Automatic detection of parallel sentences from comparable biomedical texts

Rémi Cardon, Natalia Grabar

Automatic detection of parallel sentences from comparable biomedical texts

Rémi Cardon, Natalia Grabar

CNRS, UMR 8163, F-59000 Lille, France
Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France

Abstract. Parallel sentences provide semantically similar information which can vary on a given dimension, such as language or register. Parallel sentences with register variation (like expert and non-expert documents) can be exploited for the automatic text simplification. The aim of automatic text simplification is to better access and understand a given information. In the biomedical field, simplification may permit patients to understand medical and health texts. Yet, there is currently no such available resources. We propose to exploit comparable corpora which are distinguished by their registers (specialized and simplified versions) to detect and align parallel sentences. These corpora are in French and are related to the biomedical area. Our purpose is to state whether a given pair of specialized and simplified sentences is to be aligned or not. Manually created reference data show 0.76 inter-annotator agreement. We treat this task as binary classification (alignment/non-alignment). We perform experiments on balanced and imbalanced data. The results on balanced data reach up to 0.96 F-Measure. On imbalanced data, the results are lower but remain competitive when using classification models train on balanced data. Besides, among the three datasets exploited (semantic equivalence and inclusions), the detection of equivalence pairs is more efficient.

1 Introduction

Parallel sentences provide semantically similar information which can vary on a given dimension. Typically, parallel sentences are collected in two languages and correspond to mutual translations. In the general language, the Europarl [1] corpus provides such sentences in several pairs of languages. Yet, the dimension on which the parallelism is positioned can come from other levels, such as expert and non-expert register of language. The following pair of sentences (first in expert and second in non-expert languages) illustrates this:

- *Drugs that inhibit the peristalsis are contraindicated in that situation*
- *In that case, do not take drugs intended for blocking or slowing down the intestinal transit*

Indeed, pairs of parallel sentences provide useful information on lexicon used, syntactic structures, stylistic features, etc., as well as the correspondences between the languages or registers. Hence, pairs built from different languages are

widely used in machine translation, while pairs differentiated by the register of language can be used for the text simplification. The purpose of text simplification is to provide simplified versions of texts, in order to remove or replace difficult words or information. Simplification can be concerned with different linguistic aspects, such as lexicon, syntax, semantics, pragmatics and even document structure.

Automatic text simplification can be used as a preprocessing step for NLP applications or for producing suitable versions of texts for humans. In this second case, simplified documents are typically created for children [2], for people with low literacy or foreigners [3], for people with mental or neurodegenerative disorders [4], or for laypeople who face specialized documents [5]. Our work is related to the creation of simplified medical documents for laypeople, such as patients and their relatives. It has indeed been noticed that medical and health documents contain information that is difficult to understand by patients and their relatives, mainly because of the presence of technical and specialized terms and notions. This situation has a negative effect on the healthcare process [6–8]. Hence, helping patients to better understand medical and health information is an important issue, which motivates our work.

In order to perform biomedical text simplification, we propose to collect parallel sentences, which align difficult and simple information, as they provide crucial and necessary indicators for automatic systems for the text simplification. Indeed, such pairs of sentences contain cues on transformations which are suitable for the simplification, such as lexical substitutes and syntactic modifications. Yet, this kind of resources is seldom available, especially in languages other than English. As a matter of fact, it is easier to access comparable corpora: they cover the same topics but are differentiated by their registers (documents created for medical professionals and documents created for patients). More precisely, we can exploit an existing monolingual comparable corpus with medical documents in French [9]. The purpose of our work is to detect and align parallel sentences from this comparable corpus. We also propose to test what is the impact of imbalance on categorization results: imbalance of categories is indeed the natural characteristics in textual data.

The existing work on searching parallel sentences in monolingual comparable corpora indicates that the main difficulty is that such sentences may show low lexical overlap but be nevertheless parallel. Recently, this task gained in popularity in general-language domain thanks to the semantic text similarity (STS) initiative. Dedicated *SemEval* competitions have been proposed for several years [10–12]. The objective, for a given pair of sentences, is to predict whether they are semantically similar and to assign a similarity score going from 0 (independent semantics) to 5 (semantic equivalence). This task is usually explored in general-language corpora. Among the exploited methods, we can notice:

– lexicon-based methods which rely on similarity of subwords or words from the processed texts or on machine translation [13]. The features exploited can be: lexical overlap, sentence length, string edition distance, numbers, named entities, the longest common substring [14–18];

– knowledge-based methods which exploit external resources, such as WordNet [19] or PPDB [20]. The features exploited can be: overlap with external resources, distance between the synsets, intersection of synsets, semantic similarity of resource graphs, presence of synonyms, hyperonyms or antonyms [21–23];
– syntax-based methods which exploit the syntactic modelling of sentences. The features often exploited are: syntactic categories, syntactic overlap, syntactic dependencies and constituents, predicat-argument relations, edition distance between syntactic trees [24–27];
– corpus-based methods which exploit distributional methods, latent semantic analysis (LSA), topics modelling, word embeddings, etc. [28–33].

Yet, there is no work on detection and alignment of parallel sentences in specialized areas, like biomedicine. Our work is positioned in this area.

In what follows, we first present the linguistic material used, and the methods proposed. We then present and discuss the results obtained, and conclude with directions of future work.

2 Method

We use the CLEAR comparable medical corpus [9] available online[1] which contains three comparable sub-corpora in French. Documents within these sub-corpora are contrasted by the degree of technicality of the information they contain with typically specialized and simplified versions of a given text. These corpora cover three genres: drug information, summaries of scientific articles, and encyclopedia articles.We also exploit a reference dataset with sentences manually aligned by two annotators.

2.1 Comparable Corpora

Table 1. Size of the three source corpora. Column headers: number of documents, total of occurrences (specialized and simple), total of unique lemmas (specialized and simple)

corpus	# *docs*	# occ_{sp}	# occ_{simpl}	# $lemmas_{sp}$	# $lemmas_{simpl}$
Drugs	11,800x2	52,313,126	33,682,889	43,515	25,725
Scient.	3,815x2	2,840,003	1,515,051	11,558	7,567
Encyc.	575x2	2,293,078	197,672	19,287	3,117

Table 1 indicates the size of the comparable corpus in French: number of documents, number of words (occurrences and lemmas) in specialized and sim-

[1] `http://natalia.grabar.free.fr/resources.php#clear`

plified versions. This information is detailed for each sub-corpus: drug information (*Drugs*), summaries of scientific articles (*Scient.*), and encyclopedia articles (*Encyc.*).

The *Drug* corpus contains drug information such as provided to health professionals and patients. Indeed, two distinct sets of documents exist, each of which contains common and specific information. This corpus is built from the public drug database[2] of the French Health ministry. Specialized versions of documents provide more word occurrences than simplified versions.

The *Scientific* corpus contains summaries of meta-reviews of high evidence health-related articles, such as proposed by the Cochrane collaboration [34]. These reviews have been first intended for health professionals but recently the collaborators started to create simplified versions of the reviews (*Plain language summary*) so that they can be read and understood by the whole population. This corpus has been built from the online library of the Cochrane collaboration[3]. Here again, specialized version of summaries is larger than the simplified version, although the difference is not very important.

The *Encyclopedia* corpus contains encyclopedia articles from Wikipedia[4] and Vikidia[5]. Wikipedia articles are considered as technical texts while Vikidia articles are considered as their simplified versions (they are created for children from 8 to 13 year old). Similarly to the works done in English, we associate Vikidia with Simple Wikipedia[6]. Only articles indexed in the medical portal are exploited in this work. From Table 1, we can see that specialized versions (from Wikipedia) are also longer than simplified versions.

Those three corpora have different degrees of parallelism: Wikipedia and Vikidia articles are written independently from each other, drug information documents are related to the same drugs but the types of information presented for experts and laypeople vary, while simplified summaries from the *scientific* corpus are created starting from the expert summaries.

2.2 Reference Data

The reference data with aligned sentence pairs, which associate technical and simplified contents, are created manually. We have randomly selected 2x14 *encyclopedia* articles, 2x12 *drug* documents, and 2x13 *scientific* summaries. The sentence alignment is done by two annotators following these guidelines:

1. exclude identical sentences or sentences with only punctuation and stopword difference ;
2. include sentence pairs with morphological variations (e.g. *Ne pas dépasser la posologie recommandée.* and *Ne dépassez pas la posologie recommandée.* –

[2] `http://base-donnees-publique.medicaments.gouv.fr/`
[3] `http://www.cochranelibrary.com/`
[4] `https://fr.wikipedia.org`
[5] `https://fr.vikidia.org`
[6] `http://simple.wikipedia.org`

Table 2. Size of the reference data with consensual alignment of sentences. Column headers: number of documents, sentences and word occurrences for each subset, alignment rate

		Specialized				*Simplified*				*Alignment*	
		source		*aligned*		*source*		*aligned*		*rate (%)*	
corpus	*# doc.*	*# sent.*	*# occ.*	*# pairs.*	*# occ.*	*# sent.*	*# occ.*	*# pairs.*	*# occ.*	*sp.*	*simp.*
Drugs	12x2	4,416	44,709	502	5,751	2,736	27,820	502	10,398	18	11
Scient.	13x2	553	8,854	112	3,166	263	4,688	112	3,306	20	43
Encyc.	14x2	2,494	36,002	49	1,100	238	2,659	49	853	2	21

both examples can be translated by *Do not take more than the recommended dose.*);

3. exclude sentence pairs with overlapping semantics, when each sentence brings own content, in addition to the common semantics;
4. include sentence pairs in which one sentence is included in the other, which enables many-to-one matching (e.g. *C'est un organe fait de tissus membraneux et musculaires, d'environ 10 à 15 mm de long, qui pend à la partie moyenne du voile du palais.* and *Elle est constituée d' un tissu membraneux et musculaire. – It is an organ made of membranous and muscular tissues, approximately 10 to 15 mm long, that hangs from the medium part of the soft palate.* and *It is made of a membranous and muscular tissue.*);
5. include sentence pairs with equivalent semantics – other than semantic intersection and inclusion (e.g. *Les médicaments inhibant le péristaltisme sont contre-indiqués dans cette situation.* and *Dans ce cas, ne prenez pas de médicaments destinés à bloquer ou ralentir le transit intestinal. – Drugs that inhibit peristalsis are contraindicated in that situation.* and *In that case, do not take drugs intended for blocking or slowing down the intestinal transit.*).

The judgement on semantic closeness may vary according to the annotators. For this reason, the alignments provided by each annotator undergo consensus discussions. This alignment process provides a set of 663 aligned sentence pairs. The inter-annotator agreement is 0.76 [35]. It is computed within the two sets of sentences proposed for alignment by the two annotators.

Because the three corpora vary in their capacity to provide parallel sentences, we compute their *alignment rate*. The alignment rate for a given corpus is the number of sentences that are part of an aligned pair relative to the total number of sentences. As expected, only a tiny fraction of all possible pairs corresponds to aligned sentences. We can observe that the *scientific* corpus is the most parallel with the highest alignment rate of sentences, while the two other corpora (*drugs* and *encylopedia*) contain proportionally less parallel sentences. Sentences from simplified documents in the *scientific* and *drugs* corpora are longer than sentences from specialized documents because they often add explanations for technical notions, like in this example: *We considered studies involving bulking agents (a fibre supplement), antispasmodics (smooth muscle relaxants) or antidepressants*

(drugs used to treat depression that can also change pain perceptions) that used outcome measures including improvement of abdominal pain, global assessment (overall relief of IBS symptoms) or symptom score. In the *encylopedia* corpus such notions are replaced by simpler words, or removed. Finally, in all corpora, we observe frequent substitutions by synonyms, like {*nutrition*, *food*}, {*enteral*, *directly in the stomach*}, or {*hypersensitivity*, *allergy*}. Notice that with such substitutions, lexical similarity between sentences is reduced.

The documents are pre-processed. They are segmented into sentences using strong punctuation (*i.e. .?!;:*). We removed, from each subcorpus, the sentences that are found in at least half of the documents of a given corpus. Those sentences are typically legal notices, section titles, and remainders from the conversion of the HTML versions of the documents. The lines that contain no alphabetic characters have also been removed. That reduces the total number of possible pairs for each document pair approximately from 940,000 to 590,000.

2.3 Automatic detection and alignment of parallel sentences

Automatic detection and alignment of parallel sentences is the main step of our work. The unity processed is a pair of sentences. The objective is to categorize the pairs of sentences in one of the two categories:

- alignment: the sentences are parallel and can be aligned;
- non-alignment: the sentences are non-parallel and cannot be aligned.

The reference data provide 663 positive examples (parallel sentence pairs). In order to perform the automatic categorization, we also need negative examples, which are obtained by randomly pairing all sentences from all the document pairs and removing the sentence pairs that are already found to be parallel. Approximately, 590,000 non-parallel sentences pairs are created in this way.

For the automatic alignment of parallel sentences, we first use a binary classification model that relies on logistic regression. Our goal is to propose features that can work on textual data in different languages and registers. We use several features which are mainly lexicon-based and corpus-based, so that they can be easily applied to textual data in other corpora, speacialized areas and languages or transposed on them. The features are computed on word forms (occurrences). The features are the following:

1. *Number of common non-stopwords.* This feature permits to compute the basic lexical overlap between specialized and simplified versions of sentences [28]. This feature exploits external knowledge (set of stopwords), which are nevertheless very common linguistic data;
2. *Number of common stopwords.* This feature also exploits external knowledge (set of stopwords). It concentrates on non-lexical content of sentences;
3. *Percentage of words from one sentence included in the other sentence, computed in both directions.* This features represents possible lexical and semantic inclusion relations between the sentences;

4. *Sentence length difference between specialized and simplified sentences.* This feature assumes that simplification may imply stable association with the sentence length;
5. *Average length difference in words between specialized and simplified sentences.* This feature is similar to the previous one but takes into account average difference in sentence length;
6. *Total number of common bigrams and trigrams.* This feature is computed on character ngrams. The assumption is that, at the sub-word level, some sequences of characters may be meaningful for the alignment of sentences if they are shared by them;
7. *Word-based similarity measure exploits three scores (cosine, Dice and Jaccard).* This feature provides a more sophisticated indication on word overlap between two sentences. Weight assigned to each word is set to 1;
8. *Character-based minimal edit distance* [36]. This is a classical acception of edit distance. It takes into account basic edit operations (insertion, deletion and substitution) at the level of characters. The cost of each operation is set to 1;
9. *Word-based minimal edit distance* [36]. This feature is computed with words as units within sentence. It takes into account the same three edit operations with the same cost set to 1. This feature permits to compute the cost of lexical transformation of one sentence into another.

2.4 Experimental design

The set with manually aligned pairs is divided into three subsets:

- *equivalence*: 238 pairs with equivalent semantics,
- *tech in simp*: 237 pairs with inclusion where the content of technical sentence is fully included in simplified sentence, and simplified sentence provides additional content,
- *simp in tech*: 112 pairs with inclusion where the content of simplified sentence is fully included in technical sentence, and technical sentence provides additional content.

For each subset, we perform two sets of experiments:

1. We train and test the model with balanced data (we randomly select as many non-aligned pairs as aligned pairs), and then we progressively increase the number of non-aligned pairs until we reach a ratio of 3000:1, which is close to the real data.
2. Then, for each ratio, we apply the obtained model to the whole dataset and evaluate the results.

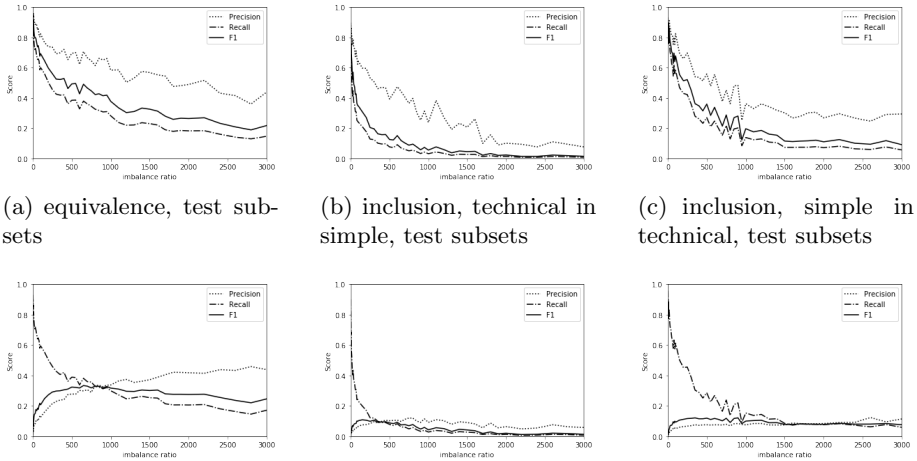
As there is some degree of variability coming with the subset of non-aligned pairs that are randomly selected for the imbalance ratio, every single one of those experiments has been performed fifty times: the results that are presented correspond to the mean values over the fifty runs.

2.5 Evaluation

For evaluating the results, in each experiment we divide the indicated datasets in two parts: two thirds for training and one third for testing. The metrics we use are Recall, Precision and F1 scores. As we are primarily focused on detection of the aligned pairs, we only report scores for that class. Another reason to exclude the negative class and the global score from the observations is that when the data are imbalanced (negative class is growing progressively), misclassifying the positive data has little influence over the global scores, which thus always appear to be high (metrics above 0.99).

3 Presentation and Discussion of Results

Fig. 1. Precision, Recall and F-1 for the various experiments and subsets.



(a) equivalence, test subsets

(b) inclusion, technical in simple, test subsets

(c) inclusion, simple in technical, test subsets

(d) equivalence, real data

(e) inclusion, technical in simple, real data

(f) inclusion, simple in technical, real data

We present the results in Figure 1: The x axis represents the growing of imbalance (the first position is 1 and corresponds to balanced data), while the y axis represents the values of Precision, Recall and F-measure. The results for three subsets are presented: equivalence (Figures 2(a) and 2(d)), inclusion of technical sentence in simple sentence (Figures 2(b) and 2(e)), and inclusion of simple sentence in technical sentence (Figures 2(c) and 2(f)). Besides, Figures 2(a), 2(b) and 2(c) present the results obtained by training and testing the model on the same with the same imbalance ratio (first set of experiments described in section 2.4). As for Figures 2(d), 2(e) and 2(f), they present the results obtained

by the models mentioned above that are applied on the whole set of manually annotated data (second set of experiments described in section 2.4).

The most visible conclusion we can draw from those experiments is that equivalent pairs (Figures 2(a) and 2(d)) are easier to classify than inclusion pairs (the rest of the Figures). Values of both, Precision and Recall, are higher on the equivalence dataset at different imbalance points. For instance, on Figure 2(a) at the starting point, we obtain 0.96 Precision, 0.93 Recall and 0.94 F-measure. This result is positive because the equivalence dataset usually provides the main and the most complete information on transformations required for the simplification. As for the inclusion relations, at the same point and experimental setting, we obtain 0.90 Precision, 0.89 Recall and 0.89 F-measure on technical in simple inclusion dataset, and 0.92 Precision, 0.93 Recall and 0.92 F-measure on simple in technical inclusion dataset. We assume that the inclusion classification models cover a large variety of situations which do not necessarily correspond to the searched information. We need to design additional filters to make the results more suitable for our purpose.

We can also observe from Figure 1 that the use of balanced data provides very high results, both for Precision and Recall, which are very close to the reference data (> 0.90 performance). This is true for the three subsets tested (equivalence and inclusions). This means that models dealing with balanced data can efficiently detect pairs of sentences with parallel contents in balanced and imbalanced datasets. As expected, when imbalance is introduced in the data, the performance of the models decreases. This means that imbalance introduces additional confusion between sentences that should be aligned and those that should not be aligned. Yet, the imbalance has greater effect on the inclusion datasets, while again the equivalence dataset resists better. We can conclude from these results that, when processing real data, it is more suitable to exploit classification models trained on balanced data. Such models show better discrimination for the detection of sentence with parallel contents.

Another interesting finding is that the values of Precision remain higher than the values of Recall. This is particularly observable with experiments using models trained on balanced data (Figures 2(a), 2(b) and 2(c)). We assume that these models can efficiently detect the positive pairs of sentences, which makes the Precision to remain high. Yet, with the increasing imbalance, additional confusion is introduced in data and the results.

Overall, we consider that the results obtained are very good when balanced data are processed. Because imbalance is a natural situation in the task we aim, as it can be observed in Table 2, our future work will concentrate in proposing additional filters to remove non-alignable sentences or to exclude pairs of sentences which should not be aligned.

4 Conclusion and Future Work

We proposed to address the task of detection and alignment of parallel sentences from monolingual comparable corpora in French. The comparable dimension is

due to the technicality of documents, which contrast technical and simplified versions of documents and sentences. We use the CLEAR corpus related to the biomedical area.

Several experiments are performed. More specifically, we work with three subsets of data (equivalence and inclusions between sentences), and with balanced and imbalanced datasets. On balanced dataset, we reach up to 0.93 F-measure, with a very good balance between Precision and Recall. On imbalanced dataset, the performance of classifiers decreases. Yet, the alignment results remain better when models trained on balanced datasets are exploited.

In future, we plan to exploit the best models generated for enriching the set of parallel sentences. The Recall scores may be the main measure for chosing the best classifier and approach. Specific attention will be paid to the filtering of the imbalanced data in order to remove non-alignable sentences and pairs. Enriching the existing reference dataset will permit to prepare data necessary for the developement of simplification methods for the medical documents in French. Other directions for future work are concerned with the exploitation of other features and approaches for the alignment of sentences. As we have seen, the lexical distance between technical and simplified sentences may be high, so the use of word embeddings or the exploitation of external knowledge may be useful to smooth lexical variation.

5 Acknowledgements

This work was funded by the French National Agency for Research (ANR) as part of the *CLEAR* project (*Communication*, *Literacy*, *Education*, *Accessibility*, *Readability*), ANR-17-CE19-0016-01.

References

1. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In: Conference Proceedings: the tenth Machine Translation Summit, Phuket, Thailand, AAMT, AAMT (2005) 79–86
2. Vu, T.T., Tran, G.B., Pham, S.B.: Learning to simplify children stories with limited data. In Springer, L.., ed.: Intelligent Information and Database Systems. (2014) 31–41
3. Paetzold, G.H., Specia, L.: Benchmarking lexical simplification systems. In: LREC. (2016) 3074–3080
4. Chen, P., Rochford, J., Kennedy, D.N., Djamasbi, S., Fay, P., Scott, W.: Automatic text simplification for people with intellectual disabilities. In: AIST. (2016) 1–9
5. Leroy, G., Kauchak, D., Mouradi, O.: A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty. Int J Med Inform **82** (2013) 717–730
6. AMA: Health literacy: report of the council on scientific affairs. Ad hoc committee on health literacy for the council on scientific affairs, American Medical Association. JAMA **281** (1999) 552–7
7. Mcgray, A.: Promoting health literacy. J of Am Med Infor Ass **12** (2005) 152–163

8. Rudd, E.: Needed action in health literacy. J Health Psychol **18** (2013) 1004–10
9. Grabar, N., Cardon, R.: CLEAR – Simple Corpus for Medical French. In: Workshop on Automatic Text Adaption (ATA). (2018) 1–11
10. Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W.: *sem 2013 shared task: Semantic textual similarity. In: *SEM. (2013) 32–43
11. Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G., Uria, L., Wiebe, J.: SemEval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In: SemEval 2015. (2015) 252–263
12. Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., Wiebe, J.: SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In: SemEval 2016. (2016) 497–511
13. Madnani, N., Tetreault, J., Chodorow, M.: Re-examining machine translation metrics for paraphrase identification. In: NAACL-HLT. (2012) 182–190
14. Clough, P., Gaizauskas, R., Piao, S.S., Wilks, Y.: METER: Measuring text reuse. In: ACL. (2002) 152–159
15. Zhang, Y., Patrick, J.: Paraphrase identification by text canonicalization. In: Australasian Language Technology Workshop. (2005) 160–166
16. Qiu, L., Kan, M.Y., Chua, T.S.: Paraphrase recognition via dissimilarity significance classification. In: Empirical Methods in Natural Language Processing, Sydney, Australia (2006) 18–26
17. Nelken, R., Shieber, S.M.: Towards robust context-sensitive sentence alignment for monolingual corpora. In: EACL. (2006) 161–168
18. Zhu, Z., Bernhard, D., Gurevych, I.: A monolingual tree-based translation model for sentence simplification. In: COLING 2010. (2010) 1353–1361
19. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to wordnet: An on-line lexical database. Technical report, WordNet (1993)
20. Ganitkevitch, J., Van Durme, B., Callison-Burch, C.: PPDB: The paraphrase database. In: NAACL-HLT. (2013) 758–764
21. Mihalcea, R., Corley, C., Strapparava, C.: Corpus-based and knowledge-based measures of text semantic similarity. In: AAAI. (2006) 1–6
22. Fernando, S., Stevenson, M.: A semantic similarity approach to paraphrase detection. In: Comp Ling UK. (2008) 1–7
23. Lai, A., Hockenmaier, J.: Illinois-LH: A denotational and distributional approach to semantics. In: Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland (2014) 239–334
24. Wan, S., Dras, M., Dale, R., Paris, C.: Using dependency-based features to take the para-farce out of paraphrase. In: Australasian Language Technology Workshop. (2006) 131–138
25. Severyn, A., Nicosia, M., Moschitti, A.: Learning semantic textual similarity with structural representations. In: Annual Meeting of the Association for Computational Linguistics. (2013) 714–718
26. Tai, K.S., Socher, R., Manning, C.D.: Improved semantic representations from tree-structured long short-term memory networks. In: Annual Meeting of the Association for Computational Linguistics, Beijing, China (2015) 1556–1566
27. Tsubaki, M., Duh, K., Shimbo, M., Matsumoto, Y.: Non-linear similarity learning for compositionality. In: AAAI Conference on Artificial Intelligence. (2016) 2828–2834
28. Barzilay, R., Elhadad, N.: Sentence alignment for monolingual comparable corpora. In: EMNLP. (2003) 25–32

29. Guo, W., Diab, M.: Modeling sentences in the latent space. In: ACL. (2012) 864–872
30. Zhao, J., Zhu, T.T., Lan, M.: ECNU: One stone two birds: Ensemble of heterogenous measures for semantic relatedness and textual entailment. In: Workshop on Semantic Evaluation (SemEval 2014). (2014) 271–277
31. Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R.S., Torralba, A., Urtasun, R., Fidler, S.: Skip-thought vectors. In: Neural Information Processing Systems (NIPS). (2015) 3294–3302
32. He, H., Gimpel, K., Lin, J.: Multi-perspective sentence similarity modeling with convolutional neural networks. In: EMNLP, Lisbon, Portugal (2015) 1576–1586
33. Mueller, J., Thyagarajan, A.: Siamese recurrent architectures for learning sentence similarity. In: AAAI Conference on Artificial Intelligence. (2016) 2786–2792
34. Sackett, D.L., Rosenberg, W.M.C., MuirGray, J.A., Haynes, R.B., Richardson, W.S.: Evidence based medicine: what it is and what it isn't. BMJ **312** (1996) 71–2
35. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20** (1960) 37–46
36. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions and reversals. Soviet physics. Doklady **707** (1966)