features for a task from raw text data (LeCun et al., 2015). This differentiates them from traditional Machine Learning methods that involve high levels of feature engineering, which can be time-consuming and error-prone. An ANN is a computational non-linear model inspired by the neural structure of the brain. It consists of artificial neurons – or processing elements – and is organised in interconnected layers: input, (one or more) hidden layer(s), and output (cf. Figure 2.4). Deep learning models involve multiple hidden layers and learn representations of data at different levels of abstraction (LeCun et al., 2015). This multi-layer architecture contributes to their high performance, since it allows them to learn complex functions from input (e.g., a Twitter post) to output (its sentiment polarity), and more complex features than shallow networks. Alongside their powerful architecture, ANNs also benefit from the abundance of training material in the form of naturally occurring text. Their success over traditional machine learning methods is undeniable, as shown by their results on the GLUE (General Language Understanding Evaluation) benchmark (Bowman et al., 2015).[1] However, it remains unclear whether they manage to learn the deep knowledge actually needed for understanding and modelling language (Tenney et al., 2019b).

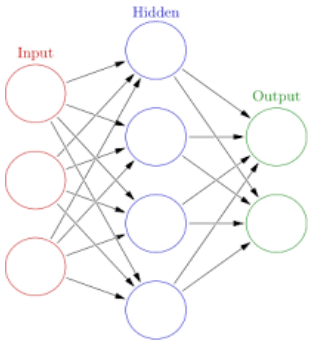


Figure 2.4: One hidden layer ANN.

State-of-the-art ANNs are actually trained on large text corpora (for example, the 1B word corpus (Chelba et al., 2013) or the English Wikipedia) and encode the observed distributional regularities in the form of complex co-occurrence statistics through "language modelling" objectives. Language modelling is the task of predicting upcoming words in a text, that is estimating how likely (probable) a particular word *w* is to occur given the previous words in a sequence. For example, when a model processes the sentence "*These flowers ___ so good*", it uses its existing (often randomly initialised) weights

[1]The GLUE benchmark includes eight tasks: MultiGenre Natural Language Inference (MNLI), Quora Question Pairs (QQP), Question NLI (QNLI), The Stanford Sentiment Treebank (SST-2), The Corpus of Linguistic Acceptability (CoLA), The Semantic Textual Similarity Benchmark (STSB), Microsoft Research Paraphrase Corpus (MRPC), Recognizing Textual Entailment (RTE).

to predict how likely each word of the language (the English vocabulary) is to follow "*flowers*". When the hidden word *smell* is revealed, the network's weights are adjusted using standard error backpropagation and gradient descent techniques (Nielsen, 2018; Rumelhart et al., 1986), such that in the future it will assign a higher probability to "*smell*" in a similar context. Gradient descent is specifically an optimisation algorithm which is used to train machine learning models and neural networks, by minimising errors between predicted and actual results.

Most powerful state-of-the-art models such as Long Short-Term Memory Recurrent Neural Networks (LSTM RNNs) (Elman, 1990; Hochreiter and Schmidhuber, 1997), bidirectional language models (Peters et al., 2018) and Transformer-based models (Vaswani et al., 2017; Radford et al., 2019; Devlin et al., 2019) rely on language modelling objectives. The representations ANNs acquire during this type of pre-training are general-purpose and can be adapted (fine-tuned) to specific downstream tasks (e.g., question answering or sentiment analysis). Task-specific datasets are typically small; hence, general knowledge about language is acquired during pre-training on large text corpora, and then the model weights are basically adjusted on datasets tailored for the task at hand, as in transfer learning (Pan and Yang, 2010). Fine-tuned models tend to deliver optimal performance, but it is unclear how much of the knowledge learned during pre-training is actually used to solve these tasks, and what knowledge is acquired during fine-tuning. Sometimes the fine-tuning process can do more harm than good, when useful information acquired during pre-training might be overwritten through exposure to new information, a process characteristically called "catastrophic forgetting" (Mccloskey and Cohen, 1989). The learning objective used for training the models has also been shown to have an important impact on the content of the acquired representations (Voita et al., 2019a; Kovaleva et al., 2019; Conneau et al., 2018).

Each model represents words either with static (word-level and out-of-context) or with contextualised (instance-based) representations. In the next sections, we present the types of word vectors that can be obtained from different models, explaining the implications that each type of representation has on the modelling of lexical semantics, in general, and the meaning of specific words, in particular.

2.5 Static Word Embeddings

2.5.1 Word Representation Learning

Word embedding models leverage neural networks to directly learn low-dimensional word representations from corpora (Bengio et al., 2003; Collobert and Weston, 2008; Collobert et al., 2011; Mikolov et al., 2013a). These "self supervision" models are trained on raw text and rely on the optimization of a specific objective. With a language modelling objective, for example, the vector estimation problem is framed directly as a supervised task, where the weights in a word vector are set to maximise the probability of the contexts where the word is observed in the training corpus.

A classical word embedding architecture is that of word2vec (Mikolov et al., 2013a) which consists of input, hidden and output layers. The input layer has the size of the vocabulary of the language and encodes the context of a target word as a combination of one-hot vector representations of the words surrounding it in texts. The output layer has the same size as the input layer and contains a vector of the target word built during the training phase. The model is trained using a predictive task which, in the case of the Continuous Bag-of-Words (CBOW) Word2Vec model, is to predict the current word using its surrounding context minimizing this loss function:

$$E = -log(p(\vec{w}_t|\vec{W}_t)) \tag{2.1}$$

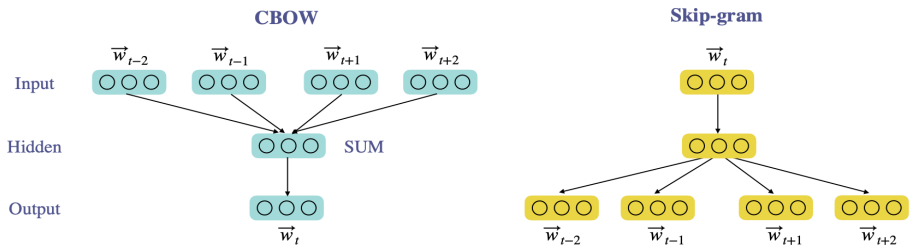


Figure 2.5: Architecture of the CBOW and Skip-gram models.

where w_t is the target word and $W_t = w_{t-n}, ..., w_t, ..., w_{t+n}$ represents the sequence of words in context. The CBOW architecture is based on the feedforward neural network language model (Bengio et al., 2003). In Word2Vec Skip-gram, on the contrary, the goal is to predict the words in the surrounding context given the target word, rather than predicting the target word itself. The difference between the two models is illustrated in Figure 2.5.

The idea underlying word embedding models is that contextual information can provide a good approximation for word meaning, since semantically similar words tend to have similar contextual distributions (Harris, 1954; Miller and Charles, 1991). This is also the guiding principle of traditional Distributional Semantic Models (DSMs) (Turney and Pantel, 2010; Erk et al., 2009; Clark, 2015), as explained in Section 2.2. Another similarity with DSMs is that self-supervised embedding learning models can be trained on raw text without need for manual annotations. As a result, they can be applied to different languages. The dimensionality of embedding vectors is, however, much lower than that of distributional vectors where dimensions correspond to words in the vocabulary (i.e. their number can easily reach hundreds of thousands or even millions). The high dimensionality of distributional vectors challenges the scalability and computational efficiency of the models. Although vector transformations can reduce their sparseness, word embedding techniques replace the heuristic stacking of vector transforms with a single supervised learning step at no manual annotation cost. Furthermore, pre-trained word embeddings outperform count-based representations in intrinsic evaluations (i.e. word similarity and relatedness tasks) (Mikolov et al., 2013c; Baroni et al., 2014b), and can be successfully integrated in downstream applications due to their high generalization potential.

2.5.2 The Meaning Conflation Problem

One of the limitations of static word embedding models, like word2vec (Mikolov et al., 2013a), GloVe (Pennington et al., 2014) and FastText (Bojanowski et al., 2017), is that they build a single representation for each word in the vocabulary of a language. **Therefore, they are by design unable to model polysemy.** On the contrary, the contextual evidence for different word meanings is conflated into a single vector. Representing a word type as a single point in the semantic space is another similarity between embedding and early distributional models (Landauer and Dumais, 1997; Lund and Burgess, 1996), and is considered as a major deficiency of these models. Not distinguishing between different meanings of a polysemous word (e.g., *plant*, *mouse*, *bug*) can negatively impact the semantic understanding of NLP systems that

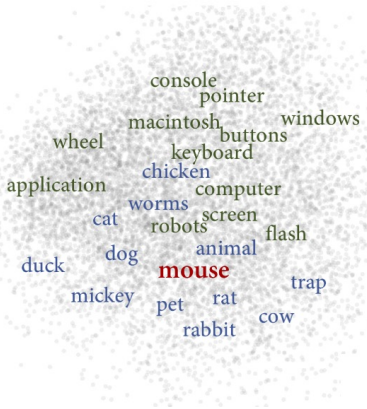


Figure 2.6: Illustration of word embeddings' meaning conflation deficiency in a 2D semantic space.

rely on these representations. The conflation of different
meanings in the same vector has consequences on the structure of the obtained semantic space and on semantic modelling accuracy (Neelakantan et al., 2014; Chen et al., 2014; Camacho-Collados and Pilevar, 2018). Representing an ambiguous word (e.g., *mouse*) as a single point in space pulls together semantically unrelated words (e.g., *rat*, *cat* and *computer*) due to their similarity to different senses of the ambiguous word, as illustrated in Figure 2.6.[2] A careful analysis shows that multiple word senses reside in linear superposition within word2vec and GloVe word embeddings, and that vectors that approximately capture the senses can be recovered using simple sparse coding (Arora et al., 2018). Furthermore, similar to distributional methods for modelling compositionality (as the ones described in Section 2.2), a simple method for generating context-specific representations for words is to aggregate over the embeddings of their co-occurrences in a sentence or a specific context window.

2.5.3 Static Embedding Evaluation

The quality of static word embeddings has traditionally been evaluated using word analogy and similarity tasks. Word analogy is usually framed as relational similarity; it models the idea that pairs of words may hold similar relations to those that exist between other pairs of words (Turney, 2006).[3] Analogies are thus perceived as equations of the form $a : b :: c : d$ (i.e. "*a is to b what c is to d*") where given the first three terms (a, b, c), the tested model needs to predict the word that stands for d. Mikolov et al. (2013a) showed that such relations are reflected in vector offsets between word pairs.[4] In the famous example "*man is to king as woman is to X*", the embedding for the word *queen* can be roughly recovered from the representations of *king*, *man* and *woman* using the following equation: $\vec{queen} \approx \vec{king} - \vec{man} + \vec{woman}$.

Word analogies became a highly popular tool for embedding evaluation, but were then discredited due to numerous concerns regarding the use of the vector offset method for solving analogies. The accuracy of this method depends on the proximity of the target vector to its source (e.g., $\vec{queen}$ and $\vec{king}$), limiting its applicability to linguistic relations that happen to be close in the vector space (Rogers et al., 2017). Reliance on cosine similarity also conflates offset consistency with largely irrelevant neighbourhood structure, while results are inconsistent when the direction of the analogy is reversed (even though the same offset is involved in both directions) (Linzen, 2016). Last but not least, linguistic relations might not always translate to linear relations between vectors but to more complex correspondence patterns (Drozd et al., 2016; Ethayarajh, 2019b). Another issue with the classical analogy task is that examples are structured such that given the first three terms, there is one specific, correct (expected) fourth term. This might be the case with factual queries involving morpho-syntactic and grammatical alternations (e.g., *high/higher*, *long/longer*), but for semantic queries there might be several equally plausible correct answers (e.g., *man:doctor :: woman:X*). These semantic analogies are more creative and various terms could be used for completion depending on the implied relation, which might be unspecified in the query (Nissim et al., 2020).

The quality of distributional and word embedding representations is also intrinsically evaluated against human similarity and relatedness judgements (Rubenstein and Goodenough, 1965; Miller and Charles, 1991; Hodgson, 1991; Finkelstein et al., 2001; Bruni et al., 2012; Jurgens et al., 2012). A high correlation between human judgements on word pairs and the cosine of the corresponding vectors, demon-

[2] Figure taken from Camacho-Collados and Pilevar (2018).

[3] As noted by Rogers et al. (2017), this conception of analogy is different from the notion of analogy in philosophy and logic. The classical analogical reasoning follows this template: objects X and Y share properties a, b and c, therefore they may also share the property d.

[4] The answer to the above question is represented by hidden vector d which is calculated as $argmax_{d \in V}(sim(d, c - a + b))$, where V is the vocabulary excluding words a, b and c, and sim is a similarity measure.

strates the quality of the constructed space. A downside of this type of evaluation is that similarity scores are given to pairs of words in isolation, and does not allow to assess the capability of the models to capture polysemy and word meaning in context.

2.6 Word Sense Representations

2.6.1 Multi-Prototype Embeddings

A solution to the meaning conflation problem of word embeddings is to generate separate vectors for different word senses. Multi-prototype methods generate such vectors for senses discovered from texts using unsupervised Word Sense Induction (WSI) methods (Reisinger and Mooney, 2010; Neelakantan et al., 2014; Huang et al., 2012). The contexts where a word occurs are clustered and a distinct prototype vector is generated from each cluster by averaging over the context vectors it contains. The method of Reisinger and Mooney (2010) is illustrated in Figure 2.7 (a).

Multi-prototype methods vary with respect to the vector representations, the clustering algorithm and the context used. Reisinger and Mooney (2010) use count-based vectors composed of features that correspond to the unigrams in a ten-word context window around a target word w_t, while Huang et al. (2012) and Neelakantan et al. (2014) use word embeddings. In terms of the clustering algorithm, Reisinger and Mooney (2010) apply a "mixture of von Mises-Fisher distributions" (movMF) clustering method, while Huang et al. (2012) use the K-means algorithm to decompose words' continuous distributed representations into multiple prototypes. The method of Neelakantan et al. (2014) is a multi-prototype extension of the Skip-gram model called Multiple-Sense Skip-Gram (MSSG), which represents the context of a target word w_t as the centroid of the vectors of the context words, and clusters them to form w_t's sense representations. Contrary to previous multi-prototype methods, clustering and sense embedding learning are performed jointly during training. The intended sense for a word is dynamically selected as the closest sense to the context and weights are updated only for that sense. Tian et al. (2014) propose a technique that significantly reduces the number of parameters in the Huang et al. (2012) model. Word embeddings in the Skip-gram model are replaced with a finite mixture model where each mixture corresponds to a prototype of the word. The multi-prototype Skip-gram model is trained using the Expectation-Maximization algorithm. In contrast to previous methods where senses were induced from words' local context, Liu et al. (2015) propose Topical Word Embeddings (TWE). Each word is allowed to have different embeddings under different topics computed globally using latent topic modelling (Blei et al., 2003b).

While offering a solution to the meaning conflation problem, multi-prototype embedding methods also face a number of challenges. In early methods, the number of clusters (or senses, k) is a parameter that needs to be pre-defined. This number is sometimes chosen arbitrarily and used for all words, independently of their polysemy (Huang et al., 2012). Moreover, these methods are generally offline and difficult to adapt to new data and domains, or to capture new senses (Chen et al., 2014). An alternative is to use a non-parametric clustering method which allows to dynamically adjust the number of senses to each word. Neelakantan et al. (2014)'s method precisely relies on the notion of "facility location" (Meyerson, 2001): A new cluster is created online during training with probability proportional to the distance λ from its context to the nearest cluster (sense).[5] The higher this distance, the higher the probability that the context describes a new sense of the word. The same idea underlies the method of Li and Jurafsky (2015) who learn embeddings for senses of a word induced using the Chinese Restaurant Processes (Blei et al., 2003a), a practical interpretation of Dirichlet Processes (Ferguson, 1973) for non-

[5] λ is a hyperparameter of the model. Its value is selected through manual exploration on a validation set.

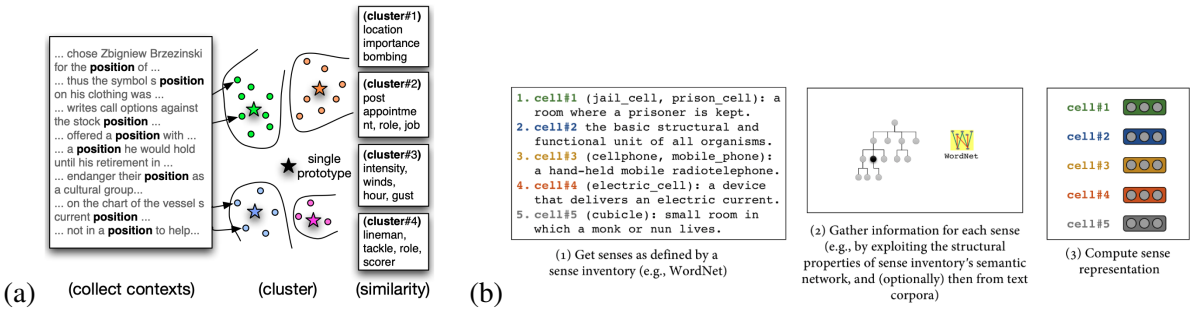


Figure 2.7: Illustration of a multi-prototype approach (a) and a sense embedding approach (b).

parametric clustering. A word is associated with a new sense vector when evidence in the context (e.g., the neighboring words) suggests that it is sufficiently different from its previously identified senses.

Other issues of multi-prototype methods are that the clusters are not always interpretable (i.e. it is difficult to determine the senses they correspond to), and the representations obtained for infrequent senses are unreliable (Pilehvar and Collier, 2016). Finally, the usefulness of using this type of sense embeddings in downstream tasks is unclear. These have been shown to outperform static embeddings in intrinsic evaluations, but when tested in real NLP applications they seem to benefit some tasks (part-of-speech tagging and semantic relation identification) and harm others (sentiment analysis and named entity extraction) (Li and Jurafsky, 2015).

2.6.2 Translation-Based Embeddings

Seeking a stable criterion for word sense identification, several studies use translations as proxies for senses. This idea dates back to the work of Gale et al. (1992) where it was put forward as a solution to the knowledge acquisition bottleneck. It has since been adopted in several word sense induction and disambiguation works (Dagan and Itai, 1994; Dyvik, 1998, 2004; Resnik and Yarowsky, 1999; Ide et al., 2002; Resnik, 2004; Diab and Resnik, 2002; Carpuat and Wu, 2007; Apidianaki, 2008b, 2009a; Lefever et al., 2011; Carpuat, 2013). The underlying assumption is that the senses of a polysemous word in a source language (w_s) are translated with different words in other languages ($T = t_1, ..., t_n$). Clustering is still relevant in this context since synonymous translations can be grouped to describe one sense of word w_s (Apidianaki, 2008c). Translation clusters can be associated to context vectors which can serve for disambiguation (Apidianaki, 2009a).

Guo et al. (2014) use translations to create sense embeddings. They project translation (English) clusters describing senses onto source language (Chinese) words in a parallel corpus, in order to create the labelled data needed for training a neural network language model that generates the sense embeddings. The obtained representations significantly outperform static and multi-prototype embeddings in a word similarity and a Named Entity Recognition task for polysemous words. The sense embedding method of Šuster et al. (2016) also exploits both monolingual and translation information. It assigns a sense to a pivot word during the encoding phase, and predicts context words based on the pivot word and its sense during a decoding (or reconstruction) phase. Parameters of encoding and reconstruction are jointly optimized, the goal being to minimize the error in recovering context words based on the pivot word and its assigned sense. The obtained sense-specific representations are shown to outperform their monolingual counterparts across a range of evaluation tasks, including in and out of context similarity estimation (Huang et al., 2012; Finkelstein et al., 2001; Rubenstein and Goodenough, 1965; Bruni et al., 2012). Finally, in the context of Neural Machine Translation, Liu et al. (2018) propose a method for captur-

ing contextual information in order to disambiguate difficult-to-translate homographs. They compute a context vector for each source word which is combined with the original word embedding to form context-aware word embeddings that improve the quality of the translations produced for homographs.

2.6.3 Sense Embeddings

Sense embedding methods offer another solution to the meaning conflation deficiency of type-level embeddings (Camacho-Collados and Pilevar, 2018). They practically produce vectors corresponding to senses found in lexicographic resources, which are more interpretable than the ones obtained through clustering. A typical sense embedding procedure is illustrated in Figure 2.7 (b).[6]

Sense embedding approaches can rely on definitions (glosses) of senses in a lexicon, or combine this knowledge with information collected from corpora. The SENSEMBED method of Iacobacci et al. (2015), for example, learns sense representations from large text corpora disambiguated and sense annotated with the knowledge-based Babelfy algorithm (Moro et al., 2014; Navigli and Ponzetto, 2010). Similarly, the "Senses and Words to Vectors" (SW2V) neural model of Mancini et al. (2017) jointly learns word and sense embeddings by exploiting knowledge from BabelNet and large text corpora. SW2V relies on the CBOW word2vec model (Mikolov et al., 2013a). The input to the model is an automatically sense-disambiguated corpus. A training instance is a sequence of words and their associated senses. The underlying assumption is that since a word is the surface form of an underlying sense, updating the embedding of the word should produce a consequent update to the embedding of the sense, and inversely. The learned embeddings are represented in the same unified vector space. The approach outperforms previous sense embeddings approaches on out-of-context word similarity tasks (SimLex-999 (Hill et al., 2015) and MEN (Bruni et al., 2012)).

Naturally, the results of these methods and the quality of the generated sense representations strongly depend on the success of the disambiguation step. If word instances are assigned the wrong senses, this has a direct impact on the quality of the representations. The method of Chen et al. (2014) alleviates this dependence by learning representations for senses using their definitions (glosses) in WordNet (Fellbaum, 1998). Each sense is represented using the average of the vectors of the content words in the gloss that are most similar to the target word. The authors modify the training objective of Skip-gram, and train word vector representations that are good at predicting not only a word's context words but also their senses. Since each sense vector corresponds to a WordNet sense, the obtained vectors can be directly used for knowledge-based WSD.

2.7 Contextualised Representations

Contextual language models constitute a novel representation paradigm where the generated embeddings encode the meaning of individual word tokens (Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019c). Contrary to static embeddings which describe word types, **contextual models assign different vectors to different instances of a word depending on the context of use** (e.g., "*There is a bug in my soup*", "*There is a bug in my code*"). These vectors are dynamic and can capture subtle meaning nuances expressed by word instances, alleviating at the same time the meaning conflation problem of static embeddings and sense embeddings' reliance on lexicographic resources.

The first contextual language model was ELMo (Embeddings from Language Models) (Peters et al., 2018). ELMo is a two-layer bidirectional LSTM (biLSTM) language model (Hochreiter and Schmid-

[6]The figure is taken from the original paper by Camacho-Collados and Pilevar (2018).

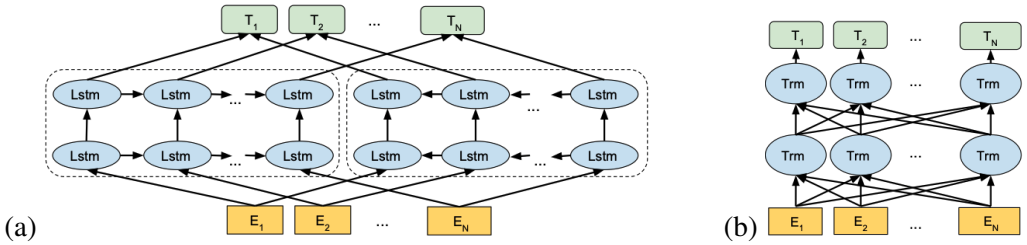


Figure 2.8: Figure (a) illustrates the architecture of the ELMo language model which relies on a bidirectional LSTM. Figure (b) describes the architecture of the Transformer-based BERT model.

huber, 1997; Graves and Schmidhuber, 2005), built over a context-independent character Convolutional Neural Network (CNN) layer. The original ELMo model was trained on the Billion Word Benchmark dataset (Chelba et al., 2013) which consists primarily of newswire text. ELMo representations are a linear combination of the internal layers of the model. Its architecture is illustrated in Figure 2.8 (a). When ELMo is integrated into task-specific architectures, the task and the linear combination of the layers are simultaneously learned in a supervised way.

Soon after ELMo appeared, the BERT Transformer model was proposed (Devlin et al., 2019). BERT has been the most influential contextual model and a high number of variants have been developed (Liu et al., 2019c; Sanh et al., 2020; Lan et al., 2020; Joshi et al., 2020). In contrast to traditional sequence models based on recurrent architectures, the Transformer model uses a fully attention-based approach. It relies on the "self-attention" mechanism where the representation of a sequence is computed by relating different words (positions) in the same sequence (Vaswani et al., 2017). The attention function can be considered as a mapping between a query and a set of key-value pairs, to an output. The output is computed as a weighted sum of the values; the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. Figure 2.9 illustrates the attention patterns produced by different attention heads in the 10th layer of BERT for inputs "*the cat sat on the mat*" (Sentence A) and "*the cat lay on the rug*" (Sentence B) (Vig, 2019).[7] The [SEP] symbol is a special separator token that indicates a sentence boundary. [CLS] is a symbol appended to the front of the input that is used for classification tasks.

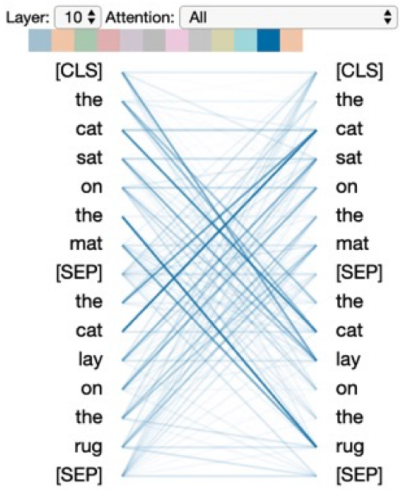


Figure 2.9: BERT multi-head attention.

The fully attention-based approach in the Transformer shows improved performance compared to traditional recurrent architectures. Apart from BERT and its variants, other high performing Transformer-based models are the OpenAI GPT-2 and GPT3 models (Radford et al., 2019) which deliver high performance on several benchmarks in a zero-shot setting. Additionally, attention is a useful interpretation tool which shows how the model assigns weight to different input elements when performing specific tasks (Raganato and Tiedemann, 2018; Voita et al., 2019a; Kovaleva et al., 2019; Rogers et al., 2020).

The power of BERT also lies in the use of a bidirectional model. Contrary to ELMo, where a forward and a backward language model are separately trained,[8] BERT jointly conditions on the left and right

[7] The figure is produced by the BertViz multiscale visualization tool for the Transformer model (Vig, 2019).

[8] A forward LM computes the probability of the sequence by modelling the probability of a given token (t_k) given the

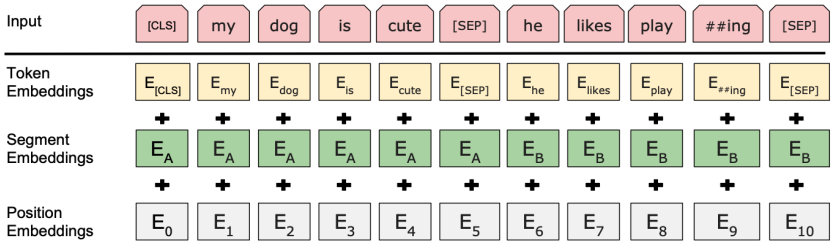


Figure 2.10: BERT input representation for the sequence "*My dog is cute. He likes playing*". The input embeddings are the sum of the token, segmentation and position embeddings.

context in all layers, as shown in Figure 2.8 (b).[9] Bidirectional representations in the BERT model are obtained with a double pre-training objective: (a) a Masked Language modelling (MLM) task similar to a Cloze task (Taylor, 1953), where a portion of the input tokens is masked at random (e.g., *The cat* [MASK] *on the mat*) and the model has to predict those tokens; and (b) a Next Sentence Prediction (NSP) task, where the model needs to predict whether two segments follow each other in the original text.

In MLM, the portion of words to mask is a parameter that needs to be set for model training. The tradeoff between masked and unmasked words is important; when too many words are masked training gets expensive, but masking very few words does not provide enough context for the model to learn good quality representations. Moreover, masking tokens in the training corpus creates a mismatch between pre-training and fine-tuning, where the [MASK] token does not appear. To mitigate this issue, only 15% of the token positions are selected for possible replacement during BERT pre-training.[10] The final hidden vector for the input token is then used to predict the original token with cross entropy loss. The NSP objective aims at improving performance on downstream tasks that require reasoning about the relationships between pairs of sentences (e.g., Natural Language Inference (NLI) or Question Answering). Positive examples (consecutive sentences in the text corpus) and negative examples (pairs of segments from different documents) are sampled with equal probability.

Two English models, $\textbf{BERT}_{\textbf{BASE}}$ and $\textbf{BERT}_{\textbf{LARGE}}$, have been trained on the BooksCorpus (800M words) (Zhu et al., 2015) and the English Wikipedia (2,500M words). The two models differ in terms of number of layers (i.e., Transformer blocks) (L=12 vs. L=24), hidden size (H=768 vs. H=1024), number of self-attention heads (A=12 vs. A=16) and total number of parameters (11M and 340M, respectively). Both models are trained with a specific kind of tokenization where some words are split into smaller units called WordPieces (Wu et al., 2016).[11] The first layer of BERT receives as input a combination of token, segment, and positional embeddings, as shown in Figure 2.10. The segment embedding shows which sentence (A or B) a token belongs to, while the position embedding shows the position of the token inside the sequence. It is also typical of BERT that the first token of every sequence is always a special classification token ([CLS]). Sentence pairs are grouped into a single sequence and separated with a special token ([SEP]). The final hidden state corresponding to the CLS token is used as the aggregate sequence representation. BERT can be fine-tuned for different tasks by simply adding

history (t_1,..., t_{k-1}). A backward LM is similar to a forward LM but runs over the sequence in reverse, predicting the previous token given the future context (Peters et al., 2018).

[9] Portion of Figure 3 in the Devlin et al. (2019) paper.

[10] When a position is chosen, the corresponding token is replaced with the special [MASK] token (80% of the time), with a random token (10% of the time), or is left unchanged (10% of the time).

[11] A 30,000 WordPiece vocabulary is generated from a training corpus by minimising the number of word splits done. This results in dedicated vocabulary units for the most common words in the corpus, while less frequent words are split into multiple wordpieces.

a classification or regression head on top of the CLS token.

BERT has set a new SOTA in numerous NLP tasks. Also, BERT models exist in different languages (e.g., (Martin et al., 2020; Le et al., 2020; Cañete et al., 2020; Koutsikakis et al., 2020; Virtanen et al., 2019)). Multilingual versions of the model also exist, covering up to 104 languages. Given that the training of BERT models is expensive, most research works use the pre-trained models. A downside of this is that it makes it hard for researchers to draw conclusions and interpret the model results by reference to the way the model was trained, since it is unclear what data the model was exposed to during training.

Among the variants of BERT that have been proposed, we find lighter models with significantly fewer parameters than BERT (e.g., DistilBert (Sanh et al., 2020) and ALBERT (A Lite BERT) (Lan et al., 2020)) but yield comparable or improved performance on most downstream tasks. Other BERT variants integrate different training procedures and objectives (RoBERTa (Liu et al., 2019c), SpanBERT (Joshi et al., 2020), AMBERT (Zhang et al., 2021)).

Although contextualised representations encode the meaning of individual instances, methods that inject sense information into them have also been proposed. In the SenseBERT model (Levine et al., 2020), this is done using an auxiliary masked word sense prediction task, alongside the usual training tasks of the contextual language model. The model that predicts the missing words' sense is trained jointly with the standard word-form level language model using information from WordNet as weak supervision for self-supervised learning: the masked word's supersenses form a set of possible labels for the sense prediction task.[12]

2.8 Conclusion

In this chapter, I described the evolution of word representations from classical distributional models to recent contextual language models. I included a thorough discussion of the meaning conflation problem which characterises both distributional approaches and static word embedding methods, and is important for lexical semantics analysis. I explained the mechanisms that have been proposed for representing specific word instances, which might be aimed at capturing compositionality or at representing the contextual meaning of word occurrences by aggregating over the vectors of their co-occurrences. I specifically described different methods that have been proposed for representing word senses by relying on automatically generated context clusters, on translations and on sense descriptions in external lexicons. I also explained the architecture of contextual language models which are currently the prevalent way for modelling lexical meaning. I clarified the differences between a bidirectional LSTM model (ELMo) and a Transformer-based model (BERT) in order to highlight what contributes to the superiority of the latter, and to their wide adoption in the community.

The goal of this introductory chapter is to illustrate the evolution of word meaning representations, and the paradigm shift attested in the computational linguistics community in the past years. In the next chapter, I will present three sets of articles which explore questions related to lexical semantics from a different standpoint, using a different methodological approach.

[12] When a single supersense is available, the network is trained to predict it given the masked word's context. When multiple supersenses are available (e.g., *bass*: noun.food, noun.animal, noun.artifact, noun.person), the model is trained to predict any of these senses, leading to a simple soft-labelling scheme.

Chapter 3

Lexical Polysemy and Clusterability

3.1 Introduction

Traditional semantic processing tasks such as word sense disambiguation and word sense induction assume that the semantic space of a polysemous word (lemma) can be partitioned into senses (Navigli, 2009; Manandhar et al., 2010). This, however, is a much easier task for some lemmas than others. In more technical distributional terms, sense identification is the process that groups the instances of an ambiguous word into clusters describing different meanings (Schütze, 1998). Naturally, instances of words that express clearly different meanings are easier to tell apart than instances which describe inter-related senses. For example, the following instances of the noun *match* describe three clearly different senses of the word: SPORTS GAME, PAIRING or LIGHTING DEVICE.

You can live stream the match via an up to date device.

They are a perfect match.

How to light a match without the box.

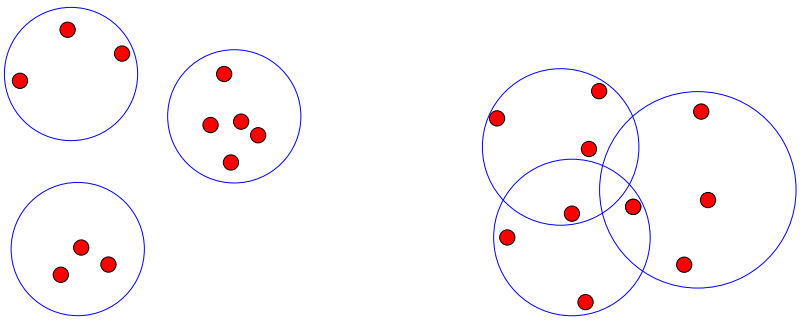
On the contrary, these instances of the noun *book*:

The book I bought the other day.
I really enjoyed reading this book.

describe closely related meanings: the CONTENT and OBJECT senses of the word.

In our McCarthy et al. (2016) paper (paper A(i) in Section 1.2), we propose to study the partitionability of lemmas into senses, that is how easy it is to decide whether different instances of a word express different meanings. Evidence about a "spectrum of partitionability" of words into senses already existed in the linguistics literature (Tuggy, 1993), however our paper was the first attempt to measure the phenomenon. To do so, we proposed to operationalize partitionability as **clusterability, a measure of how easy the occurrences of a lemma are to cluster**.

We specifically approached clusterability using metrics from the machine learning literature which explore the general clusterability of a data set (Ackerman and Ben-David, 2009), contrary to clustering quality metrics which instead measure the goodness of particular clusterings. A data set is considered to be more clusterable if the partitions of the data points it contains are easier to make. Figure 3.1(a) illustrates a clusterable data set, while Figure 3.1(b) illustrates a data set where the partitions are not easy to make since the data points are grouped into overlapping clusters. In our study, the data points corresponded to individual instances of a word, which can be grouped into more or less distinct clusters



(a) More clusterable data (b) Less clusterable data

Figure 3.1: A more clusterable dataset compared to a less clusterable one.

depending on how similar the meanings they express are. In the case of a highly clusterable word (e.g., *match*) different instances describe distinct meanings and can be grouped into interpretable clusters; on the contrary, in the case of a word with low clusterability (e.g., *book* or *paper*), different instances are expected to describe more or less inter-related meanings.

3.2 Annotation-based Word Clusterability Estimation

An important step in our methodology is to measure the similarity of different word instances. This is also a common step in Word Sense Induction (WSI) methods which aim at identifying the senses of words from raw data (Manandhar et al., 2010; Jurgens et al., 2012).[1] We adopt a graph-based approach to WSI. Traditional graph-based algorithms reveal the senses of a word by partitioning a co-occurrence graph built from its contexts into vertex sets grouping semantically related co-occurrences (Véronis, 2004; Di Marco and Navigli, 2013). In our experiments, we instead rely on substitute and translation annotations of word instances as a proxy for their meaning in context. We use data from the SemEval 2007 Lexical Substitution (LEXSUB) dataset (McCarthy and Navigli, 2007) and the SemEval 2010 Cross-lingual Lexical Substitution (CLLS) dataset (Mihalcea et al., 2010). In our setting, the similarity of individual word instances is reflected in, and can be measured through, the similarity of their annotations. Example annotations for instances of the adjective *clear* are given in Table 3.2. The first two instances have similar meaning, as shown by their shared annotations (*obvious*, *claro*, *obvio*, *evidente*), which differs from the meaning expressed by the last instance (with substitutes *clean*, *fresh* and *pure*).

In our paper, we propose two approaches for measuring the clusterability of a word:

- The first approach considers the distances of the data points that correspond to the instances of the word. We implement this approach using clusterability metrics from the machine learning literature which aim to measure the goodness of optimal k-means clusterings (Ackerman and Ben-David, 2009). We call these **intra-clustering** (intra-clust) metrics;
- The second approach considers the similarity of the clusterings obtained based on the substitute and translation annotations. It relies on the idea that if a lemma is more clusterable, two clusterings based on two different "views" of the same data points will be more congruent. We call this our **inter-clustering** (inter-clust) method.

For both methods, we first need to obtain a clustering of the instances for each word.

[1] Such methods are also commonly described as word sense identification or discrimination methods.

Instances	Substitutes	Translations
In this regard , neither ink appears to have a clear advantage over the other.	distinct (3), obvious (2), unambiguous (1), definite (1)	claro (2), obvio (2), evidente (1), perceptible (1)
I made that point perfectly clear to those of you who I discussed that with on Thursday.	obvious (3), understandable (1), distinctly (1), evident (1), unambiguous (1)	claro (4), evidente (2), obvio (2), sin lugar a duda (1), conciso (1), preciso (1)
After opening a few files in the sample project the concept of projects and file usage is clear.	obvious (3), plain (1), unmistakable (1), evident (1), apparent (1)	claro (4), entendido (1), esclarecido (1), comprendido (1), obvio (1)
The clear cool breeze is on your face first thing in the morning , and the children have new energy and interest in the season.	clean (3), fresh (2), pure (1)	claro (4), entendido (1), esclarecido (1), comprendido (1), obvio (1)

Figure 3.2: Instances of the adjective *clear* from the LEXSUB and CLLS datasets with more or less similar senses, as reflected in their English substitutes and Spanish translations.

3.2.1 An Intra-Clustering Clusterability Approach

Clusterability Metrics. We follow Ackerman and Ben-David (2009) and measure the general clusterability of a data set. A data set is considered to be more clusterable if the partitions of the data points it contains are easier to make. The notions of clusterability considered are all based on the k-means algorithm, and involve optimum clusterings for a fixed k. Let X be a set of data points, then a k-means k-clustering of X is a partitioning of X into k sets. We consider three clusterability measures:

- **Variance ratio** (VR): The intuition underlying VR is that in a good clustering, points should be close to the centroid of their cluster, and clusters should be far apart (Zhang, 2001).
- **Worst pair ratio** (WPR): This metric relies on a similar intuition as VR, in that it, too, considers a ratio of a within-cluster measure and a between-cluster measure. A difference lies in the focus on "worst pairs" (Epter et al., 1999), i.e. the closest pair of points that are in different clusters, and the most distant points that are in the same cluster.
- **Separability** (SEP): SEP measures the improvement in clustering (in terms of the k-means loss function) when we move from $(k-1)$ clusters to k clusters (Ostrovsky et al., 2006).

For VR and WPR, higher values indicate better clusterability, but the opposite is true for SEP. Lower separability values signal a larger drop in k-means loss when moving from $(k-1)$ to k clusters. We use an external method to determine k. We approximate k-means optimality by performing many clusterings of the same data set with different random starting points, and using the clustering with minimal k-means loss L.

Instance Similarity for k-Means Clustering. In order to measure the similarity between the instances $\mathbb{I}$ of a lemma l, we turn the substitute annotations (S) for an instance $i \in \mathbb{I}$ into a vector (v_{Si}). Each possible substitute $s \in S$ for l over all its ten LEXSUB instances becomes a dimension d in v_{Si}, with as value the number of annotators who proposed it. The resulting vector has as many dimensions as the number of annotations ($|S|$) across all ten instances of l. Dimensions that correspond to substitutes which were not proposed for an instance get a value of zero. For example, the vector for the first example given for *clear* in Table 3.2 has an entry of 3 in the dimension *distinct*, an entry of 2 in the dimension *obvious*, a value of 1 in the dimensions *unambiguous* and *definite*, and zero in all other dimensions (*unmistakable*, *evident*, *fresh*, *pure*, etc). The translations (T) for an instance in the CLLS data are turned into a vector in the same way. This procedure results in vectors of the same dimensionality

for all instances of the same lemma. The distance (d_{vec}) between two instances i, i' of a lemma ℓ is calculated as the Euclidean distance between their vectors.

Number of Clusters for k-means. The number of clusters (k) needed by the clustering algorithm is determined using a simple graph-based approach which groups instances with a minimum number of shared substitutes. We build two undirected graphs for each lemma in LEXSUB and CLLS. The two datasets contain the same word instances as Usim (Erk et al., 2009, 2013), which serves to evaluate the clusterability metrics. For a given lemma l, each instance $i \in I$ is represented by a vertex in the graph and is associated with its set of substitutes (S) or translations (T). Two vertices (instances) are linked by an edge if their distance is found to be low enough, as defined by a similarity measure which considers their unique and shared annotations (Goldberg et al., 2010). Practically, the distance (d_{node}) between two instances (nodes) i and i' with substitute sets S and S' corresponds to the number of moves necessary to convert S into S'. We use the metric proposed by Goldberg et al. (2010) which considers the elements that are shared by, and are unique to, each of the sets.

$$d_{node}(S, S') = |S| + |S'| - 2|S \cap S'| \tag{3.1}$$

The graph built for l is partitioned into non-overlapping clusters (connected components) which describe the senses of that lemma. Two instances belong to the same component if there is a path between their vertices.

3.2.2 An Inter-Clustering Clusterability Approach

The basic assumption behind this method is that if a lemma l is highly clusterable, then two clusterings based on different "views" of the same data points (instances) should be relatively similar. We derive two clusterings of the same set of instances from the graphs constructed as explained in Section 3.2.1 using substitutes and translations, which we consider as two different ways (or views) to describe the meaning of a lemma in the Usim dataset (cf. Table 3.2). We compare the two clustering solutions using measures from the SemEval 2010 Word Sense Induction task (Manandhar et al., 2010):

- **V-measure** (V) is the harmonic mean of homogeneity and completeness (Rosenberg and Hirschberg, 2007). Homogeneity refers to the degree that each cluster consists of data points primarily belonging to a single gold-standard class, while completeness refers to the degree that each gold-standard class consists of data points primarily assigned to a single cluster. V depends on both entropy and number of clusters, since systems that provide more clusters do better.
- **Paired F score** (pF) is the harmonic mean of precision and recall (Artiles et al., 2009). Precision is the number of common instance pairs between clustering solution and gold-standard classes, divided by the number of pairs in the clustering solution. Recall is the same numerator but divided by the total number of pairs in the gold-standard. pF penalizes a difference in number of clusters to the gold-standard in either direction.

Since both measures use the harmonic mean, we can alternatively use CLLS or LEXSUB as gold standard. The harmonic mean of homogeneity and completeness, or precision and recall, is the same regardless which clustering solution is considered as 'gold'.

Gold partitionability estimates		Clusterability measures	
Umid:	↘	VR:	↗
Uiaa:	↗	WPR:	↗
		SEP:	↘
		V:	↗
		pF:	↗
		nc_s:	↘

Table 3.1: Directions of partitionability estimates and clusterability measures: ↗ means that high values denote high partitionability, and ↘ means that high values denote low partitionability.

3.2.3 Overlap-based Baseline Method

Word sense induction has often involved a hard partitioning of usages into senses, but it can also be viewed in a graded (or soft clustering) fashion (Erk et al., 2009, 2013; Jurgens and Klapaftis, 2013). As opposed to our inter-clustering and intra-clustering methods which are applicable to hard clusterings, our baseline method relies on a simple criterion for estimating clusterability, the overlap that exists between clusters. The idea for proposing this baseline is that if the amount of cluster overlap indicates clusterability, then soft clustering would be a simple and useful tool for identifying lemmas with clear cut senses. These would have little or no overlap between clusters, as depicted in Figure 3.1(b). The clusters in this case correspond to a soft grouping solution of the nodes in the graphs described in Section 3.2.1. A cluster (called CLIQUE) consists of a maximal set of nodes that are pairwise adjacent. These are typically finer-grained than the partitions in the hard clustering solution, since there might be vertices in a component that have a path between them without being adjacent.

3.2.4 Gold Standard Partitionability

We compare the clusterability ratings obtained with the measures described in the previous sections to two gold standard partitionability estimates, derived from the usage similarity (Usim) dataset (Erk et al., 2009, 2013). Usim contains ten instances for 56 target words (nouns, adjectives, verbs and adverbs). Word instances are manually annotated with pairwise graded similarity scores on a scale from 1 (completely different) to 5 (same meaning). Each sentence pair was rated by multiple annotators and the average judgement for a pair was retained. In order to use the Usim data as a reference, we model partitionability as:

(a) **inter-tagger agreement** (**Uiaa**), i.e. the average pairwise Spearman's correlation between the ranked judgements of the annotators for a lemma;

(b) **proportion of mid-range judgements** over all instances for a lemma and all annotators (**Umid**). Mid-range judgements are between 2 and 4, i.e. not 1 (completely different usages) or 5 (the same usage).

3.2.5 Evaluation

The partitionability estimates and the clusterability measures vary in their directions. In some cases, high values denote high partitionability (WPR and VR); in other cases (SEP), high values indicate low partitionability. We expect WPR and VR to positively correlate with Uiaa and negatively with Umid, and the direction of correlation to be reversed for SEP.

Our clustering evaluation metrics (V and pF) provide correlations with the gold standards in the same

direction as WPR and VR. High congruence between the two solutions obtained for a lemma from different annotations of the same sentences should be indicative of higher clusterability and, consequently, higher values of Uiaa and lower values of Umid. As regards the overlap-based baseline approach, since we assume that lemmas that are harder to partition will have higher values of nc_s, high values of nc_s should be positively correlated with Umid and negatively correlated with Uiaa (like SEP). Table 3.1 gives an overview of the expected directions.

We measure the Spearman's ρ correlation between a ranking of gold partitionability estimates and a ranking produced by clusterability predictions. Given that polysemy can influence clusterability, we control for polysemy by grouping lemmas into polysemy bands depending on their number of senses, and measure the correlation for lemmas in the same band. The number of senses (k) for lemma l corresponds to the number of its clusters, i.e. the number of hard clusters (components) for VR, WPR and SEP), and the number of soft clusters (cliques) for the baseline. For the cluster congruence metrics (V and pF), k is the average number of clusters for l in LEXSUB and CLLS. Three polysemy bands are defined:[2]

- low: $2 \leq k < 4.3$
- mid: $4.3 \leq k < 6.6$
- high: $6.6 \leq k < 9$

We hereby present the correlation experiments which demonstrate the usefulness of the proposed clusterability metrics. A second set of experiments is presented in the paper which performs linear regression to link partitionability to clusterability, using the degree of polysemy k as an additional independent variable.

We calculated Spearman's correlation coefficient (ρ) for the two gold-standards (Uiaa and Umid) and all our clusterability measures: intra-clust (VR, WPR and SEP), inter-clust (V and pF) and the baseline nc_s. For all these measures (except for inter-clust), we calculate ρ using LEXSUB and CLLS separately as our clusterability measure input. For the inter-clust measures, LEXSUB and CLLS are the two views of the data. We calculate the correlation for lemmas in the polysemy bands (low, mid and high) where there are at least five lemmas. Table 3.2 shows the average Spearman's ρ over all trials for each clusterability measure.

We observe that all average ρ scores are in the anticipated direction, specified in Table 3.1; SEP and nc_s are positively correlated with Umid and negatively with Uiaa, whereas for all other measures the direction of correlation is reversed. Some of the metrics show a promising level of correlation. WPR, on the contrary, is quite weak possibly because it only considers the worst pair rather than all data points. The baseline (nc_s) is also particularly weak, suggesting that the amount of overlap is not a strong indication of clusterability. The inter-clust measures (pF and V) have a stronger correlation with Uiaa, whereas the machine learning measures are more strongly correlated with Umid. Results for many individual trials do not give significant results, as shown in the two last columns of the table, because controlling for polysemy leaves less data (lemmas) for each correlation. However, all significant correlations are in the anticipated direction.

From the machine learning metrics, VR (with a higher proportion of successful trials) and SEP (with the highest average correlations) are most consistent in indicating partitionability. While there are some success trials for the inter-clust approaches, the results are not consistent and only one trial showed a highly significant correlation. Similarly, the overlap-based baseline approach has only one significant

[2]In cases where the number of COMPS is one, the lemma is excluded from analysis, since the clustering algorithm itself decides that the instances are not easy to partition.

measure type	measure	average ρ Umid	average ρ Uiaa	prop. $\rho > 0.4$* or ** Umid	prop. $\rho > 0.4$* or ** Uiaa
intra-clust	VR	-0.4827	0.3651	2/3	2/3
	SEP	0.5694	-0.3895	2/3	1/3
	WPR	-0.3221	0.2097	1/3	0/3
inter-clust	*pF*	-0.3183	0.5398	0/2	1/2
	V	-0.1228	0.4925	0/2	0/2
baseline	nc_s	0.0525	-0.1641	0/6	1/6

Table 3.2: The macro-averaged correlation of each clusterability metric with the Usim gold-standard rankings Uiaa and Umid: All correlations are in the expected direction. We also give, the proportion (prop.) of trials with moderate or stronger correlation in the correct direction with a statistically significant result.

result, while in 4 out of 12 trials the correlation was in the non-anticipated direction. All other individual results were in the anticipated direction, except from one result for WPR in the non-anticipated direction, and one result for *V* on the fence. We made similar observations when controlling for polysemy, which could be an important confounder. SEP and VR were again the most promising metrics, while the inter-clust metrics (*V* and *pF*) and the baseline are overall not as consistent.

Clusterability metrics can serve to estimate the cost and effort needed in annotation projects, to determine the appropriate representation (clusters or per-instance vectors) for a lemma, and to identify words on which disambiguation efforts should be focused (for example, for query expansion). We have shown that clusterability metrics from machine learning are particularly relevant for lexical semantic analysis. A limitation of this study is that it relies on manually produced annotations. An obvious perspective for future work involves applying the measures to automatically generated word usage annotations, and to word embedding representations of the instances. This would allow us to measure clusterability over a larger vocabulary. In our recent work with Aina Garí Soler (my former PhD student in the MULTISEM project) which is presented in the next section (Garí Soler and Apidianaki, 2021a), we perform this type of clusterability analysis using contextual embeddings produced by neural language models (Devlin et al., 2019).

3.3 Lexical Semantic Analysis with Language Models

In the previous section, we showed that the theoretical notion of clusterability from machine learning (Ackerman and Ben-David, 2009) is relevant for lexical semantic analysis and can serve to reveal the lemmas' degree of partitionability into senses (Tuggy, 1993; Erk et al., 2013). We demonstrated how clustering of word instances based on in-context substitute and translation annotations can be used with clusterability measures to estimate how easily a word's usages can be partitioned into discrete senses, and have operationalized clusterability as consistency in clustering across these two information sources.

In more recent work (Garí Soler and Apidianaki, 2021a), we also address the questions of polysemy and sense partitionability, but this time using language model (LM) representations. In a series of experiments conducted in English, French, Spanish and Greek, we show that LMs encode rich knowledge about lexical polysemy which serves to tell polysemous from monosemous words, and to rank them according to their polysemy level. Additionally, we demonstrate that by leveraging LM representations, it is possible to scale up clusterability estimation to an open vocabulary. This overcomes the limitations

inherent to the use of manual annotations in the McCarthy et al. (2016) study, which constrained clusterability estimation to words in the LEXSUB and CLLS annotated data sets. Clusterability experiments are conducted in English due to the lack of evaluation data in the other three languages.

3.3.1 Probing for Semantic Information

The success of pre-trained LMs in numerous natural language understanding tasks (Devlin et al., 2019; Peters et al., 2018) has motivated a large number of studies exploring what these models actually learn about language (Voita et al., 2019a; Clark et al., 2019; Voita et al., 2019b; Tenney et al., 2019a). The bulk of this interpretation work relies on probing tasks which serve to predict linguistic properties from the representations generated by the models (Linzen, 2018; Rogers et al., 2020). The focus was initially put on structural linguistic aspects pertaining to grammar and syntax (Linzen et al., 2016; Hewitt and Manning, 2019; Hewitt and Liang, 2019). The first probing tasks addressing semantic knowledge explored phenomena in the syntax-semantics interface, such as semantic role labelling and coreference (Tenney et al., 2019a; Kovaleva et al., 2019), and the symbolic reasoning potential of LM representations (Talmor et al., 2019).

Language model representations have also been probed for lexical meaning. First, it was shown that they can successfully leverage sense annotated data (from Wikipedia and the SemCor corpus (Miller et al., 1993)) for disambiguation. Wiedemann et al. (2019) and Reif et al. (2019) specifically show that BERT can organise word usages in the semantic space in a way that reflects the meaning distinctions present in the data. These works address the disambiguation capabilities of the model but do not show what BERT actually knows about words' polysemy. In an exploration of word meaning representation in context, Aina et al. (2019) explore the interplay between word type and token-level information in the hidden representations of LSTM LMs. They probe the hidden representations of a bidirectional (bi-LSTM) LM for lexical (type-level) and contextual (token-level) information. They specifically train diagnostic classifiers on the tasks of retrieving the input embedding of a word (Adi et al., 2017; Conneau et al., 2018), and a representation of its contextual meaning as reflected in its lexical substitutes. The results show that the information about the input word that is present in LSTM representations is not lost after contextualisation.

The work of Ethayarajh (2019a) explores the similarity estimates that can be drawn from contextualised representations without directly addressing word meaning. This study provides valuable observations regarding the impact of context on the representations, without explicitly addressing the semantic knowledge encoded by the models. Through an exploration of BERT, ELMo and GPT-2 (Radford et al., 2019), the author highlights the highly distorted similarity of the obtained contextualised representations which is due to the anisotropy of the vector space built by each model. This issue affects all tested models and is particularly present in the last layers of GPT-2, resulting in highly similar representations even for random words. Although addressing word representation in context, this work does not address the question of meaning, making it hard to draw any conclusions about lexical polysemy.

Vulić et al. (2020b) probe BERT representations for lexical semantics, but they do so using "static" word embeddings derived from contextualised representations. These are obtained through pooling over several contexts, or by extracting representations for words in isolation and from BERT's embedding layer before contextualisation. Naturally, these representations are evaluated on tasks traditionally used for assessing the quality of static embeddings, such as out-of-context word similarity and analogy (Drozd et al., 2016; Hill et al., 2015; Vulić et al., 2020a) and the bilingual lexicon induction task (Artetxe et al., 2020), which are not well-suited for addressing lexical polysemy.

Our proposed experimental setup is aimed at investigating the information about polysemy that is encoded in the representations built at different layers of deep pre-trained LMs. This question is also addressed by Pimentel et al. (2020), who adopt an information theoretic perspective to measuring lexical ambiguity using BERT embeddings. The assumption underlying their method is that the contexts in which a word appears are systematically adapted to enable disambiguation. Consequently, the lexical ambiguity of a word should negatively correlate with its contextual uncertainty. Work by Xypolopoulos et al. (2021) explores ELMo's (Peters et al., 2018) knowledge about lexical polysemy based on the geometry of the space built for different instances of words. Their approach builds multiresolution grids in the contextual embedding space based on the assumption that the volume covered by the cloud of points corresponding to different instances of a word is representative of its polysemy. Specifically, they construct a hierarchical discretization of the space where, at each level, the same number of bins are drawn along each dimension. Each level corresponds to a different resolution. The polysemy score for a word is based on the volume (i.e. the proportion of bins) covered by its vectors at each level.[3]

Our approach basically relies on the similarity of contextualised representations, which amounts to word usage similarity estimation as in the paper of Ethayarajh (2019a). This is a classical task in lexical semantics which precisely involves predicting the similarity of word instances in context without use of sense annotations (Erk et al., 2009; Huang et al., 2012; Erk et al., 2013). BERT has been shown to be particularly good at this task (Garí Soler et al., 2019a; Pilehvar and Camacho-Collados, 2019). Our experiments allow to explore and understand what this ability is due to.

3.3.2 Lexical Polysemy Detection

Our approach to investigate the knowledge that LMs encode about lexical polysemy and sense partitionability follows a rigorous experimental protocol proper to lexical semantic analysis, which involves the use of datasets carefully designed to reflect different sense distributions. This allows us to investigate the knowledge models acquire during training, and the influence of context variation on token representations. Our investigation encompasses monolingual models in different languages (English, French, Spanish and Greek) (Le et al., 2020; Cañete et al., 2020; Koutsikakis et al., 2020) and the language specific parts of multilingual BERT (Devlin et al., 2019). We demonstrate that contextualised representations generated by these models encode an impressive amount of knowledge about polysemy, and are able to distinguish monosemous (`mono`) from polysemous (`poly`) words in a variety of settings and configurations. Additionally, we show that BERT representations can serve to determine how easy it is to partition a word's semantic space into senses. They thus provide a way to scale up the McCarthy et al. (2016) study to an open vocabulary and to new languages where contextual language models are available.

Mono and Poly Sentence Pools. We build the English dataset for our experiments using SemCor 3.0 (Miller et al., 1993), a corpus manually annotated with WordNet senses (Fellbaum, 1998). It is important to note that the annotations present in the corpus do not serve for training or evaluating any of the models. They, instead, only serve to control the composition of the sentence pools that are used for generating contextualised representations, and to analyse the results. We form sentence pools for monosemous (`mono`) and polysemous (`poly`) words that occur at least ten times in SemCor. For each

[3]The binning strategy is deemed preferable to a clustering-based approach because of the non-uniform distribution of word representations in the embedding space (Ethayarajh, 2019a). According to the authors, clustering would assign vectors lying in the same dense area of the space to the same cluster, and outliers lying in the same but sparser area to many different small clusters. This would make the number of clusters an unreliable indicator of the space a word covers.

mono word, we randomly sample ten of its instances in the corpus. For each poly word, we form three sentence pools of size ten reflecting different sense distributions:

- **Balanced** (poly-bal). We sample a sentence for each sense of the word in SemCor until a pool of ten sentences is formed. This pool contains a balanced distribution of the word's senses.
- **Random** (poly-rand). We randomly sample ten poly word instances from SemCor. We expect this pool to be highly biased towards a specific sense due to the skewed frequency distribution of word senses (Kilgarriff, 2004; McCarthy et al., 2004). This configuration is closer to the expected natural occurrence of senses in a corpus, it thus serves to estimate the behaviour of the models in a real-world setting.
- **Same sense** (poly-same). We sample ten sentences illustrating only one sense of the poly word. Although the composition of this pool is similar to that of the mono pool (i.e. all instances describe the same sense) we call it poly-same because it describes one sense of a polysemous word.[4] Specifically, we want to explore whether BERT representations derived from these instances can distinguish mono from poly words, even though there is no variation inside the respective pools.

The controlled composition of the poly sentence pools allows us to investigate the behaviour of the models when they are exposed to instances of polysemous words describing the same or different senses. There are 1,765 poly words in SemCor with at least 10 sentences available.[5] We randomly subsample 418 from these in order to balance the mono and poly classes. Our English dataset is composed of 836 mono and poly words, and their instances in 8,195 unique sentences. For French, Spanish and Greek, we retrieve sentences from the Eurosense corpus (Delli Bovi et al., 2017) which contains texts from the Europarl corpus automatically annotated with word senses from the multilingual semantic network BabelNet (Navigli and Ponzetto, 2012).[6] We extract sentences from the high precision version[7] of the corpus, and create sentence pools in the same way as in English, by balancing the number of monosemous and polysemous words (418). The number of senses for a word is defined as the number of its Babelnet senses that are mapped to a WordNet sense.

Contextualised Word Representations. In our Garí Soler and Apidianaki (2021a) paper, we experiment with representations generated by three English models: BERT (Devlin et al., 2019)[8], ELMo (Peters et al., 2018), and context2vec (Melamud et al., 2016). We present here the results obtained by the best-performing model, BERT. We use the bert-base-uncased and bert-base-cased models, pre-trained on the BooksCorpus (Zhu et al., 2015) and English Wikipedia.[9] For French, Spanish and Greek, we use BERT models specifically trained for each language: Flaubert (flaubert_base_uncased) (Le et al., 2020), BETO (bert-base-spanish-wwm-uncased) (Cañete et al., 2020), and Greek BERT (bert-base-greek-uncased-v1) (Koutsikakis et al., 2020). We also use the mBERT model (bert-base-multilingual-cased) for each of the four languages. mBERT was trained on Wikipedia data of 104 languages. All BERT models generate 768-*d* representations.

The Self-Similarity Metric. All models used in this study produce representations that describe word meaning in specific contexts of use. For each instance i of a target word w in a sentence, we extract its representation from each of the 12 layers of a BERT-type model. These models are trained with a

[4] The polysemous words are the same as in poly-bal and poly-rand.
[5] We use sentences of up to 100 words.
[6] BabelNet was built from lexicographic and encyclopedic resources such as WordNet and Wikipedia.
[7] Disambiguation in this version of Eurosense is more accurate than in the high coverage version of the corpus.
[8] We use Huggingface transformers (Wolf et al., 2019).
[9] The training and architecture of the BERT model are explained in detail in Section 2.7.

specific kind of tokenization where some words are split into smaller units called WordPieces (WPs) (Wu et al., 2016). When a word is split into multiple WPs, we obtain its representation by averaging the WPs.

Our tool for measuring the similarity of the representations generated for different word instances is self-similarity (*SelfSim*), one of the three measures of contextuality proposed by Ethayarajh (2019a) for measuring how contextualised a representation is.[10] They define *SelfSim* as follows: Let w be a word that appears in sentences $\{s_1, ..., s_n\}$ at indices $\{i_1, ..., i_n\}$ respectively, such that $w = s_1[i_1] = ... = s_n[i_n]$. Let $f_l(s,i)$ be a function that maps $s[i]$ to its representation in layer l of model f. The self similarity of w in layer l is given by Equation 3.2.

$$SelfSim_l(w) = \frac{1}{|n|^2 - |n|} \sum_j \sum_{\substack{k \in I \\ k \neq j}} cos(f_l(s_j, i_j), f_l(s_k, i_k)) \tag{3.2}$$

We calculate *SelfSim* for a word w in a sentence pool p and a layer l of a model by taking the average of the pairwise cosine similarities of the representations of its instances in l. We report the average *SelfSim* for all w's in a pool p. *SelfSim* is in the range [-1, 1]. **We expect the average *SelfSim* for monosemous words and words with low polysemy to be higher than that of highly polysemous words.** We also expect the `poly-same` pool, which contains instances of the same sense of a `poly` word, to have a higher average *SelfSim* than the other `poly` pools which contain instances of different senses.

Ethayarajh (2019a) has shown that contextualisation has a strong impact on *SelfSim* since it introduces variation in the token-level representations, making them more dissimilar. The *SelfSim* value for a word would be 1 with non-contextualised (or static) embeddings, as all its instances would be assigned the same vector. In contextual models, *SelfSim* is lower in layers where the impact of the context is stronger. It is, however, important to note that contextualisation in BERT models is not monotonic, as shown by previous studies of the models' internal workings (Voita et al., 2019a; Ethayarajh, 2019a). Our experiments provide additional evidence in this respect.

3.3.3 Results and Analysis

Figure 3.3 shows the average *SelfSim* obtained for each sentence pool (`mono`, `poly-same`, `poly-rand`, `poly-bal`) with representations produced by BERT models in the four languages of the study. The numbers on the x axis of the plots (1, ..., 12) correspond to the layer of the tested model. In the upper left plot, which shows results obtained with the English BERT model, the thin lines illustrate the average *SelfSim* score obtained using representations from the uncased English BERT model, while the thicker lines correspond to scores obtained with the cased model. **We observe a clear distinction of words according to their polysemy**: *SelfSim* is higher for `mono` than for `poly` words across all layers and sentence pools.

A highly important and clear distinction seen in the plots is the one between the `mono` and `poly-same` pools, which contain instances of only one sense. This distinction suggests that **BERT encodes information about a word's monosemous or polysemous nature regardless of the contexts where the word is seen**, and which serve to derive the representations. BERT produces less similar representations for word instances in the `poly-same` pool compared to `mono`, reflecting that `poly` words can have different meanings. We also observe **a clear ordering of the three `poly` sentence pools**: Average

[10] The other two are intra-sentence similarity and maximum explainable variance.

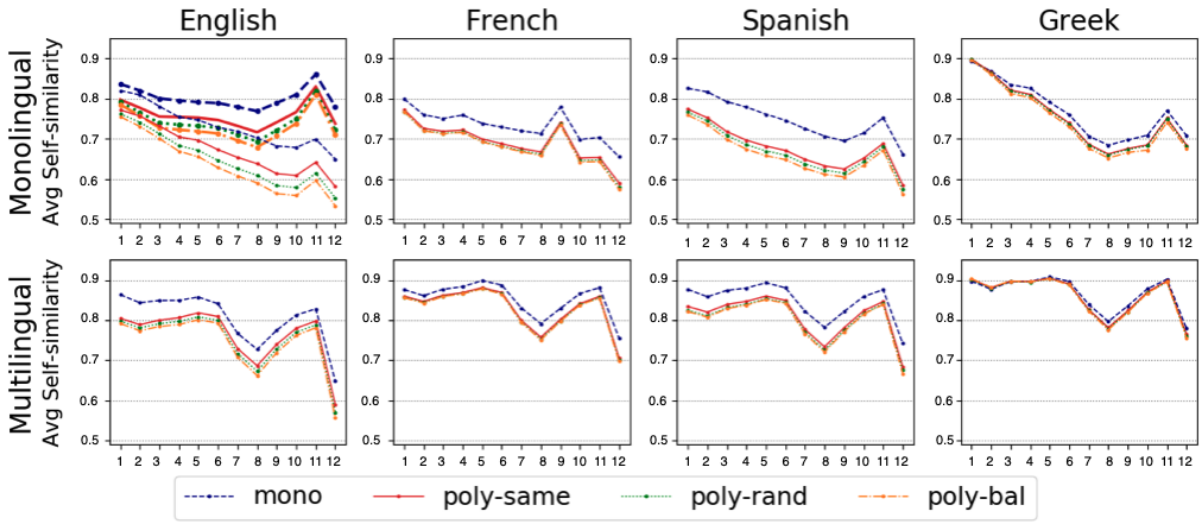


Figure 3.3: Average *SelfSim* obtained with monolingual BERT models (top row) and mBERT (bottom row) across all layers of the models (horizontal axis). In the first plot, thick lines correspond to the `cased` model.

SelfSim is higher in the `poly-same` pool, which only contains instances of one sense, followed by mid-range values in `poly-rand`, and gets its lowest values in the balanced setting (`poly-bal`). This is noteworthy given that `poly-rand` contains a mix of senses but with a stronger representation of *w*'s most frequent sense than in `poly-bal` (71% of the instances vs. 47% in `poly-bal`).[11]

Our results demonstrate that BERT representations encode two types of lexical semantic knowledge:

- Information about the polysemous nature of words acquired through pre-training, as reflected in the distinction between `mono` and `poly-same`;
- information from the particular instances of a word used to create the contextualised representations, as shown by the finer-grained distinctions between different `poly` settings (`poly-same`, `poly-bal`, `poly-rand`).

BERT's knowledge about polysemy can be due to differences in the types of context where words of different polysemy levels are used. We expect `poly` words to be seen in more varied contexts than `mono` words, reflecting their different senses. BERT encodes this variation with the LM objective through exposure to large amounts of data, and this is reflected in the representations. The same ordering pattern is observed with mBERT (lower part of Figure 3.3).

Using the `bert-base-cased` model leads to an overall increase in *SelfSim* and to smaller differences between bands, as shown by the thick lines in the upper left plot of Figure 3.3. Our explanation for the lower distinction ability of the cased model is that it encodes sparser information about words than the `uncased` model, since it was trained on a more diverse set of strings which included capitalised and non-capitalised forms. Surprisingly, in spite of that, the cased model has a smaller vocabulary size (29K WPs) than the uncased model (30.5K). Averaging the WPs of words might also have an impact on the resulting representations, since the cased model contains a higher number of WPs corresponding to word parts, compared to `bert-base-uncased` (6,478 vs. 5,829).

The decreasing trend in *SelfSim* observed for BERT in Figure 3.3, and the peak in layer 11, confirm the

[11] Numbers are macro-averages for words in the pools.

phases of context encoding and token reconstruction observed by Voita et al. (2019a).[12] In earlier layers, context variation makes representations more dissimilar and *SelfSim* decreases. In the last layers, information about the input token is recovered for LM prediction and similarity scores are boosted. Our results show clear distinctions across BERT layers. This suggests that lexical information is spread throughout the layers of the models, and contributes new evidence to the discussion on the localisation of semantic information (Rogers et al., 2020; Vulić et al., 2020b).

The average *SelfSim* scores obtained for words in our pools using monolingual French, Spanish and Greek models are shown in the top row of Figure 3.3. Flaubert, BETO and Greek BERT representations clearly distinguish monosemous from polysemous words, but average *SelfSim* values for different `poly` pools are much closer than in English. The results obtained with mBERT representations are shown in the lower part of the figure. The model assigns highly similar average *SelfSim* values to different `poly` pools, making distinction harder than with monolingual models. We test the statistical significance of the `mono`/`poly-rand` distinction.[13] The results show that differences are significant across all layers of the models in English ($\alpha = 0.01$). In the other languages, the difference between *SelfSim* values in `mono` and `poly-rand` is significant in all layers of the monolingual models, and with mBERT for Spanish and French. In mBERT for Greek, the difference is significant in ten layers, but the magnitude of the difference is smaller compared to the other models.

We also conduct a classification experiment where contextualised representations are probed to test their ability to guess whether a word is polysemous, and which `poly` band it belongs to. We use a binary logistic regression classifier for the `mono-poly` distinction, and a multi-class classifier for predicting the polysemy level (`poly-band`), using *SelfSim* as feature. Details about the experimental setup are given in the paper. The results show that BERT achieves good accuracy (higher than mBERT) in both the binary and multiclass settings. BERT embeddings can thus be used to determine whether a word has multiple meanings, and to provide a rough indication of its polysemy level. A simple frequency-based classifier performs on par with mBERT in the binary setting, highlighting that frequency information is highly relevant for the `mono-poly` distinction. Results in the other three languages are not as high as those obtained in English, but most models perform better than a frequency-based classifier.[14]

We believe that the lower quality results obtained with mBERT are partly due to the fact that the multilingual WP vocabulary is mostly English-driven, resulting in arbitrary partitionings of words in the other languages. This word splitting procedure must have an impact on the quality of the lexical information in mBERT representations. The nature and quantity of the training data must also have a strong impact on the quality of the embedding space built by each model. Nevertheless, since we are using the pre-trained models, it is not possible to draw conclusions and interpret the results based on the quality of the training data.

Anisotropy Analysis. In order to better understand the differences between monolingual and multilingual models, we analyse their anisotropy using the method proposed by Ethayarajh (2019a). We form pairs of random words,[15] calculate the cosine similarity between two random instances of the words in a pair, and take the average over all pairs. We call this score *RandSim*. The results are given

[12] They study the information flow in the Transformer estimating the MI between representations at different layers.

[13] We use unpaired two-samples t-tests when the normality assumption is met (as determined with Shapiro Wilk's tests). Otherwise, we run a Mann Whitney U test, the non-parametrical alternative of this t-test. In order to lower the probability of type I errors (false positives) that increases when performing multiple tests, we correct p-values using the Benjamini–Hochberg False Discovery Rate (FDR) adjustment (Benjamini and Hochberg, 1995).

[14] Only exceptions are Greek mBERT in the multi-class setting, and Flaubert in both settings.

[15] 2,183 in English and 1,318 in each of the other languages.

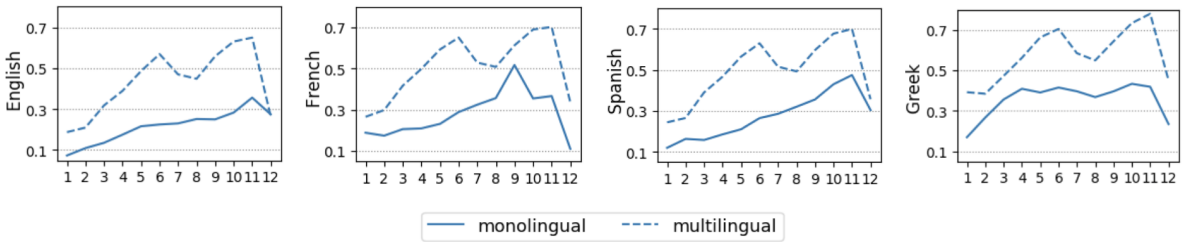


Figure 3.4: Similarity between random word instances in monolingual models and mBERT.

in Figure 3.4. We observe a clear difference in the scores obtained by monolingual models (solid lines) and mBERT (dashed lines). Clearly, mBERT assigns higher similarities to random words, an indication that its semantic space is more anisotropic than the one built by monolingual models. High anisotropy means that representations occupy a narrow cone in the vector space (Ethayarajh, 2019a). This also indicates that the vectors of random words are found close to each other which is not a desirable property of semantic spaces, since it results in lower quality similarity estimates and in the model's limited potential to establish clear semantic distinctions.

In a quality semantic space, we would also expect the similarity between same word instances (*SelfSim*) to be much higher than similarity of random words (*RandSim*). A second experiment described in the paper shows that the difference is smaller in the space built by mBERT, and becomes very low in the last layers of the model. This means that the multilingual model is less capable of distinguishing instances of the same and different words than monolingual models and, consequently, that it is more anisotropic than monolingual spaces.

3.3.4 Polysemy Level Prediction

We consider that the polysemy level of words can have an impact on the results. Distinguishing a monosemous word from a polysemous word with a high number of senses should be an easier task than distinguishing it from a polysemous word with fewer senses. In a second set of experiments, we explore the impact of words' degree of polysemy on the representations. We control for this factor by grouping words into three polysemy bands, as in our McCarthy et al. (2016) work, which correspond to a specific number of senses (k): `low`: $2 \leq k \leq 3$, `mid`: $4 \leq k \leq 6$, `high`: $k > 6$.[16]

In Figure 3.5, we compare `mono` words with lemmas in each polysemy band, in terms of their average *SelfSim*. Values for `mono` words are taken from Section 3.3.2. For `poly` words, we use representations from the `poly-rand` sentence pool which better approximates natural word occurrence in a corpus. For comparison, we report results obtained in English using sentences from the English `poly-same` and `poly-bal` pools in Figure 3.6. In English, the pattern is clear in all plots: *SelfSim* is higher for `mono` than for `poly` words in any band, confirming that BERT is able to distinguish `mono` from `poly` words at different polysemy levels. The range of *SelfSim* values for a band is inversely proportional to its k: Words in `low` get higher values than words in `high`. The results denote that the meaning of highly polysemous words is more variable (lower *SelfSim*) than the meaning of words with fewer senses. As expected, scores are higher and inter-band similarities are closer in `poly-same` (cf. Figure 3.6 (b)) compared to `poly-bal` and `poly-rand`, where distinctions are clearer. These results confirm that BERT can predict the polysemy level of words, even from instances describing the same sense.

[16] In English, the bands contain 551, 663 and 551 words, respectively. In the other languages, they contain 300 words each.

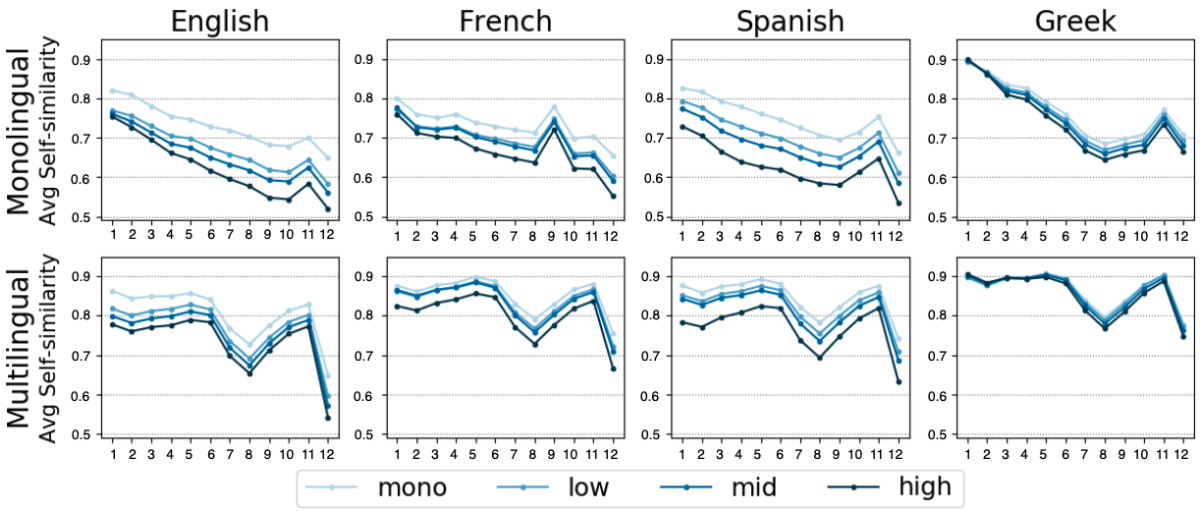


Figure 3.5: Polysemy level prediction results. We show the average *SelfSim* obtained with monolingual BERT models (top row) and mBERT (bottom row) for `mono` and `poly` lemmas in different polysemy bands. Representations are derived from sentences in the `poly-rand` pool.

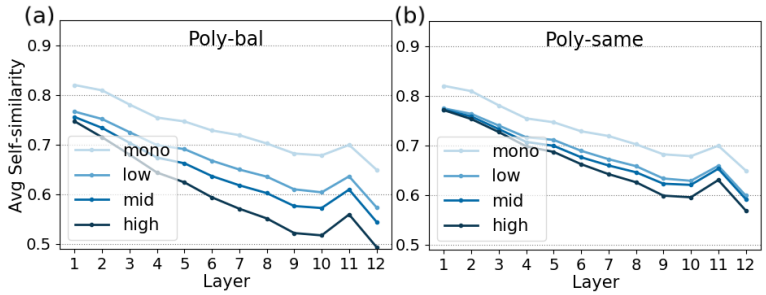


Figure 3.6: Comparison of BERT average *SelfSim* for `mono` and `poly` lemmas in different polysemy bands in the `poly-same` and `poly-bal` sentence pools.

The inter-band differences detected in `poly-rand` are significant in all but the first layer of BERT.[17] In the other languages, the bands are also correctly ranked but with smaller inter-band differences than in English.[18] This variation across languages can be explained to some extent by the quality of the automatic EuroSense annotations, which has a direct impact on the quality of the sentence pools.[19] The mBERT model makes less clear distinctions than the monolingual models, as shown in the lower row of Figure 3.5. Still, inter-band differences are significant in most layers of mBERT and the monolingual French, Spanish and Greek models.[20]

Controlling for frequency and part-of-speech category. Given the strong correlation between word frequency and number of senses (Zipf, 1945), we explore how frequency influences the obtained results. We use frequency information from Google Ngrams (Brants and Franz, 2006) for English, and gather frequency counts from the OSCAR corpus (Suárez et al., 2019) for the other languages. In an initial experiment, we examine whether BERT can rely on *SelfSim* to distinguish words by frequency. We find that *SelfSim* can indeed serve to produce a clear ranking, with less frequent words having higher values

[17] We use the same approach as in Section 3.3.2.

[18] In Greek, clear distinctions are only made in a few middle layers.

[19] The WSD precision is ten points higher in English (81.5) and Spanish (82.5) than in French (71.8) (Delli Bovi et al., 2017). The Greek portion has not been evaluated.

[20] With the exception of `mono`→`low` in mBERT for Greek and `low`→`mid` in Flaubert and mBERT for French.

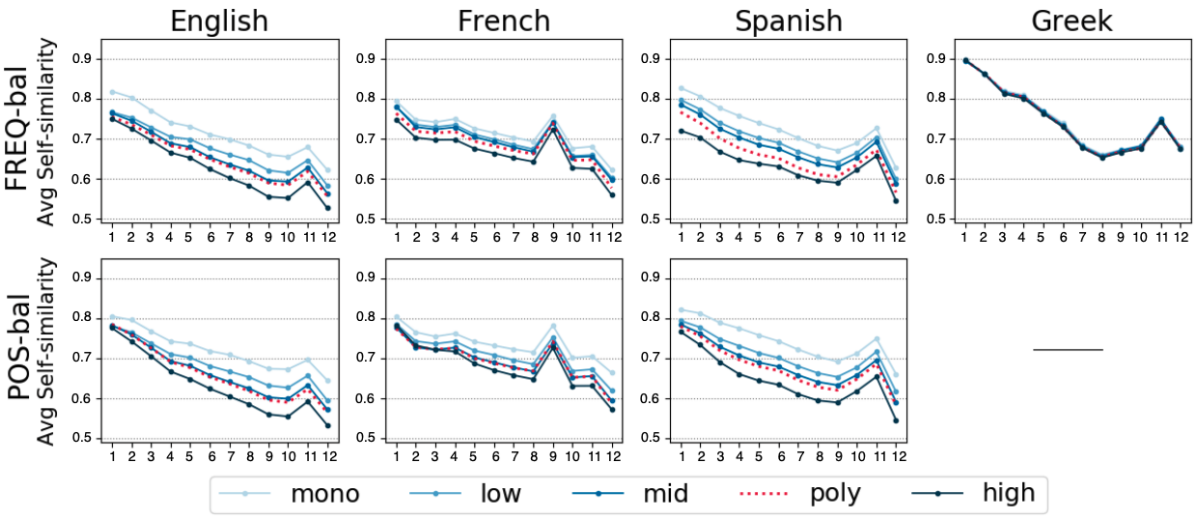


Figure 3.7: Average *SelfSim* inside the `poly` bands balanced for frequency (FREQ-bal) and part of speech (POS-bal). *SelfSim* is calculated using representations generated by monolingual BERT models from sentences in each language-specific pool. We do not balance the Greek dataset for PoS because it only contains nouns.

than more frequent ones. The words can also be distinguished by part of speech. Verbs, which are often highly polysemous,[21] have the lowest *SelfSim* values. The same trend is observed for monolingual models in the other three languages. More details about these two experiments are given in the paper.

In order to test whether grammatical category and word frequency are to some extent responsible for the good polysemy detection results, we create two new settings (called POS-bal and FREQ-bal) where we control for the composition of the `poly` bands in terms of these two factors. We examine the average *SelfSim* values obtained for words in each band in `poly-rand`. Figure 3.7 shows the results for monolingual models. We observe a similar ordering of polysemy bands as in Figure 3.5, although inter-band distinctions become less clear. These results indicate that BERT's polysemy predictions do not rely on frequency or part of speech. Statistical tests show that all inter-band distinctions are still significant in most layers of the BERT model. For French and Spanish, all distinctions in POS-bal are significant in at least one layer of the models. The same applies to the `mono`→`poly` distinction in FREQ-bal but finer-grained distinctions get lost. Greek BERT cannot establish correct inter-band distinctions when the influence of frequency is neutralised in the FREQ-bal setting.

3.4 Clusterability Estimation with Language Model Representations

3.4.1 Measuring Sense Partitionability with Contextualised Vectors

We have shown that representations from pre-trained LMs can serve to predict words' degree of polysemy. In this section, we explore whether they can also point to the clusterability of polysemous words. This study extends our McCarthy et al. (2016) work by replacing the manual annotations with contextualised representations. The need for manual annotations constrained the initial study to a small set of words found in the Usim, LEXSUB and CLLS datasets. The use of contextualised embeddings allows us to scale up the approach to an open vocabulary. This experiment is conducted in English due to the availability of evaluation data (Erk et al., 2013) and also for comparison with previous results. We specifically test the ability of contextualised representations to estimate how easily the instances

[21] Only 10.8% of monosemous words in our `mono` pool are verbs.

of a polysemous word can be grouped into interpretable clusters, Following McCarthy et al. (2016), we use the clusterability metrics proposed by Ackerman and Ben-David (2009) to measure the ease of clustering word instances into senses.

In order to make comparison with previous results possible, we run our experiments on the usage similarity (Usim) dataset (Erk et al., 2013). We represent target word instances in Usim in two ways: using **contextualised representations** generated by BERT, and using **automatically generated substitute annotations**. The substitute-based approach makes possible a direct comparison with our annotation-based method in the McCarthy et al. (2016) paper. In this initial experiment, we represented each instance i of a word w in Usim as a vector $\vec{i}$, where each substitute s assigned to w over all its instances ($i \in I$) becomes a dimension (d_s). For a given i, the value for each d_s is the number of annotators who proposed substitute s. d_s contains a zero entry if s was not proposed for i. We refer to this type of representation as `Gold-SUB`.

We generate our substitute-based representations with BERT using the simple "word similarity" approach in Zhou et al. (2019). For an instance i of word w in context C, we rank a set of candidate substitutes $S = \{s_1, s_2, ..., s_n\}$ based on the cosine similarity of the BERT representations for i and for each substitute $s_j \in S$ in the same context C.[22] As candidate substitutes, we use the paraphrases of w in the Paraphrase Database (PPDB) XXL package (Ganitkevitch et al., 2013; Pavlick et al., 2015).[23] For each instance i of w, we obtain a ranking R of all substitutes in S, and remove low-quality substitutes (i.e., noisy paraphrases or substitutes referring to a different sense of w) using a the filtering approach we proposed in Garí Soler et al. (2019a).[24] We build vectors as in McCarthy et al. (2016) using the cosine similarity assigned by BERT to each substitute as a value. We call this representation BERT-SUB.

Clusterability metrics. We use the two best performing metrics from McCarthy et al. (2016): Variance Ratio (VR) (Zhang, 2001) and Separability (SEP) (Ostrovsky et al., 2006). We also experiment with the Silhouette coefficient (SIL) (Rousseeuw, 1987) as a clusterability metric, as it can assess cluster validity. For VR and SIL, a higher value indicates higher clusterability. The inverse applies to SEP.

We calculate Spearman's ρ correlation between the results of each clusterability metric and the two gold standard measures from McCarthy et al. (2016): **Uiaa**, the inter-annotator agreement for a lemma in terms of average pairwise Spearman's correlation between annotators' judgements; and **Umid**, the proportion of mid-range judgements (between 2 and 4) assigned by annotators to all sentences of a target word. Higher Uiaa values indicate higher clusterability, meaning that sense partitions are clearer and easier to agree upon. Therefore, higher Umid values indicate lower clusterability, since it indicates how often usages do not have identical (5) or completely different (1) meaning.

Sense Clustering. In order to estimate the clusterability of a lemma l, we need to first cluster its instances in the data. For this, we use the k-means algorithm. We define the optimal number of clusters k for a lemma using the Silhouette coefficient (SIL).[25] For a data point i, SIL compares the intra-cluster distance (i.e., the average distance from i to every other data point in the same cluster) with the average distance of i to all points in its nearest cluster. The SIL value for a clustering is obtained by averaging SIL for all data points, and it ranges from -1 to 1. We cluster each type of representation for w using

[22] We use representations from the last layer of the model.

[23] We use PPDB to reduce variability in our substitute sets, compared to the ones that would be proposed by looking at the whole vocabulary. PPDB can be found at: `http://www.paraphrase.org`

[24] This filtering ensures that substitutes are both a good fit in the context and semantically related in the PPDB.

[25] We do not use McCarthy et al.'s graph-based approach because it is not compatible with all our representation types.

Gold	Metric	BERT-REP	BERT-SUB	Gold-SUB	BERT-AGG	Gold-AGG
Uiaa	SEP ↘	-0.48*$_{10}$	-0.03	-0.20	-0.48*$_{11}$	–
	VR ↗	0.17$_{12}$	0.09	0.34*	0.33*$_{12}$	–
	SIL ↗	0.61*$_{11}$	0.10	0.32*	**0.69***$_{10}$	**0.80***
Umid	SEP ↗	0.43*$_{9}$	0.05	0.16	0.43*$_{9}$	–
	VR ↘	-0.24$_{9}$	-0.15	-0.24	-0.32*$_{5}$	–
	SIL ↘	**-0.46***$_{10}$	-0.11	-0.38*	-0.44*$_{8}$	**-0.48***

Table 3.3: Spearman's ρ correlation between automatic metrics and gold standard clusterability estimates. The arrows indicate the expected direction of correlation for each metric. Subscripts indicate the BERT layer that achieved best performance.

k-means with a range of k values ($2 \leq k \leq 10$), and retain the k of the clustering with the highest mean SIL. Additionally, since BERT representations' cosine similarity correlates well with usage similarity (Garí Soler et al., 2019a), we experiment with Agglomerative Clustering with average linkage directly on the cosine distance matrix obtained with BERT representations (BERT-AGG). For comparison, we also use Agglomerative Clustering on the gold usage similarity scores from Usim, transformed into distances (Gold-AGG).

3.4.2 Evaluation

The results of the clusterability experiment are given in Table 3.3. We show the Spearman's ρ correlation between automatic metrics and gold standard clusterability estimates. Significant correlations (where the null hypothesis $\rho = 0$ is rejected with $\alpha < 0.05$) are marked with *.

Best results on the Uiaa evaluation are given by Agglomerative Clustering on the gold Usim similarity scores (Gold-AGG) in combination with the SIL clusterability metric ($\rho = 0.80$). This is unsurprising, since Umid and Uiaa are derived from the same Usim scores. From the automatically generated representations, the strongest corrselation with Uiaa (0.69) is obtained with BERT-AGG and the SIL clusterability metric. The SIL metric also works well with BERT-REP achieving the strongest correlation with Umid (-0.46). It constitutes, thus, a good alternative to the SEP and VR metrics used in previous studies.

An interesting point is that the correlations obtained using raw BERT representations are much higher than the ones observed with McCarthy et al. (2016)'s representations which relied on manual substitutes (Gold-SUB): These were in the range of 0.20-0.34 for Uiaa and 0.16-0.38 for Umid. These results demonstrate that BERT representations offer good estimates of the partitionability of words into senses. As expected, the substitution-based approach performs better with clean manual substitutes (Gold-SUB) than with automatically generated ones (BERT-SUB).

3.5 Conclusion

The studies presented in this chapter demonstrate how the clusterability of words can been estimated using meaning annotations and a graph-based methodology, but also – and more efficiently – by directly using the representations that are generated by contextual language models. The findings from this series of experiments demonstrate that contextualised BERT representations encode rich information about lexical polysemy. Our experimental results suggest that the knowledge that allows BERT to detect polysemy in different configurations is acquired during the pre-training phase. Our findings hold for monolingual BERT models in all four languages addressed in the study, and to a lesser extent for

multilingual BERT.

These results open up new avenues for research in multilingual semantic analysis, with multiple theoretical and application-oriented extensions. The polysemy and sense-related knowledge revealed by the models can serve to develop novel methodologies for improved cross-lingual alignment of embedding spaces and cross-lingual transfer (Ruder et al., 2019; Liu et al., 2019b), pointing to more polysemous words for which transfer might be harder. Predicting the polysemy level of words can also be useful for determining the context needed for acquiring representations that properly reflect the meaning of word instances in running text, and for identifying words which can be used as safe (unambiguous) cues for disambiguation (Leacock et al., 1998; Mihalcea, 2002; Agirre and Martinez, 2004; Loureiro and Camacho-Collados, 2020). From a more theoretical standpoint, we expect this work to be useful for studies on the organisation of the semantic space in different languages and on lexical semantic change (Kutuzov et al., 2018; Schlechtweg et al., 2019; Giulianelli et al., 2020).

Chapter 4

In-Context Lexical Substitution

4.1 Introduction

The lexical substitution (LexSub) task involves selecting meaning-preserving substitutes for words in context (cf. Table 4.1). The task was initially proposed as a testbed for word sense disambiguation (WSD) systems (McCarthy and Navigli, 2007). Contrary to other WSD evaluation settings, LexSub does not rely on pre-defined sense inventories (like WordNet). This permits the evaluation of different types of systems and makes cross-system comparison easier. In more recent works, LexSub is mainly seen as a way for evaluating the in-context lexical inference capability of vector-space models (Kremer et al., 2014; Melamud et al., 2015). Lexical substitution is also useful for language learners and can assist writers and translators in finding alternative lexicalisations. It can also serve in applications that involve re-writing, such as text simplification and summarization methods, as well as in text generation and machine translation evaluation metrics.

In this section, I again adopt a contrastive perspective and present two studies which address the question of in-context lexical substitution using different methodology. **The main idea behind LexSub models is the contextualisation of the vector representation that is available for a word**, i.e. its adaptation to each individual context of use. This is generally done by combining a target word's basic vector with the vectors of words found in the surrounding context (for example, in the same sentence or in a specific context window). Alternatively, words that are related to the target with a specific syntactic relation can be used. Appropriate substitutes for a specific target word instance are its paraphrases or synonyms that are similar to this contextualised representation. These are both semantically similar to the target and a good fit in the context. Hence, a substitution method needs to consider both the semantics of the candidate substitutes, as well as the context where the substitution will take place for prediction.

The early paper that I will present appeared at the Empirical Methods for Natural Language Processing (EMNLP) conference in 2016, and the most recent one at the International Conference on Computational Semantics (IWCS) in 2019.

(i) Apidianaki (2016): Marianna Apidianaki, "*Vector-space models for PPDB paraphrase ranking in context*", Proceedings of EMNLP 2016, p. 2028–2034.

(ii) Garí Soler et al. (2019b): Garí Soler, Aina and Cocos, Anne and Apidianaki, Marianna and Callison-Burch, Chris, "*A Comparison of Context-sensitive Models for Lexical Substitution*", Proceedings of the 13th International Conference on Computational Semantics - Long Papers, p. 271–282.

Lemma	Substitutes	Sentences
bright	intelligent (3), clever (3)	He was **bright** and independent and proud .
	shining (2), deep (1), vivid (1), luminous (1), vibrant (1)	Snow covered areas appear **bright** blue in the image which was taken in early spring and shows deep snow cover .
mood	humour (2), temperament (1), disposition (1), feeling (1), vibe (1), state of mind (1)	Trying to take away my good **mood** .
	atmosphere (4), ambience (1), feeling (1), tone (1)	In the room was perhaps 10 teachers , all friends and colleagues , when the director walked in the **mood** stiffened .

Table 4.1: Examples of in-context substitutes for the adjective *bright* and the noun *mood* from the SemEval-2007 LexSub dataset. Numbers in brackets indicate the number of annotators who proposed each substitute.

The EMNLP 2016 paper is a short paper where a syntax-based distributional model (Thater et al., 2011) is used to rank the unigram paraphrases (synonyms) of words in the Paraphrase Database (Ganitkevitch et al., 2013; Pavlick et al., 2015) according to new contexts of use. I show that good out-of-context paraphrases might not be adequate for in-context substitution. They need to be ranked according to how well they fit specific contexts of use. I use a syntax-based distributional model for this ranking which generates a context-specific representation for each target word, and uses this vector to find the best candidates for substitution (Thater et al., 2011). In our Garí Soler et al. (2019b) paper, we present more recent neural lexical substitution models which explicitly model the context of substitution and the semantics of individual lexical items. Before ending the chapter, I also briefly discuss even more recent substitution approaches which rely on the capability of BERT-like contextualised models to perform cloze-style slot filling.

4.2 A Syntax-based Lexical Substitution Model

Paraphrases are alternative ways to convey the same information and can improve natural language processing by making systems more robust to language variability and unseen words. The paraphrase database (PPDB) (Ganitkevitch et al., 2013; Pavlick et al., 2015) contains millions of lexical, phrasal and syntactic paraphrases in 21 languages, associated with features that serve to their ranking. These were automatically acquired by applying bi- and multi-lingual pivoting on parallel corpora (Bannard and Callison-Burch, 2005). In PPDB's release 2.0, such features include natural logic entailment relations, distributional and word embedding similarities, formality and complexity scores, and paraphrase quality scores assigned by a supervised ranking model (Pavlick et al., 2015). These features serve to identify good candidate paraphrases but do not say much about their substitutability in context.

In order to judge the adequacy of paraphrases for specific word instances, the surrounding context needs to be considered. This can be done using vector-space models of semantics which calculate the meaning of word occurrences in context based on distributional representations (Mitchell and Lapata, 2008; Erk and Padó, 2008; Dinu and Lapata, 2010; Thater et al., 2011). In the EMNLP 2016 paper (Apidianaki, 2016), I consider as candidate substitutes the unigram paraphrases of words in PPDB (i.e. their synonyms). I propose to use vector-space distributional semantic models for LexSub in order to select PPDB paraphrases that preserve the meaning of specific text fragments. This is a ranking variant of the LexSub task where systems are not expected to identify substitutes from the whole vocabulary, but rather to estimate the suitability of items in a specific pool of substitutes and rank them accordingly (Kremer et al., 2014). I used the vector space models of Thater et al. (2011) to rank the PPDB para-

phrases in context, and retain the ones that preserve the semantics of specific sentences. I evaluated the vector-based paraphrase ranking on data hand-annotated with lexical substitutes from the CoInCo corpus (Kremer et al., 2014).[1] I also compared the obtained ranking to out-of-context confidence estimates available in the PPDB, in order to show the importance of context filtering for paraphrase selection and substitution.

Vector-based models of semantic composition. Vector-based models of meaning determine a gradual concept of semantic similarity which does not rely on a fixed set of dictionary senses. They have been used for word sense discrimination and induction (Schütze, 1998; Turney and Pantel, 2010) and can capture the contextualised meaning of words and phrases (Mitchell and Lapata, 2008; Erk and Padó, 2008; Dinu and Lapata, 2010; Thater et al., 2011). Vector composition methods build representations that go beyond individual words to obtain word meanings in context. They specifically represent the contextualised meaning of a target word w_t in context c through vector composition, by creating a vector which combines the vectors of w_t and of the words $\{w_1, ..., w_n\}$ in c using some operation such as component-wise multiplication or addition.

Some models use explicit sense representations. In the method of Dinu and Lapata (2010), for example, word meaning is represented as a probability distribution over a set of latent senses reflecting the out-of-context likelihood of each sense, and the contextualised meaning of a word is modelled as a change in the original sense distribution.[2] Thater et al. (2011), on the contrary, use no explicit sense representation, but rather modify the basic meaning vector of a target word by reweighting its components on the basis of the context of occurrence. Paraphrase candidates for a target word are then ranked according to the cosine similarity of their basic vector representation to the contextualised vector of the target.

Paraphrase Ranking. I test whether vector-based models can select appropriate paraphrases for words in context. Given an instance of a target word w and a set of paraphrases P from PPDB, the task is to rank the elements in P according to their adequacy as paraphrases of w in the given context. I use instances from the CoInCo corpus (Kremer et al., 2014) because the provided annotations can be used for evaluation.

PPDB	Instances	Lemmas	Avg $\|P\|$
S	2,146	560	2.67
M	3,716	855	2.92
L	6,228	1,394	3.57
XL	13,344	2,822	10.33
XXL	14.507	3,308	185.09

Table 4.2: Number of CoInCo instances and unique lemmas covered by PPDB.

For each annotated English target word (noun, verb, adjective or adverb) in CoInCo, I collect its lexical paraphrases ($P = p_1, p_2, ..., p_n$) from each PPDB package (from S to XXL).[3] Table 1 shows the number of CoInCo tokens and lemmas with more than one paraphrases in a PPDB package, and the average size of the retained paraphrase sets. The larger the size of the resource, the greater the coverage of target words in CoInCo, and the noisier the retained set of paraphrases.[4] I test three flavours of the Thater et al. (2011) model:

(a) a syntactically structured model (`Syn.Vec`) which uses vectors recording co-occurrences based on dependency triples, explicitly recording syntactic role information within the vectors;

[1] CoInCo is a subset of the "Manually Annotated Sub-Corpus" MASC (Ide et al., 2008). It comprises more than 15K word instances manually annotated with single and multi-word substitutes.

[2] The latent senses are induced using non-negative matrix factorization (NMF) (Lee and Seung, 2000) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003b).

[3] PPDB comes in packages of different sizes.

[4] In the paper, I also report statistics and results for words with one paraphrase in PPDB. These practically evaluate the extent to which this unique candidate fits in different contexts where the word occurs.

PPDB	Bow.Vec	Syn.Vec	Filter.Vec	Google	AGiga	Ppdb1	Ppdb2	Parprob	Random (5)
S	**0.91**	**0.91**	**0.91**	0.78	0.86	0.66	0.83	0.66	0.78
M	0.91	0.91	**0.92**	0.79	0.87	0.68	0.84	0.68	0.79
L	0.90	0.90	**0.91**	0.78	0.85	0.66	0.83	0.66	0.77
XL	0.78	**0.79**	**0.79**	0.58	0.67	0.44	0.66	0.43	0.58
XXL	0.53	0.56	0.57	0.27	0.36	0.12	**0.58**	0.12	0.27

Table 4.3: Average GAP scores obtained by the contextual models, the paraphrase adequacy methods and the random baseline on the CoInCo dataset.

(b) a syntactically filtered model (Filter.Vec) which uses dependency-based cooccurrence information without explicitly representing the syntactic role in the vector representations, as in (Padó and Lapata, 2007));

(c) a bag of words model (Bow.Vec) which uses a window of $\pm$ 5 words.

Co-occurrence counts are extracted from the English Gigaword corpus[5] analysed with Stanford dependencies (de Marneffe et al., 2006). The syntactic model vectors are based on dependency triples that occur at least five times in the corpus and which have a PMI score of at least 2. For the bag of words model, two words are retained as co-occurrences if they have been observed at least five times in the same context window. The task of the vector-space models for each target word instance is to rank the contents of the corresponding PPDB paraphrase set so that the actual substitutes are ranked higher than the rest. For example, *newspaper*, *manuscript* and *document* are candidate paraphrases for *paper* but we would expect a good ranking model to order *newspaper* higher than the other two candidates in the sentence: "*the paper's local administrator*". In the `Bow.Vec` model, the context is made up of five words before and after the target while in the syntactic models, it corresponds to the target's direct syntactic dependents. The contextualised vector for w is obtained through vector addition.

Paraphrase candidates are ranked according to the cosine similarity between the contextualised vector of the target word and the basic meaning vectors of the candidates. Following Kremer et al. (2014), I compare the resulting ranked list of paraphrases to the CoInCo gold standard annotations using Generalised Average Precision (GAP) (Kishida, 2005) and annotation frequency as weights. GAP score is in the range [0,1]. A score of 1 indicates a perfect ranking in which all correct substitutes precede all incorrect ones, and correct high-weight substitutes precede low-weight ones.[6]

Evaluation. The average GAP scores obtained by the three vector-space models for words with more than one paraphrase in the PPDB are shown in Table 4.3.[7] These are compared to five out-of-context rankings reflecting paraphrase quality in the PPDB (Pavlick et al., 2015), and to a random baseline.

- **AGigaSim** captures the distributional similarity of a phrase p and its paraphrases $\{p_1, ...p_n\} \in P$, as computed according to contexts observed in the Annotated Gigaword corpus (Napoles et al., 2011);
- **GoogleNgramSim** reflects the distributional similarity of p and its paraphrases in P, as computed according to contexts observed in the Google Ngram corpus (Brants and Franz, 2006);
- **ParProb**: the paraphrase probability of every paraphrase $\{p_1, ...p_n\} \in P$ given the original phrase p (Bannard and Callison-Burch, 2005);

[5] http://catalog.ldc.upenn.edu/LDC2003T05

[6] In order to calculate the GAP score, we assign a very low score (0.001) to paraphrases that are not present in CoInCo for a target word (i.e. which were not proposed by the annotators).

[7] The paper contains results for all words, including the ones with a single paraphrase.

- **Ppdb1**: the heuristic scoring used for ranking in the original release of the PPDB (Ganitkevitch et al., 2013);
- **Ppdb2**: the improved ranking of English paraphrases in PPDB 2.0 using a supervised scoring model trained on human judgements of paraphrase quality (Pavlick et al., 2015);
- **Random**: A baseline where the paraphrases of p are randomly ranked. The reported figures are PPDB package-specific since a different paraphrase set is retained from each package. They correspond to averages over five runs. The quality of the ranking produced by the baseline clearly decreases as the size of the PPDB resource increases due to the higher number of retained paraphrases which makes ranking harder.

The results show that the vector-space models provide a better ranking than the PPDB paraphrase quality estimates and largely outperform the random baseline. The three models perform similarly on this ranking task according to the average GAP score with the syntactically-informed models getting slightly higher scores. Differences between `Syn.Vec` and `Filter.Vec`, as well as between `Bow.Vec` and the syntactic models, are highly significant in the XL and XXL packages (p-value < 0.001) as computed with approximate randomisation. In the L package, the difference between `Syn.Vec` and `Filter.Vec` is significant ($p < 0.05$) and the one between `Bow.Vec` and `Filter.Vec` is highly significant. Finally, in the M package, only the difference between `Bow.Vec` and `Filter.Vec` is significant ($p < 0.05$), while `Syn.Vec` and `Filter.Vec` seem to deal similarly well with the contents of this package.

From the PPDB ranking methods, **AGigaSim** and **Ppdb2** obtain good results. This is probably due to the natural skewed distribution towards predominant senses in CoInCo where whole documents were annotated (Kremer et al., 2014). This distribution of senses favours non-contextualised baseline models since the most frequent sense is the one often attested. The good performance of **Ppdb2** is due to the use of a supervised scoring model. Human judgements of paraphrase quality were used to fit a regression to the features available in PPDB 1.0 plus numerous other features including cosine word embedding similarity, lexical overlap features, WordNet features and distributional similarity features. The small difference observed between **Ppdb2** and the syntactic models score in the XXL package is highly significant. However, **Ppdb2** scores are available only for English, while the vector-space methodology is unsupervised and can be easily applied to other languages. The performance of the models remains high with the XL package which ensures a high coverage since it contains paraphrase sets of reasonable size (about 10 paraphrases per word), and lowers in XXL which contains 185 paraphrases in average per word (cf. Table 1). To use this package more efficiently, one could initially reduce the number of erroneous paraphrases on the basis of the Ppdb2 score which provides a good ranking of the XXL package contents before applying the vector-based models.

To conclude, vector-based models of semantics can be successfully applied to the in-context ranking of PPDB paraphrases. Allowing for better context-informed substitutions, they can be used to filter PPDB paraphrases on the fly and select variants preserving the correct semantics of words and phrases in texts. This processing would be beneficial to numerous applications that need paraphrase support (e.g., summarization, query reformulation and language learning) as it provides a practical means for exploiting the extensive knowledge present in the multilingual PPDB resource. It can also be useful in evaluation settings (as in Machine Translation and summarization evaluation) where there is need to understand whether the rephrasings proposed by a model are semantically plausible (Denkowski and Lavie, 2014; Marie and Apidianaki, 2015). Although tested on English, the proposed methodology can be applied to all languages in the PPDB even to those that do not dispose of a dependency parser, as shown by the high performance of the `Bow.Vec` models.

4.3 Neural Models for Substitution

Word embedding models provide good estimates of word meaning, they would thus be a good fit for the LexSub task. In our work presented at the International Conference on Computational Semantics (IWCS 2019) Garí Soler et al. (2019b), we compare different types of representations on this task. Specifically, we compare static embeddings (GloVe (Pennington et al., 2014) and FastText (Bojanowski et al., 2017)) and contextualised (ELMo) representations (Peters et al., 2018), with models specifically designed for lexical substitution (Melamud et al., 2015, 2016). We evaluate all models on the SemEval 2007 LexSub task test set (McCarthy and Navigli, 2007). Given an instance of a target word t and a set of candidate substitutes ($S = \{s_1, s_2, ..., s_n\}$), we use each model to get a ranking of the substitutes in S depending on how well they describe the meaning of t. Higher ranked substitutes are both good paraphrases of the target and a good fit in the context. Similar to our experiments presented in Section 4.2, we consider as candidate substitutes $S = \{s_1, s_2, ..., s_n\}$ for t its paraphrases in the Paraphrase Database (PPDB) XXL package (Pavlick et al., 2015)[8] that are also present in the gold standard SemEval 2007 LexSub annotations.

4.3.1 Lexical Substitution Methods

Target-to-substitute (tTs) ELMo embedding similarity. ELMo representations are contextualised, so the embedding for a token is a function of the entire sentence in which it appears. We apply the ELMo vectors for the first time to the lexical substitution task. The proposed substitute ranking method uses target-to-substitute (*tTs*) similarity, as measured by the cosine similarity of the corresponding ELMo representations. Given a new sentence C with an instance of the target word t to be substituted, we first obtain a representation for t. We experiment with the top layer (ELMo-top) and the average of the three layers of the model (ELMo-avg).[9] We then replace t with all its potential substitutes $s_t \in S_t$ one by one, and obtain the ELMo vector for each substitute in the same context. Substitutes are then ranked by the cosine similarity of their vector with that of t in the same context.

The AddCos method. This method has been proposed by Melamud et al. (2015) and is based on the skip-gram word embedding model. The method explicitly leverages the context embeddings generated within skip-gram, which are generally considered as internal and discarded at the end of the learning process. The proposed substitutability measures capture two types of similarity:

(a) *target-to-substitute*, showing how similar a potential substitute is to the target word;

(b) *target-to-context*, reflecting the substitute's compatibility with a given sentential context.

Similarities are estimated using the cosine distance between the skip-gram word and context embeddings. The proposed measures differ in the way they combine the score elements together, using either an arithmetic or geometrical mean. We choose the more flexible additive approach which (contrary to the multiplicative variants) does not require high similarities in all elements of the product to highly rank a substitute, but can yield a high score even if one of the elements in the sum is zero. The *Add* measure equation (1),[10] estimates the substitutability of a substitute s with the target word t in context C, where C corresponds to the set of words in the sentence, and c corresponds to an individual context word.

[8] http://paraphrase.org

[9] Averaging across layers has been shown to improve performance in several syntax and semantics-related tasks compared to using the top LSTM layer.

[10] We call this method *AddCos* because of the Cosine function applied to the vector representations of words and contexts.

$$AddCos(t,s,C) = \frac{cos(s,t) + \sum_{c \in C} cos(s,c)}{|C|+1} \tag{4.1}$$

We use 300-dimensional skip-gram word and context embeddings trained on the 4B words of the Annotated Gigaword corpus (Napoles et al., 2012).[11] We also apply this method to ELMo embeddings. We extract the vector for the target and for each substitute in the same sentence.

The *context2vec*-based model. The *context2vec* (*c2v*) model of (Melamud et al., 2016) jointly learns context and word embeddings using a bidirectional LSTM. The model is based on word2vec's CBOW architecture (Mikolov et al., 2013a) but replaces its naive context modelling (of averaged word embeddings in a fixed window) with a full-sentence neural representation of context. Words and contexts are embedded in the same low-dimensional space which allows for calculating target-to-context (*t2c*), context-to-context (*c2c*) and target-to-target (*t2t*) similarities. A score for a candidate substitute is computed using the following formula:

$$\text{c2v\_score} = \frac{cos(s,t)+1}{2} \times \frac{cos(s,C)+1}{2} \tag{4.2}$$

where t and s are the word embeddings of the target and the substitute, and C is the *c2v* context vector of the sentence with an empty slot at the target's position. We use the 600-dimensional c2v embeddings released by Melamud et al. (2016). We also use Equation 4.2 (hereafter called *c2vf*) with standard ELMo (instead of skip-gram) vectors. We obtain the ELMo embedding of the target word and its substitutes in the same context. The context vector (C) is the average of the ELMo embeddings of all words in the context.

Baselines. The context-unaware baseline models solely rely on the target-to-substitute similarity of pre-trained word embeddings: 300-dimensional GloVe vectors (Pennington et al., 2014)[12] and FastText vectors, both trained on Common Crawl (Mikolov et al., 2018).[13] Since these embeddings are uncontextualised, the proposed substitute ranking is the same for all contexts. We also test an "enriched" version of the baseline models which integrates a simple representation of the context consisting of the average of the embeddings of words contained in it. We then compare target and substitute vectors to the generated context vector using the context2vec formula (Equation 4.2).

4.3.2 Evaluation

We compare the models by testing them in a substitute ranking task. We use 158 target words from the SemEval-2007 Lexical Substitution (LexSub) task dataset (McCarthy and Navigli, 2007) which have more than one substitutes ($S = \{s_1, s_2, ..., s_n\}$) in PPDB 2.0 XXL.[14] The substitutes in S need to be ranked according to their suitability in context. For example, the noun *way* has the following candidate paraphrases in PPDB XXL: *sense, means, aspect, technique, passage, respect, direction, characteristic, journey, method, route, practice, fashion, manner*. These need to be ranked for every instance in a new

[11] The vectors used by the original *Add* method are syntax-based embeddings (Levy and Goldberg, 2014). We use the lighter adaptation proposed by Apidianaki et al. (2018) which circumvents the need for syntactic analysis.

[12] https://nlp.stanford.edu/projects/glove

[13] https://fasttext.cc/docs/en/english-vectors.html

[14] There are 1,584 sentences for these words. The full dataset contains 201 target words and 2,010 sentences extracted from the English Internet Corpus (Sharoff, 2006) and annotated with substitutes by five native English speakers.

sentence, for example: "on the way out of the parking lot johnny felt a thump". A model performs well if it ranks the gold substitutes for this instance *route* (3), *passage* (1), *journey* (1)[15] higher than the other paraphrases (synonyms) of this noun.

Similar to the experiments in Section 4.2, we measure the quality of the automatic ranking by comparing it to the gold ranking using Generalized Average Precision (GAP) (Kishida, 2005). The results for different model and vector combinations are given in Table 4.4. We observe that context2vec outperforms other methods. This must be due to the model's training objective which makes it highly suited for the LexSub task: context2vec is explicitly trained with pairs of target words and sentential contexts, optimising the similarity of context vectors and potential fillers. In contrast, ELMo is trained as a general language model to predict the immediate next tokens. The ELMo-avg configuration gives slightly better performance than ELMo-top. FastText embeddings perform slightly better than GloVe in this task, while both models benefit from adding context. For the AddCos method, one word around the target word (c=1) is more effective than a bigger context window (c=4).

Method	Vectors	GAP
AddCos (c=1 or c=4)	Skip-gram	0.527 (0.520)
	ELMo-avg	0.527 (0.498)
	ELMo-top	0.513 (0.476)
c2vf	UkWac c2v	**0.587**
	ELMo-avg	0.529
	ELMo-top	0.516
tTs	ELMo-avg	0.534
	ELMo-top	0.531
Glove + context	Glove	0.467
Fasttext + context	Fasttext	0.491
Baselines	Glove	0.465
	Fasttext	0.485

Table 4.4: The GAP score of substitute ranking methods. The two scores for AddCos are for different window sizes (c =1 and c = 4).

4.3.3 BERT as LexSub model

More recently, the use of BERT for lexical substitution has also been considered, but there are several issues to be considered regarding the straightforward application of this model to this task. One problem is that the proposed substitutes – which are often obtained using masking, as in cloze-task probing studies – might not preserve the semantics of the original text. For example, if *cat* is masked in the sentence "*I love this cat*", BERT would propose substitutes that fit this context without preserving the semantics of the original sentence (e.g., *dog*, *food*, *movie*, *restaurant*). To address this problem, Li et al. (2020) proposed to generate examples for their BERT-ATTACK method without using the `[MASK]` token. They instead query BERT with the whole sentence (without masking). This method poses other problems such as the probability mass going to the original (unmasked) target word, making it hard to choose good candidates from the remaining probability space. Zhou et al. (2019) address this issue by partially masking the target word. This is done by applying embedding dropout and having BERT propose substitutes for that position. This is a mid-way solution between target word masking, which can generate semantically irrelevant substitutes and unmasking which would put about 99.99% of the predicted probability distribution into the target word.

4.4 Conclusion

This chapter explains how in-context lexical substitution was performed with distributional (syntax-based and bag-of-words) models, and how this is now done with neural model representations. I present a comparison of the performance of static and contextualised embeddings on this task. I also compare models that have been trained with a language modelling objective (like ELMo) with models specifi-

[15]The numbers denote how many annotators proposed each substitute.

cally trained for the LexSub task (like context2vec). The results of this comparison show that powerful contextualised word representations, which give high performance in several semantics-related tasks, deal less well with the subtle in-context similarity relationships that need to be considered for substitution. This is better handled by models trained with a specific lexical substitution objective (Melamud et al., 2015, 2016), where the inter-dependence between word and context representations is explicitly modelled during training.

Chapter 5

Adjective Intensity Detection

5.1 Introduction

Adjectives like *pretty*, *beautiful*, and *gorgeous* all describe appearance in positive terms but differ in intensity. Understanding these differences between adjectives is necessary for reasoning about natural language. Modelling this distinction is also important for language understanding tasks such as sentiment analysis (Pang et al., 2008), question answering (de Marneffe et al., 2010), and textual inference (Dagan et al., 2006). Information on the relative intensities of adjectives, however, is not present in existing lexico-semantic resources such as WordNet (Miller, 1995; Fellbaum, 1998).

We have approached the question of scalar adjective intensity using two approaches. **The first relies on paraphrases** extracted from an automatically built large-scale multilingual database, and combines them with patterns extracted from texts and with information from semantic lexicons (Cocos et al., 2018). **Our second approach leverages the semantic space built by contextual language models** (Garí Soler and Apidianaki, 2020). We specifically show that abstract semantic notions such as intensity can be identified in the space constructed by these models and can serve to rank words with different intensity in different languages. In what follows, we present the details of the two approaches and a comparison in intrinsic and extrinsic evaluation settings.

Previous approaches to adjective intensity detection gathered evidence from large text corpora using patterns. These included lexical (Sheinman and Tokunaga, 2009; de Melo and Bansal, 2013; Sheinman et al., 2013) and syntactic (Shivade et al., 2015) patterns indicative of an intensity relationship. For example, the patterns "X, but not Y" (e.g., *good but not great*), "not just X but Y" (*not just good but great*), "X though not Y" (*good though not great*) provide evidence that X is an adjective less intense than Y ($X < Y$). These methods are precise but have relatively small coverage of comparable adjectives, even when using web-scale corpora (de Melo and Bansal, 2013; Ruppenhofer et al., 2014).

Lexicon-based approaches employ resources that map an adjective to real-valued scores encoding both sentiment polarity and intensity (Hatzivassiloglou and McKeown, 1993). For example, *good* might map to 1 and *phenomenal* to 5, both positive scores reflecting the respective intensity of the words; while *bad* maps to negative score -1 and *awful* to -3. The lexicon used in these methods might be compiled automatically – for example, from analyzing adjectives' appearance in star-valued product or movie reviews (de Marneffe et al., 2010; Rill et al., 2012; Sharma et al., 2015; Ruppenhofer et al., 2014) – or manually (e.g., the SO-CAL lexicon (Taboada et al., 2011)). Similar to pattern-based approaches, these methods are highly precise but have low coverage, since they are constrained by the limited coverage of the used lexicons.

5.2 Paraphrase-based Adjective Ranking

5.2.1 Our Intensity Detection Method

Our method for automatically learning the relative intensity relation that holds between scalar adjectives presented in Cocos et al. (2018) relies on paraphrases. The proposed method provides increased coverage compared to pattern- and lexicon-based approaches. We demonstrate how paraphrases can be useful in this scenario, and also how the three information sources (paraphrases, patterns and lexicons) can be combined for effectively detecting intensity in adjective scales.

The novelty of our method is its reliance on paraphrases as a source of evidence for intensity identification. A paraphrase is a pair of words or phrases with approximately similar meaning, such as *really great* $\leftrightarrow$ *phenomenal*. Adjectival paraphrases of the form RB $JJ_u \leftrightarrow JJ_v$ where one phrase is comprised of an adjective (JJ_u) modified by an intensifying adverb (RB) and the other is a single-word adjective (JJ_v), provides evidence about their relative intensity, i.e. that the first adjective is less intense than the second ($JJ_u < JJ_v$). For example, the paraphrase rules "*really great* $\leftrightarrow$ *phenomenal*" and "*very pleasant* $\leftrightarrow$ *delightful*" provide evidence that "*great* $<$ *phenomenal*", and "*pleasant* $<$ *delightful*". The proposed relationships indeed hold between these adjectives.

We extract this type of knowledge from the Paraphrase Database (PPDB) (Ganitkevitch et al., 2013; Pavlick et al., 2015), which contains over 36k adjectival pairs.[1] A few more examples illustrating our basic assumption that the presence of an intensifying adverb in a paraphrase rule involving two adjectives is indicative of an intensity relationship between them are given below.

$$\textit{particularly pleased} \leftrightarrow \textit{ecstatic} \Rightarrow \textit{pleased} < \textit{ecstatic}$$
$$\textit{quite limited} \leftrightarrow \textit{restricted} \Rightarrow \textit{limited} < \textit{restricted}$$
$$\textit{rather odd} \leftrightarrow \textit{crazy} \Rightarrow \textit{odd} < \textit{crazy}$$
$$\textit{so silly} \leftrightarrow \textit{dumb} \Rightarrow \textit{silly} < \textit{dumb}$$
$$\textit{completely mad} \leftrightarrow \textit{crazy} \Rightarrow \textit{mad} < \textit{crazy}$$
$$\textit{really beautiful} \leftrightarrow \textit{gorgeous} \Rightarrow \textit{beautiful} < \textit{gorgeous}$$

We use these paraphrases to build a graph (called JJGRAPH) where nodes represent adjectives, and directed edges represent the intensifying adverbs (e.g., *particularly*, *rather*, *really*) linking the two adjectives.[2] We identify the intensifying adverbs needed for constructing the graph using the following bootstrapping approach.

1. **Round 1**: We start with a small seed set of adjective pairs (JJ_u,JJ_v) having a known intensity relationship, such as the base-form adjective and its comparative or superlative form (e.g., *very hard* $\leftrightarrow$ *harder*, *so hard* $\leftrightarrow$ *harder*). Since we know that $JJ_u < JJ_v$ in these cases, we infer that adverb RB is intensifying. All such RB's are added to an initial adverb set R_1.
2. **Round 2**: The process continues by extracting paraphrases (RB $JJ_{u'} \leftrightarrow JJ_{v'}$) with RB $\in R_1$, indicating additional adjective pairs ($JJ_{u'}$, $JJ_{v'}$) with intensity direction inferred by RB.
3. **Round 3**: Finally, the adjective pairs extracted in this second iteration are used to identify additional intensifying adverbs R_3, which are added to the final set $R = R_1 \cup R_3$.

This process generates a set of 610 adverbs, many of which are intensifying (e.g., *very*, *truly*, *abundantly*, *particularly*). A few de-intensifying adverbs (e.g. *hardly*, *kind of*) and adverbs that do not

[1] www.paraphrase.org

[2] Intensifying adverbs (e.g., *very*, *totally*) strengthen the adjectives they modify in contrast to de-intensifying adverbs (e.g., *slightly*, *somewhat*) which weaken the intensity of the adjectives.

express intensity (e.g. *ecologically*) are also included due to the bootstrapping process, and the noise in the automatically-compiled PPDB resource. The JJGRAPH is built using all 36,756 adjectival paraphrases in PPDB of the specified form $RB\ JJ_u \leftrightarrow JJ_v$, where the adverb belongs to R. The resulting JJGRAPH has 3,704 unique adjective nodes. It can be used for pairwise intensity prediction, since the directed adverb edges between two adjectives JJ_u and JJ_v provide evidence about the intensity relationship between them.

The proposed paraphrase approach can be combined with pattern- and lexicon-based approaches. This allows us to benefit from the different approaches, and to improve the quality of the results both in terms of precision and coverage. We use pattern-based evidence gathered using the de Melo and Bansal (2013) approach which relies on intensity patterns from Google *n*-Grams (Brants and Franz, 2006). These comprise "weak-strong" patterns (e.g., *"X, but not Y"*) and "strong-weak" patterns (e.g., *"not X, but still Y"*) which serve as evidence about the directionality of the intensity relationship. Given an adjective pair (JJ_u,JJ_v) (e.g., *good*, *great*), an overall pattern-based "weak-strong" score is calculated based on the number of times where the adjectives occur in patterns corresponding to the "weak-strong" and "strong-weak" intensity relations (e.g., *"good, but not great"*, *"not only good but great"*, *"not great, just good"*). Additionally, we use evidence from the manually-compiled SO-CAL[3] lexicon (Taboada et al., 2011) for inferring intensity. SO-CAL includes integer weights in the range $[-5,5]$ for 2,826 adjectives. The sign of the weight encodes sentiment polarity (positive or negative) and the value encodes intensity (e.g., *phenomenal*: 5; *unlikable*: -3). We derive a pairwise intensity prediction score for adjectives having the same polarity by subtracting their scores.

We combine two metrics **x** and **y** to generate a score for an adjective pair (JJ_u, JJ_v), by simply using the first metric **x** if it can be reliably calculated for the pair, and backing off to metric **y** otherwise. For pattern-based evidence, a score is reliably calculated for a pair of adjectives when they appear together in an intensity pattern. For lexicon-based evidence, there is confidence in a prediction when both adjectives are in the lexicon and have the same polarity (positive/negative). For paraphrase-based evidence, when the two adjectives are directly connected in JJGRAPH.

5.2.2 Intrinsic Evaluation

We evaluate our method against three manually created datasets. We use each type of evidence (paraphrases, patterns, lexicon) separately, and their combination, for ranking the scalar adjectives in these resources.

- The de Melo and Bansal (2013) dataset (**deMelo**)[4] contains 87 adjective sets extracted from WordNet 'dumbbell' structures (Gross and Miller, 1990) which are partitioned into half-scale sets based on their pattern-based evidence in Google N-Grams (Brants and Franz, 2006), and manually annotated for intensity relations ($<$, $>$, and $=$).
- The Wilkinson and Oates (2016) dataset (**Wilkinson**) contains 12 full adjective scales annotated for intensity through crowdsourcing.
- **Crowd** is a crowdsourced dataset of 79 adjective scales with high coverage of the PPDB vocabulary (293 adjective pairs).

We measure the agreement between the gold standard ranking of adjectives along each scale in these datasets and the ranking predicted by each of our methods. We consider intensity predictions as high quality if they agree with the reference ranking. We calculate the **pairwise accuracy of the predictions**

[3] https://github.com/sfu-discourse-lab/SO-CAL
[4] http://demelo.org/gdm/intensity/

Test Set	Score Type	Coverage	Pairwise Accuracy
deMelo	$score_{pat}$	0.48	**0.844**
	$score_{pp}$	0.33	0.458
	$score_{socal}$	0.28	0.546
	$score_{pat+socal}$	0.61	0.757
	$score_{pat+socal+pp}$	**0.70**	0.722
Crowd	$score_{pat}$	0.11	**0.784**
	$score_{pp}$	0.74	0.676
	$score_{socal}$	0.35	0.757
	$score_{pat+socal}$	0.81	0.687
	$score_{pat+socal+pp}$	**0.82**	0.694
Wilkinson	$score_{pat}$	0.44	0.852
	$score_{pp}$	0.80	0.753
	$score_{socal}$	0.31	**0.895**
	$score_{pat+socal}$	**0.89**	0.833
	$score_{pat+socal+pp}$	**0.89**	0.833

Table 5.1: Pairwise relation prediction for each score type in isolation, and for the best-scoring combinations of two or three score types on each dataset. We also report the coverage of each method.

against the gold standard. For each pair of adjectives along the same scale, we compare the predicted to the gold-standard ordering for the pair. We report the overall accuracy of the pairwise predictions in Table 5.1. We report the results for each score type in isolation, and for the best-scoring combinations of two or three score types on each dataset. $score_{pat}$ shows the results for the pattern-based method, $score_{pp}$ for the paraphrase-based method, and $score_{socal}$ shows the performance of the lexicon-based method that relies on SO-CAL (Taboada et al., 2011). We also report the performance of the combined methods $score_{pat+socal}$ and $score_{pat+socal+pp}$.

The pairwise accuracy scores for the pattern-based and the lexicon-based methods are higher than those of the paraphrase-based methods for all datasets, but their coverage is relatively limited. One exception is the deMelo dataset where the pattern-based method has high coverage, because the dataset was also compiled by finding adjective pairs that matched lexical patterns in the corpus. For all datasets, highest coverage is achieved using one of the combined metrics that incorporates paraphrase-based evidence.

5.2.3 Extrinsic Evaluation

We also evaluate our method on an Indirect Question-Answering task (de Marneffe et al., 2010). This task involves polar (*yes*/*no*) questions for which the answers often do not contain an explicit *yes* or *no*. They rather give information that the hearer can use to infer such an answer in a context with some degree of certainty. Interpreting the answer is straightforward in some cases (Q: Was it bad? A: It was terrible.) but in other cases the answer is unclear (Q: Was it good? A: It was provocative.).

The de Marneffe et al. (2010) dataset is focused on the interpretation of answers to polar questions where the main predication involves a gradable modifier (e.g., *highly unusual*, *not good*, *little*) and the answer either involves another gradable modifier or a numerical expression. In such cases, the implied answer depends on the relative intensity of adjective modifiers in the question and answer. For example, in the exchange:

Q: Was he a successful ruler?
A: Oh, a tremendous ruler.

the implied answer is "*yes*". This answer is inferred because *successful* is less intense than *tremendous*. Inversely, in the exchange:

Q: Does it have a large impact?
A: It has a medium-sized impact.

the implied answer is "*no*" because *large* is less intense than *medium − sized*.

Interpreting such question–answer pairs requires dealing with modifier meanings, specifically, learning context-dependent scales of expressions (Fauconnier, 1975) that determine how, and to what extent, the answer as a whole resolves the issue raised by the question (de Marneffe et al., 2010).

For this evaluation, we use the dataset that was compiled by de Marneffe et al. (2010), which contains 123 examples of indirect question-answer pairs (IQAP) extracted from dialogue corpora and annotated (with *yes* or *no*) through crowdsourcing. For each QA pair, the implied answer depends on the relative intensity relationship between the modifiers present in the question and answer texts.

As in the intrinsic evaluation described in the previous section, we test the quality of the predictions made using the paraphrase-based evidence with predictions made using pattern-based, lexicon-based, and combined scoring metrics. In order to use the pairwise scores for inference, we employ a decision procedure nearly identical to that of de Marneffe et al. (2010). If the adjective in the question (j_q) and the adjective in the answer (j_a) are scorable,[5] then $j_q \leq j_a$ implies the answer is "*yes*", and $j_q > j_a$ implies the answer is "*no*". If the score of the pair of adjectives is undefined, then the prediction is *no* (they could be antonyms or unrelated). If either adjective is missing from the scoring vocabulary, then the prediction is *uncertain* because they are impossible to compare. Table 5.2 shows the results of this evaluation. We report the accuracy, the macro-averaged precision, the recall, and the F1-score for each of the tested methods, as well as the percentage of pairs with one or two out-of-vocabulary (OOV) adjectives. We compare to an "all-YES" baseline which predicts all answers to be "YES", and to the result of the original method of de Marneffe et al. (2010), which used an automatically compiled lexicon to make polarity predictions for each indirect QA pair.

Method	% OOV	Acc.	P	R	F
all-"YES"	.00	**.691**	.346	.500	.409
(de Marneffe et al., 2010)	.02	.610	.597	.594	.596
score$_{socal}$	.33	.504	**.710**	.481	.574
score$_{pp}$	.09	.496	.568	.533	.550
score$_{pat}$	.07	.407	.524	.491	.507
score$_{socal+pp}$	.09	.634	.690	.663	.676
score$_{socal+pat+pp}$	.06	.642	.684	**.683**	**.684**

Table 5.2: Results of the evaluation on the Indirect Question Answering dataset.

The simple "all-YES" baseline gets highest accuracy in this imbalanced test set, but all score types perform better than this baseline in terms of precision and F1-score. score$_{socal}$, which was derived from a manually-compiled lexicon, is very precise and scores higher than score$_{pp}$ and score$_{pat}$ with a Precision of .710. However, due to its low coverage of the IQAP vocabulary, it mispredicts 33% of the pairs as *uncertain*. score$_{pp}$ has relatively high coverage and a mid-level F1-score, while score$_{pat}$ scores poorly on this dataset due to its sparsity.[6] The paraphrase-based and lexicon-based evidence is complementary. Thus, the combined score$_{socal+pp}$ and score$_{socal+pat+pp}$ produce significantly better accuracy than any score in isolation (McNemar's test, $p < .01$), and outperform the original ranking method of de Marneffe et al. (2010). In later work, Kim and de Marneffe (2013) report a higher F1-score (.7058) on this dataset using word embeddings and a vector offset method (Mikolov et al., 2013c).

[5] Two adjectives are scorable if they have an intensity relationship along the same half-scale

[6] All modifiers in the IQAP dataset are in the Google N-grams vocabulary but most of them do not have observed patterns. Therefore, score$_{pat}$ returns "NO" predictions.

5.3 Adjective Intensity in Contextual Language Models

The paraphrase, pattern-based and lexicon-based methods presented in the previous section complement each other and achieve fair results in both intrinsic and extrinsic evaluations. As has been shown, they all involve different processing steps and engineering choices (fixing weights, choosing features for graph construction, selecting patterns, etc) which condition their success. In this section, I present our recent work on leveraging the representations of contextual language models for scalar adjective intensity identification.

Interpretability studies which explore the information encoded in neural language model representations, such as the ones presented in Section 3.3.1, have shown that these encode rich linguistic, commonsense and world knowledge. Inspired by the findings of these studies and by works which showed that semantic notions such as gender are encoded in the space constructed by word embedding models (Bolukbasi et al., 2016; Dev and Phillips, 2019), **we set to explore whether intensity is also encoded in these representations.** In our paper Garí Soler and Apidianaki (2020), we investigate the knowledge that the pre-trained BERT model (Devlin et al., 2019) encodes about the intensity of the emotion expressed on an adjective scale, without access to any external resources such as lexicons or paraphrases. We consider the contextualised representations produced by BERT to be a good fit for this task, since the scalar relationship between adjectives is context-dependent (Kennedy and McNally, 2005). We view intensity as a direction in the semantic space which, once identified, can serve to detect the intensity relationship of new adjectives on the fly. This work was done in collaboration with my former PhD student Aina Garí Soler, in the MULTISEM project.

5.3.1 Contextualised Adjective Representations

In order to explore the knowledge that BERT encodes about relationships in an adjective scale s (e.g., *pretty* $\rightarrow$ *beautiful* $\rightarrow$ *gorgeous*), we generate a contextualised representation for each adjective $a \in s$ occurring in the same context. This ensures that variation in the representations reflects differences in the semantics of the adjectives, instead of context variation (Ethayarajh, 2019b; Mickus et al., 2020). Since it is difficult, or impossible, to find such examples in running text, we construct sentence sets satisfying this condition using the ukWaC corpus (Baroni et al., 2009)[7] and the Flickr 30K dataset (Young et al., 2014).[8] We use the adjective scales in the three datasets (D) described in Section 5.2.2: demelo (de Melo and Bansal, 2013), Wilkinson (Wilkinson and Oates, 2016), and Crowd (Cocos et al., 2018).

For every scale $s \in D$, and for each adjective $a \in s$, we collect 1,000 instances (sentences containing that adjective) from each corpus.[9] We substitute each instance i of $a \in s$, with each $b \in s$ where $b \neq a$, creating $|s| - 1$ new sentences. For example, we substitute *beautiful* in the sentence "*This beach is beautiful*" with the other adjectives in the scale "*pretty* $\rightarrow$ *beautiful* $\rightarrow$ *gorgeous*", creating: "*This beach is pretty*" and "*This beach is gorgeous*". We filter out sentences where substitution should not take place, such as cases of specialisation between a hypernym and its hyponym, or instantiation (IS-A relations). For example, the sentences that would be created by substituting *deceptive* with *fraudulent* in "*Viruses and other deceptive software*" or in "*Deceptive software such as viruses*" would not be valid. We identify cases of specialization and instantiation by parsing the sentences[10] to reveal their

[7] http://u.cs.biu.ac.il/~nlp/resources/downloads/context2vec/

[8] Flickr contains crowdsourced captions for 31,783 images describing everyday activities, events and scenes.

[9] ukWaC has perfect coverage. Flickr 30K covers 96.56% of DEMELO scales and 86.08% of CROWD scales. A scale s is not covered when no $a \in s$ is found in a corpus.

[10] We use `stanza` (Qi et al., 2020).

dependency structure, and then identifying the ones that describe *is-a* relations between nouns using Hearst lexico-syntactic patterns (Hearst, 1992).

We use additional criteria to ensure the quality of the generated sentences. We filter the substitutes using language modelling criteria, since adjectives that belong to the same scale might not be replaceable in all contexts. Polysemy can also influence their substitutability; for example, a *hot drink* is *warm*, but a *hot topic* is *interesting*. In order to select contexts where all adjectives in a scale ($\forall a \in s$) fit, we measure the fluency of the generated sentences using a score assigned by the context2vec substitution model (Melamud et al., 2016). This score reflects for an $a \in s$ how well it fits a context by measuring the cosine similarity between a and the context representation.[11] We calculated the context2vec score for all sentences generated for a scale s through substitution, and kept the ten with the lowest standard deviation (STD). Low STD for a sentence means that $\forall a \in s$ are reasonable choices in this context. For comparison, we also randomly sampled ten sentences from all the ukWaC sentences collected for each scale. We call these sets of sentences SENT-SETs.

We extracted the contextualised representation for each $a \in s$ in the ten sentences retained for scale s, using the pre-trained `bert-base-uncased` model.[12] This results in $|s| * 10$ BERT representations for each scale. We repeated the procedure for each BERT layer.

5.3.2 Adjective Ranking with a Reference Point

In a first experiment, we explored whether BERT encodes knowledge about adjective intensity. We used as reference point the adjective with the highest intensity (a_{ext}) in a scale s. We rank $\forall a \in s$ where $a \neq a_{ext}$ by intensity by measuring the cosine similarity of their representation to that of a_{ext} in the ten ukWaC sentences retained for s, and in every BERT layer. For example, to rank [*pretty, beautiful, gorgeous*] we measure the similarity of the representations of *pretty* and *beautiful* to that of *gorgeous*. We then average the similarities obtained for each a and use these values for ranking.

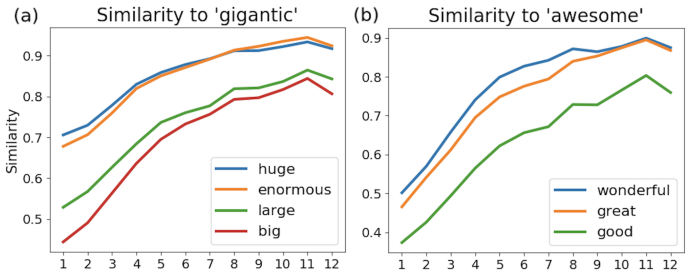


Figure 5.1: Examples of BERTSIM ranking predictions for two adjective scales from WILKINSON.

We evaluate the quality of the ranking obtained for a scale by measuring its correlation with the gold standard ranking in the corresponding dataset D. We use Kendall's tau and Spearman's rho correlation coefficients. We report correlations as a weighted average using the number of adjective pairs in a scale as weights. We also measure a model's pairwise accuracy (P-ACC) which shows whether it correctly predicted the relative intensity ($<$, $>$, $=$) for each pair a_i-$a_j \in s$ with $i \neq j$. [13] We compare the BERTSIM method to two baselines which rank adjectives by frequency (FREQ) and number of senses (SENSE). Our assumption is that mild words (e.g., *good, old*) are more frequent and polysemous than their extreme counterparts on the same scale (e.g., *awesome, ancient*). We rank adjectives by their frequency counts in Google Ngrams (Brants and Franz, 2006). In SENSE, adjectives are ranked according to their number

[11] This score was shown to work better on a development set that we constructed from CoInCo, where it was compared with a BERT-based perplexity score, and with the probability assigned to each adjective in a slot filling task.

[12] When an adjective is split into multiple wordpieces (Wu et al., 2016), we average them to obtain its representation.

[13] We exclude scales where there is only one adjective ($|s| = 1$) apart from a_{ext} (26 out of 79 scales in CROWD; 9 out of 21 scales in WILKINSON).

of senses in WordNet. Accuracy and correlation are moderate to high in the three datasets, showing that the similarity of scalar adjectives' BERT representations reflects the notion of intensity. The good results obtained by the FREQ and SENSE baselines (especially on CROWD) highlight the relevance of frequency and polysemy for scalar adjective ranking.

The results of this evaluation are presented in Table 5.3. Overall, similarities derived from BERT representations encode the notion of intensity, as shown by the moderate to high accuracy and correlation in the three datasets. The good results obtained by the FREQ and SENSE baselines (especially on CROWD) highlight the relevance of frequency and polysemy for scalar adjective ranking.

Dataset	Metric	BERTSIM	FREQ	SENSE
DEMELO	P-ACC	**0.591**$_{11}$	0.571	0.493
	τ	**0.364**$_{11}$	0.304	0.192
	ρ_{avg}	**0.389**$_{11}$	0.309	0.211
CROWD	P-ACC	**0.646**$_{11}$	0.608	0.570
	τ	**0.498**$_{11}$	0.404	0.428
	ρ_{avg}	0.494$_{11}$	0.499	**0.537**
WILKINSON	P-ACC	**0.913**$_{9}$	0.739$_{9}$	0.739$_{9}$
	τ	**0.826**$_{9}$	0.478	0.586
	ρ_{avg}	**0.724**$_{9}$	0.345	0.493

Table 5.3: BERTSIM results using contextualised representations from ukWaC. Subscripts denote the best-performing BERT layer.

Figure 5.1 shows BERT ranking predictions across layers for two adjective scales from WILKINSON:[14] (a) [*big* → *large* → *enormous* → *huge* → *gigantic*], (b) [*good* → *great* → *wonderful* → *awesome*]. Adjectives are ranked according to their similarity to the extreme adjectives in these scales, *gigantic* and *awesome*. Predictions are generally stable and reasonable across layers. We observe that *huge* and *enormous*, which have similar intensity, are inverted in some layers, but are not confused with adjectives further down the scale (*large, big*). Same for *wonderful* and *great* in figure (b), which are very close in the last layers of the model but are not confused with *good*.

5.3.3 Identifying an Intensity Direction in Vector Space

In real life scenarios, however, no concrete reference points (e.g., a_{ext}) are available and scalar adjective interpretation needs to be performed on the fly. We need, for example, to recognise that a *great book* is better than a *well-written* one, without necessarily detecting the relationship of these two adjectives to *brilliant*. Our proposed adjective ranking method does not need reference points and allows to perform this type of inference.

The method draws inspiration from word analogies in gender bias work, where a gender subspace is identified in word-embedding space by calculating the main directions spanned by the differences between vectors of gendered word pairs (e.g., $\overrightarrow{he}$ - $\overrightarrow{she}$, $\overrightarrow{man}$ - $\overrightarrow{woman}$) (Bolukbasi et al., 2016; Dev and Phillips, 2019; Ravfogel et al., 2020; Lauscher et al., 2020a). **Our proposition is to obtain an intensity direction by subtracting the representation of a mild intensity adjective a_{mild} from that of an extreme adjective a_{ext} on the same scale.** Given, for example, *pretty* and *gorgeous* which express a similar core meaning (they are both on the BEAUTY scale) but with different intensity, we expect the embedding that would result from their subtraction to represent this notion of intensity or degree. We call this embedding $\overrightarrow{dVec}$.

$$\overrightarrow{dVec} = \overrightarrow{gorgeous} - \overrightarrow{pretty}$$

We can then compare other adjectives' representations to $\overrightarrow{dVec}$, and rank them according to their similarity to this intensity vector: Our assumption is that the closer an adjective is to $\overrightarrow{dVec}$, the more intense it is. We calculate the $\overrightarrow{dVec}$ for each scale $s \in D$ (a dataset from Section 5.2.2) using the most extreme (a_{ext}) and the mildest (a_{mild}) words in s, and subtracting their contextualised vectors. We experiment

[14] The representations are obtained from ukWaC sentences.

			DEMELO (DM)			CROWD (CD)			WILKINSON (WK)		
		Method	P-ACC	τ	ρ_{avg}	P-ACC	τ	ρ_{avg}	P-ACC	τ	ρ_{avg}
BERT	ukWaC	DIFFVEC-DM	-	-	-	**0.739**$_{12}$	**0.674**$_{12}$	**0.753**$_{12}$	0.918$_{6}$	0.836$_{6}$	0.839$_{6}$
		DIFFVEC-CD	0.646$_{8}$	0.431$_{8}$	**0.509**$_{8}$	-	-	-	0.869$_{11}$	0.738$_{11}$	0.829$_{11}$
		DIFFVEC-WK	0.584$_{9}$	0.303$_{9}$	0.313$_{10}$	0.706$_{10}$	0.603$_{9}$	0.687$_{9}$	-	-	-
	Flickr	DIFFVEC-DM	-	-	-	0.730$_{12}$	0.667$_{12}$	0.705$_{10}$	**0.934**$_{9}$	**0.869**$_{9}$	**0.871**$_{9}$
		DIFFVEC-CD	0.620$_{10}$	0.377$_{10}$	0.466$_{10}$	-	-	-	0.902$_{7}$	0.803$_{7}$	0.798$_{7}$
		DIFFVEC-WK	0.579$_{1}$	0.294$_{1}$	0.321$_{1}$	0.702$_{8}$	0.608$_{8}$	0.677$_{8}$	-	-	-
	Random	DIFFVEC-DM	-	-	-	**0.739**$_{12}$	0.673$_{12}$	0.743$_{12}$	0.918$_{6}$	0.836$_{6}$	0.839$_{6}$
		DIFFVEC-CD	0.626$_{8}$	0.388$_{8}$	0.466$_{8}$	-	-	-	0.836$_{12}$	0.672$_{12}$	0.790$_{10}$
		DIFFVEC-WK	0.557$_{9}$	0.246$_{9}$	0.284$_{6}$	0.703$_{8}$	0.598$_{8}$	0.676$_{8}$	-	-	-
word2vec		DIFFVEC-DM	-	-	-	0.657	0.493	0.543	0.787	0.574	0.663
		DIFFVEC-CD	0.633	0.398	0.444	-	-	-	0.803	0.607	0.637
		DIFFVEC-WK	0.593	0.323	0.413	0.618	0.413	0.457	-	-	-
Baseline		FREQ	0.575	0.271	0.283	0.606	0.386	0.452	0.754	0.508	0.517
		SENSE	0.493	0.163	0.165	0.658	0.498	0.595	0.721	0.586	0.575
		Cocos et al. '18	**0.653**	**0.633**	-	0.639	0.495	-	0.754	0.638	-

Table 5.4: Results of our DIFFVEC adjective ranking method on the DEMELO, CROWD, and WILKINSON datasets. We report results with contextualised (BERT) representations obtained from different SENT-SETs (ukWaC, Flickr, Random) and with static (word2vec) vectors. We compare to the frequency (FREQ) and number of senses (SENSE) baselines, and to results from previous work (Cocos et al., 2018). Results for a dataset are missing (-) when the dataset was used for building the $\overrightarrow{dVec}$ vector.

with BERT embeddings from the SENT-SETs generated through substitution as described in the previous section, where the adjectives occur in the same context.

We build a $\overrightarrow{dVec}$ from every sentence c in the set of 10 sentences C for a scale s by subtracting the BERT representation of a_{mild} in c from that of a_{ext} in c. For example, we subtract the vector of *beautiful* in "*This beach is beautiful*" from that of *gorgeous* in the same context after substitution: "*This beach is gorgeous*". The same from all other sentences collected for that scale ("*You look beautiful/gorgeous today*", etc), until we obtain ten $\overrightarrow{dVec}$ vectors. We then average the ten $\overrightarrow{dVec}$'s obtained for s to construct an intensity vector specific for that scale. Subsequently, we construct a global $\overrightarrow{dVec}$ for dataset D by averaging the vectors of $\forall s \in D$. We then use the $\overrightarrow{dVec}$ obtained from a dataset for adjective ranking. For a fair evaluation of the contribution of this vector to this task, we perform a lexical split in the data used for deriving $\overrightarrow{dVec}$ and the data used for testing. Hence, when evaluating on CROWD, we calculate a $\overrightarrow{dVec}$ on DEMELO (DIFFVEC-DM) and one on WILKINSON (DIFFVEC-WK) omitting all scales where a_{ext} or a_{mild} are present in CROWD. We perform similar splits for the other datasets.

We also compare with results obtained using static word2vec embeddings (Mikolov et al., 2013a) trained on Google News.[15] We obtain the $\overrightarrow{dVec}$ for a scale s by simply calculating the difference between the word2vec embeddings of a_{ext} and a_{mild} in s. We also compare our results to the FREQ and SENSE baselines, and to the best results obtained in our previous study (Cocos et al., 2018) where we used information from lexico-syntactic patterns, a SO-CAL intensity-annotated lexicon (Taboada et al., 2011), and paraphrases from PPDB. We show the results of our experiments in Table 5.4. The DIFFVEC method gets remarkably high performance compared to previous results, especially when $\overrightarrow{dVec}$ is calculated with BERT embeddings. With the exception of Kendall's tau and pairwise accuracy on the DEMELO dataset, DIFFVEC outperforms results from previous work and the baselines across the board. We believe the lower correlation scores on the DEMELO dataset to be due to the large amount of ties present in this dataset: 44% of scales in DEMELO contain ties, versus 30% in CROWD and 0%

[15] We use the `magnitude` library (Patel et al., 2018).

			DEMELO			CROWD		
		# Scales	P-ACC	τ	ρ_{avg}	P-ACC	τ	ρ_{avg}
BERT	ukWaC	1 (+)	0.653_{9}	0.438_{9}	0.489_{11}	$\mathbf{0.709}_{12}$	$\mathbf{0.611}_{12}$	$\mathbf{0.670}_{12}$
		1 (−)	0.611_{10}	0.350_{10}	0.424_{11}	0.648_{10}	0.477	0.507_{10}
		5	0.650_{10}	0.430_{10}	0.514_{10}	0.700_{11}	0.595_{10}	0.673_{10}
	Flickr	1 (+)	0.656_{8}	0.449_{8}	0.504_{8}	0.676_{12}	0.552_{8}	0.612_{8}
		1 (−)	0.600_{3}	0.324_{3}	0.375_{5}	0.641_{9}	0.470_{9}	0.502_{9}
		5	0.647_{12}	0.426_{12}	0.498_{11}	0.692_{11}	0.587_{11}	0.640_{11}
	Random	1 (+)	$\mathbf{0.659}_{11}$	$\mathbf{0.451}_{11}$	0.493_{11}	0.691_{11}	0.570_{11}	0.658_{11}
		1 (−)	0.608_{12}	0.340_{12}	0.421_{10}	0.655_{10}	0.490_{10}	0.514_{12}
		5	0.653_{11}	0.442_{11}	$\mathbf{0.538}_{10}$	0.694_{11}	0.582_{11}	0.653_{11}
word2vec		1 (+)	0.602	0.334	0.364	0.624	0.419	0.479
		1 (−)	0.613	0.359	0.412	0.661	0.506	0.559
		5	0.641	0.415	0.438	0.688	0.559	0.601

Table 5.5: Results of DIFFVEC on DEMELO and on CROWD using a single positive (1 (+)) or negative (1 (−)) $a_{ext} - a_{mild}$ pair, and five pairs (5).

in WILKINSON, where we obtain better results. Our models cannot easily predict ties using similarities which are continuous values. The composition of the SENT-SETs used for building BERT representations plays a role on model performance as well. Overall, the selection method described in Section 5.3.1 offers a slight advantage over random selection, with ukWaC and Flickr sentences doing better on different datasets. Overall, the best-performing BERT layers are situated in the upper half of the Transformer network. The only exception is DIFFVEC-WK with the Flickr SENT-SET on DEMELO, where all layers perform similarly. Overall, the FREQ and SENSE baselines get lower performance than our method with BERT embeddings. SENSE manages to give results comparable to DIFFVEC with static embeddings and to previous work (Cocos et al., 2018) in one dataset (CROWD), but is still outperformed by DIFFVEC with contextualised representations.

Given the high performance of the DIFFVEC method in the ranking task, we carry out additional experiments to explore the impact that the choice of scales and sentences has on the quality of the intensity vector. We test the method with a $\overrightarrow{dVec}$ built from a single $a_{ext} - a_{mild}$ pair of either positive (*awesome-good*) or negative (*horrible-bad*) polarity, that we respectively call DIFFVEC-1 (+)/(−). We also experiment by varying the number of scales and the number of sentences used to extract the representations. We specifically add three more scales: *ancient-old*, *gorgeous-pretty* and *hideous-ugly*) to form DIFFVEC-5. The scales are from WILKINSON, so we exclude this dataset from the evaluation.

Results are given in Table 5.5. We observe that a small number of word pairs is enough to build a $\overrightarrow{dVec}$ with competitive performance. Interestingly, DIFFVEC-1 (+) with random sentences obtains the best pairwise accuracy on DEMELO. **The fact that the method performs so well with just a few pairs is very encouraging, making our approach easily applicable to other datasets and languages.** A larger number of scales seems to be beneficial for the method with static word2vec embeddings, which seem to better capture intensity on the negative scale. For BERT, intensity modeled using a positive pair (DIFFVEC (+)) gives best results across the board. The use of five pairs of mixed polarity improves results over a single negative pair, and has comparable performance to the single positive one.

Finally, we compare the performance of DIFFVEC-1 (+)/(−) and DIFFVEC-5 when the contextualised representations are extracted from a single sentence instead of ten. Our main observation is that reducing the number of sentences harms performance, especially when the sentence used is randomly selected.

5.3.4 Multilingual Adjective Ranking

The knowledge-lean adjective ranking method presented in the previous section performs well even with an intensity vector built from a single adjective scale. Given that it is so lightweight, it is easily extendable to new languages. In our Garí Soler and Apidianaki (2021b) paper, we show that vectors representing intensity can be built in other languages for which language models are available. We address French, Spanish and Greek, and build scalar adjective resources in these languages by translating the deMelo and Wilkinson datasets. We make this new dataset, called MULTI-SCALE, available for research purposes. Table 5.6 shows examples of original English scales and their French, Spanish and Greek translations. Table 5.7 contains statistics on the composition of the translated datasets.

	DEMELO
EN	**dim** < **gloomy** < **dark** < **black**
FR	terne < sombre < foncé < noir
ES	sombrío < tenebroso < oscuro < negro
EL	αμυδρός || αχνός < μουντός < σκοτεινός < μαύρος
	WILKINSON
EN	**bad** < **awful** < **terrible** < **horrible**
FR	mauvais < affreux < terrible < horrible
ES	malo < terrible < horrible < horroroso
EL	κακός < απαίσιος < τρομερός < φρικτός

Table 5.6: Example translations from each dataset. The symbol "||" indicates ties.

		# unordered pairs	# adjectives
DEMELO	EN	548 (524)	339 (293)
	FR	590 (567)	350 (303)
	ES	448 (431)	313 (275)
	EL	557 (535)	342 (295)
WILKINSON	EN	61 (61)	59 (58)
	FR	67 (67)	61 (60)
	ES	59 (59)	58 (56)
	EL	68 (68)	61 (58)

Table 5.7: Composition of the translated datasets with the number of unique adjectives and pairs in parentheses.

In order to test contextual models on the ranking task, we collect sentences containing the adjectives from the OSCAR corpus (Suárez et al., 2019), a multilingual corpus derived from CommonCrawl. French, Spanish and Greek are morphologically rich languages where adjectives need to agree in number and gender with the noun they modify. In order to keep the method resource-light, we gather sentences that contain the adjectives in their unmarked (non-inflected) form. For each scale s, we randomly select ten sentences from OSCAR where adjectives from s occur. Then, we generate additional sentences through lexical substitution. Specifically, for every sentence (context) c that contains an adjective a_i from scale s, we replace a_i with $\forall\ a_j \in s$ where $j = 1...|s|$ and $j \neq i$. This process results in a total of $|s|$ * 10 sentences per scale and ensures that $\forall\ a \in s$ is seen in the same ten contexts.

We conduct experiments with state-of-the-art contextual language models and several baselines on the MULTI-SCALE dataset. We use the pre-trained `cased` and `uncased` multilingual BERT model (Devlin et al., 2019) and report results of the best variant for each language. We also report results obtained with four monolingual models: `bert-base-uncased` (Devlin et al., 2019), `flaubert_base_uncased` (Le et al., 2020), `bert-base-spanish-wwm-uncased` (Cañete et al., 2020), and `bert-base-greek-uncased-v1` (Koutsikakis et al., 2020). We compare to results obtained using fastText static embeddings in each language (Grave et al., 2018). For a scale s, we feed the corresponding set of sentences to a model and extract the contextualised representations for $\forall\ a \in s$ from every layer. When an adjective is split into multiple BPE units, we average the representations of all wordpieces (we call this approach "WP") or all pieces but the last one ("WP-1"). The intuition behind excluding the last WP is that the ending of a word often corresponds to a suffix with morphological information.

The DIFFVEC method. We apply the adjective ranking method that we proposed in our EMNLP 2020 paper (Garí Soler and Apidianaki, 2020) to the MULTI-SCALE dataset. The method relies on an inten-

sity vector (called $\overrightarrow{dVec}$) built from BERT representations and yields state-of-the art results with very little data. This makes it easily adaptable to new languages. We build a sentence specific intensity representation ($\overrightarrow{dVec}$) in each language by subtracting the vector of a mild intensity adjective, a_{mild} from that of a_{ext}, an extreme adjective on the same scale in the same context. We create a *dVec* representation from every sentence available for these two reference adjectives, and average them to obtain the global $\overrightarrow{dVec}$ for that pair.

In Garí Soler and Apidianaki (2020), we showed that a single positive adjective pair (DIFFVEC-1 $(+)$) is enough for obtaining highly competitive results in English. We apply this method to the other languages using the translations of a positive English (a_{mild}, a_{ext}) pair from the CROWD dataset. We select the pair (*perfect*,*good*). Its translations are (*parfait*,*bon*) in French, (*perfecto*,*bueno*) in Spanish, and (τέλειος,καλός) in Greek. Additionally, we learn two dataset specific representations: one by averaging the $\overrightarrow{dVec}$'s of all (a_{ext}, a_{mild}) pairs in WILKINSON that do not appear in DEMELO (DIFFVEC-WK), and another one from pairs in DEMELO that are not in WILKINSON (DIFFVEC-DM). We rank adjectives in a scale by their cosine similarity to each $\overrightarrow{dVec}$: The higher the similarity, the more intense the adjective is.

Baselines. We compare our results to a frequency and a polysemy baseline (FREQ and SENSE). The idea is that low intensity words (e.g., *nice*, *old*) are more frequent and polysemous than their extreme counterparts (e.g., *awesome*, *ancient*). Extreme adjectives often limit the denotation of a noun to a smaller class of referents than mild intensity adjectives (Geurts, 2010). For example, an "awesome view" is more rare than a "nice view". This assumption has been confirmed for English in Garí Soler and Apidianaki (2020). FREQ orders words in a scale according to their frequency. Words with higher frequency have lower intensity. Given the strong correlation between word frequency and number of senses (Zipf, 1945), we also expect highly polysemous words (which are generally more frequent) to have lower intensity. This is captured by the SENSE baseline which orders the words according to their number of senses; words with more senses have lower intensity.

Frequency is taken from Google Ngrams for English, and from OSCAR for the other three languages. The number of senses is retrieved from WordNet for English, and from BabelNet (Navigli and Ponzetto, 2012) for Spanish and French.[16] For adjectives that are not present in BabelNet, we use a default value which corresponds to the average number of senses for adjectives in the dataset (DEMELO or WILKINSON) for which this information is available. We omit the SENSE baseline for Greek due to low coverage.[17]

Adjective Ranking Results. We use the same evaluation metrics as in previous work (de Melo and Bansal, 2013; Cocos et al., 2018; Garí Soler and Apidianaki, 2020). We compare the predicted ordering for every adjective pair in a scale ($<$, $>$, $=$) to the gold ordering in a dataset using pairwise accuracy (P-ACC). We evaluate the ordering for full scales with Kendall's τ and Spearman's ρ correlation. The results are given in Table 5.8. The reported correlation values are a weighted average of the correlations obtained for each scale in a dataset (DM: deMelo, WK: Wilkinson) with the number of adjective pairs in each scale as weights. Monolingual models perform consistently better than the multilingual model, except in French. We report the best wordpiece approach for each model: WP-1 works better with all monolingual models and the multilingual model for English. Using all wordpieces (WP) is a better choice for the multilingual model in other languages. We believe the lower performance of WP-1

[16] We omit Named Entities from BabelNet entries (e.g., names of TV shows or locations).

[17] Only 47% of the Greek adjectives have a BabelNet entry, compared to 95.7% and 88.9% for Spanish and French. All English adjectives are present in WordNet.

		EN			FR			ES			EL		
		Mono WP-1			**Mono WP-1**			**Mono WP-1**			**Mono WP-1**		
		P-ACC	τ	ρ_{avg}	P-ACC	τ	ρ_{avg}	P-ACC	τ	ρ_{avg}	P-ACC	τ	ρ_{avg}
DM	DV-1 (+)	**.651**$_9$	**.435**$_9$	**.496**$_9$	**.610**$_3$	**.369**$_3$	**.396**$_3$	.658$_9$	.381$_9$	**.407**$_9$	.564$_2$	.238$_1$	.271$_2$
DM	DV-WK	.586$_6$	.267$_6$	.300$_6$	.515$_1$	.167$_1$	.166$_7$	**.670**$_7$	**.404**$_7$	**.407**$_7$	.589$_2$	.294$_2$	.325$_2$
WK	DV-1 (+)	.852$_1$	.705$_1$	.802$_1$	.612$_6$	.257$_6$	.215$_6$	**.814**$_7$	**.627**$_7$	**.803**$_9$	.618$_8$	.282$_8$	.256$_8$
WK	DV-DM	**.918**$_{10}$	**.836**$_{10}$	**.859**$_{10}$	.642$_7$	.322$_2$	.392$_2$	.780$_6$	.559$_6$	.684$_6$	**.750**$_{10}$	**.564**$_{10}$	**.586**$_{10}$
		Multi WP-1			**Multi WP**			**Multi WP**			**Multi (unc) WP**		
DM	DV-1 (+)	.609$_4$	.346$_4$	.389$_4$	.559$_7$	.260$_7$	.311$_7$	.614$_3$	.291$_3$	.268$_5$	.517$_9$	.139$_9$	.163$_9$
DM	DV-WK	.544$_3$	.208$_3$	.241$_4$	.517$_{10}$	.170$_{10}$	.179$_{10}$	.618$_{12}$	.301$_{12}$	.303$_{12}$	.539$_9$	.181$_9$	.207$_9$
WK	DV-1 (+)	.836$_6$	.672$_6$	.717$_6$	.672$_3$	.382$_3$	.380$_3$	.797$_3$	.593$_3$	.639$_3$	.662$_{10}$	.388$_9$	.423$_9$
WK	DV-DM	.836$_7$	.672$_7$	.766$_7$	**.701**$_6$	**.441**$_6$	**.476**$_2$	.695$_{10}$	.390$_{10}$	.511$_{10}$	.691$_5$	.447$_5$	.502$_5$
		Static models and baselines											
DM	DV-1 (+)	.637	.407	.458	.573	.288	.275	.656	.383	**.421**	.575	.266	.273
DM	DV-WK	.599	.330	.406	.454	.033	-.006	.616	.298	.315	.549	.205	.217
DM	FREQ	.575	.271	.283	.602	.346	.345	.585	.227	.239	**.596**	**.306**	**.334**
DM	SENSE	.493	.163	.165	.512	.229	.185	.516	.139	.151	-	-	-
WK	DV-1 (+)	.787	.574	.663	.582	.197	.152	.695	.390	.603	.706	.464	.566
WK	DV-DM	.852	.705	.783	.642	.325	.280	.712	.424	.547	.691	.447	.451
WK	FREQ	.754	.508	.517	.567	.167	.148	.576	.153	.382	.676	.417	.427
WK	SENSE	.721	.586	.575	.567	.255	.340	.644	.411	.456	-	-	-

Table 5.8: Results of the DIFFVEC (DV) method in English, French, Spanish and Greek with monolingual (Mono) and multilingual (Multi) contextual models, and static embeddings. Subscripts denote the best layer. The best result obtained for each dataset in each language is indicated in boldface.

in these settings to be due to the fact that the multilingual BPE vocabulary is mostly English-driven. This naturally results in highly arbitrary partitionings in these languages (e.g., ES: *fantástico* → fant-ástico; EL: γιγάντιος (*gigantic*)→γ-ι-γ-άν-τιος). On the contrary, the tokenisers of the monolingual models tend to split words in a way that more closely reflects the morphology of the language (e.g., ES: *fantástico* → fantás-tico; EL: γιγάντιος→γιγά-ντι-ος.

We observe that DIFFVEC-1 (+) yields comparable and sometimes better results than DIFFVEC-DM and DIFFVEC-WK, which are built from multiple pairs. This is important especially in the multilingual setting, since it shows that just one pair of adjectives is enough for obtaining good results in a new language. The best layer varies across models and configurations. The monolingual French and Greek models, and the multilingual model for English, generally obtain best results in earlier layers. For the other models performance improves in the upper half of the Transformer network (layers 6-12). This shows that the semantic information relevant for adjective ranking is not situated at the same level of the Transformer in different languages. We plan to investigate this finding further in future work.

The lower results in French can be due to the higher amount of ties present in the datasets compared to other languages.[18] The baselines obtain competitive results showing that the underlying linguistic intuitions hold across languages. The best models beat the baselines in all configurations except for Greek on the DEMELO dataset, where FREQ and static embeddings obtain higher results. Overall, results are lower than those reported for English. This shows that there is room for improvement in new languages.

We proposed a new multilingual benchmark for scalar adjective ranking, and set performance baselines on it using monolingual and multilingual contextual language model representations. Our results show that adjective intensity information is present in the contextualised representations in the studied lan-

[18] 58% of the French DEMELO scales contain a tie, compared to 45% in English.

guages. We have made the scalar adjective datasets in the three languages, and the sentence contexts, available to promote future research on scalar adjectives detection and analysis in different languages.

5.4 Conclusion

This section has presented two methods to adjective intensity detection which have been applied to the scalar adjective ranking task. The first method combines information from a large paraphrase resource with information acquired from corpora using patterns, and information found in polarity lexicons. The combination of these sources of information allows to leverage the strengths of the paraphrase-based approach, namely its high coverage, as well as the high precision guaranteed by the pattern and lexicon-based approaches. The second paper presented in this section demonstrates how adjective intensity can be captured and described in the space built by contextual language models. We have shown that a method which relies on simple calculations in the generated vector space obtains similar performance in the scalar adjective ranking task as the resource-rich approaches previously used. These results highlight the richness of the lexical semantic information that is encoded in contextualised word representations. We demonstrate an efficient way to retrieve this information in a controlled experimental setting which allows to reduce the strong impact of context variation and to reason about the meaning of words.

We have extended our method to French, Spanish and Greek, using the BERT-like models that are available in these languages. In order to evaluate the performance of the method in these languages, we translated two of our evaluation datasets, which we made available for future research. The results from these experiments showed that the vector spaces built by contextual models in the other languages also encode the notion of intensity, which can be retrieved and used for ranking words along this dimension.

It is our belief that these results open up avenues for future work on other abstract semantic notions which might be encoded in the vector space of contextual models (Edmonds and Hirst, 2002; Allaway and McKeown, 2021). We would, for example, expect to find traces of other types of connotations, such as formality and complexity, but also politeness and register, as well as connotations which might indicate a specific ideological or political stance (Webson et al., 2020; Romanov et al., 2019).

Chapter 6

Conclusion and Perspectives for Future Research

6.1 Illustrating the Paradigm Shift in Lexical Semantics

This synthesis document presents a selection of articles that have been published in the past ten years, adopting a comparative perspective which juxtaposes older and more recent research methodologies used in the field of lexical semantics. I selected a sample of papers which address specific research questions following a different approach and experimental paradigm. In each set of papers, the earlier study involves the use of external knowledge sources or annotations. In our clusterability study, for example, presented in Section 3.2, the analysis relies on manual substitute and translation annotations (McCarthy et al., 2016). For our study of lexical substitution (Apidianaki, 2016) presented in Section 4.2, and our work on adjective intensity detection in Section 5.2 (Cocos et al., 2018), our main source of knowledge has been the unigram paraphrases of words in the Paraphrase Database (PPDB) (Ganitkevitch et al., 2013; Pavlick et al., 2015). In the latter study, these were combined with information extracted from large corpora using patterns, and with emotion intensity scores from semantic orientation lexicons (Taboada et al., 2011).

In contrast, in the more recent papers present in the three sets, we leverage the knowledge that is encoded in neural language models to perform the tasks at hand (Garí Soler et al., 2019b; Garí Soler and Apidianaki, 2020, 2021a,b). We show that information about polysemy is encoded in the BERT model during pre-training where it is exposed to massive amounts of text data, and is complemented with information from new contexts of use. Additionally, we demonstrate that due to the rich lexical knowledge that is encoded, BERT representations can serve to estimate the clusterability of lemmas without need for manual annotations. Notably, these representations yield even better results on this task than the approach which relied on manually defined in-context substitutes (McCarthy et al., 2016). The sole reliance on words' contextualised representations permits to scale the method up in order to cover a much larger vocabulary than previous studies, and to extend its applicability to other languages where language models are available.

In our earlier work on lexical substitution, we demonstrated how syntax-based distributional models performed on this task. Our later study presents a comparison of static (GloVe and FastText) and contextualised (ELMo) embedding representations, which we also compare to representations generated by dedicated lexical substitution models (Melamud et al., 2015, 2016). Our results show that models that were specifically trained with a lexical substitution objective outperform all other models and types

of representations on this task.

Finally, in our Cocos et al. (2018) study on adjective intensity, we demonstrated that paraphrases can be used as a proxy for identifying this dimension of meaning, possibly combined with pattern-based information and polarity lexicons. In our Garí Soler and Apidianaki (2020) work, we showed that it is possible to identify a dimension in the vector space built by BERT which describes this abstract semantic notion. This can be easily discovered through simple calculations in the vector space, provided that a few seed examples illustrating this relation are available. In the case of scalar adjectives, a single pair of adjectives with different intensity was enough for obtaining a vector describing this dimension. The intensity vector served for scalar adjective ranking through similarity calculations in the BERT vector space, outperforming previous pattern- and resource-based methodologies. We also demonstrated the straightforward application of the method to new languages, using both monolingual language models specifically trained for the language and multilingual models (Garí Soler and Apidianaki, 2021b). Our adjective ranking experiments in French, Spanish and Greek showed that although intensity differences are less pronounced in multilingual models, they are still identifiable and can be used for reasoning about adjectives' relationships.

The comparisons presented in this report serve to illustrate the paradigm shift that has occurred in the field of lexical semantics in particular and, more generally, in computational linguistics. In each of the three sets of papers, the latter work which involves the use of neural language models outperforms the earlier work which involves the use of distributional methods and external resources. These promising findings highlight the superiority of neural models compared to traditional methods for lexical semantic analysis. Due to the rich information they encode about words and their meanings, their representations are more efficient for performing semantic tasks. Importantly, as soon as such language models are available, the application of the proposed methodology to new languages is straightforward without need for additional semantic resources.

In what follows, we present some shortcomings of using neural models for lexical semantic analysis and discuss perspectives for future work.

6.2 Semantic Knowledge in Neural Language Models

6.2.1 On the Systematicity of the Encoded Knowledge

We have shown that the representations that are generated by contextual language models encode rich information about lexical polysemy. Our results revealed clear patterns in English, as well as in morphologically-richer (and resource-poorer) languages. Other studies have shown that contextualised representations also encode a good amount of encyclopedic knowledge (Petroni et al., 2019; Bouraoui et al., 2020) which is acquired by the models during their training on data encoding such knowledge (for example, Wikipedia). The results from experiments that explore other types of semantic knowledge have not, however, been that encouraging. Interestingly, it has been shown that even though a model might succeed in a probing task, this does not mean that it "understands" the concept it is being queried for, or that it encodes systematic knowledge about it.

In a compelling study, Ravichander et al. (2020) probe BERT representations for hypernymy and show that although the model is able to retrieve hypernyms in cloze tasks, this does not correspond to systematic knowledge about this relationship. They propose a set of diagnostics to examine how systematically this knowledge generalises, where they evaluate the model's ability to consistently answer queries reflecting the understanding of a concept. They specifically combine related zero-shot probes based on

the assumption that if a model succeeds on one probe and is drawing on a systematic general ability, then it should also succeed on the paired probe.[1] They also examine the model's ability to recognise the correct abstraction for a word (hyponym) in context.[2] Their results show that high probing accuracy for a particular competence does not necessarily follow that BERT understands a concept, and it cannot be expected to systematically generalise across applicable contexts.

Ettinger (2020) also demonstrates that although a model might encode encyclopedic knowledge (for example, about US Presidents and their biography), it does not succeed in recognising negation in simple sentences. This failure has severe consequences on the model's reasoning capabilities. **It is important to propose diagnostic devices which test the systematicity of different types of knowledge in neural models, in order to be able to draw safe conclusions about the types of information these models encode, and their actual understanding of linguistic phenomena.** This will help to better explain their behaviour and also to develop methods for enhancing the quality of the encoded knowledge, when needed, and to determine when it would be necessary to complement this with information from other modalities.

Lastly, **it is important to ensure that the proposed diagnostic devices are applicable to different types of models.** Probing and other interpretation methods are often proposed with respect to a specific model and a specific architecture (they might, for example, analyse the observed attention patterns) and cannot thus be applied to other types of models. So, alongside the development of diagnostic suites and controls, we need to ensure that the proposed probing methodology is generalisable to different types of models and across different implementations of the same model (Sellam et al., 2022), and that it can allow cross-comparison between them. **Determining the right prompts to use for extracting the information of interest is also a challenge.** These can be manually crafted, as in the studies cited above, or automatically generated. The latter can be discrete (or "hard") (Gao et al., 2021; Jiang et al., 2020; Shin et al., 2020) or continuous (or "soft") (Li and Liang, 2021; Lester et al., 2021; Zhong et al., 2021), each type exhibiting specific advantages and shortcomings. Continuous prompts, in particular, might present generalisation issues since the embedding spaces of different models might not be aligned.

6.2.2 Abstract Semantic Notions in Vector Space

We believe that the results obtained so far regarding the linguistic knowledge that is encoded in neural models are promising. Apart from information about linguistic structure, lexical semantic phenomena and common-sense knowledge, we have shown that it is possible to identify dimensions in the vector space built by contextual language models which describe abstract semantic notions such as intensity. **We believe that this type of investigation could be extended and applied to other semantic notions such as formality** (e.g., *get/acquire/obtain/snag*, *wealthy/rich*), **complexity** (e.g., *medical practitioner/doctor*, *prevalent/very common*) **and politeness.** These can serve to analyse the emotional load of texts, the ideological position of the author, as well as situational factors such as the context of the communication and the relationship of the participants.

By applying the methodology that we proposed for analysing scalar adjectives (Garí Soler and Apidianaki, 2020) to these notions, it would be worth investigating if they are also encoded in the vector space, in the same way as intensity. Illustrating these notions and identifying the relevant dimensions in the vector space using a few seed examples (as in the case of scalar adjectives) would be highly useful for semantic processing since it would permit to analyse the connotations (or subtle meanings)

[1] For example, a model which understands hypernymy is expected to answer queries about this relationship which contain nouns in singular as well as in plural form (e.g., *A robin is a* [MASK] / *robins are* [MASK]).

[2] For example, *A robin perches in its nest* → *A* [MASK] *perches in its nest*, where the hypernym *bird* is acceptable.

conveyed by words apart from their denotation (literal meaning). These are relevant for numerous tasks since they express authors' ideological attitudes and stance, as well as their cultural and emotional perspectives (Clark, 1992; Edmonds and Hirst, 2002; Webson et al., 2020; Allaway and McKeown, 2021). It has, for example, been shown that the vocabulary used in a text can, for example, reflect the authors' positioning towards a specific political situation (e.g., *undocumented workers* vs. *illegal aliens*) (Webson et al., 2020).

6.3 Probing and Explainability

6.3.1 Use of the Encoded Knowledge for Prediction

The majority of interpretability studies rely on probing in order to explain what the high performance of artificial neural networks is due to. They investigate the knowledge encoded inside these blackboxes and their internal workings, with the goal to determine whether neural models acquire the abstractions we intuitively believe are important for common-sense reasoning. Although probing results have revealed the rich linguistic and world knowledge that is encoded in language model representations (Linzen et al., 2016; Hewitt and Manning, 2019; Tenney et al., 2019a; Petroni et al., 2019; Vulić et al., 2020b; Ravichander et al., 2020; Ettinger, 2020; Garí Soler and Apidianaki, 2021a), it is hard to say whether this knowledge is actually used by the models to perform specific tasks. Consequently, it is hard to establish a causality relation between the encoded knowledge and downstream performance, which casts doubt on the value of probing as a tool for explaining model behaviour.

Interestingly, recent studies provide increasing evidence that ANNs do not really use the encoded knowledge for performing the tasks they are trained for (Elazar et al., 2021; Ravichander et al., 2021). Identifying the knowledge that is actually used for reasoning is difficult with the existing interpretation methodology which mainly reveals correlations between the encoded properties (Belinkov, 2021; Feder et al., 2021). Probing can serve to indicate such correlations between model representations and linguistic properties, but it does not tell us whether these are actually involved in the predictions made by the original model (Feder et al., 2021). This uncertainty is accentuated by the disconnect between the probing classifier used and the original model which are separately trained (Belinkov, 2021). Importantly, results of adversarial studies show that state-of-the-art models are highly vulnerable to slight perturbations of the processed data, a behaviour that would not be expected if the models could understand language and make informed decisions (Jin et al., 2020).

6.3.2 Counterfactual Methods

Counterfactual methods aim at estimating whether the knowledge encoded in model representations is actually used for prediction. The Amnesic Probing method proposed by Elazar et al. (2021), for example, uses counterfactual representations derived from pre-trained model representations where the property of interest (e.g., number agreement or syntactic dependencies information) has been removed. The change in model behaviour that occurs after this neutralisation procedure allows to measure how useful the information was for the task at hand.[3] Feder et al. (2021) also argue in favour of the use of counterfactual examples which reveal the causal effect of a concept of interest on the performance of a given model, rather than simply showing correlations between features and predictions.

These counterfactual methods mainly address grammatical and structural linguistic aspects (e.g., part-

[3]The idea is similar to that underlying ablation studies where some component is being removed and the influence of the intervention on the result is being measured.

of-speech, tense or number agreement) (Elazar et al., 2021; Ravfogel et al., 2021). In our study (Celikkanat et al., 2020), we used the Amnesic Probing method for detecting the traces of passivisation and negation in contextualised representations, in a controlled experimental setting.

In future work, such methods should be extended and applied to other types of linguistic (structural and semantic) phenomena. This would make it possible to perform a more thorough analysis of the impact of the rich knowledge that is encoded in language models on results in different tasks. Given that most neural models are blackboxes, interpretability constitutes a real and very important challenge. It is also a crucial step for system development, since it provides the possibility to decipher the complex decision mechanisms inside the models, and the reasons that lead to certain predictions. Finally, it is essential for increasing the trustworthiness of the models and, consequently, the users' confidence in them, with real impact in their everyday life.

6.3.3 Adversarial Methods

Another indication that high performing models might not rely on the knowledge that is encoded in their representations for performing natural language understanding tasks, is that they are highly vulnerable to adversarial attacks. This means that they can be fooled when exposed to adversarial examples that have imperceptible alterations from their original counterparts, a behaviour that would not be expected if they a real "understanding" of the processed text and the task was taking place.

The Suzuki DR-Z400 is a dual purpose motorcycle manufactured by Suzuki. Suzuki DR-Z400

(a) Original text sample.

bgaaeaTheaSuzukicDR-Z400aisaaadualepurposecmotorcyclecmanufacturedabyaSuzuki.bSuzukiaDR-Z400

(b) Adversarial sample generated with FGSM directly.

The Suzuki DR-Z400 is a*f* dual pur*b*o*c*e m*a*torcycle*b*ma*a*ufacture*a* by Suzuki. Suzuki DR-Z400

(c) Manipulating top characters (shown in ***red italic***).

Figure 6.1: Adversarial text generated using FGSM.

Adversarial methodology is a useful tool for evaluating model robustness (Goodfellow et al., 2015; Li et al., 2017; Alzantot et al., 2018; Jin et al., 2020). The idea is to impose minor data perturbations which preserve the meaning of the original text, and to evaluate model performance on the generated examples. If the models rely on their encoded semantic and common-sense knowledge for reasoning, they are expected to continue making correct predictions on the altered data. Specifically, they would be able to identify the similarity of the adversarial and the original examples, and would not alter their predictions (i.e. after-attack accuracy would be the same as original accuracy). Results from adversarial studies show that this is not generally the case, and that drops in performance when models are exposed to perturbated data are huge.[4]

Adversarial attacks initially became popular in the field of image processing where example generation is simple, since small perturbations to many pixels might not be perceptible to a human viewer (Szegedy et al., 2014; Goodfellow et al., 2015). Generating adversarial examples for text data is challenging since due to the discrete nature of word tokens (as opposed to the continuous nature of image pixel values), changes are perceptible (Alzantot et al., 2018). Liang et al. (2018) show that applying a gradient-based method such as the ones used in the image domain to text data generates unintelligible text. An example of this manipulation is given in Figure 6.1 (b) where the Fast Gradient Sign Method (FGSM) of Goodfellow et al. (2015) has been applied. Manipulating only a few characters with the highest gradient magnitude still generates unnatural text with noticeable perturbations (Figure 6.1 (c)). Given

[4]Indicatively, Jin et al. (2020) report a reduction in BERT prediction accuracy by about 5-7 times on the Yelp reviews sentiment classification task (Zhang et al., 2015), and a drop from 89.4% to 4.0% on the SNLI dataset (Bowman et al., 2015) for the same model.

the quality of the altered text, a model's incapability to properly manipulate it would not reveal much about its understanding of language. A natural language attacking system is, instead, expected to satisfy three utility-preserving properties (Jin et al., 2020):

1. **human prediction consistency**: human predictions on the altered text should remain unchanged;
2. **semantic similarity**: the crafted example should bear the same meaning as the source;
3. **language fluency**: the generated examples should look natural and grammatical.

Manually changing every example of interest by introducing, or removing, a concept of choice is costly, time-consuming, and therefore implausible for large data sets (Feder et al., 2021). Heuristic attack methods for textual data involve simulating typos by randomly replacing characters with their nearby key on the keyboard (Belinkov and Bisk, 2018); misspelling words and adding punctuation between the letters (Hosseini et al., 2017); word scrambling, insertion and deletion (Li et al., 2017; Alzantot et al., 2018; Liang et al., 2018; Li et al., 2019). However, with such heuristic strategies, it still remains challenging to find the optimal solution in the massive space of possible combinations of insertions, replacements and deletions, while preserving semantic consistency and language fluency (Li et al., 2020). Furthermore, word deletion and insertion might result in unnatural sentences, where syntactic and semantic plausibility are compromised. An alternative is to use a lexical substitution method, which guarantees higher semantic similarity of the original and generated texts than alternative procedures (e.g., word deletion). A good substitution system would also preserve the fluency of the generated text. In existing methods (Alzantot et al., 2018; Tsai et al., 2019; Jin et al., 2020), substitution candidates for a target word are often its nearest neighbours in the used embedding space. These might, however, be antonyms of the target or words of a different grammatical category which happen to be close in space; these substitutes would, however, compromise both fluency and semantic consistency. Furthermore, the use of static embeddings does not guarantee that the selected substitutes are good fit in context due to the meaning conflation problem which influences similarity.

The potential of using contextual language models as lexical substitution tools (as discussed in Section 4) would permit to address these issues. Lexical substitution models, precisely, address both the in-context suitability of candidate substitutes, preserving fluency, and their similarity to the target (the word to be substituted) (Melamud et al., 2015, 2016; Li et al., 2020). Consequently, they satisfy the above constraints of adversarial methods to a larger degree than alternative approaches, ensuring the fluency of the generated text and its similarity to the original.

Adversarial methods constitute a valuable tool for assessing the knowledge language models encode, their understanding of it and the extent to which it is used for accomplishing specific tasks. **Future work should focus on developing adversarial methods which satisfy the three utility-preserving constraints (human prediction consistency, similarity and fluency), in order to increase the usefulness of such methods.** An interesting research direction would be to propose an adversarial framework where the candidate substitutes describe the concepts of interest. By implementing controlled data perturbations (for example, along specific axes of meaning variation), such an adversarial framework would allow to investigate the impact of the concepts of interest on model behaviour. We could, for example, restrict the candidates to words with varying (increasing or decreasing) intensity (e.g., *good* → *great* → *wonderful* → *awesome*; *love* → *adore*; *hate* → *despise*; *secretly* → *confidentially*) (Garí Soler and Apidianaki, 2020), and measure the impact of intensity on prediction (for example, in a sentiment classification task). This could be extended to other concepts and abstract semantic notions that may be modelled in the vector space built by contextual models. We could, for example, explore the impact of the graded notion of lexical complexity (Shardlow, 2013; Paetzold and Specia, 2016; Kriz et al.,

2018) using a model that ranks words according to this dimension (e.g., *detention* → *imprisonment* → *incarceration*) and a lexical substitution method for generating text at different complexity levels (Kriz et al., 2018). Similarly for formality (e.g., *swing* → *oscillate*, *clean* → *immaculate*) and the impact of this stylistic dimension on model decisions (Pavlick and Tetreault, 2016).

6.4 Exploration of Other Types of Semantic Knowledge

6.4.1 Regular Meaning Alternations

Encoding rich information about word usage, contextualised representations can open up new avenues for Formal Semantics research. Thanks to their great sensitivity to syntactic variation and the surrounding context, they can provide representations of the descriptive content of words and also reflect minor meaning shifts. It has been shown that they can also serve to describe types of polysemy that are central in Formal Semantics but difficult to handle with traditional DSM models. These involve related (or complementary) senses which are less sensible to contextual priming than contrastive senses, and thus more difficult to disambiguate (Haber and Poesio, 2020, 2021).

Modelling regular polysemy, and other types of meaning alternations (such as metonymy or metaphor), using contextualised representations is a research question worth exploring in future work. The term regular polysemy describes types of polysemy which are not word specific, but are rather instances of general sense alternations such as ANIMAL/MEAT (*lamb*, *chicken*), FOOD/EVENT (*lunch*, *dinner*), CONTAINER/CONTENT (*glass*, *bottle*, *cup*), PHYSICAL/INFORMATION/ORGANISATION (*newspaper*, *magazine*), BUILDING/PUPILS/DIRECTORATE/INSTITUTION (*school*, *university*) (Boleda et al., 2012a; Haber and Poesio, 2021). A large number of linguistics and cognitive science studies have provided evidence about these meaning variation regularities which are due to general analogical processes (Apresjan, 1974; Lakoff and Johnson, 1980; Copestake and Briscoe, 1995; Pustejovsky, 1995).

Regular alternations of meaning are rather common in languages, however lexical semantic analysis methods often analyse the semantics of individual words separately, ignoring the similarities of their meaning alternations. Boleda et al. (2012a) proposed to address this phenomenon using a distributional model which assesses how well a "meta alternation" (e.g., ANIMAL-FOOD) explains a pair of senses of a lemma (e.g., *lamb* or *chicken*).[5] In their model, "meta senses" are represented by a vector which is defined as the centroid (average vector) of the monosemous words instantiating it. Meta alternations are then represented by the centroids of their meta senses' vectors. An advantage of this method is that the lemmas do not need to be disambiguated, but meta alternation and polysemous words are represented as simple centroid vectors. The method does not involve a word sense induction step (Schütze, 1998; Pantel and Lin, 2002) which, as pointed out by Boleda et al. (2012a), would allow for more flexible and realistic models.

Regarding coercion and logical metonymy, these have been traditionally approached in the distributional semantics literature using an approach to thematic fit modelling which considers for each verb role a prototype vector (which averages the vectors of its most typical fillers) and its similarity to the candidate fillers (Baroni and Lenci, 2010; Lenci, 2011). More recently, logical metonymy, and complement coercion in general, can be regarded as an instance of argument complexity caused by the effort required to repair the violation of the verb selectional preferences (Chersoni et al., 2021).

[5]The meta senses can be defined a priori, or induced from the data. They are cross-word senses, since they describe the meaning alternations present in different lemmas.

Current contextual language models offer great potential for studying these questions since their representations capture the meaning of specific word instances. This allows to generate representations corresponding to specific senses (for example, by selecting the contexts or instances that will be used to derive the vectors) and to avoid the noise that might be introduced by polysemy. In future work, I plan to explore how these representations can reveal semantic alternations by their proximity to the vectors of monosemous words representing meta senses, as in the above cited distributional study. Furthermore, it would be interesting to investigate whether these alternations can be identified and located in the vector space constructed by contextual language models, in a similar way as abstract semantic notions like intensity which can be described by its own vector representation.

6.4.2 Semantic Properties of Concepts and Entities

6.4.2.1 Language Models' Knowledge of Noun Properties

Neural language model representations encode rich linguistic (grammatical, syntactic and semantic) and world knowledge. There are, however, types of knowledge that are more difficult to identify and retrieve. This is, for example, the case with visual or common-sense knowledge about entities. This difficulty might not be due to the models themselves and how advanced they are, but rather to some types of information being rarely stated in texts. This is described in the literature as **the "reporting bias" phenomenon** (Gordon and Van Durme, 2013) which poses challenges for knowledge extraction. According to this phenomenon, the frequency with which people write about actions and properties is not necessarily a reflection of real-world frequencies, or of the degree to which a property is characteristic of a class of individuals. Hence, rare actions or properties are over-represented in texts at the expense of trivial ones. For example, we would expect to find more mentions in a text to "a man who is jumping" than to "a man who is breathing", since the first describes an exceptional action. As far as entity properties are concerned, *bananas* or *strawberries* would more often be described as *ripe* or *tasty*, than as *yellow* or *red* since these are prototypical properties of the objects referred to by these nouns. These are obvious and already known by the participants in the communication, and do not bring in new information. Interestingly, Shwartz and Choi (2020) show that the impressive generalisation capability of pre-trained language models allows them to better estimate the plausibility of frequent but unspoken actions, outcomes and properties than previous models, but that they also tend to overestimate that of the very rare, amplifying the bias towards rare (non prototypical) events that already exists in their training corpus.

In our work (Apidianaki and Garí Soler, 2021), we used probing methodology to explore whether retrieving knowledge about noun properties constitutes a challenge for these models. We probed BERT for the properties of English nouns as expressed by adjectives that do not restrict the reference scope of the noun they modify (as in *red car*), instead emphasise some inherent aspect (*red strawberry*). Adjectival modification is one of the main types of composition in natural language (Baroni and Zamparelli, 2010; Guevara, 2010). Adjectives (As) in attributive position[6] usually have a restrictive role on the reference scope of the noun they modify, limiting the set of things it refers to (e.g., *white rabbits* $\sqsubset$ *rabbits*). This property of adjectives has interesting entailment implications, generally leading to adjective-noun (AN) constructions where the entailment relationship with the head noun holds (AN $\models$ N) (Baroni et al., 2012).[7] When A is prototypical of the N it modifies (as in *soft silk, red lobster, small blueberry*), its

[6] Adjectives that appear immediately before the noun (N) they modify and form part of the noun phrase (e.g., *white rabbit*), as opposed to adjectives in predicative position that occur after the noun (e.g., this *rabbit* is *white*).

[7] Entailment is directional (*white rabbit* $\models$ *rabbit* but *rabbit* $\not\models$ *white rabbit*) (Kotlerman et al., 2010), unless modification is not restrictive.

insertion does not reduce the scope of N or add new information, but rather emphasises some inherent property (Pavlick and Callison-Burch, 2016a). In these cases, N and AN denote the same set; they are in an equivalence relation (*red lobster* = *lobster*) and entailment is symmetric.

Adjective prototypicality has been understudied in the computational linguistics literature, as opposed to prototypicality relationships between nouns[8] (Roller and Erk, 2016; Vulić et al., 2017). This linguistic property is also absent from lexico-semantic resources such as WordNet (Fellbaum, 1998) and HyperLex (Vulić et al., 2017). Alongside its theoretical interest – given its impact on the entailment properties of AN constructions – knowledge about adjective prototypicality has interesting practical implications. It can serve to retrieve information about the general concept (e.g., *silk, blueberry*) for queries containing AN phrases where the modifier does not restrict the denotation of the noun (e.g., *soft silk, small blueberry*). It can also serve to discard adjectives that do not add new information about the noun for sentence compression and summarization.

In our Apidianaki and Garí Soler (2021) paper, we proposed to study the prototypicality of noun properties as expressed by modifiers in AN constructions. We used the psycholinguistics datasets that were compiled by McRae et al. (2005) and Herbelot and Vecchi (2015), which describe noun properties as association norms and capture the association strength between nouns and their semantic features using quantifiers.[9] Using these two datasets, we probed BERT for its knowledge of noun properties and their prevalence, using cloze tasks and in a classification setting. We found that the model has marginal knowledge of these features and their prototypicality, as expressed in the datasets used for evaluation. This can be due to the model's limited knowledge of these properties or to the absence of this knowledge from the training data due to reporting bias (Gordon and Van Durme, 2013; Shwartz and Choi, 2020). In order to more thoroughly investigate the knowledge that is encoded by the models, **alternative evaluation scenarios should be considered where the quality of the actual model output would be assessed**. This type of evaluation would circumvent the constraints related to the coverage of existing resources and to the nature of their content, which has been collected using specific experimental protocols.

6.4.2.2 A Multimodal Approach to Noun Property Detection

It would be possible to alleviate the issues posed by reporting bias to knowledge modelling and extraction through access to different modalities. In recent work (Yang et al., 2022), we specifically explore the alternative of using images. Our assumption is that reporting bias mainly affects perceptual properties which are well-known to the speakers of a language (e.g., *cute cat*, *small blueberry*, *soft silk*) and are, consequently, not often stated in the communication. We have attempted to extract these properties from images and to use them in an ensemble model in order to complement the information predicted by language models. This idea follows up on previous models which combine different modalities for providing a sort of grounding for different types of common-sense knowledge

The originality of our approach is that we use property concreteness (Brysbaert et al., 2014) as a lever to calibrate the contribution of each source (text vs. images). We consider visual (or perceptual) properties (e.g., *red*, *round*) as more concrete than abstract attributes (e.g., *interesting*, *flawless*). Concreteness is a graded notion that strongly correlates with the degree of imageability (Friendly et al., 1982; Byrne,

[8]For example, *cat* is a more prototypical animal than *snake*, and *basketball* is a more prototypical sport than *wrestling*.

[9]The Herbelot and Vecchi (2015) dataset adds an extra layer of quantification annotations to the norms in the McRae et al. (2005) dataset (e.g., ALL *guitars are musical instruments*, but SOME *guitars are electric*). Quantification is important for semantic inference since it serves to understand set relations (such as synonymy and hyponymy), and to derive logically entailed sentences.

1974). Furthermore, concrete words generally tend to refer to tangible objects that the senses can easily perceive (Paivio et al., 1968). We extend this idea to noun properties and hypothesize that vision models would have better knowledge of perceptual, and more concrete, properties (e.g., *red*, *flat*, *round*) than text-based language models, which would better capture abstract properties (e.g., *free*, *inspiring*, *promising*).

We evaluate our ensemble model (which has access to both text and images) in a ranking task, where the actual properties of a noun need to be ranked higher than other non-relevant properties. Candidate properties are taken from the McRae et al. (2005) dataset, from the MEMORY COLORS dataset (Norlund et al., 2021), and from the CSLB (Centre for Speech, Language and Brain) dataset (Devereux et al., 2014). Our results show that the proposed combination of text and images greatly improves noun property prediction compared to powerful text-based language models like BERT-LARGE (Devlin et al., 2019), ROBERTA-LARGE (Liu et al., 2019c), GPT2-LARGE (Radford et al., 2019) and GPT3-DAVINCI. Our ensemble model also outperforms the powerful CLIP vision-language model in the property ranking task (Radford et al., 2021).[10] This shows that concreteness is a useful lever for calibrating the contribution of texts and images which might be useful for exploring other types of semantic knowledge.

Still, even though our results demonstrate the power of combining text and images, model performance in the property ranking task remains low. Top-1 Accuracy on the McRae et al. (2005) dataset only reaches 40.1 with our ensemble model and gold concreteness scores,[11] and 37.9 with GPT3-DAVINCI. Accuracy at 5 is higher (76.2 for the ensemble model with gold scores and 61.5 for GPT3), while the recall at 5 is 40 for the ensemble model (31.8 for GPT-3). These results show that there is still room for improvement for this challenging task.

6.4.3 Semantic Scope

In formal semantics, the scope of a semantic operator (e.g., negation or quantifier) is the semantic object to which it applies. For example, in the sentence "*Mary doesn't like mozzarella but she loves pecorino*", the proposition "*Mary doesn't like mozzarella*" occurs within the scope of negation, while "*she loves pecorino*" does not. Scope also determines the semantic order of operations. This is not always defined by the position of the operators and the syntactic structure of the sentence (Nakov and Hearst, 2005; Campbell, 2002), but is also strongly influenced by the semantics of the words contained in it (Kamp and Partee, 1995). Scope is also relevant for the interpretation of noun phrases involving multiple modifiers. The semantics of simple modifier-noun constructions have been modelled using first-order logic (McCrae et al., 2014), linear mapping methodology[12] (Baroni and Zamparelli, 2010), and other explicit compositional operations such as weighted addition and multiplication (Boleda et al., 2012b, 2013). Other works address the semantics of these constructions using an inference-based approach (Pavlick and Callison-Burch, 2016b,a; Apidianaki and Garí Soler, 2021).

In recent work (Lyu et al., 2022), we explore the semantics of recursive noun phrases (NPs) which involve more than one modifier (e.g., *the so-called imminent danger*, *an arguable perfect solution*). We show that interpreting the meaning of these constructions is a real challenge for language models. **Although the scope of the modifiers can be useful, it cannot always explain the meaning of the NPs because of the influence from modifier semantics.** Hence, two NPs with the same syntactic structure:

[10] CLIP which is pre-trained on 400M image-caption pairs. It integrates a text encoder f_T and a visual encoder f_V, and is trained to align the embedding spaces learned from images and text, using contrastive loss as a learning objective.

[11] 39.9 with scores predicted by a regression model (Charbonnier and Wartena, 2019) with FastText embeddings.

[12] The adjective is seen as a linear function from the noun vector to the AN representation.

a [big [fake gun]] and *a [big [black gun]]*, can have entirely different inference patterns, (i.e., the latter is a gun while the former is not). In order to interpret these NPs and accurately determine the modifiers' scope, subtle knowledge about their semantics is also needed which language models do not seem to capture. For example, *former* negates the noun *diplomat* in the sentence "*A former American diplomat*" (the person is probably still American), and the adjective *beginner* in "*A former beginner drummer*" (the person is probably still a drummer).

In our work, we adopt an inference-based approach which explores the entailment properties of recursive NPs (e.g., "does *a fake fur* entail *fur*?", "does *a tall basketball player* entail *a tall man*?"), and an event plausibility comparison task. We find that the interpretation of recursive noun phrases is a real challenge for large language models, independent of the semantic category of the modifiers (intersective, privative, subsective, nonsubsective) (Kamp and Partee, 1995; Partee, 1995), which can however be learned to some extent if the models are exposed to a small amount of data from the challenge set (a process called "inoculation") (Liu et al., 2019a). Our results demonstrate that there is still room for improving models' understanding of noun phrase constructions, and more generally of semantic scope, as well as the reasoning processes that rely on their successful interpretation.

Bibliography

Ackerman, M. and Ben-David, S. (2009). Clusterability: A theoretical study. In *Artificial Intelligence and Statistics*, volume 5, pages 1–8, Florida, US.

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., and Goldberg, Y. (2017). Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks. In *Proceedings of ICLR*, Toulon, France.

Agirre, E. and Martinez, D. (2004). Unsupervised WSD based on automatically retrieved examples: The importance of bias. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Barcelona, Spain. Association for Computational Linguistics.

Aina, L., Gulordava, K., and Boleda, G. (2019). Putting Words in Context: LSTM Language Models and Lexical Ambiguity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3348, Florence, Italy. Association for Computational Linguistics.

Allaway, E. and McKeown, K. (2021). A Unified Feature Representation for Lexical Connotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2145–2163, Online. Association for Computational Linguistics.

Alzantot, M., Sharma, Y., Elgohary, A., Ho, B.-J., Srivastava, M., and Chang, K.-W. (2018). Generating Natural Language Adversarial Examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Apidianaki, M. (2006). Traitement de la polysémie lexicale dans un but de traduction. In *Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 53–62, Leuven, Belgique. ATALA.

Apidianaki, M. (2007). Repérage de sens et désambiguïsation dans un contexte bilingue. In *Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 195–204, Toulouse, France. ATALA.

Apidianaki, M. (2008a). Acquisition automatique de sens pour la désambiguïsation et la sélection lexicale en traduction. PhD Thesis. University Paris 7, Denis Diderot.

Apidianaki, M. (2008b). Translation-oriented word sense induction based on parallel corpora. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Apidianaki, M. (2008c). Translation-oriented Word Sense Induction Based on Parallel Corpora. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 3269–3275, Marrakech, Morocco.

Apidianaki, M. (2009a). Data-driven Semantic Analysis for Multilingual WSD and Lexical Selection in Translation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, pages 77–85, Athens, Greece. Association for Computational Linguistics.

Apidianaki, M. (2009b). La place de la désambiguïsation lexicale dans la traduction automatique statistique. In *Actes de la 16ème conférence sur le Traitement Automatique des Langues Naturelles. Prise de position*, pages 7–12, Senlis, France. ATALA.

Apidianaki, M. (2011). Unsupervised cross-lingual lexical substitution. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 13–23, Edinburgh, Scotland. Association for Computational Linguistics.

Apidianaki, M. (2012). Measuring the adequacy of cross-lingual paraphrases in a machine translation setting. In *Proceedings of COLING 2012: Posters*, pages 63–72, Mumbai, India. The COLING 2012 Organizing Committee.

Apidianaki, M. (2016). Vector-space models for PPDB paraphrase ranking in context. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2028–2034, Austin, Texas. Association for Computational Linguistics.

Apidianaki, M. and Garí Soler, A. (2021). ALL dolphins are intelligent and SOME are friendly: Probing BERT for nouns' semantic properties and their prototypicality. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 79–94, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Apidianaki, M. and Gong, L. (2015). LIMSI: Translations as source of indirect supervision for multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 298–302, Denver, Colorado. Association for Computational Linguistics.

Apidianaki, M. and He, Y. (2010). An algorithm for cross-lingual sense-clustering tested in a MT evaluation setting. In *Proceedings of the 7th International Workshop on Spoken Language Translation: Papers*, pages 219–226, Paris, France.

Apidianaki, M., Ljubešić, N., and Fišer, D. (2012a). Disambiguating vectors for bilingual lexicon extraction from comparable corpora. In *Proceedings of the Eighth Language Technologies Conference*, pages 10–15, Ljubljana, Slovenia.

Apidianaki, M., Ljubešić, N., and Fišer, D. (2013). Vector Disambiguation for Translation Extraction from Comparable Corpora. *Informatica*, 37:193–201.

Apidianaki, M. and Sagot, B. (2012). Applying cross-lingual WSD to wordnet development. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 833–840, Istanbul, Turkey. European Language Resources Association (ELRA).

Apidianaki, M., Verzeni, E., and McCarthy, D. (2014). Semantic clustering of pivot paraphrases. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4270–4275, Reykjavik, Iceland. European Language Resources Association (ELRA).

Apidianaki, M., Wisniewski, G., Cocos, A., and Callison-Burch, C. (2018). Automated paraphrase lattice creation for HyTER machine translation evaluation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 480–485, New Orleans, Louisiana.

Apidianaki, M., Wisniewski, G., Sokolov, A., Max, A., and Yvon, F. (2012b). WSD for n-best reranking and local language modeling in SMT. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 1–9, Jeju, Republic of Korea. Association for Computational Linguistics.

Apresjan, I. D. (1974). Regular polysemy. *Linguistics*, (142):5–32.

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. (2018). Linear Algebraic Structure of Word Senses, with Applications to Polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495.

Arora, S., Liang, Y., and Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of ICLR*, Toulon, France.

Artetxe, M., Ruder, S., and Yogatama, D. (2020). On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Artiles, J., Amigó, E., and Gonzalo, J. (2009). The role of named entities in Web People Search. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*, pages 534–542, Singapore.

Asher, N. (2011). *Lexical Meaning in Context - A Web of Words*. Cambridge University Press.

Asher, N., Cruys, T. V. d., Bride, A., and Abrusán, M. (2016). Integrating Type Theory and Distributional Semantics: A Case Study on Adjective–Noun Compositions. *Computational Linguistics*, 42(4):703–725.

Bannard, C. and Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, Michigan. Association for Computational Linguistics.

Baroni, M., Bernardi, R., Do, N.-Q., and Shan, C.-c. (2012). Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 23–32, Avignon, France. Association for Computational Linguistics.

Baroni, M., Bernardi, R., and Zamparelli, R. (2014a). Frege in Space: A Program of Compositional Distributional Semantics. *Linguistic Issues in language technology (LILT)*, 9(6):241–346.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Journal of Language Resources and Evaluation*, 43(3):209–226.

Baroni, M., Dinu, G., and Kruszewski, G. (2014b). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland. Association for Computational Linguistics.

Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Baroni, M. and Zamparelli, R. (2010). Nouns are Vectors, Adjectives are Matrices: Representing Adjective-Noun Constructions in Semantic Space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, Cambridge, MA. Association for Computational Linguistics.

Belinkov, Y. (2021). Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, pages 1–13.

Belinkov, Y. and Bisk, Y. (2018). Synthetic and Natural Noise Both Break Neural Machine Translation. In *International Conference on Learning Representations*.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3(null):1137–1155.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Blei, D. M., Jordan, M. I., Griffiths, T. L., and Tenenbaum, J. B. (2003a). Hierarchical Topic Models and the Nested Chinese Restaurant Process. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, page 17–24, Cambridge, MA, USA. MIT Press.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003b). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6(1):213–234.

Boleda, G., Baroni, M., Pham, T. N., and McNally, L. (2013). Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 35–46, Potsdam, Germany. Association for Computational Linguistics.

Boleda, G. and Herbelot, A. (2016). Formal distributional semantics: Introduction to the special issue. *Computational Linguistics*, 42(4):619–635.

Boleda, G., Padó, S., and Utt, J. (2012a). Regular polysemy: A distributional model. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 151–160, Montréal, Canada. Association for Computational Linguistics.

Boleda, G., Vecchi, E. M., Cornudella, M., and McNally, L. (2012b). First order vs. higher order modification in distributional semantics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1223–1233, Jeju Island, Korea. Association for Computational Linguistics.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems 29*, pages 4349–4357, Barcelona, Spain.

Bommasani, R., Davis, K., and Cardie, C. (2020). Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

Bouraoui, Z., Camacho-Collados, J., and Schockaert, S. (2020). Inducing Relational Knowledge from BERT. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7456–7463, New York, NY, USA. AAAI Press.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Brants, T. and Franz, A. (2006). Web 1T 5-gram Version 1. In *LDC2006T13*, Philadelphia, Pennsylvania. Linguistic Data Consortium.

Bruni, E., Boleda, G., Baroni, M., and Tran, N.-K. (2012). Distributional Semantics in Technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145, Jeju Island, Korea. Association for Computational Linguistics.

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Byrne, B. (1974). Item concreteness vs spatial organization as predictors of visual imagery. *Memory & Cognition*, 2(1):53–59.

Camacho-Collados, J. and Pilevar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788.

Campbell, R. (2002). Computation of modifier scope in NP by a language-neutral method. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Carpuat, M. (2013). NRC: A Machine Translation Approach to Cross-Lingual Word Sense Disambiguation (SemEval-2013 Task 10). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 188–192, Atlanta, Georgia, USA. Association for Computational Linguistics.

Carpuat, M. and Wu, D. (2007). Improving Statistical Machine Translation using Word Sense Disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 61–72, Prague, Czech Republic.

Cañete, J., Chaperon, G., Fuentes, R., and Pérez, J. (2020). Spanish Pre-Trained BERT Model and Evaluation Data. In *to appear in PML4DC at ICLR 2020*.

Celikkanat, H., Virpioja, S., Tiedemann, J., and Apidianaki, M. (2020). Controlling the Imprint of Passivization and Negation in Contextualized Representations. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 136–148, Online. Association for Computational Linguistics.

Charbonnier, J. and Wartena, C. (2019). Predicting word concreteness and imagery. In *Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, pages 176–187. Association for Computational Linguistics.

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., and Robinson, T. (2013). One billion word benchmark for measuring progress in statistical language modeling. In *INTERSPEECH*, Lyon, France.

Chen, X., Liu, Z., and Sun, M. (2014). A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035, Doha, Qatar. Association for Computational Linguistics.

Chersoni, E., Santus, E., Lenci, A., Blache, P., and Huang, C.-R. (2021). Not all arguments are processed equally: a distributional model of argument complexity. *Language Resources and Evaluation*, 55 (4):873–900.

Clark, E. V. (1992). Conventionality and contrast: Pragmatic principles with lexical consequences. In *Frames, fields, and contrasts: New essays in semantic and lexical organization*, pages 171–188. Lawrence Erlbaum Associates, Inc.

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Clark, S. (2015). Vector Space Models of Lexical Meaning. In *The Handbook of Contemporary Semantic Theory*, pages 493–522. John Wiley Sons, Ltd.

Cocos, A., Apidianaki, M., and Callison-Burch, C. (2017). Word sense filtering improves embedding-based lexical substitution. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 110–119, Valencia, Spain.

Cocos, A., Wharton, S., Pavlick, E., Apidianaki, M., and Callison-Burch, C. (2018). Learning Scalar Adjective Intensity from Paraphrases. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1752–1762, Brussels, Belgium. Association for Computational Linguistics.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 160–167, New York, NY, USA. Association for Computing Machinery.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 999888:2493–2537.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. (2018). What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Copestake, A. and Briscoe, T. (1995). Semi-productive polysemy and sense extension. *Journal of Semantics*, 12:15–67.

Dagan, I., Glickman, O., and Magnini, B. (2006). The PASCAL Recognising Textual Entailment Challenge. In Quiñonero-Candela, J., Dagan, I., Magnini, B., and d'Alché Buc, F., editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment: First PASCAL Machine Learning Challenges Workshop (MLCW 2005), Southampton, UK, April 11-13, 2005, Revised Selected Papers*, pages 177–190. Springer Berlin Heidelberg, Berlin, Heidelberg.

Dagan, I. and Itai, A. (1994). Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

de Marneffe, M.-C., Manning, C. D., and Potts, C. (2010). Was it good? It was provocative. Learning the meaning of scalar adjectives. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 167–176, Uppsala, Sweden.

de Melo, G. and Bansal, M. (2013). Good, great, excellent: Global inference of semantic intensities. *Transactions of the Association for Computational Linguistics*, 1:279–290.

Delli Bovi, C., Camacho-Collados, J., Raganato, A., and Navigli, R. (2017). EuroSense: Automatic Harvesting of Multilingual Sense Annotations from Parallel Text. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 594–600, Vancouver, Canada. Association for Computational Linguistics.

Denkowski, M. and Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Dev, S. and Phillips, J. M. (2019). Attenuating Bias in Word Vectors. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, Naha, Okinawa, Japan.

Devereux, B. J., Tyler, L. K., Geertzen, J., and Randall, B. (2014). The Centre for Speech, Language and the Brain (CSLB) concept property norms. *Behavior research methods*, 46(4):1119–1127.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Di Marco, A. and Navigli, R. (2013). Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction. *Computational Linguistics*, 39(3):709–754.

Diab, M. and Resnik, P. (2002). An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 255–262, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Dinu, G. and Lapata, M. (2010). Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1162–1172, Cambridge, MA. Association for Computational Linguistics.

Drozd, A., Gladkova, A., and Matsuoka, S. (2016). Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3519–3530, Osaka, Japan. The COLING 2016 Organizing Committee.

Dyvik, H. (1998). Translations as Semantic Mirrors. In *Proceedings of the W13 Workshop: Multilinguality in the lexicon II, at the 13th biennial European Conference on Artificial Intelligence ECAI'98*, pages 24–44, Brighton, UK.

Dyvik, H. (2004). Translations as semantic mirrors: From parallel corpus to WordNet. *Language and Computers*, 49:311–326.

Edmonds, P. and Hirst, G. (2002). Near-Synonymy and Lexical Choice. *Computational Linguistics*, 28(2):105–144.

Elazar, Y., Ravfogel, S., Jacovi, A., and Goldberg, Y. (2021). Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2):179–211.

Epter, S., Krishnamoorthy, M., and Zaki, M. (1999). Clusterability Detection and Initial Seed Selection in Large Data Sets. Technical Report 99-6, Rensselaer Polytechnic Institute, Computer Science Department.

Erk, K. (2012). Vector Space Models of Word Meaning and Phrase Meaning: A Survey. *Language and Linguistics Compass*, 6:635–653.

Erk, K., McCarthy, D., and Gaylord, N. (2009). Investigations on Word Senses and Word Usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 10–18, Suntec, Singapore. Association for Computational Linguistics.

Erk, K., McCarthy, D., and Gaylord, N. (2013). Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554.

Erk, K. and Padó, S. (2008). A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 897–906, Honolulu, Hawaii. Association for Computational Linguistics.

Ethayarajh, K. (2019a). How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Ethayarajh, K. (2019b). Rotate King to get Queen: Word Relationships as Orthogonal Transformations in Embedding Space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3503–3508, Hong Kong, China. Association for Computational Linguistics.

Ettinger, A. (2020). What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Fauconnier, G. (1975). Pragmatic scales and logical structure. *Linguistic Inquiry*, 6:335–375.

Feder, A., Oved, N., Shalit, U., and Reichart, R. (2021). CausaLM: Causal Model Explanation Through Counterfactual Language Models. *Computational Linguistics*, 47(2):333–386.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. The MIT Press.

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2001). Placing search in context: The concept revisited. *ACM Transactions on Information Systems - TOIS*, 20:406–414.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*. Philological Society, Oxford. reprinted in Palmer, F. (ed. 1968) Selected Papers of J. R. Firth, Longman, Harlow.

Friendly, M., Franklin, P. E., Hoffman, D., and Rubin, D. C. (1982). The Toronto Word Pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation*, 14(4):375–399.

Gale, W. A., Church, K. W., and Yarowsky, D. (1992). Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Montréal, Canada.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, pages 758–764, Atlanta, Georgia.

Gao, T., Fisch, A., and Chen, D. (2021). Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Garí Soler, A. and Apidianaki, M. (2020). BERT knows Punta Cana is not just beautiful, it's gorgeous: Ranking scalar adjectives with contextualised representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7371–7385, Online. Association for Computational Linguistics.

Garí Soler, A. and Apidianaki, M. (2021a). Let's Play Mono-Poly: BERT Can Reveal Words' Polysemy Level and Partitionability into Senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.

Garí Soler, A. and Apidianaki, M. (2021b). Scalar Adjective Identification and Multilingual Ranking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4653–4660, Online. Association for Computational Linguistics.

Garí Soler, A., Apidianaki, M., and Allauzen, A. (2019a). Word Usage Similarity Estimation with Sentence Representations and Automatic Substitutes. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 9–21, Minneapolis, Minnesota. Association for Computational Linguistics.

Garí Soler, A., Cocos, A., Apidianaki, M., and Callison-Burch, C. (2019b). A comparison of context-sensitive models for lexical substitution. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 271–282, Gothenburg, Sweden. Association for Computational Linguistics.

Geurts, B. (2010). *Quantity implicatures*. Cambridge University Press.

Giulianelli, M., Del Tredici, M., and Fernández, R. (2020). Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Goldberg, M. K., Hayvanovych, M., and Magdon-Ismail, M. (2010). Measuring Similarity between Sets of Overlapping Clusters. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing (SocialCom/PASSAT)*, pages 303–308, USA. IEEE Computer Society.

Goodfellow, I., Shlens, J., and Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*.

Gordon, J. and Van Durme, B. (2013). Reporting Bias and Knowledge Acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, page 25–30, New York, NY, USA. Association for Computing Machinery.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5):602–610. IJCNN 2005.

Gross, D. and Miller, K. J. (1990). Adjectives in wordnet. *International Journal of lexicography*, 3(4):265–277.

Guevara, E. (2010). A Regression Model of Adjective-Noun Compositionality in Distributional Semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 33–37, Uppsala, Sweden. Association for Computational Linguistics.

Guo, J., Che, W., Wang, H., and Liu, T. (2014). Learning sense-specific word embeddings by exploiting bilingual resources. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 497–507, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Haber, J. and Poesio, M. (2020). Assessing polyseme sense similarity through co-predication acceptability and contextualised embedding distance. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 114–124, Barcelona, Spain (Online). Association for Computational Linguistics.

Haber, J. and Poesio, M. (2021). Patterns of polysemy and homonymy in contextualised language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2663–2676, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Harris, Z. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Hatzivassiloglou, V. and McKeown, K. R. (1993). Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics (ACL)*, pages 172–182, Columbus, Ohio.

Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics*.

Herbelot, A. and Vecchi, E. M. (2015). From concepts to models: some issues in quantifying feature norms. *Linguistic Issues in Language Technology (LiLT)*, 2(4).

Hewitt, J. and Liang, P. (2019). Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Hewitt, J. and Manning, C. D. (2019). A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Hill, F., Reichart, R., and Korhonen, A. (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-term Memory. *Neural computation*, 9:1735–80.

Hodgson, J. M. (1991). Informational constraints on pre-lexical priming. *Language and Cognitive Processes*, 6:169–205.

Hosseini, H., Kannan, S., Zhang, B., and Poovendran, R. (2017). Deceiving Google's Perspective API Built for Detecting Toxic Comments. *CoRR*, abs/1702.08138.

Huang, E., Socher, R., Manning, C., and Ng, A. (2012). Improving Word Representations via Global Context and Multiple Word Prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882, Jeju Island, Korea. Association for Computational Linguistics.

Iacobacci, I., Pilehvar, M. T., and Navigli, R. (2015). SensEmbed: Learning Sense Embeddings for Word and Relational Similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 95–105, Beijing, China. Association for Computational Linguistics.

Ide, N., Baker, C., Fellbaum, C., Fillmore, C., and Passonneau, R. (2008). MASC: the Manually Annotated Sub-Corpus of American English. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Ide, N., Erjavec, T., and Tufis, D. (2002). Sense discrimination with parallel corpora. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 61–66. Association for Computational Linguistics.

Ide, N. and Wilks, Y. (2007). Making Sense About Sense. In Agirre, E. and Edmonds, P., editors, *Word Sense Disambiguation, Algorithms and Applications*, pages 47–73. Springer.

Jiang, Z., Xu, F. F., Araki, J., and Neubig, G. (2020). How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Jin, D., Jin, Z., Zhou, J. T., and Szolovits, P. (2020). Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8018–8025.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020). SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Jurgens, D. and Klapaftis, I. (2013). SemEval-2013 task 13: Word Sense Induction for Graded and Non-Graded Senses. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 290–299, Atlanta, Georgia, USA.

Jurgens, D., Mohammad, S., Turney, P., and Holyoak, K. (2012). SemEval-2012 task 2: Measuring degrees of relational similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 356–364, Montréal, Canada. Association for Computational Linguistics.

Kamp, H. and Partee, B. (1995). Prototype theory and compositionality. *Cognition*, 57(2):129–191.

Kennedy, C. and McNally, L. (2005). Scale Structure and the Semantic Typology of Gradable Predicates. *Language*, 81:345–381.

Kilgarriff, A. (1998). 'I don't believe in word senses'. *Computers and the Humanities*, 31(2):91–113.

Kilgarriff, A. (2004). How Dominant Is the Commonest Sense of a Word? Lecture Notes in Computer Science (vol. 3206), Text, Speech and Dialogue, Sojka Petr, Kopeček Ivan, Pala Karel (eds.), pages 103–112. Springer, Berlin, Heidelberg.

Kim, J.-K. and de Marneffe, M.-C. (2013). Deriving adjectival scales from continuous space word representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1625–1630, Seattle, Washington.

Kishida, K. (2005). Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments. Technical report, NII-2005-014E.

Kotlerman, L., Dagan, I., Szpektor, I., and Zhitomirsky-Geffet, M. (2010). Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389.

Koutsikakis, J., Chalkidis, I., Malakasiotis, P., and Androutsopoulos, I. (2020). GREEK-BERT: The Greeks visiting Sesame Street. In *11th Hellenic Conference on Artificial Intelligence*, pages 110–117.

Kovaleva, O., Romanov, A., Rogers, A., and Rumshisky, A. (2019). Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Kremer, G., Erk, K., Padó, S., and Thater, S. (2014). What Substitutes Tell Us - Analysis of an "All-Words" Lexical Substitution Corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 540–549, Gothenburg, Sweden. Association for Computational Linguistics.

Kriz, R., Miltsakaki, E., Apidianaki, M., and Callison-Burch, C. (2018). Simplification Using Paraphrases and Context-Based Lexical Substitution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 207–217, New Orleans, Louisiana. Association for Computational Linguistics.

Kutuzov, A., Øvrelid, L., Szymanski, T., and Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Lakoff, G. and Johnson, M. (1980). *Metaphors we Live by*. University of Chicago Press, Chicago.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

Landauer, T. K. and Dumais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104(2):211–240.

Lauscher, A., Glavaš, G., Ponzetto, S. P., and Vulić, I. (2020a). A General Framework for Implicit and Explicit Debiasing of Distributional Word Vector Spaces. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York City, NY, USA.

Lauscher, A., Vulić, I., Ponti, E. M., Korhonen, A., and Glavaš, G. (2020b). Specializing Unsupervised Pretraining Models for Word-Level Semantic Similarity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1371–1383, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020). FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Leacock, C., Chodorow, M., and Miller, G. A. (1998). Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–165.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521:436–44.

Lee, D. D. and Seung, H. S. (2000). Algorithms for non-negative matrix factorization. In Leen, T. K., Dietterich, T. G., and

Tresp, V., editors, *NIPS*, pages 556–562. MIT Press.

Lefever, E., Hoste, V., and De Cock, M. (2011). ParaSense or how to use parallel corpora for word sense disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 317–322, Portland, Oregon, USA. Association for Computational Linguistics.

Lenci, A. (2011). Composing and updating verb argument expectations: A distributional semantic model. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 58–66, Portland, Oregon, USA. Association for Computational Linguistics.

Lester, B., Al-Rfou, R., and Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Levin, B. (1993). *English Verb Classes and Alternations A Preliminary Investigation*. University of Chicago Press, Chicago and London.

Levine, Y., Lenz, B., Dagan, O., Ram, O., Padnos, D., Sharir, O., Shalev-Shwartz, S., Shashua, A., and Shoham, Y. (2020). SenseBERT: Driving Some Sense into BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667, Online. Association for Computational Linguistics.

Levy, O. and Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 302–308.

Li, J., Ji, S., Du, T., Li, B., and Wang, T. (2019). Textbugger: Generating adversarial text against real-world applications. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society.

Li, J. and Jurafsky, D. (2015). Do multi-sense embeddings improve natural language understanding? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1722–1732, Lisbon, Portugal. Association for Computational Linguistics.

Li, J., Monroe, W., and Jurafsky, D. (2017). Understanding Neural Networks through Representation Erasure. *arXiv pre-print*, 1612.08220.

Li, L., Ma, R., Guo, Q., Xue, X., and Qiu, X. (2020). BERT-ATTACK: Adversarial Attack Against BERT Using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

Li, X. L. and Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Liang, B., Li, H., Su, M., Bian, P., Li, X., and Shi, W. (2018). Deep text classification can be fooled. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 4208–4215. AAAI Press.

Linzen, T. (2016). Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18, Berlin, Germany. Association for Computational Linguistics.

Linzen, T. (2018). What can linguistics and deep learning contribute to each other? *CoRR*, abs/1809.04179.

Linzen, T., Dupoux, E., and Goldberg, Y. (2016). Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Liu, F., Lu, H., and Neubig, G. (2018). Handling homographs in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1336–1345, New Orleans, Louisiana. Association for Computational Linguistics.

Liu, N. F., Schwartz, R., and Smith, N. A. (2019a). Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2171–2179, Minneapolis, Minnesota. Association for Computational Linguistics.

Liu, Q., McCarthy, D., and Korhonen, A. (2020). Towards Better Context-aware Lexical Semantics: Adjusting Contextualized Representations through Static Anchors. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4066–4075, Online. Association for Computational Linguistics.

Liu, Q., McCarthy, D., Vulić, I., and Korhonen, A. (2019b). Investigating cross-lingual alignment methods for contextualized embeddings with token-level evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 33–43, Hong Kong, China. Association for Computational Linguistics.

Liu, Y., Liu, Z., Chua, T.-S., and Sun, M. (2015). Topical word embeddings. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2418–2424.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019c). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loureiro, D. and Camacho-Collados, J. (2020). Don't Neglect the Obvious: On the Role of Unambiguous Words in Word Sense Disambiguation. *arXiv preprint:2004.14325*.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic space from lexical co-occurence. *Behavior Research Methods Instruments & Computers*, 28:203–208.

Lyu, Q., Hua, Z., Li, D., Zhang, L., Apidianaki, M., and Callison-Burch, C. (2022). Is "my favorite new movie" my favorite movie? probing the understanding of recursive noun phrases. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5286–5302, Seattle, United States. Association for Computational Linguistics.

Manandhar, S., Klapaftis, I., Dligach, D., and Pradhan, S. (2010). Semeval-2010 task 14: Word sense induction and disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation(SemEval)*, pages 63–68, Uppsala, Sweden. Association for Computational Linguistics.

Mancini, M., Camacho-Collados, J., Iacobacci, I., and Navigli, R. (2017). Embedding Words and Senses Together via Joint Knowledge-Enhanced Training. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 100–111, Vancouver, Canada. Association for Computational Linguistics.

Marie, B. and Apidianaki, M. (2015). Alignment-based sense selection in METEOR and the RATATOUILLE recipe. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 385–391, Lisbon, Portugal. Association for Computational Linguistics.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., and Sagot, B. (2020). CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

McCarthy, D., Apidianaki, M., and Erk, K. (2016). Word Sense Clustering and Clusterability. *Computational Linguistics*, 42(2):245–275.

McCarthy, D., Koeling, R., Weeds, J., and Carroll, J. (2004). Finding Predominant Word Senses in Untagged Text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 279–286, Barcelona, Spain.

McCarthy, D. and Navigli, R. (2007). SemEval-2007 Task 10: English Lexical Substitution Task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic. Association for Computational Linguistics.

Mccloskey, M. and Cohen, N. J. (1989). Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. *The Psychology of Learning and Motivation*, 24:104–169.

McCrae, J. P., Quattri, F., Unger, C., and Cimiano, P. (2014). Modelling the semantics of adjectives in the ontology-lexicon interface. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)*, pages 198–209, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

McNally, L. (2016). Scalar alternatives and scalar inference involving adjectives: A comment on van Tiel, et al. 2016. In Ostrove, J., Kramer, R., and Sabbagh, J., editors, *Asking the Right Questions: Essays in Honor of Sandra Chung*, pages 17–28.

McNally, L. and Boleda, G. (2004). Relational adjectives as properties of kinds. Colloque de Syntaxe et Sémantique à Paris.

McRae, K., Cree, G., Seidenberg, M., and Mcnorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37:547–59.

Melamud, O., Goldberger, J., and Dagan, I. (2016). context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany. Association for Computational Linguistics.

Melamud, O., Levy, O., and Dagan, I. (2015). A simple word embedding model for lexical substitution. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 1–7, Denver, Colorado.

Meyerson, A. (2001). Online facility location. In *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science*, FOCS '01, page 426, USA. IEEE Computer Society.

Mickus, T., Paperno, D., Constant, M., and van Deemter, K. (2020). What do you mean, BERT? Assessing BERT as a Distributional Semantics Model. In *Proceedings of the Society for Computation in Linguistics*, volume 3.

Mihalcea, R., Sinha, R., and McCarthy, D. (2010). SemEval-2010 Task 2: Cross-Lingual Lexical Substitution. In *Proceedings of the Fifth International Workshop on Semantic Evaluation (SemEval-2010)*, pages 9–14, Uppsala, Sweden.

Mihalcea, R. F. (2002). Bootstrapping large sense tagged corpora. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint:1301.3781v3*.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, pages 52–55, Miyazaki, Japan.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013c). Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6:1–28.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. (1993). A Semantic Concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Mitchell, J. and Lapata, M. (2008). Vector-based Models of Semantic Composition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL'08: HLT)*, pages 236–244, Columbus, Ohio, USA.

Montague, R. (1973). The Proper Treatment of Quantification in Ordinary English. In *Approaches to Natural Language*, volume 49. Springer, Dordrecht.

Moro, A., Raganato, A., and Navigli, R. (2014). Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Nakov, P. and Hearst, M. (2005). Search engine statistics beyond the n-gram: Application to noun compound bracketing. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 17–24, Ann Arbor, Michigan. Association for Computational Linguistics.

Napoles, C., Gormley, M., and Van Durme, B. (2012). Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 95–100.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2).

Navigli, R. and Ponzetto, S. P. (2010). BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Neelakantan, A., Shankar, J., Passos, A., and McCallum, A. (2014). Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069, Doha, Qatar. Association for Computational Linguistics.

Nielsen, M. A. (2018). Neural networks and deep learning. Determination Press.

Nissim, M., van Noord, R., and van der Goot, R. (2020). Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497.

Norlund, T., Hagström, L., and Johansson, R. (2021). Transferring knowledge from vision to language: How to achieve it and how to measure it? In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–162, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ostrovsky, R., Rabani, Y., Schulman, L. J., and Swamy, C. (2006). The Effectiveness of Lloyd-Type Methods for the k-Means Problem. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 165–176, Berkeley, CA.

Padó, S. and Lapata, M. (2007). Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, 33(2):161–199.

Paetzold, G. and Specia, L. (2016). SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Paivio, A., Yuille, J. C., and Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology*, 76(1p2):1.

Pan, S. J. and Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*,

2(1–2):1–135.

Pantel, P. and Lin, D. (2002). Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–619, New York, NY, USA. ACM.

Partee, B. (1995). Lexical semantics and compositionality. In Gleitman, L. R. and Liberman, M., editors, *Language: An invitation to cognitive science*, page 311–360. The MIT Press.

Patel, A., Sands, A., Callison-Burch, C., and Apidianaki, M. (2018). Magnitude: A Fast, Efficient Universal Vector Embedding Utility Package. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 120–126, Brussels, Belgium. Association for Computational Linguistics.

Pavlick, E. and Callison-Burch, C. (2016a). Most "babies" are "little" and most "problems" are "huge": Compositional entailment in adjective-nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2164–2173, Berlin, Germany. Association for Computational Linguistics.

Pavlick, E. and Callison-Burch, C. (2016b). So-called non-subsective adjectives. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 114–119, Berlin, Germany. Association for Computational Linguistics.

Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2015). PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification . In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430, Beijing, China. Association for Computational Linguistics.

Pavlick, E. and Tetreault, J. (2016). An Empirical Analysis of Formality in Online Communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (2019). Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Pilehvar, M. T. and Camacho-Collados, J. (2019). WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Pilehvar, M. T. and Camacho-Collados, J. (2020). *Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, Cambridge, MA.

Pilehvar, M. T. and Collier, N. (2016). De-Conflated Semantic Representations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1680–1690, Austin, Texas. Association for Computational Linguistics.

Pimentel, T., Hall Maudslay, R., Blasi, D., and Cotterell, R. (2020). Speakers Fill Lexical Semantic Gaps with Context. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4004–4015, Online. Association for Computational Linguistics.

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press, Cambridge, MA.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *arXiv preprint arXiv:2003.07082*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. Technical report, OpenAi.

Raganato, A. and Tiedemann, J. (2018). An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium. Association for Computational Linguistics.

Rajana, S., Callison-Burch, C., Apidianaki, M., and Shwartz, V. (2017). Learning antonyms with paraphrases and a

morphology-aware neural network. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 12–21, Vancouver, Canada. Association for Computational Linguistics.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., and Goldberg, Y. (2020). Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. *arXiv preprint arXiv:2004.07667*.

Ravfogel, S., Prasad, G., Linzen, T., and Goldberg, Y. (2021). Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 194–209, Online. Association for Computational Linguistics.

Ravichander, A., Belinkov, Y., and Hovy, E. (2021). Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

Ravichander, A., Hovy, E., Suleman, K., Trischler, A., and Cheung, J. C. K. (2020). On the systematicity of probing contextualized word representations: The case of hypernymy in BERT. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 88–102, Barcelona, Spain (Online). Association for Computational Linguistics.

Reif, E., Yuan, A., Wattenberg, M., Viegas, F. B., Coenen, A., Pearce, A., and Kim, B. (2019). Visualizing and Measuring the Geometry of BERT. In *Advances in Neural Information Processing Systems*, pages 8592–8600, Vancouver, Canada.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Reisinger, J. and Mooney, R. J. (2010). Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117, Los Angeles, California. Association for Computational Linguistics.

Resnik, P. (2004). Exploiting hidden meanings: Using bilingual text for monolingual annotation. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, pages 283–299, Berlin, Heidelberg. Springer Berlin Heidelberg.

Resnik, P. and Yarowsky, D. (1999). Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5:113–133.

Rill, S., vom Scheidt, J., Drescher, J., Schütz, O., Reinel, D., and Wogenstein, F. (2012). A generic approach to generate opinion lists of phrases for opinion mining applications. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM)*, Beijing, China.

Rogers, A., Drozd, A., and Li, B. (2017). The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 135–148, Vancouver, Canada. Association for Computational Linguistics.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Roller, S. and Erk, K. (2016). Relations such as Hypernymy: Identifying and Exploiting Hearst Patterns in Distributional Vectors for Lexical Entailment. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2163–2172, Austin, Texas. Association for Computational Linguistics.

Romanov, A., Rumshisky, A., Rogers, A., and Donahue, D. (2019). Adversarial decomposition of text representation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 815–825, Minneapolis, Minnesota. Association for Computational Linguistics.

Rosenberg, A. and Hirschberg, J. (2007). V-measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10):627–633.

Ruder, S., Vulić, I., and Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65(1):569–630.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning Representations by Back-propagating Errors. *Nature*, 323(6088):533–536.

Ruppenhofer, J., Wiegand, M., and Brandes, J. (2014). Comparing methods for deriving intensity scores for adjectives. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Gothenburg, Sweden.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv pre-print*, 1910.01108.

Schlechtweg, D., Hätty, A., Del Tredici, M., and Schulte im Walde, S. (2019). A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. Association for Computational Linguistics.

Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97–123.

Sellam, T., Yadlowsky, S., Tenney, I., Wei, J., Saphra, N., D'Amour, A., Linzen, T., Bastings, J., Turc, I. R., Eisenstein, J., Das, D., and Pavlick, E. (2022). The multiBERTs: BERT reproductions for robustness analysis. In *International Conference on Learning Representations*.

Shardlow, M. (2013). The CW Corpus: A New Resource for Evaluating the Identification of Complex Words. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 69–77, Sofia, Bulgaria. Association for Computational Linguistics.

Sharma, R., Gupta, M., Agarwal, A., and Bhattacharyya, P. (2015). Adjective intensity and sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods for Natural Language Processing (EMNLP)*, Lisbon, Portugal.

Sharoff, S. (2006). Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11(4):435–462.

Sheinman, V., Fellbaum, C., Julien, I., Schulam, P., and Tokunaga, T. (2013). Large, huge or gigantic? Identifying and encoding intensity relations among adjectives in WordNet. *Language resources and evaluation*, 47(3):797–816.

Sheinman, V. and Tokunaga, T. (2009). AdjScales: Visualizing Differences between Adjectives for Language Learners. *IEICE Transactions on Information and Systems*, 92-D:1542–1550.

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., and Singh, S. (2020). AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Shivade, C. P., de Marneffe, M.-C., Fosler-Lussier, E., and Lai, A. M. (2015). Corpus-based discovery of semantic intensity scales. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT)*, pages 483–493, Denver, Colorado.

Shwartz, V. and Choi, Y. (2020). Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Suárez, P. J. O., Sagot, B., and Romary, L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Šuster, S., Titov, I., and van Noord, G. (2016). Bilingual learning of multi-sense embeddings with discrete autoencoders. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1346–1356, San Diego, California. Association for Computational Linguistics.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks. In *International Conference on Learning Representations*.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Talmor, A., Elazar, Y., Goldberg, Y., and Berant, J. (2019). oLMpics – On what Language Model Pre-training Captures. arXiv preprint arXiv:1912.13283v1.

Taylor, W. L. (1953). "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Tenney, I., Das, D., and Pavlick, E. (2019a). BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Durme, B. V., Bowman, S., Das, D., and Pavlick, E. (2019b). What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Thater, S., Fürstenau, H., and Pinkal, M. (2011). Word meaning in context: A simple and effective vector model. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1134–1143, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Tian, F., Dai, H., Bian, J., Gao, B., Zhang, R., Chen, E., and Liu, T.-Y. (2014). A probabilistic model for learning multi-prototype word embeddings. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 151–160, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Tsai, Y.-T., Yang, M.-C., and Chen, H.-Y. (2019). Adversarial attack on sentiment classification. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 233–240, Florence, Italy. Association for Computational Linguistics.

Tuggy, D. H. (1993). Ambiguity, polysemy and vagueness. *Cognitive linguistics*, 4(2):273–290.

Turney, P. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Van de Cruys, T. and Apidianaki, M. (2011). Latent Semantic Word Sense Induction and Disambiguation. In *Proceedings of the 49th Annual Meeting of the ACL*, pages 1476–1485, Portland, Oregon, USA.

van der Plas, L. and Apidianaki, M. (2014). Cross-lingual word sense disambiguation for predicate labelling of French. In *Proceedings of TALN 2014 (Volume 1: Long Papers)*, pages 46–55, Marseille, France. Association pour le Traitement Automatique des Langues.

van der Plas, L., Apidianaki, M., and Chen, C. (2014). Global methods for cross-lingual semantic role and predicate labelling. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1279–1290, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Van Tiel, B., Van Miltenburg, E., Zevakhina, N., and Geurts, B. (2016). Scalar Diversity. *Journal of semantics*, 33(1):137–175.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Véronis, J. (2004). HyperLex: Lexical Cartography for Information Retrieval. *Computer Speech & Language*, 18(3):223–252.

Vig, J. (2019). A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., and Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. *arXiv*.

Voita, E., Sennrich, R., and Titov, I. (2019a). The Bottom-up Evolution of Representations in the Transformer: A Study with Machine Translation and Language Modeling Objectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4396–4406, Hong Kong, China. Association for Computational Linguistics.

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., and Titov, I. (2019b). Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Vulić, I., Baker, S., Ponti, E. M., Petti, U., Leviant, I., Wing, K., Majewska, O., Bar, E., Malone, M., Poibeau, T., Reichart, R., and Korhonen, A. (2020a). Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897.

Vulić, I., Gerz, D., Kiela, D., Hill, F., and Korhonen, A. (2017). HyperLex: A Large-Scale Evaluation of Graded Lexical Entailment. *Computational Linguistics*, 43(4):781–835.

Vulić, I., Ponti, E. M., Litschko, R., Glavaš, G., and Korhonen, A. (2020b). Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Webson, A., Chen, Z., Eickhoff, C., and Pavlick, E. (2020). Are "undocumented workers" the same as "illegal aliens"? Disentangling denotation and connotation in vector spaces. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4090–4105, Online. Association for Computational Linguistics.

Wiedemann, G., Remus, S., Chawla, A., and Biemann, C. (2019). Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 161–170, Erlangen, Germany.

Wilkinson, B. and Oates, T. (2016). A gold standard for scalar adjectives. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia.

Wittgenstein, L. (1953). *Philosophical Investigations*. Basil Blackwell, Oxford.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint:1910.03771*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J.

(2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint:1609.08144*.

Xypolopoulos, C., Tixier, A., and Vazirgiannis, M. (2021). Unsupervised Word Polysemy Quantification with Multiresolution Grids of Contextual Embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3391–3401, Online. Association for Computational Linguistics.

Yang, Y., Panagopoulou, A., Apidianaki, M., Yatskar, M., and Callison-Burch, C. (2022). Visualizing the Obvious: A Concreteness-based Ensemble Model for Noun Property Prediction. *arXiv preprint arXiv:2210.12905*.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Zanzotto, F. M., Korkontzelos, I., Fallucchi, F., and Manandhar, S. (2010). Estimating linear models for compositional distributional semantics. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1263–1271, Beijing, China. Coling 2010 Organizing Committee.

Zhang, B. (2001). Dependence of Clustering Algorithm Performance on Clustered-ness of Data. Technical report, Hewlett-Packard Labs.

Zhang, X., Li, P., and Li, H. (2021). AMBERT: A Pre-trained Language Model with Multi-Grained Tokenization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 421–435, Online. Association for Computational Linguistics.

Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Zhong, Z., Friedman, D., and Chen, D. (2021). Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

Zhou, W., Ge, T., Xu, K., Wei, F., and Zhou, M. (2019). BERT-based Lexical Substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373, Florence, Italy. Association for Computational Linguistics.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV'15)*, page 19–27, Santiago, Chile. IEEE Computer Society.

Zipf, G. K. (1945). The meaning-frequency relationship of words. *Journal of General Psychology*, 33(2):251–256.