Direct vs. indirect evaluation of distributional thesauri

Vincent Claveau, Ewa Kijak

Direct vs. indirect evaluation of distributional thesauri

Vincent Claveau
IRISA - CNRS
Campus de Beaulieu
35042 Rennes, France
vincent.claveau@irisa.fr

Ewa Kijak
IRISA - Univ. of Rennes 1
Campus de Beaulieu
35042 Rennes, France
ewa.kijak@irisa.fr

Abstract

With the success of word embedding methods in various Natural Language Processing tasks, all the fields of distributional semantics have experienced a renewed interest. Beside the famous word2vec, recent studies have presented efficient techniques to build distributional thesaurus; in particular, Claveau et al. (2014) have already shown that Information Retrieval (IR) tools and concepts can be successfully used to build a thesaurus. In this paper, we address the problem of the evaluation of such thesauri or embedding models. Several evaluation scenarii are considered: direct evaluation through reference lexicons and specially crafted datasets, and indirect evaluation through a third party tasks, namely lexical subsitution and Information Retrieval. For this latter task, we adopt the query expansion framework proposed by Claveau and Kijak (2016). Through several experiments, we first show that the recent techniques for building distributional thesaurus outperform the word2vec approach, whatever the evaluation scenario. We also highlight the differences between the evaluation scenarii, which may lead to very different conclusions when comparing distributional models. Last, we study the effect of some parameters of the distributional models on these various evaluation scenarii.

1 Introduction

For years, distributional semantic has aimed at building thesauri (or lexicons) automatically from text corpora. For a given input (ie. a given word), these thesauri identify semantically similar words based on the assumption that they share a distribution similar to the input word's one. In practice, this distributional assumption is set such that two words would be considered close if their occurrences share similar contexts. These contexts are typically co-occurring words in a limited window around the considered words, or words syntactically linked. Recently, many studies have explored new techniques to represent the word (or phrase or document) through embeddings: most often, words are thus represented in a vector space, such that two words with close meanings are close in this space. One of the most popular approach is the famous word2vec technique (Mikolov et al., 2013). Of course, such embedding techniques rely on the same assumption than "traditional" distributional semantics (although many studies do not acknowledge it clearly).

Evaluating these thesauri or embeddings remains a crucial point to assess the quality of the construction methods and parameters used. In this article, we propose to examine this evaluation problem with different evaluation protocols. Indeed, a commonly used approach is to compare the generated thesauri to one or several reference lexicons. This evaluation procedure, called 'intrinsic', has the advantage of being straightforward and simple as it aims at estimating the quality and completeness of the generated thesaurus. However, it is based on reference lexicons whose own completeness, quality, or simply their availability for the considered domain/language/genre are not always granted. Here, we propose to examine the impact of these problems by also using other evaluation protocols, based on other types of reference datasets, and by third-party tasks. In particular, following the work of Claveau and Kijak (2016), we rely on information retrieval (IR) as a realistic evaluation use case. The comparison of the results obtained with these different protocols/datasets should help us to judge the relevance of these assessment scenarios.

After a review of related work (next section), the article addresses the aforementioned subjects: the aspects related to the construction of thesauri are presented in Section 3, while those about the evaluation

with specially crafted resources or by IR are respectively discussed in Section 4 and Section 5. Finally, we present some conclusions and perspectives about this work in the last section.

2 Related work

2.1 Building distributional thesauri

Distributional thesauri rely on the famous formula of Firth (Firth, 1957): *"You should know a word by the company it keeps"*. In such thesauri, each word is semantically characterized by all the contexts in which it appears. The semantic neighbors of an entry word are then words whose contexts are similar to that of the entry. Since the pioneering work of Grefenstette (1994) and Lin (1998), many studies have examined distributional thesaurus building. The semantic link considered between an entry word and its neighbors is varied: synonyms, hyperonymy, hyponymy or another (Budanitsky and Hirst, 2006; Adam et al., 2013, for a discussion). Despite their diversity, these links are interesting for many applications related to Natural Language Processing. The various aspects of the thesaurus building is therefore a research field still very active.

One first step concerns the definition of the distributional context of a given word. The graphical contexts simply consider the words appearing around the occurrences of the target word, while syntactic contexts are formed with the syntactic predicates and arguments of the occurrences of the target word. The latter, if it is considered more accurate, requires a prior parsing step which (i) is not always available, (ii) may be misleading.

There are many relationships between distributional semantics and IR. For example, vectorial representations of the contexts are often used (Turney and Pantel, 2010), but unrelated with weighting schemes and relevance functions used in IR (with the exception of Vechtomova and Robertson (2012) in the slightly different context of computing similarities between named entities). However, the weighting of contexts provides more relevant neighbors. Broda et al. (2009) thereby proposed to consider the ranks rather than directly the weights of contexts to overcome the influence of weighting functions. Some bootstrap methods were also proposed to modify the weight of the contexts of a word, by taking into account its semantic neighbors (Zhitomirsky-Geffet and Dagan, 2009; Yamamoto and Asakura, 2010). Another example of relationship between distributional semantics and IR is the use of search engines to collect co-occurrence information or contexts on the web (Turney, 2001; Bollegala et al., 2007; Sahami and Heilman, 2006; Ruiz-Casado et al., 2005). Finally, given that the "traditional" distributional representation of contexts is sparse and redundant (Hagiwara et al., 2006), several methods for dimension reduction were tested: from Latent Semantic Analysis (Landauer and Dumais, 1997b; Padó and Lapata, 2007; Van de Cruys et al., 2011) to *Random Indexing* (Sahlgren, 2001), through factorization by non-negative matrices (Van de Cruys, 2010).

The problem of the construction of distributional thesaurus may be expressed in a simple way as a conventional IR problem (Claveau et al., 2014). All contexts of a target word are then represented as a document (or query), and distributional neighbors of the target word are the sets of similar contexts. This formulation of distributional neighbors search process offers interesting research tracks and easily accessible tools.

2.2 Tested models

In this paper, several distributional models are tested. A first group of models, that may be called traditional distributional models are considered. In the following sub-section we report results obtained by a state-of-the art approach, hereafter denoted *base*, that uses a cosine similarity and weighting by mutual information (Ferret, 2013), an improved version (*rerank*) which uses machine learning technique to rerank neighbors (Ferret, 2013), and another version (*synt*) based on syntactic contexts (Ferret, 2014) rather than graphic ones are also tested.

As explained in Claveau et al. (2014), the problem of building a distributional thesaurus can be translated into a problem of searching similar documents and can therefore be carried out with IR techniques. In this context, all the contexts of a given word in a corpus are collected; this set of contexts forms what is considered as a document. Building an entry in the thesaurus, ie. finding the closest words (in a dis-

tributional sense) of a word w_i, is thus equivalent to finding documents (contexts) close to the document representing the contexts of w_i (seen as a query in the IR system). This has led to new distributional models, that differ from the previous ones by the way the similarity between two words (or their contexts) is computed. We also report the results of systems based on this IR approach. Several variants, each using a different way to compute the similarity between the sets of context, are considered: namely TF-IDF/cosine and Okapi-BM-25 (Robertson et al., 1998). In previous work, we also proposed an adjusted versions of them *adjusted-TF-IDF, adjusted-Okapi BM25*, in which the influence of the document size is reinforced in order to give more importance to the most discriminating context words (Claveau et al., 2014). Other IR systems are tested; they are based on probabilistic language modeling (denoted LM), with both Dirichlet smoothing (varying the values of the parameter μ) and Hiemstra smoothing (smoothing with the probabilities of occurrence of words throughout the collection; with different values of λ) (Claveau and Kijak, 2016). All these very classical IR models are not detailed further here; the interested reader will find the concepts and useful details (such as the role of the parameters) in the cited references or IR surveys (Manning et al., 2008, for example).

As a last group of models, we consider approaches based on dimensionality reduction. In such models, the data is represented in dense vector space, usually of small dimensionality (for instance, $\mathbb{R}^{500}$). We report results yielded by models based on usual dimension reduction techniques (LSI, LDA, Random projections (RP)), with different numbers of dimensions. In this group, we also consider the very popular embedding approaches, namely Word2Vec (Mikolov et al., 2013). Indeed, they have been shown to be equivalent to standard distributional models with an additional dimensionality reduction step (Levy and Goldberg, 2014). For comparison purpose, we also indicate the results of a word2vec model pre-trained on the Google News corpus (which is significantly larger than AQUAINT-2); it is freely available at `https://code.google.com/p/word2vec/`.

2.3 Evaluating distributional thesauri

As already mentioned, the evaluation of a thesaurus can be done in two ways: (i) the intrinsic evaluation consists in comparing the produced thesaurus with a reference resource; (ii) the extrinsic evaluation assesses the thesaurus through its use in a given task.

The intrinsic evaluation requires to have reference lexicons. Usual lexicons used as references are WordSim 353 (Gabrilovich and Markovitch, 2007), WordNet 3.0 (Miller, 1990) or Moby (Ward, 1996). The two latter exploit larger resources (as synonyms) and are those used in this work for the intrinsic evaluation, as in Ferret (2013). Other data sets, as the set of synonyms from the TOEFL test (Landauer and Dumais, 1997a) or the semantic relationships in BLESS (Baroni and Lenci, 2011), are not directly lexicons, but can also be used for direct evaluation. Given a reference lexicon, it is then easy to compute recall, accuracy or any other measure of quality.

If the intrinsic evaluation is simple, its relevance depends on the adequacy of the lexicons used as references. For this reason, several studies have suggested extrinsic evaluation through a task, such as the lexical substitution task proposed at SemEval 2007 (McCarthy and Navigli, 2009). The goal is to replace a word in a sentence by a neighbor (given by the evaluated thesaurus) and verify that it did not change the meaning of the sentence, by comparing the obtained results to the substitutions proposed by humans. In such a task, the exact synonyms are favored over other types of semantic relationships.

Several studies use distributional information within an IR framework (Besançon et al., 1999; Billhardt et al., 2002), like recent lexical representations such as word2vec (Huang et al., 2012; Mikolov et al., 2013). The aim is to improve the representation of documents and/or the Relevance Status Value function (RSV, ie. the function used in IR systems to rank the answers to a query according to their supposed relevance), by exploiting the similarities between word contexts. Nevertheless, the process of creating the distributional thesaurus is not dissociated from the IR process in these studies, which makes the evaluation of the only distributional information contribution impossible. Recently, we have proposed to specifically evaluate the distributional thesauri in IR by using semantic neighbors to expand the queries (Claveau and Kijak, 2016). In this paper, we adopt the same framework in Section 5.

3 Intrinsic Evaluation of distributional models

3.1 Principles and material

For the sake of comparison with published results, the data used for our experiments are those used in several studies. The corpus used to collect the contexts is AQUAINT-2 (Vorhees and Graff, 2008); it is composed of articles in English containing a total of 380 millions of words. The words considered for our thesaurus entries are common nouns occurring at least 10 times in the corpus, that is 25 000 different nouns. The contexts of all occurrences of these words are collected; in the experiments reported below, contexts are formed by the two words at the right and two words at the left of the target noun, along with their position. For example, in the sentence "... all forms of restriction on freedom of expression, threats ..." the words restriction-2, on-1, of+1, expression+2 are added to the set of contexts of freedom.

As we mentioned earlier, we use WordNet (WN) and Moby for intrinsic assessment of generated thesauri. These two resources have different, additional characteristics: WN identifies strong semantic links (synonyms or quasi-synonyms) while Moby identifies a greater variety of links (hypernyms, meronyms, co-hyponymy...). WN offers on average 3 neighbors for 10 473 nouns of AQUAINT-2, and Moby contains on average 50 neighbors of 9 216 nouns. Together, these resources cover 12 243 nouns of the corpus with 38 neighbors on average. These resources are used as reference for the evaluation; more details about the semantic links considered by these resources and their use for distributional thesaurus evaluation can be found in the literature (Ferret, 2013; Claveau et al., 2014). The number of nouns and the variety of semantic relations that they contain make these references a comprehensive evaluation data set, compared with other existing benchmarks (such as WordSim 353 (Gabrilovich and Markovitch, 2007) for instance).

3.2 Intrinsic evaluation results

Figure 1 presents the results obtained by different thesaurus building systems presented in Section 2.2, applied to the AQUAINT-2 corpus. The performance measures used to compare the generated thesauri with the reference (WordNet + Moby, denoted by WN+M) are those typically used for this task: precision at different levels (on the top 5, 10, 50, 100 neighbors), MAP (Mean Average Precision) and R-precision, expressed as a percentage, averaged on the 12 243 nouns in the WN+M reference. For all these distributional models, among all the parameters tested, we only report some of the best performing ones in terms of MAP (number of dimensions, size of the context window...).

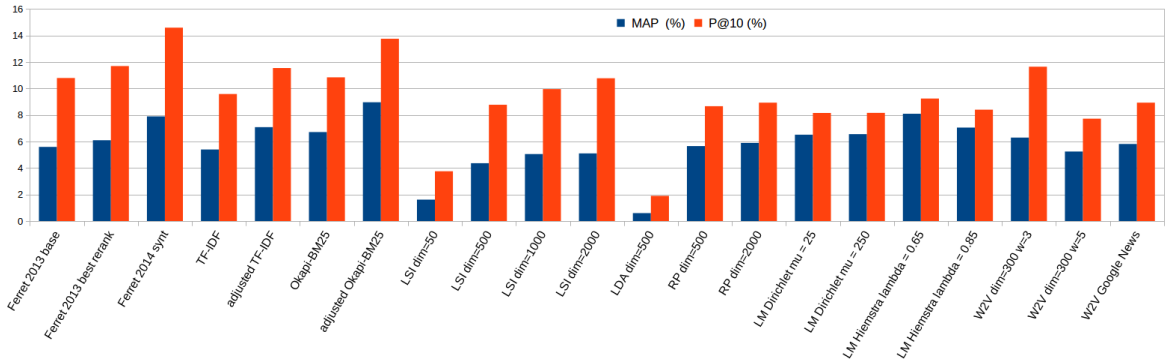

Figure 1: Performance of various distributional and embedding models for building distributional thesauri over the WN+M reference

As already mentioned in the literature, this kind of evaluation with reference lexicons leads to very severe conclusions about the quality of the evaluated distributional thesauri. For instance, the P@10 score states that, in average, 9 out the 10 first neighbors of an entry word are errors, whatever the model. Nonetheless, it is worth noting that some traditional models, and in particular recent IR-based ones such as adjusted Okapi-BM25, obtain better results than the popular word2vec ones. Overall, dimension reduction techniques yields low results: The lower the number of dimensions considered, the worse the

Method	Pearson's r	Spearman's ρ	Kendal's τ
adjusted Okapi	-0.0009 (p=0.9723)	0.3148 (p=5.0e-32)	0.2093 (p=2.5e-30)
W2V dim=300 w=3	0.0027 (p=0.9314)	0.2913 (p=1.4e-21)	0.1944 (p=9.7e-21)

Table 1: Correlation coefficients (with their p-values) between the Average Precision (AP) of each word on the WN+M dataset and its frequency

Method	Pearson's r	Spearman's ρ	Kendal's τ
adjusted Okapi	-0.0923 (p=0.0007)	0.0604 (p=0.0274)	0.0465 (p=0.0111)
W2V dim=300 w=3	-0.1218 (p=9.01e-05)	0.0657 (p=0.0352)	0.0479 (p=0.0213)
W2V GoogleNews	-0.1609 (p=5.08e-09)	-0.1848 (p=1.77e-11)	-0.1374 (p=1.07e-13)

Table 2: Correlation coefficients (with their p-values) between the Average Precision (AP) of each word on the WN+M dataset and its polysemy

results. This negative result is in line with some conclusions of previous work (Van de Cruys, 2010). The occurrence of certain very specific contextual words is indeed a strong indicator of the semantic proximity of words. Aggregation of different words into a single dimension is then detrimental to distinguish the semantic neighbors. This is also confirmed by the fact that within a model family, the parameter settings leading to the best results are those which give more weight to discriminating words: squared IDF for Okapi, very few smoothing for language modeling (ie. low values of μ and λ).

3.3 Influence of data characteristics

It is interesting to examine how some characteristics of the data may influence the results. For instance, it has already been noted (Ferret, 2013) that the frequency of words for which we try to find the neighbors has a great influence on the final quality. This is easily explained by the fact that the more frequent the nouns are, the more contexts they have to describe them; and finally, the better the results are. In order to verify what is the actual effect on our results, given our data and methods, we estimate the correlation between the entry quality in the distributional thesaurus, measured by the Average Precision on the WN+M dataset, and the its frequency in the AQUAINT-2 corpus; results are given in Table 1.

As expected, these results confirm the effect of the frequency on the entry quality. Yet, it should be noted that this correlation is not perfect; other characteristics may interfere on the results.

Another data property that may influence the results is polysemy. Indeed, all the methods evaluated in Section 3.2 consider that all the occurrences of a word (or lemma) should result in one thesaurus entry, that is, no disambiguation is performed. Thus, one could suspect that the list of distributional neighbors of a polysemic word would be impacted. As we have done before, to verify this supposition, we estimate the correlation between the entry quality (AP on the WN+M reference) and the number of senses for that entry as encoded in WordNet. Results are given in Table 2. For comparison purpose, we also report the results obtained with the word2vec GoogleNews model.

The effect of polysemy on the thesaurus results is not obvious. On the one hand, one can observe that there is no correlation between the results and the number of word senses for the models trained on our corpus. On the other hand, the word2vec GoogleNews model shows a statistically significant negative correlation; it means that the less polysemic words tends to yield the best results.

4 Evaluating with specially crafted resources

The evaluation through reference lexicons, as presented above, has several shortcomings already mentioned in Section 2. In addition to these, it is worth noting that resources such as WordNet or Moby were not initially designed for evaluation, and choices made for their building are not necessarily adequate for our evaluation task. For instance, there is no graduation: for a given semantic relation two words are either related or not. In order to provide more relevant evaluation resources, other direct evaluation of distributional thesauri were proposed. In this section, we consider SimLex999 (Hill et al., 2014) that was specially developed to evaluate distributional models. One of its most interesting particularities is that,

according to the authors: "it explicitly quantifies similarity rather than association or relatedness, so that pairs of entities that are associated but not actually similar have a low rating".

4.1 Experimental setting

SimLex999 is a resource in which pairs of words (nouns, verbs, adjectives) are given a score according to their association strength. This score was given by a group of annotators with light instructions on what is to be considered as "association" (Hill et al., 2014, Sec. 3.3). This is intended to reflect the light formalization of the semantic links captured by distributional models.

The evaluation based on this resource aims at comparing the word pair list sorted by association strength and the word pair sorted according to the distributional score. In practice, the Spearman's ρ rank correlation is used. A review of the most recent results obtained with this dataset can be found at `http://www.cl.cam.ac.uk/~fh295/simlex.html`.

4.2 Results

Table 3 shows the SimLex999 results of the best traditional model (adjusted Okapi-BM25) and the best word2vec model. For comparison purpose, we also report the results obtained with the freely available word2vec model trained on the GoogleNews corpus. One can observe a very important difference between the Okapi model and the word2vec trained on the AQUAINT-2 corpus. The GoogleNews model also yields a good score, but one has to keep in my mind that it was trained on a larger corpus than ours.

Method	Spearman's ρ
adjusted Okapi	0.4511
W2V dim=300 w=3	0.3691
W2V GoogleNews	0.4419

Table 3: Correlation coefficients (with their p-values) of distributional models on the SimLex999 dataset

4.3 Influence of data characteristics

As done for the previous evaluation scenario, it is interesting to examine how the SimLex999 results are impacted by the data characteristics. In Figure 2, we report the evolution of Spearman's ρ according to the frequencies of the words considered; ρ is computed on subsets of the SimLex999 pairs whose frequencies are both under a given threshold. The model used here is word2vec trained on the AQUAINT-2 corpus with dim=300 and w=3. One can observe that there are some variations when considering the lowest frequencies; they are due to the small number of SimLex pairs which make the correlation computation very sensitive. Beside that, overall, the frequency does not seem to play an active role in the performance, which seems to contradict what was observed in Section 3.3. Yet, it is worth noting that the less frequent SimLex999 words are not rare in our corpus (more than 2,000 occurrences for the rarest one). SimLex999 contains only frequent or very frequent words and thus does not adequately evaluate the capacity of the model to handle rarer words.

Figure 3 adopts the same setting and presents the impact of polysemy on the SimLex999 performance: ρ is computed on subsets of the SimLex999 pairs whose sum of senses, as encoded in WordNet, is under a given threshold. Here, the effect of polysemy appears very clearly: polysemic words have a negative impact on the capacity of the model to encode the similarity between words as measured by the SimLex999 dataset. This result is not surprising, but it did not appeared in the previous lexicon-based evaluation scenario due to the protocol used.

Finally, these two direct evaluation scenarii, using reference lexicon or specially crated datasets, give dissimilar results, and are not impacted by the same data characteristics.

5 Indirect evaluation through query expansion

Following our previous work (Claveau and Kijak, 2016), in this section we use an IR task as a way to evaluate distributional models. More precisely, we use distributional thesauri to expand queries: for each

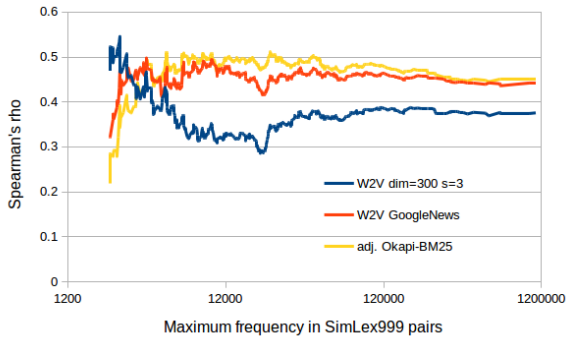

Figure 2: Performance on the SimLex999 dataset according to the frequency of the words considered; log-scale

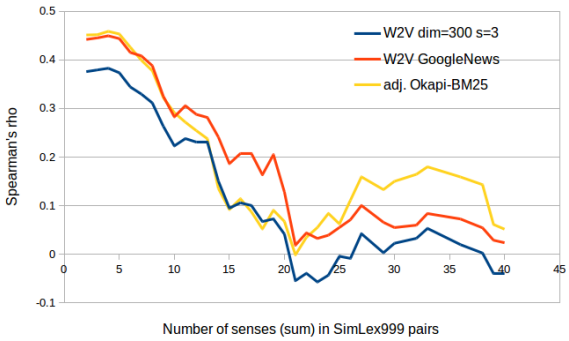

Figure 3: Performance on the SimLex999 dataset according to the sum of the sense of the words in the considered pairs

query noun, its neighbors found in the considered thesaurus are added to the query. The experimental framework and the results obtained are successively presented below.

5.1 Experimental setting

The IR collection used in the experiments comes from the Tipster project and was used as part of TREC. It contains more than 170 000 documents and 50 queries in English (the queries are structured: the query itself, a narrative field detailing the criteria of relevance; in the experiments reported below, we only use the query field). This collection is particularly suited since its documents come from the *Wall Street Journal* and are similar to those of AQUAINT-2.

The IR system used is Indri (Metzler and Croft, 2004; Strohman et al., 2005). This probabilistic system implements a combination of language modeling (Ponte and Croft, 1998) (as the ones used in Sect. 3) and inference networks (Turtle and Croft, 1991); it is known to provide state-of-the-art results. In the experiments reported below, we use its standard settings, ie. Dirichlet smoothing (with $\mu = 2500$ as recommended). In our case, Indri offers an additional advantage: it has a complex query language that allows us to include the words of the distributional thesaurus by making best use of the inference network model; in practice, the dedicated operator `'#syn'` is used to aggregate the counts of the words indicated as synonyms (see Indri documentation for details). To remove the effects of inflection on the results, the plural and singular forms of nouns of the queries are added, either in the non-extended, original queries or those extended with the semantic neighbors. As example, we give below a sample query, with its non-expanded form and its expanded form (adjusted Okapi top 5) using the inference network operators of Indri:

– query : `coping with overcrowded prisons`
– normal form : `#combine( coping with overcrowded #syn( prisons prison ) )`
– expanded form : `#combine( coping with overcrowded #syn( prisons prison inmate inmates jail jails detention detentions prisoner prisoners detainee detainees ) )`

The performance for this IR task is typically measured by precision at different thresholds (P@x), R-precision, and MAP. Therefore, to evaluate the thesaurus, we measure the gains in terms of precision, MAP, etc. between the results without and with expansion. We also indicate the average of the AP (Average Precision) gain by query, noted AvgGainAP (not be confused with the gain of MAP, which is the gain calculated from the AP averages over the query). Non statistically significant results (Wilcoxon and t-test with $p < 0.05$) are in italics.

5.2 Expansion results

Table 4 presents the performance gains achieved by expanding the queries with the words collected in the thesaurus (the ones used in the previous experiments). Here we only report results when expanding

with the top 10 nearest neighbors (other settings lead to similar conclusions ; see (Claveau and Kijak, 2016)). We also show the results obtained by expanding the queries with the reference lexicons (WN alone and WN+M).

Expansion	MAP	AvgGainAP	R-Prec	P@5	P@10	P@50	P@100
without	21.78	-	30.93	92.80	89.40	79.60	70.48
with WN	+12.44	+36.3	+7.01	+4.31	+7.16	+7.60	+10.87
with WN+M	+11.00	+28.33	+7.78	*+3.02*	+5.37	+6.53	+9.17
with adjusted Okapi top 10	+13.80	+24.36	+9.58	*+2.16*	+4.03	+5.58	+8.26
with W2V dim=300 s=3 top 10	+5.20	+17.83	+4.75	*+1.29*	*+2.68*	+4.32	+5.16
with W2V GoogleNews top 10	+13.70	+30.52	+9.52	*+3.02*	*+3.58*	+8.14	+10.19

Table 4: Relative gain of performance (%) when expanding queries with different thesauri

First, we note that for any thesaurus used, the query expansion brings a significant gain in performance. By the way, it contradicts the conclusions of (Voorhees, 1994) about the alleged lack of interest in using WN to expand queries. The most notable fact here is the excellent results obtained with the thesaurus built automatically (traditional or word2vec), that even exceed those of the reference lexicons. While its precision on the first 10 neighbors was evaluated under 14% in Section 3, the adjusted Okapi-BM25 and word2vec thesauri generate expansions yielding the better MAP gains. The average AP gain (AvgGainAP) also provides interesting information: it is maximum with WN, which therefore provides a stable improvement (gain for most queries). This is due to the fact that the words added to the query by WN are synonyms, which presents a low risk. This stability is lower with other thesauri; as the MAP gain remains generally good, it indicates that only certain queries benefit from significant absolute gains.

6 Comparing evaluation results

6.1 Overview

The results of the previous experiences raise questions about the consistency between the lexicon based, similarity-based and task-based evaluations. We want to know, for example, if the P@10 on the WN-based evaluation between two thesaurus construction methods, even if stated as statistically significant, is sensible when now using WN+M as a reference dataset, or in IR, or in the SimLex999 evaluation. We also add the results obtained on the lexical substitution framework proposed in SemEval 2007 (McCarthy and Navigli, 2009). In this task, the distributional thesauri are evaluated on their ability to provide words that are judged as similar in a specific context. In the experiments reported below, we use a very simple strategy: the 10 closest neighbors are proposed, whatever the context. The results are evaluated in terms of precision (referred as the out-of-ten precision in SemEval 2007).

Let us consider Figure 4. The performance on the different evaluation protocols is reported for four distributional models used before: the adjusted Okapi-BM25, word2vec (dim=300, w=3), word2vec (dim=300, w=9) and the GoogleNews word2vec model. Several things are worth noting from this figure. Some models (adjusted Okapi-BM25 for instance) outperform others (for instance, word2vec) on every evaluation protocol. Yet, large difference for one evaluation protocol does not lead to large margin in another evaluation (consider for instance P@10 on the IR task versus the SimLex999 correlation scores). When considering models with more similar results, one can even see that one model can be better at a task and worse at another, sometimes with significant margins (for instance, consider the two word2vec models with s=3 and s=9).

This latter point raises questions on the validity of certain evaluation protocols. In previous work (Claveau and Kijak, 2016), we have shown that the precision of the thesaurus, as measured by a comparison with exiting lexicons, is largely under-estimated when considering the IR-based evaluation. It is mainly caused by the incompleteness of the lexicons used in the intrinsic evaluation. Yet, in this previous work, there was a correlation between these two evaluation protocols (models obtaining the best results for IR yielded the best results for the lexicon-based evaluation, etc.). As seen in Figure 4 with the adjusted Okapi and GoogleNews models, this is not the case in general.

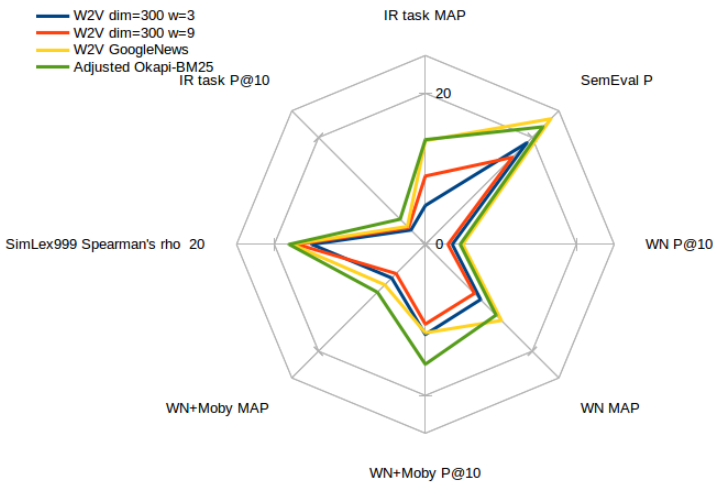

Figure 4: Performance of the adjusted Okapi and word2vec models trained on the AQUAINT-2 corpus; for comparison purpose, the results of the GoogleNews word2vec model are also reported.

6.2 Model parameters

One could argue that the previous results are due to the different approaches (or training corpus in the case of the GoogleNews model) used by these models. In order to confirm or infirm that, we study the evolution of all our evaluation scores according to one specific parameter for a given model (here, word2vec). Figure 5 shows the performance evolution on the different evaluation scenarii according to the dimensionality (respectively the size of the context window). Several points are worth noting. First, this figure highlights here again the fact that the best model (or set of parameters) for one evaluation scenario is not necessarily the best for another. A small context window yields the best results for SimLex999, while a maximum is reached at s=3 when evaluating with WN (strict synonymy) or WN+M (large range of semantic relations). Concerning the IR task, the window size has few effect on P@10 but has a great impact on the MAP. This is due to the fact that the terms used as expansion barely modify the 10 first documents retrieved by the search engines, but have a greater impact on the whole list of retrieved documents (expansion helps finding documents that share no common words with the original queries). Large windows help identifying words appearing in the same documents than the query words (much like LSI would do) and thus benefits to the MAP.
Using a small dimension is detrimental to every task; The resulting vector space cannot represent adequately the word semantics. Conversely, a large number of dimensions does not allow the necessary generalization needed for retrieving new documents (IR MAP) or the large variety of relations used in WN+M. Thus, for these evaluation scenarii, a dimension between 300 and 500 appears as a good compromise. But for the SimLex999 and the WN-based evaluation, a large number of dimension will allow a more precise description of the word: such setting is more suited to detect synonymy or close semantic links.

7 Conclusion

In this article, following the work of Claveau and Kijak (2016), we compared different scenarii to evaluate distributional and embedding models. Beside the intrinsic evaluation through reference lexicons (here, WordNet and WordNet+Moby) and specialy crafted data (SimLex999), we also rely on an extrinsic evaluation with an IR task, as initially proposed by Claveau and Kijak (2016).

In this work, several conclusions are worth noting. The main one is certainly that direct, or intrinsic, evaluation (be it with reference lexicons or specially crafted data) should be avoided if possible. Indeed, the thesaurus characteristics they evaluate are unclear and may be very different from one's specific need. This is particularly obvious when comparing the IR task results with those of the WN+Moby evaluation (Section 5 and see also (Claveau and Kijak, 2016) for further discussion). Indeed, the very weak results

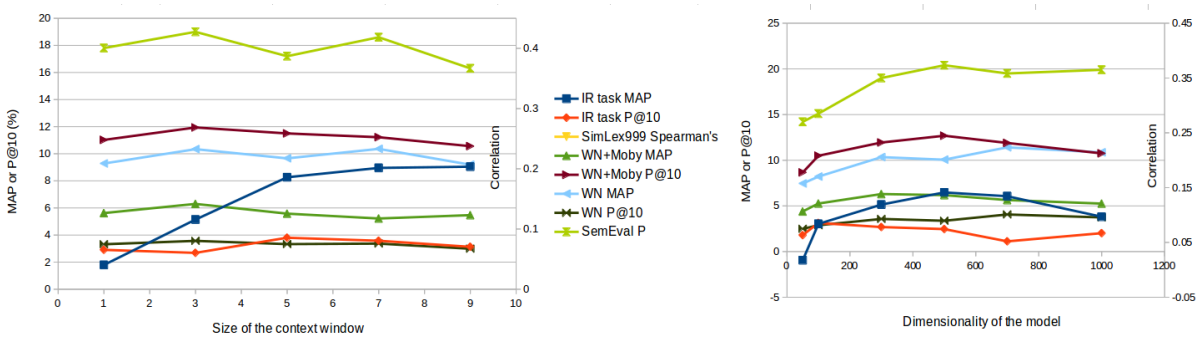

Figure 5: Influence of context size and dimensionality in word2vec models on the results of different evaluation protocols. On the left,the size of the context window is set to 3; on the right, the dimensionality is set to 300.

of the generated thesaurus at the lexicon-based evaluations are not confirmed in the third-party evaluation framework (in our case, query expansion for IR).

Beside that, the evaluation resources (WordNet, Moby or SimLex999) are not complete enough to provide reliable results. For instance, by showing that the thesaurus generated with our models obtains extrinsic results at least as good as the reference lexicons (WN and Moby) used for the intrinsic evaluation, we question previous conclusions of many studies only based on intrinsic evaluation. This incompleteness problem also exists with SimLex999 since it focuses on frequent words (cf. Section 4.3), which are known to be the easiest to model with distributional approaches, but are not necessarily the most interesting for one's needs. Indirect, or task-based, evaluation (lexical substitution or IR) seems conceptually more grounded, but the datasets used may suffer from the same problem of incompleteness or non-representativity. Another related conclusion remark, is that the model parameters have very different effects depending on the considered evaluation scenario (cf. Section 6.2). It is thus important to fine tune with respect to the final task rather than on unrelated datasets.

As a side results, all the experiments presented here confirm the interest of using the IR approaches for building distributional thesaurus (Claveau et al., 2014). In particular, the adjusted Okapi-BM25 model offers significant gains of performance over word2vec for most of the evaluation scenarii considered.

Acknowledgements

This work was partly funded by French government supports granted to the FUI project NexGenTV and to the CominLabs excellence laboratory and managed by the National Research Agency in the "Investing for the Future" program under reference ANR-10-LABX-07-01.

References

Clémentine Adam, Cécile Fabre, and Philippe Muller. 2013. Évaluer et améliorer une ressource distributionnelle : protocole d'annotation de liens sémantiques en contexte. *TAL*, 54(1):71–97.

M. Baroni and A. Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10.

Romaric Besançon, Martin Rajman, and Jean-Cédric Chappelier. 1999. Textual similarities based on a distributional approach. In *in Proceedings of the Tenth International Workshop on Database and Expert Systems Applications (DEXA'99)*, pages 180–184.

Holger Billhardt, Daniel Borrajo, and Victor Maojo. 2002. A context vector model for information retrieval. *J. Am. Soc. Inf. Sci. Technol.*, 53(3):236–249, February.

D. Bollegala, Y. Matsuo, and M. Ishizuka. 2007. Measuring semantic similarity between words using web search engines. In *Proceedings of WWW'2007*.

Bartosz Broda, Maciej Piasecki, and Stan Szpakowicz. 2009. Rank-Based Transformation in Measuring Semantic Relatedness. In *22nd Canadian Conference on Artificial Intelligence*, pages 187–190.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Vincent Claveau and Ewa Kijak. 2016. Distributional thesauri for information retrieval and vice versa. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Vincent Claveau, Ewa Kijak, and Olivier Ferret. 2014. Improving distributional thesauri by exploring the graph of neighbors. In *International Conference on Computational Linguistics, COLING 2014*, Dublin, Irlande, August.

Olivier Ferret. 2013. Identifying bad semantic neighbors for improving distributional thesauri. In *51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 561–571, Sofia, Bulgaria.

Olivier Ferret. 2014. Typing relations in distributional thesauri. In N. Gala, R. Rapp, and G. Bel, editors, *Advances in Language Production, Cognition and the Lexicon*. Springer.

John R. Firth, 1957. *Studies in Linguistic Analysis*, chapter A synopsis of linguistic theory 1930-1955, pages 1–32. Blackwell, Oxford.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *20th International Joint Conference on Artificial Intelligence (IJCAI 2007)*, pages 6–12.

Gregory Grefenstette. 1994. *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers.

Masato Hagiwara, Yasuhiro Ogawa, and Katsuhiko Toyama. 2006. Selection of effective contextual information for automatic synonym acquisition. In *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, pages 353–360, Sydney, Australia.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Evaluating semantic models with (genuine) similarity estimation.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *50th Annual Meeting of the Association for Computational Linguistics (ACL'12)*, pages 873–882.

Thomas Landauer and Susan Dumais. 1997a. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Thomas K. Landauer and Susan T. Dumais. 1997b. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211–240.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2177–2185. Curran Associates, Inc.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (ACL-COLING'98)*, pages 768–774, Montréal, Canada.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Diana McCarthy and Roberto Navigli. 2009. The English lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159.

D. Metzler and W.B. Croft. 2004. Combining the language model and inference network approaches to retrieval. *Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval*, 40(5):735–750.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*, pages 746–751, Atlanta, Georgia.

George A. Miller. 1990. WordNet: An On-Line Lexical Database. *International Journal of Lexicography*, 3(4).

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

J. M. Ponte and W. B. Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR '98)*, pages 275–281.

Stephen E. Robertson, Steve Walker, and Micheline Hancock-Beaulieu. 1998. Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive. In *Proc. of the 7th Text Retrieval Conference, TREC-7*, pages 199–210.

M Ruiz-Casado, E. Alfonseca, and P. Castells. 2005. Using context-window overlapping in synonym discovery and ontology extension. In *Proceedings of RANLP-2005*, Borovets, Bulgarie.

M. Sahami and T.D. Heilman. 2006. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of WWW'2006*.

Magnus Sahlgren. 2001. Vector-based semantic analysis: Representing word meanings based on random labels. In *ESSLLI 2001 Workshop on Semantic Knowledge Acquisition and Categorisation*, Helsinki, Finland.

T. Strohman, D. Metzler, H. Turtle, and W.B. Croft. 2005. Indri: A language-model based search engine for complex queries (extended version). Technical report, CIIR.

P. Turney and P. Pantel. 2010. From frequency to meaning : Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

P.D. Turney. 2001. Mining the web for synonyms: Pmiir versus lsa on toefl. *Lecture Notes in Computer Science*, 2167:491–502.

H. Turtle and W.B. Croft. 1991. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information System*, 9(3):187–222.

T. Van de Cruys, T. Poibeau, and A. Korhonen. 2011. Latent vector weighting for word meaning in context. In Association for Computational Linguistics, editor, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1012–1022.

Tim Van de Cruys. 2010. *Mining for Meaning. The Extraction of Lexico-semantic Knowledge from Text*. Ph.D. thesis, University of Groningen, The Netherlands.

Olga Vechtomova and Stephen E. Robertson. 2012. A domain-independent approach to finding related entities. *Information Processing and Management*, 48(4):654–670.

Ellen M. Voorhees. 1994. Query expansion using lexical-semantic relations. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 61–69, New York, NY, USA. Springer-Verlag New York, Inc.

Ellen Vorhees and David Graff. 2008. Aquaint-2 information-retrieval text research collection.

Grady Ward. 1996. Moby thesaurus. Moby Project.

Kazuhide Yamamoto and Takeshi Asakura. 2010. Even unassociated features can improve lexical distributional similarity. In *Second Workshop on NLP Challenges in the Information Explosion Era (NLPIX 2010)*, pages 32–39, Beijing, China.

Maayan Zhitomirsky-Geffet and Ido Dagan. 2009. Bootstrapping Distributional Feature Vector Quality. *Computational Linguistics*, 35(3):435–461.