Textual Similarities based on a Distributional Approach

Romaric Besançon, Martin Rajman, Jean-Cédric Chappelier
Swiss Federal Institute of Technology, Artificial Intelligence Laboratory,
Lausanne, Switzerland
{Romaric.Besançon,Martin.Rajman,Jean-Cedric.Chappelier}@epfl.ch

Abstract

The design of efficient textual similarities is an important issue in the domain of textual data exploration. Textual similarities are for example central in document collection structuring (e.g. clustering), or in Information Retrieval (IR) which relies on the computation of textual similarities for measuring the adequacy between a query and documents.

The objective of this paper is to present and compare several textual similarity measures in the framework of the Distributional Semantics (DS) model for IR. This model is an extension of the standard Vector Space model, which further takes the co-frequencies between the terms in a given reference corpus into account. These co-frequencies are considered to provide a distributional representation of the "semantics" of the terms. The co-occurrence profiles are used to represent the documents as vectors.

Practical retrieval experiments using DS-based similarity models have been conducted in the framework of the AMARYLLIS evaluation campaign. The results obtained are presented. They indicate significant improvement of the performance in comparison with the standard approach.
Keywords *:Textual similarity, Information Retrieval, Distributional Semantics.*

1. Introduction

The increasing amount of textual data available in electronic form is an important motivation for the search of efficient techniques in the general field of textual data exploration. For most of the domains in this area (Information Retrieval, Textual Database Clustering, Topic Detection) efficient textual similarity computation is a central point.

Structuring large textual databases, for example, can be achieved by clustering close documents, on the basis of a well-chosen textual similarity model, semantically grounded. Similarly, the objective of Information Retrieval (IR) is to efficiently identify relevant documents in a database, satisfying an information need expressed by a user in a form of a textual query. This problem is strongly related to the notion of textual similarity since it can be viewed as the search, in a given space to be defined, of the documents the most similar to the query. Such a search can be carried out through the computation of textual similarities between the query and each of the documents in the database.

The standard models used in Information Retrieval are based on a vector representation of the documents associated with a similarity measure on the underlying vector space. The model proposed in this paper refines the standard model by introducing a representation of the document that contains more "semantic" information, using a distributional representation of the semantics of the words.

Section 2 first presents the standard Vector Space (VS) model [8], and then the extension of this model using distributional information, along with some further developments.

Section 3 presents practical experiments conducted with the DS model in the framework of the AMARYLLIS evaluation campaign for Information Retrieval systems for French and the promising results obtained in this context.

2. Distributional textual similarity

2.1. Vector Space model

2.1.1. Document representation

The standard VS model uses statistical information, in particular the distribution of terms extracted from the collection to represent the documents and the queries. More precisely, in the SMART model [8], each document d_n is represented by a vector $(w_{n1}, \ldots, w_{nM})$, where w_{nk} is the *weight* (or importance) of the *term* t_k in the document d_n, and M is the size of the indexing term set. The vector $(w_{n1}, \ldots, w_{nM})$ is called the *lexical profile* of the document.

A term is a chosen "semantic" textual unit. It can be a word, a stem, a lemma, a compound. The terms used to

index the documents are chosen to be as discriminative as possible. Salton & al. [9] showed that selecting the terms with *document frequency* (*i.e.* the number of documents in which a term occurs) between N/100 and N/10 keeps terms with a good discriminative power.

The weight of a term in a document is often simply the number of occurrences of the term in the document (*occurrence frequency*). However, a weighted scheme taking into account the term importance within the entire collection improves the retrieval performance [9]. More weight should be given to terms that rarely occur within the collection (terms that are used in many documents are more general and less useful for discrimination that the ones that appear in very few documents). Such a goal can be achieved by using the occurrence frequency of the term weighted by the inverted document frequency factor ($\log(\frac{N}{df_i})$, where N is the number of documents and df_i the document frequency of a term t_i).

A lot of other weighting schemes have been proposed over the years [1, 8], using functions depending on occurrence frequencies (to reduce the range of frequencies) or factors introducing document length normalisation (to reduce the otherwise systematic advantage of long documents over short ones [11]).

Within the VS framework, a collection of N documents is then represented by a $N \times M$ *occurrence matrix* F, each row being the lexical profile of a document:

$$F = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{pmatrix} = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1M} \\ w_{21} & w_{22} & \dots & w_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & \dots & w_{NM} \end{pmatrix}$$

2.1.2. Textual similarity and document retrieval

For the queries, a similar vector representation is applied: each query q is represented by a vector $(q_1, q_2, \dots, q_M)$ where q_i is the weight of term i in the query.

Retrieval is achieved by measuring the similarity between a document and a query in the underlying vector space. A usual similarity measure is the cosine of the angle between the two vectors.

$$\text{sim}_{\cos}(d_n, q) = \frac{d_n \cdot q}{||d_n||\ ||q||} = \frac{\sum_{i=1}^{M} w_{ni}\ q_i}{\sqrt{\sum_{i=1}^{M} w_{ni}^2 \ \sum_{i=1}^{M} q_i^2}}$$

Several other similarity measures can also be used for textual information retrieval (χ^2 distance, Kullback-Leibler Divergence). Their properties have been presented in [2, 5].

2.2. Distributional Semantics model

2.2.1. The DS concept

The Distributional Semantics (DS) model assumes that there exists a strong correlation between the observable distributional characteristics of a word and its meaning, *i.e.* the semantics of a word is related to the set of contexts in which that word appears [4]. For instance, given the following contexts:

(1) *The X sleeps near the wooden fence.*
(2) *The X chews the grass in the meadow.*
(3) *The farmer shears the white X.*

The set of words {*sleep, fence, chew, grass, meadow, farmer, shear, white*} that constitutes a simplified representation of the cumulated contexts of *X*, gives sufficient information (at least for a human) to identify *X* as a "sheep".

Therefore, the DS hypothesis can be rephrased as: two words are semantically similar to the extent that their contexts are similar. Following this, a possible operational implementation of the above hypothesis is that two documents are similar if the average context[1] of the terms they contain are similar.

Distributional Semantics applied to Information Retrieval has been studied in several systems (DSIR [6, 7], Co-occurrence based IR [10]).

2.2.2. Document representation

The notion of word context is defined within a co-occurrence model. The co-occurrence frequency (or *co-frequency*) between two words is defined as the frequency of both words occurring within a given textual unit. Typical examples of textual units can be k words windows, sentences, paragraphs, sections, or whole documents.

For a given term, the *co-occurrence profile* is defined by the vector of the co-frequencies between this term and each element of an *a priori* chosen set of terms, referred to as the set of *indexing features*. If P denotes the size of the indexing feature set, the co-occurrence profile of a term t_i can be written $c_i = (c_{i1}, \dots, c_{iP})$.

For a set of M terms, a $M \times P$ *co-occurrence matrix* C is built, each row representing the co-occurrence profile of a term:

$$C = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_M \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1P} \\ c_{21} & c_{22} & \dots & c_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ c_{M1} & c_{M2} & \dots & c_{MP} \end{pmatrix}$$

[1] the notion of "average context" is clarified in the next section

A whole document is represented by the weighted average of the co-occurrence profiles of the terms it contains: $d_n = \sum_{i=1}^{M} w_{ni}\, c_i$

The weight w_{ni} given to each co-occurrence profile c_i is the same as the one given to the term t_i in the VS model. The document collection can then be represented by the product matrix: $D = FC$

Notice that the dimension of the vector representation of a document is the size of the indexing feature set, which may be much smaller than the size of term set of the VS model. However, the DS model actually takes as many terms as the VS model into account, because all the terms in the term set (even the ones not present in the indexing feature set) are represented in the weighted average through their co-occurrence profile on the indexing features. The DS model can then be seen as a good way to reduce the dimensionality of the vector representation, from the whole term set to the indexing feature set, using distributional information connecting the two sets, *i.e.* the distributions of co-frequencies between the terms and the indexing features.

The problem of the choice of indexing features with good discriminative power still remains. Like for the terms in the VS model (see section 2.1.1, page 1), the indexing features can be chosen in the term set on the basis of their document frequencies.

The textual similarity for document retrieval in the DS model is similar to the one used in the VS model (a similarity measure based on the vector representation of the documents and queries), for instance the cosine similarity. The difference between the VS model and the DS model does therefore not originate in different similarity measures, but rather in a different document vector representation.

In order not to lose the direct information about the terms that are effectively contained in the documents, an hybrid representation of the documents, combining the VS approach with the DS approach, can be used.

If we call F' the VS weighted occurrence matrix restricted only to terms that also are indexing features, and α a scalar hybridation parameter ($0 \leq \alpha \leq 1$), the hybrid document representation is: $D = (1-\alpha)F' + \alpha FC$

2.3. DS refinements

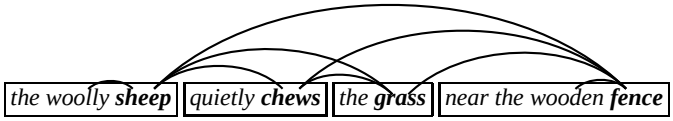
Several refinements for the co-occurrence matrix computation have been considered. As far as semantics is concerned, the simple co-occurrence model (based on words co-occuring within a given textual unit, with no additional restriction) can take inappropriate co-occurrences into account. For example, let us consider the following sentence : "*the woolly sheep quietly chews the grass near the wooden fence.*"

Taking the whole sentence as a textual unit, the *wooden-fence* or *woolly-sheep* co-occurrences seem to be relevant, whereas the *woolly-fence*, *wooden-sheep*, *woolly-wooden* co-occurrences seem spurious (or at least not suggested by the meaning of this sentence).

To solve this problem, more syntactic information is integrated in the process of co-occurrence computation [3, 6]. The co-occurrences are computed only

- between the words within the same syntactic group[2]
- between the heads[2]of different syntactic groups

In the previous example, considering only co-occurrences between nouns, verbs and adjectives and representing in bold the heads of the syntactic groups, we obtain:



i.e. the effective computed co-occurrences will be

within syntactic group	between heads	
woolly-sheep	*sheep-chew*	*chew-grass*
wooden-fence	*sheep-grass*	*chew-fence*
	sheep-fence	*grass-fence*

Notice that spurious co-occurrences have been eliminated.

This technique therefore allows to improve the relevance and the discriminative power of the co-frequencies in the co-occurrence matrix. Moreover, since less co-occurrences are considered, this technique further allows, on a practical level, to use a more important set of indexing features.

3. Experiments

Practical experiments using the DS-based model have been conducted in the framework of the AMARYLLIS evaluation campaign for Information Retrieval systems for French.

3.1. Data and parameters

The AMARYLLIS campaign is organised in two steps: a training phase on reference document collections, and a test phase on another document collections. Since the results of the test phase have not been released yet, only the results obtained during the training phase are presented in this section.

The data for the training phase is composed of three reference corpora:

[2] obtained by a syntactic analysis

- LRSA: a set of 502 documents extracted from books on Melanesia, along with 15 search themes associated with 423 relevant documents[3];
- OFIL: a set of 11016 articles from a newspaper (Le Monde), along with 26 search themes associated with 587 relevant documents[3];
- INIST: a set of 163308 documents from bibliographical notes, along with 30 search themes associated with 1407 relevant documents[3].

The documents and queries were first analysed by a syntactic parser , in order to find the part-of-speech tags and lemmas of the words, and to identify the syntactic groups along with their heads. A set of 62895 terms (lemmas of nouns, verbs and adjectives) were extracted from the documents and queries.

Indexing features for the co-occurrences matrices are chosen in the term set according to their document frequency. Three kinds of hybrid model have been considered, depending on the sets of indexing features and the computation method for co-occurrences:

	nb indexing features	document frequencies	uses syntactic dependencies
Hyb1	2382	[450, 1500]	no
Hyb2	2832	[450, 1500]	yes
Hyb3	6131	[200, 5000]	yes

These three kinds of model have been tested for a hybridation parameter $\alpha = 0$ (corresponding to the standard VS model), $\alpha = 0.5$ and $\alpha = 1$ (DS model). 250 documents were retrieved for each query, and ordered according to their similarity to the query.

3.2. Results

The evaluation of the IR systems is usually done with the standard measures of *precision* (P) and *recall* (R), where:

$$P = \frac{\textit{number of relevant documents retrieved}}{\textit{total number of documents retrieved}}$$

$$R = \frac{\textit{number of relevant documents retrieved}}{\textit{total number of relevant documents}}$$

Table 1 presents the average precision, the R-precision[4] and precisions $P(n)$ for n retrieved documents, averaged on the three corpora, as well as the total number of relevant documents retrieved, cumulated on the three corpora.

On one hand, these results show a significant improvement in the precision score when using a DS-based model ($\alpha > 0$) rather than the standard VS model ($\alpha = 0$).

[3] as judged by human experts

[4] the R-precision is the precision obtained at a number of retrieved documents corresponding to the actual number of relevant documents. So in that particular case, precision equals recall.

On another hand, as far as the model for co-occurrence computation is concerned, results are not really significant. With the same size of indexing feature set (Hyb1 and Hyb2), the results appear to be worse when using the syntactic dependencies for the co-occurrences computation (the matrix is much sparser so that the remaining discriminative power does not seem to be sufficient to balance the decrease of information). However, a slight improvement of the performance is noticed when using the syntactic dependencies for a larger set of indexing features (Hyb3), that could not otherwise be used (the computation of all co-occurrences for such an indexing feature set would not be tractable).

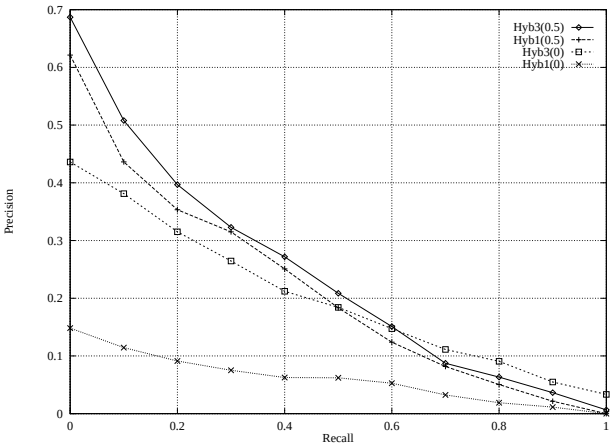


Figure 1. Precision/Recall of the different systems for the OFIL corpus

In order to illustrate further the comparison between the hybrid and VS model, and the improvement obtained by enlarging the indexing feature set, Figures 1 and 2 present the precision/recall graphs for the OFIL and INIST corpora, for Hyb1 and Hyb3 with $\alpha = 0$ and $\alpha = 0.5$.

These results show that the improvement of the DS approach is very significant for a low recall, and less significant (or even slightly worse) for high values of recall (notice that the results for the high values of recall are not really significant, since the precision/recall values are computed for until 1000 documents retrieved and the systems retrieve only 250 documents).

The precision of the system for low/high values of recall must be estimated according to the purpose of information retrieval systems. Usually, a system with good precision for low values of recall will be favoured in cases where there is a need for strongly relevant documents in the first documents retrieved (for instance, in a search engine on the Internet), whereas a system with good precision for high values of recall will be preferred if all the relevant documents present in the database have to be identified (for instance, in a search for legal precedents for jurisprudence).

	$\alpha = 0$		$\alpha = 0.5$			$\alpha = 1$		
	Hyb1 Hyb2	Hyb3	Hyb1	Hyb2	Hyb3	Hyb1	Hyb2	Hyb3
nb docs retrieved	904	1312	1322	1280	**1335**	1262	904	1200
Avg. Precision	0.1320	0.2202	0.2544	0.2413	**0.2688**	0.2393	0.1320	0.2380
R-Precision	0.1503	0.2450	0.2799	0.2716	**0.3078**	0.2774	0.1503	0.2760
Prec. at 5 docs	0.2113	0.3690	0.4639	0.4552	**0.4959**	0.4561	0.4492	0.4554
Prec. at 10 docs	0.1942	0.3408	0.4159	0.3807	**0.4277**	0.3905	0.3770	0.3895
Prec. at 15 docs	0.1803	0.2994	0.3564	0.3471	**0.3815**	0.3609	0.3507	0.3572
Prec. at 20 docs	0.1666	0.2762	0.3218	0.3129	**0.3435**	0.3150	0.3119	0.3172
Prec. at 30 docs	0.1434	0.2358	0.2724	0.2668	**0.2929**	0.2617	0.2590	0.2685
Prec. at 100 docs	0.0919	0.1362	0.1458	0.1401	**0.1471**	0.1385	0.1356	0.1349
Prec. at 200 docs	0.0654	0.0898	0.0924	0.0899	**0.0932**	0.0873	0.0866	0.0848

Table 1. Precision values averaged on the three corpora (best results are in bold)

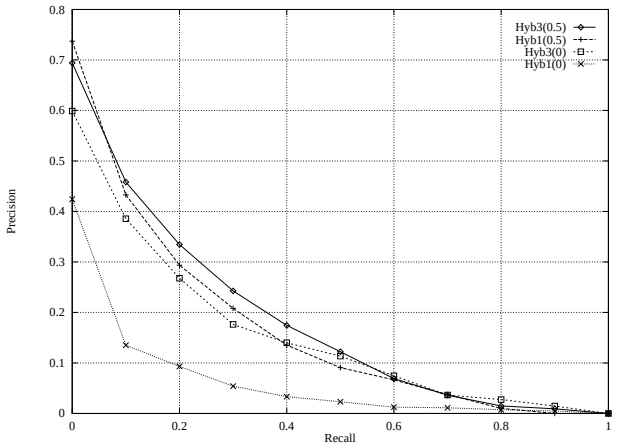


Figure 2. Precision/Recall of the different systems for the INIST corpus

4. Conclusion

This paper presents a textual similarity model based on a distributional representation of the semantics of a term, using Natural Language Processing tools to integrate restricted semantic information derived from a syntactic analysis of the sentences. This distributional semantic information is used to reduce the dimension of the representation space of the documents. The validation of the model can be achieved in several application domains for textual similarities that provide methodologies and metrics for evaluation.

In particular, the model has been tested for an Information Retrieval task on reference data in the framework of the AMARYLLIS evaluation campaign. The results obtained so far are promising. Further validations of the DS-based approach of textual similarity will be conducted on other tasks, like document collection structuring (where textual similarity can be used to group the documents into clusters), word sense disambiguation (where textual similarity can be used to evaluate the relevance of a definition according to a certain context), or novelty detection (where textual similarities can be used to judge whether a document brings new information according to the textual data that have been already processed).

References

[1] J. Lee. Combining multiple evidence from different properties of weighting schemes. In F. E.A., editor, *Proceedings of the 18th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, USA, 1995.

[2] L. Lee. *Similarity-Based Approaches to Natural Language Processing*. PhD thesis, Harvard University, 1997.

[3] M. Rajman. *Apports d'une approche à base de corpus aux techniques de traitement automatique de langage naturel*. PhD thesis, ENST, Paris, 1995.

[4] M. Rajman and A. Bonnet. Corpora-base linguistics: new tools for natural language processing. In *1st Annual Conference of the Association for Global Strategic Information*, Bad Kreuznach, Germany, 1992.

[5] M. Rajman and L. Lebart. Similarités pour données textuelles. In *4^{th} International Conference on Statistical Analysis of Textual Data (JADT'98)*, Nice, February 1998.

[6] A. Rungsawang. *Recherche Documentaire à base de sémantique distributionnelle*. PhD thesis, ENST, Paris, 1997.

[7] A. Rungsawang and M. Rajman. Textual information retrieval based on the concept of distributional semantics. In *proc. of JADT'95 (3rd International Conference on Statistical Analysis of Textual Data)*, Rome, 1995.

[8] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:513–523, 1988.

[9] G. Salton, C. Yang, and C. Yu. A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 1975.

[10] H. Schütze and J. Pedersen. A co-occurrence-based thesaurus and two applications to information retrieval. In *Proceedings of the RIAO'94*, pages 266–274, Rockefeller University, New York, 1994.

[11] A. Singhal, G. Salton, M. Mitra, and C. Buckley. Document length normalization. Technical report, Department of Computer Science, Cornell University, 1995.