Distributional Thesauri for Information Retrieval and vice versa

Vincent Claveau, Ewa Kijak

Distributional Thesauri for Information Retrieval and vice versa

Vincent Claveau[1], Ewa Kijak[2]
[1] IRISA-CNRS, [2] IRISA-University of Rennes 1
Campus de Beaulieu, 35042 Rennes, France
vincent.claveau@irisa.fr ewa.kijak@irisa.fr

Abstract

Distributional thesauri are useful in many tasks of Natural Language Processing. In this paper, we address the problem of building and evaluating such thesauri with the help of Information Retrieval (IR) concepts. Two main contributions are proposed. First, following the work of (Claveau et al., 2014), we show how IR tools and concepts can be used with success to build a thesaurus. Through several experiments and by evaluating directly the results with reference lexicons, we show that some IR models outperform state-of-the-art systems. Secondly, we use IR as an applicative framework to indirectly evaluate the generated thesaurus. Here again, this task-based evaluation validates the IR approach used to build the thesaurus. Moreover, it allows us to compare these results with those from the direct evaluation framework used in the literature. The observed differences bring these evaluation habits into question.

Keywords: distributional semantics, IR models, query expansion

1. Introduction

Distributional semantics aims 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 similar distribution than 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.

Evaluating these thesauri remains a crucial point to assess the quality of the construction methods used. 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 allows to estimate 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.

In this article[1], we propose to examine those two aspects – the construction and the evaluation of distributional thesauri – by using information retrieval (IR) both as a set of techniques and as a use case. Concerning the construction, recent work (Claveau et al., 2014) showed that IR systems could advantageously be used to implement distributional analysis systems. We propose in this paper to further explore this IR approach to build thesauri. We examine the interest of various classic models of IR for distributional analysis and compare them with the state-of-the-art.

Regarding the evaluation, we offer an extrinsic evaluation of the generated thesauri through a conventional IR task. We are then able to compare these results with those of the intrinsic evaluation, and therefore to judge the relevance of these assessment scenarios.

After a state-of-the art (next section), the article addresses these two contributions successively: the aspects related to the construction of thesauri are presented in Section 3., while those about the evaluation by IR are in Section 4.. Finally, we present some conclusions and perspectives about this work in the last section.

2. Related work

2.1. Building distributional thesauri

Building distributional thesauri has been the subject of many studies, including the pioneering work of (Grefenstette, 1994) and (Lin, 1998). All these works are based on the distributional assumption (Firth, 1957) summarized by the famous formula: *"You should know a word by the company it keeps"*. It is therefore considered that each word is semantically characterized by all the contexts in which it appears. For an entry word in a thesaurus, words that share similar contexts are proposed; these are called semantic neighbors thereafter. In the studies, the nature of the semantic link between an entry and its neighbors is variable; the neighbors can be synonyms of the entry, hypernyms, hyponyms or other types of semantic links (Budanitsky and Hirst, 2006; Adam et al., 2013, for a discussion)). These semantic links, even if they are very diverse, are nevertheless useful for many applications related to Natural Language Processing. This explains why this field of research is still very active, with contributions on various aspects related to the construction of the thesaurus.

First, different options of what should be considered as a distributional context has been explored. One usually distinguishes between graphical contexts and syntactic contexts. The former are simply the words appearing around the occurrences of a target word. The second are the syntactic predicates and arguments of the occurrences of the target word. The second approach is often considered more accurate, but it is based on a prior parsing step which is not always possible and can even be inaccurate and misleading. There are many connections between distributional semantics and IR. Several researchers have used search engines

[1]This work was partly funded via the BigClin and LIMAH CominLabs excellence laboratory projects financed by the French National Research Agency under reference ANR-10-LABX-07-01.

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). The vector representations of the contexts are also often used in different ways (Turney and Pantel, 2010), but without the usual 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). Yet, several studies have examined the problem of weighting contexts to get more relevant neighbors. For example, (Broda et al., 2009) proposed to not consider directly the weight of contexts, but their ranks in order to overcome the influence of weighting functions. Considering the semantic neighbors of a word, others suggested bootstrap methods to change the weight of its contexts (Zhitomirsky-Geffet and Dagan, 2009; Yamamoto and Asakura, 2010). Moreover, many studies are based on the fact that the "traditional" distributional representation of contexts is very sparse and redundant, as illustrated by (Hagiwara et al., 2006). In this context, several dimension reduction methods also used in IR were tested: from Latent Semantic Indexing (Landauer and Dumais, 1997b; Padó and Lapata, 2007; Van de Cruys et al., 2011) to *Random Indexing* (Sahlgren, 2001), through the non-negative matrix factorization (Van de Cruys, 2010). Recently, (Claveau et al., 2014) proposed to make a deeper analogy between the research on distributional neighbors and a conventional IR problem. All contexts of all the occurrences of a word can indeed be represented as one document or a query, allowing to easily find similar words, or more precisely sets of similar contexts. While sharing many similarities with the state of the art, this simple way to address the problem of building distributional thesauri offers interesting research avenues and easily accessible tools. In this paper, we also adopt this approach which is described with further details in Section 3.1..

2.2. Evaluating distributional thesauri

As mentioned previously, the evaluation of generated thesauri is either intrinsic, by comparison with a reference resource, or extrinsic, through their use in a specific task.

In the case of intrinsic assessment, reference lexicons are needed. It is then easy to calculate precision, recall or any other measure of quality of the generated distributional thesaurus. This approach was used in numerous previous studies. Among the lexicons regularly used as references, let us cite WordSim 353 (Gabrilovich and Markovitch, 2007), or those used by (Ferret, 2013) that exploits larger resources, ie. synonyms for WordNet 3.0 (Miller, 1990) and the Moby thesaurus (Ward, 1996). In this paper, we also use these two resources for our intrinsic assessment; see below for a presentation. Other resources are not directly lexicons, but data sets that can be used for direct assessment, as the set of synonyms from the TOEFL test (Landauer and Dumais, 1997a) or the semantic relationships in BLESS (Baroni and Lenci, 2011).

Direct assessment is appealing for its simplicity, but it raises the question of the adequacy of lexicons used as references. Therefore, several studies have proposed indirect assessments through a task requiring the generated thesauri. One well known task is the lexical substitution as proposed at SemEval 2007 (McCarthy and Navigli, 2009). Given a word in a sentence, the goal is to replace this word by one of its neighbors and to check that this does not alter the meaning of the sentence. The results obtained are then compared to the substitutions proposed by humans. This task therefore focus on exact synonyms to the detriment of other types of semantic relationships.

To our knowledge, the evaluation of distributional thesaurus through IR tasks has not been explored. Of course, the use of information that can be called distributional within an IR framework has been the subject of several studies (Besançon et al., 1999; Billhardt et al., 2002). It continues today by the work on lexical representations learned by neural networks (Huang et al., 2012; Mikolov et al., 2013). In every case, these studies aim at taking advantage of similarities between word contexts to improve the representation of documents and/or the Relevance Status Value function (RSV). However, these studies do not separate the process of creating the distributional thesaurus from the IR process, which makes impossible the evaluation of the contribution of the only distributional information. In our case, the extrinsic IR evaluation we propose (see Section 4.) is simply based on the use of semantic neighbors to expand queries; the rest of the IR system is standard. This allows us to easily assess the quality of the generated thesauri.

3. IR Models for distributional analysis

3.1. Principles and material

As explained in the introduction, the problem of building a distributional thesaurus can be viewed as a search problem of similar documents and can therefore be carried out with IR techniques. In this context, all contexts of a given word in a corpus are collected and compiled. This set of contexts forms what is considered as a document. Building an entry in the thesaurus, ie. finding the closest words (in a distributional 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).

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; 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...). A detailed description of the semantic links considered by these re-

sources is given in (Ferret, 2013; Claveau et al., 2014). 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. The number of nouns and the variety of semantic relations that they contain make this reference a comprehensive evaluation data set, compared with other existing benchmarks (e.g. WordSim 353).

3.2. Test of the IR models

Table 1 presents the results obtained by different thesaurus building systems, applied to the AQUAINT-2 corpus. The performance measures used to compare the generated thesauri with the reference (WordNet + Moby) 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+Moby reference.

For comparative purposes, we report the results obtained under the same experimental conditions with (i) a state-of-the art approach, denoted *base*, that uses a cosine similarity and weighting by mutual information (Ferret, 2013), (ii) an improved version (*rerank*) which uses machine learning technique to rerank neighbors (Ferret, 2013), and (iii) another version (*synt*) based on syntactic contexts (Ferret, 2014) rather than graphic ones. We also report the results of the systems already tested by (Claveau et al., 2014), based on TF-IDF/cosine and Okapi-BM-25 (Robertson et al., 1998). These authors also proposed an adjusted version of the latter called *adjusted-Okapi BM25*, in which the influence of the document size is reinforced by taking $b = 1$ and by the IDF squared, in order to give more importance to the most discriminating context words. We also apply this strategy to get an adjusted version of the TF-IDF/cosine taking the IDF squared.

In addition to these models, we test other IR systems 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). We also test the dimension reduction techniques (LSI, LDA, Random projections (RP)), with different numbers of dimensions. These classical IR models are not detailed further here (Manning et al., 2008, for further details).

First, one can observe the difficulty of the task, since in every case, the precision of the generated thesauri are very low according to this intrinsic evaluation process. The comparison with the reference lexicons therefore leads to a very severe conclusion about the supposed quality of these thesauri. Yet some IR models perform particularly well compared to the state-of-the-art, such as models based on Okapi, or on language modeling. On the contrary, dimension reduction techniques yields low results: The lower the number of dimensions considered, the worse the 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. Frequency analysis

Some authors noted that the frequency of words for which we try to find the neighbors has a great influence on the final quality (Ferret, 2013). With the state-of-the-art techniques, the more frequent the nouns are, the more contexts they have to describe them; and finally, the better the results are. In the following experiment, we check whether the use of IR methods leads to the same observation. In order to do this, we take the previous experimental framework and the adjusted-Okapi model, but here, the frequency of the entry word is taken into account: words having the highest frequencies ($>$1000), those with the lowest frequency ($<$100) and the remaining third with medium frequencies. These results are shown in Table 2. Again, we also report the state-of-the-art results (Ferret, 2013) for comparison.

It appears that the IR approach has a much more stable behavior on the frequencies that the system of (Ferret, 2013). In particular, adjusted-Okapi provides relatively good results for low-frequency words. Since word frequency is directly related to the size of contexts sets, it indicates the importance of normalization according to the size of the documents used in this IR approach.

3.4. Limits of the analogy with IR

The analogy between similar document search and search for distributional neighbors yields good results, but it should however be pointed some limits of this analogy. Indeed, the collection of contexts, which are considered as documents, have substantially different properties than actual documents. To illustrate this, we represent in Figure 1 the distribution of the size (number of words) of standard document (they are those of AQUAINT corpus, ie., newspaper articles) and the distribution of the size of the context collections. One can observe a much larger range of sizes in the case of sets of contexts. It seems therefore important to take this into account to adapt the length normalization part of the RSV functions of the models (ie. the similarity function used by in IR models).

The word distribution is also quite different from that found in an actual document collection. This is illustrated in Figure 2 in which we give the distribution of document frequency (DF), compared again with the one of the original AQUAINT corpus. Context words usually occur in many more contexts than this is the case for real document. For example, the number of words appearing in 1 over 10 000 documents (DF = 0.0001) is nearly 100 times higher than for real documents. As we have already observed in the previous experiments, this phenomenon deserves a specific consideration of in the models (through smoothing in language models or through IDF in vector models for example, or even by inventing new weighting schemes).

Method	MAP	R-Prec	P@1	P@5	P@10	P@50	P@100
Ferret 2013 *base*	5.6	7.7	22.5	14.1	10.8	5.3	3.8
Ferret 2013 *best rerank*	6.1	8.4	24.8	15.4	11.7	5.7	3.8
Ferret 2014 *synt*	7.9	10.7	29.4	18.9	14.6	7.3	5.2
TF-IDF	5.40	7.28	21.73	13.74	9.59	5.17	3.49
adjusted TF-IDF	7.09	9.02	24.68	15.13	11.55	5.96	4.31
Okapi-BM25	6.72	8.41	24.82	14.65	10.85	5.16	3.66
adjusted Okapi-BM25	8.97	10.94	31.05	18.44	13.76	6.46	4.54
LSI dim=50	1.62	2.86	5.00	4.12	3.76	2.78	2.35
LSI dim=500	4.37	6.27	16.00	10.76	8.78	4.61	3.45
LSI dim=1000	5.06	6.87	21.09	13.20	9.96	5.39	4.02
LSI dim=2000	5.11	6.86	23.11	14.34	10.78	5.12	3.72
LDA dim=500	0.60	1.25	2.17	2.21	1.90	1.29	1.13
RP dim=500	5.66	6.48	27.3	12.85	8.67	3.04	1.86
RP dim=2000	5.90	7.04	27.13	13.71	8.94	3.21	1.96
LM Dirichlet $\mu = 25$	6.52	7.56	23.46	11.88	8.16	2.99	1.89
LM Dirichlet $\mu = 250$	6.56	7.43	23.08	12.31	8.17	2.77	1.73
LM Dirichlet $\mu = 2500$	5.83	6.77	23.28	12.06	8.00	2.98	1.81
LM Hiemstra $\lambda = 0.45$	5.41	6.79	25.09	12.07	8.17	3.05	1.90
LM Hiemstra $\lambda = 0.65$	8.10	8.98	27.06	13.35	9.25	3.41	2.13
LM Hiemstra $\lambda = 0.85$	7.06	7.88	25.28	12.44	8.41	3.04	1.89
LM Hiemstra $\lambda = 0.95$	6.49	7.64	27.21	13.62	9.17	3.28	2.06

Table 1: Performance of IR models for building distributional thesauri over the WN+Moby reference

Freq.	Method	MAP	R-Prec	P@1	P@5	P@10	P@50	P@100
high	Ferret 2013 *base*	6.5	11.0	41.3	26.8	20.8	-	7.3
	adjusted Okapi	7.21	10.73	39.78	24.8	19.31	9.16	5.99
medium	Ferret 2013 *base*	7.4	9.3	20.9	12.3	9.3	-	3.2
	adjusted Okapi	9.85	11.32	30.58	16.19	11.85	5.19	3.55
low	Ferret 2013 *base*	2.4	2.1	3.3	1.7	1.5	-	0.7
	adjusted Okapi	6.93	6.79	9.88	4.83	3.84	1.97	1.49

Table 2: Performance for building distributional thesauri over the WN+Moby reference according to the frequency of the words considered

4. Evaluating through an IR task

To assess the contribution of a distributional thesaurus in a classic task of IR, we use it to expand queries. For each query noun, its neighbors found in the considered thesaurus are added to the query. We describe below our experimental context, and the results obtained. We then propose to draw a parallel between this indirect assessment and the results of the intrinsic evaluation seen in the previous section.

4.1. Experimental setting

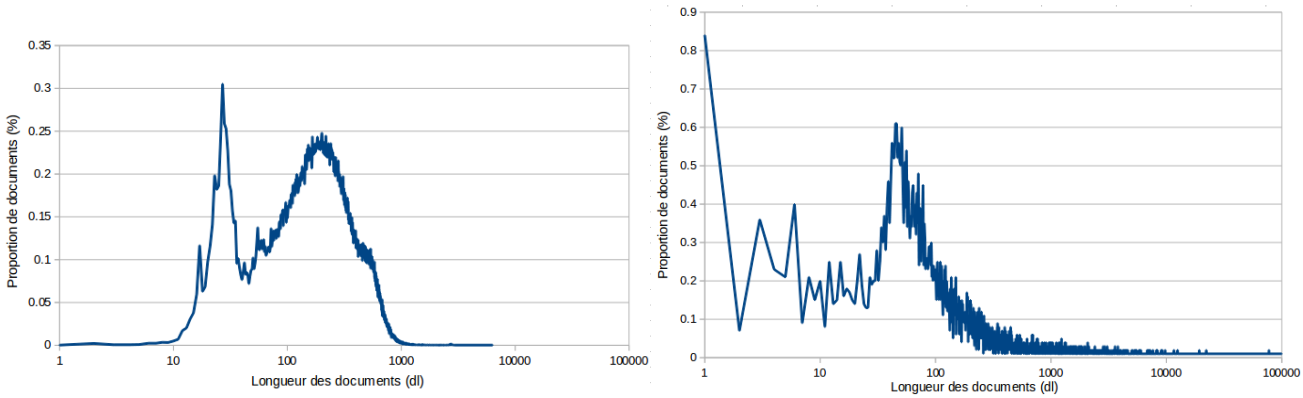
The IR collection used in the experiments is the one developed for the Tipster project and used as part of TREC. It contains more than 170 000 documents and 50 queries. These queries are composed of several fields (the query itself, a narrative field detailing the criteria of relevance); in the experiments reported below, we only use the query field. This collection is particularly suited since it is composed of English documents of the same nature as the AQUAINT-2 corpus (articles of the *Wall Street Journal*) from which the distributional thesaurus was built.

The IR system we use is Indri (Metzler and Croft, 2004; Strohman et al., 2005), known for offering state-of-the-art performance. This probabilistic system implements a combination of language modeling (Ponte and Croft, 1998) and inference networks (Turtle and Croft, 1991). In the experiments reported below, we use it with standard settings, ie. Dirichlet smoothing (with $\mu = 2500$ as recommended). In our case, this IR system offers the additional advantage of having 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, we use the dedicated operator `'#syn'` to aggregate the counts of the words indicated as synonyms (see Indri documentation for details). To remove the effects of flexion (plural) 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.

The performance for this IR task is typically measured by precision at different thresholds (P@x), R-precision, and MAP (Mean Average Precision). 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). In the tables below, non statistically significant results (Wilcoxon and t-test with $p < 0.05$) are in italics.

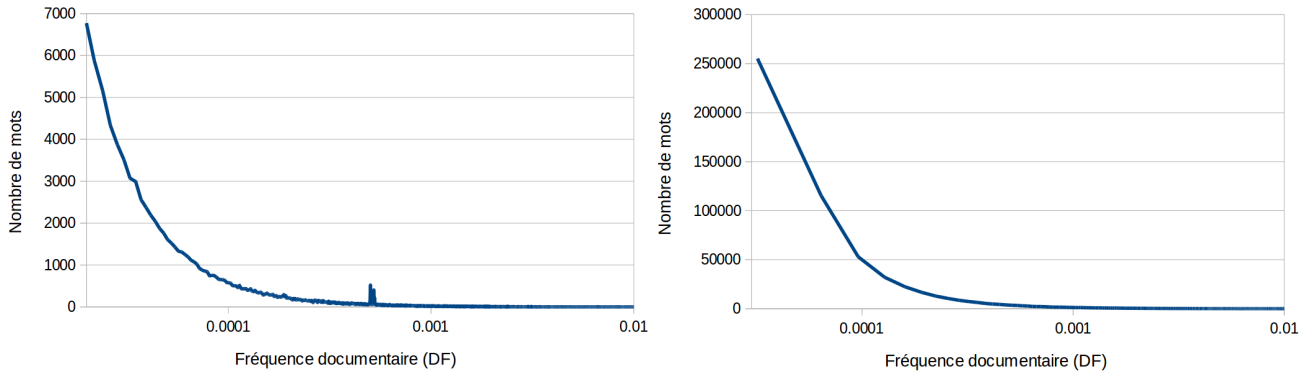
4.2. Expansion results

Table 3 presents the performance gains achieved by expanding the queries with the words collected in the thesaurus. We choose the thesaurus with the best intrinsic results, that is, the one built with the adjusted Okapi method. Since this thesaurus orders the neighbors by proximity with



(a) (b)

Figure 1: Distributions of the size of documents with standard documents (a) and with sets of contexts (b); log. scale



(a) (b)

Figure 2: Distribution of document frequencies (DF) with standard documents (a) and with set of contexts (b); log. scale

the entry-noun, we test different scenarios: for each noun of the query, we only keep the 5, 10 or 50 nearest neighbors. For comparison purposes, we also show the results obtained by expanding the queries with the reference lexicons WN alone and WN+Moby. Here is 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 ) )`

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, that even exceed those of the reference lexicons. While its precision on the first 10 neighbors was evaluated under 14% in Section 3., this thesaurus generates expansions yielding the best MAP gain. The average AP gains (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 queries neighbors added by WN are very close semantically (exact synonyms). This stability is lower with other thesauri, and is the lowest with the expansions by the 50 nearest neighbors from the thesaurus generated with adjusted Okapi model. As the MAP gain remains generally good, it indicates that only certain queries benefit of significant absolute gains.

4.3. Intrinsic vs. extrinsic evaluation

The results of the previous experiences raise questions about the consistency between intrinsic and extrinsic evaluations. We want to know if the gain of precision between two thesaurus construction methods, even if stated as statistically significant, is sensible in IR. In order to answer this question, we propose additional experiments comparing intrinsic precision with extrinsic performance. Figure 3 reports the results of query expansion with the first 10 neighbors of several thesauri generated with various IR models, according to their intrinsic P@10. It shows that the preci-

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 5	+13.14	+29.99	+11.17	+3.45	+5.15	+9.40	+12.43
with adjusted Okapi top 10	+13.80	+24.36	+9.58	*+2.16*	+4.03	+5.58	+8.26
with adjusted Okapi top 50	+10.02	+17.99	+8.82	+3.45	+3.36	+3.72	+5.36

Table 3: Relative gain of performance (%) when expanding queries with different thesauri

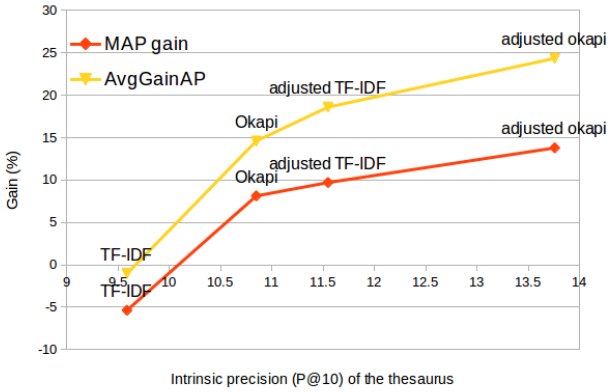


Figure 3: MAP gains et AvgGainAP for query expansion with thesauri generated by various models according to their intrinsic P@10

sion measured with the direct evaluation is related to query expansion gains since the order is respected: The best thesaurus according to intrinsic evaluation (best P@10) gets the best MAP gain at the IR task, etc. Yet, the correlation is not linear as it might be expected. Moreover, statistically significant differences in the intrinsic evaluation (as between adjusted TF-IDF and adjusted Okapi) do not necessarily result in statistically significant differences in the expansion task. Among the false positives according to intrinsic evaluation (words detected as close by the retrieval method but absent in reference lexicons), some seem more or less harmful to expand queries.

It is interesting to explore further the effect of these false positives. Again, we examine the evolution of the performance of the IR task depending on the intrinsic quality of the neighboring lists used to expand queries, but this time, neighbor lists with more or less noise are generated from the reference thesaurus. We control the amount of noise by replacing neighbors with words randomly chosen from the vocabulary. Therefore, it is possible to produce neighbor lists with a controlled intrinsic precision and to evaluate their performance for expanding queries. Figure 4 shows the evolution of MAP and AvgGainAP according to precision of artificially noisy neighbors lists generated from WN alone and WN+Moby. For comparison purposes, we report the MAP obtained from the top 5, 10 and 50 of the adjusted-Okapi thesaurus.

As expected, the two performance metrics fall when intrinsic precision of the neighbor lists decreases. Yet, no performance gains on the IR task are observed for lists with precision of 50%, and below this precision, expansions degrade the results. So, there is indeed a correlation between the accuracy of the lists measured by intrinsic evaluation and performance as measured by extrinsic evaluation, at least when the false positives are random. But in the case of the generated thesaurus, the IR performance obtained is comparable to lists with an intrinsic precision between 70 and 100% (depending on the cases), while the actual intrinsic precision of the thesaurus ranged between 10 and 20%. More than the severity of the intrinsic evaluation, this highlights the weakness of the intrinsic evaluation based on references whose completeness cannot be taken for granted: some neighbors considered as false positives, because not listed in the references, are actually good candidates.

To illustrate this last point, we report in Table 4 the performance obtained by the adjusted Okapi thesaurus when expanding queries with the first 10 neighbors of each noun, but excluding those who are listed as neighbors in WN or WN + Moby. In other words, we only keep the neighbors judged as false positives by the intrinsic evaluation. Clearly, the results obtained suggest that these alleged false positives are semantically related to the entry. For the word `jail` seen in the previous query, among the top 10 neighbors, those absent from WN + Moby are: `award`, `abuse`, `detainee`, `guard`, `custody`, `defendant`, `inmate`, `prisoner`. They actually seem semantically related to *jail*.

5. Conclusion

In this article, we explored the use of IR both to build and to evaluate a distributional thesaurus. We firstly used the similarities of models developed in IR on the contexts of words which allows us, for a given word, to find those sharing a contextual similarity, and hence semantics. Moreover, through classical task of IR, we offer an application framework for an indirect evaluation of the thesaurus.

In this work, two major conclusions emerge. By extending the approach proposed by (Claveau et al., 2014), we confirmed the validity of the IR approach for building distributional thesaurus. More specifically, we have shown the importance of taking into account the discriminating words in different models (through specific weights with IDF or smoothing). We also shed light on the advantage of IR models over conventional methods when dealing with words with few occurrences. Of course, there are also limits to the analogy between IR and distributional analysis: The sets of contexts have statistical properties (size, word frequency...) that are very different from 'real' documents. This argues for the establishment of weighting and RSV functions adapted to this reality and therefore opens poten-

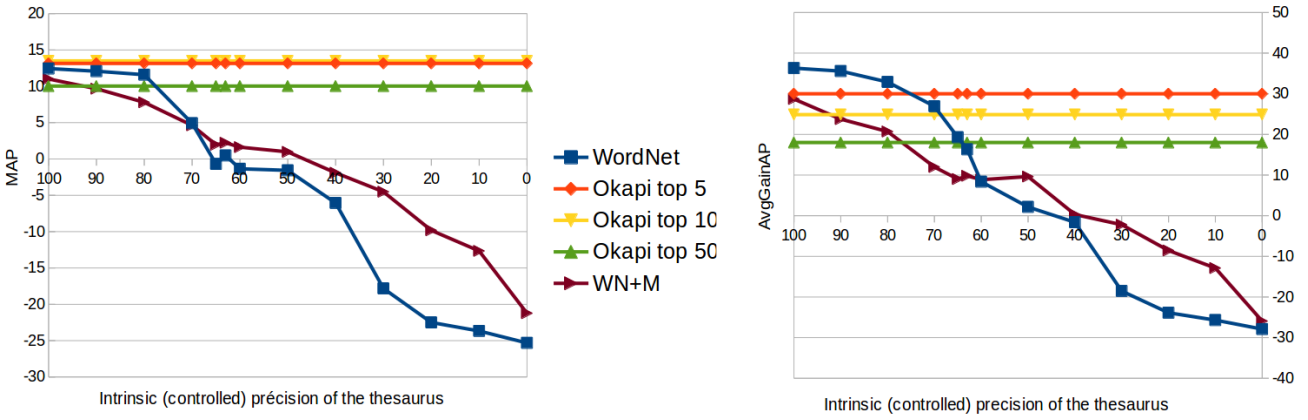


Figure 4: MAP gains (left) and AvgGainAP (right) of expanded queries according to the controlled precision of the thesaurus used for query expansion

Expansion with adjusted Okapi	MAP	AvgGainAP	R-Prec	P@5	P@10	P@50	P@100
top 10 but WN	+11.80	+21.60	+8.37	*+2.16*	+3.58	+5.08	+6.87
top 10 but WN+M	+9.36	+19.22	+6.41	*+3.02*	+3.36	*+3.17*	+5.73

Table 4: Relative gains of performance (%) when expanding queries with neighbors that are considered as false positives in the intrinsic evaluation

tial avenues for improvement. Other perspectives on this part concern the use of recent techniques of IR for the construction of thesauri (learning to rank, continuous representations...).

The other major conclusion of this article is about the reliability of the intrinsic evaluation. 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. Indeed, the very weak results of the generated thesaurus at the intrinsic evaluations are not confirmed in the third-party evaluation framework (in our case, query expansion for IR). Of course, these conclusions should be put into perspective: our IR task may be less sensitive to expansions that are loosely related to the entry. Other tasks, such as lexical substitution, more focused on exact synonyms, might give different results. Therefore, an interesting perspective would be to measure the correlation between the intrinsic and extrinsic evaluation scores in different tasks and scenarios to better help choose the most suitable distributional method for a given task.

6. References

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

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

Besançon, R., Rajman, M., and Chappelier, J.-C. (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.

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

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

Broda, B., Piasecki, M., and Szpakowicz, S. (2009). Rank-Based Transformation in Measuring Semantic Relatedness. In *22nd Canadian Conference on Artificial Intelligence*, pages 187–190.

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

Claveau, V., Kijak, E., and Ferret, O. (2014). Improving distributional thesauri by exploring the graph of neighbors. In *International Conference on Computational Linguistics, COLING 2014*, Dublin, Irlande, August.

Ferret, O. (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.

Ferret, O. (2014). Typing relations in distributional thesauri. In N. Gala, et al., editors, *Advances in Language Production, Cognition and the Lexicon*. Springer.

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

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

Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers.

Hagiwara, M., Ogawa, Y., and Toyama, K. (2006). Selection of effective contextual information for automatic synonym acquisition. In *21st International Conference on Computational Linguistics and 44th Annual Meet-*

ing of the Association for Computational Linguistics (COLING-ACL 2006), pages 353–360, Sydney, Australia.

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (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.

Landauer, T. and Dumais, S. (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.

Landauer, T. K. and Dumais, S. T. (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.

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

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

McCarthy, D. and Navigli, R. (2009). The english lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159.

Metzler, D. and Croft, W. (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.

Mikolov, T., Yih, W.-t., and Zweig, G. (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.

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

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

Ponte, J. M. and Croft, W. B. (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.

Robertson, S. E., Walker, S., and Hancock-Beaulieu, M. (1998). Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive. In *Proc. of the 7^{th} Text Retrieval Conference, TREC-7*, pages 199–210.

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

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

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

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

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

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

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

Van de Cruys, T., Poibeau, T., and Korhonen, A. (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.

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

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

Voorhees, E. M. (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.

Ward, G. (1996). Moby thesaurus. Moby Project.

Yamamoto, K. and Asakura, T. (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.

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