Stochastic Distributional Models for Textual Information Retrieval

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

Abstract

The objective of this paper is to present a textual similarity model for Information Retrieval (IR) based on the Distributional Semantic (DS) model. This model is an extension of the standard Vector Space model, which further takes into account the co-frequencies between the terms in a given reference corpus, that are considered to provide a distributional representation of the "semantics" of the terms. Practical retrieval experiments using DS-based similarity models have been conducted in the framework of the AMARYLLIS evaluation campaign. The results obtained are presented, and 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 and, in particular, Information Retrieval (IR).

The main objective of IR is to efficiently identify relevant documents in a database, satisfying an information need expressed by a user in a form of a query.

This task becomes more difficult as the size of the searched databases increases and approaches aiming at reducing the size of the search space by structuring the document collections are now also considered.

In the domains of IR and textual database structuring (*e.g.* clustering), the design of efficient textual similarities is a central issue. in IR, the goal can be viewed as the search, in a given semantic space, of the documents the most similar to the query. And the search can therefore be carried out through the computation of textual similarities between the query and each of the documents in the database. Document collection structuring can also be achieved by clustering the documents that appear close according to a well-chosen, semantically grounded, textual similarity.

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

The necessary validation of a textual similarity model has to be realised in specific application domains that provide methodologies and metrics for evaluation. The validation framework chosen for the DS model presented in this paper is the AMARYLLIS evaluation campaign for Information Retrieval systems for French which provide reference data on which quantitative performance measures can be carried out.
In section 2, we first focus on the description of the elements (representations and metrics) of the DS model. Then, in section 3, we present the results obtained on the AMARYLLIS reference data.

2. The DS Model

2.1. Document representation

The standard Vector Space (VS) model uses statistical information, in particular the distributions of terms extracted from the collection to represent the documents and the queries. More precisely, in the SMART model (Salton Buckley 1988), each document d_n, is represented by the vector $(w_{n1},\ldots,w_{nM})$, where w_{nk} is the *weight* (or importance) of the *term* t_k in the document d_n (M being the size of the indexing term set). In our work, the weights are always considered as normalised (*i.e.* $\sum_i w_{ni} = 1$). 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 or a compound. The terms used to index the documents are chosen to be as discriminative as possible (as it can be measured, for instance, on the basis of their document frequency (Salton al. 1975)).
The weight of a term in a document is often simply the number of occurrences of the term in the document or *occurrence frequency*. This frequency can be weighted by different factors to take into account the term importance within the entire collection (Salton Buckley 1988), or the document length normalisation (Singhal al. 1995). A collection of N documents is then represented by a $N \times M$ *occurrence matrix* F, in which each row corresponds to the lexical profile of a document.
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.* that the semantics of a word is related to the set of contexts in which that word appears (Rajman Bonnet 1992, Rungsawang 1997).
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 X.*

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

Practically, the context of a term t_i is represented by a co-occurrence profile $c_i=(c_{i1},\ldots,c_{iP})$ defined as the vector of the co-frequencies c_{ik} between the term t_i and each of the P terms t_k of an *a priori* chosen set of terms, referred to as the set of *indexing features*. The co-frequency between two terms is further defined as the frequency of both terms occurring within a given textual unit (typically a k word window, a sentence, a paragraph, a section or a whole document).

For any set of M terms, a $M\times P$ *co-occurrence matrix* C, in which each row represents the co-occurrence profile of a term, can be built to represent the distributional semantics of the M terms.

A document is then represented by the weighted average of the co-occurrence profiles of the terms it contains: $d_n = \sum w_{ni} c_i$, where 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 $D=FC$.

One of the important properties of the DS model is to allow using the whole term set (instead of a reduced number of it as in the VS model) to build the representations of the documents. The dimension reduction is realised in the feature set that is related to the term set through the available distributional information (*i.e.* the distributions of co-frequencies between the terms and the indexing features). As in the VS model, the indexing features can be chosen in the term set on the basis of their discriminative power as measured by their document frequency.

2.2 Similarity Metrics

Retrieval is achieved by measuring the similarity between a document and a query in the underlying vector space. Several similarity measures can be used to evaluate textual similarities (Rajman & Lebart 1998). Motivations underlying the choice of a specific similarity measure often rely on the interpretation given to the vector representation of the documents.

The direction of the vectors representing the documents can for example be viewed as "meanings" in a given "semantic space": the directions of the co-occcurrence profiles of the terms are therefore considered as the "meaning" of the terms (in the "semantic space" of the indexing features space), and the interpretation of the DS model is that the semantic content of a document is the weighted average of the "meanings" (*i.e.* directions) of the terms it contains. In this case, a quite natural similarity measure between a document d_n and a query q is the cosine of the angle between the directions representing the document and the query: $\cos(d_n, q) = \frac{d_n \cdot q}{\|d_n\| \|q\|}$

Another possible interpretation is to consider the vectors representing the documents as stochastic distributions of the meaning of the documents over the terms: the co-occurrence profiles are then interpreted as the distributions of the meaning of the terms over the indexing features, and the interpretation of the DS model is that the semantic content of a document is the distribution corresponding to the weighted average of the "meanings" (*i.e.* distributions) of the terms it contains. In this case, a usual measure for the dissimilarity between distributions is the *Kullback-Leibler (KL) divergence,* or *relative entropy* (Cover and Thomas, 1991), defined for two distributions q and r as: $D(q \parallel r) = \sum_{y \in Y} q(y) \log \frac{q(y)}{r(y)}$

But as the KL divergence is not symmetric, we prefer to use the *Total Divergence to the Mean* (Dagan al. 1997) defined as: $A(q \parallel r) = D(q \parallel \frac{q+r}{2}) + D(r \parallel \frac{q+r}{2})$

$A(q||r)$ is symmetric and has the following interpretation: if q and r are two empirical frequency distributions, then $A(q||r)$ can be used as a test statistic for the hypothesis that q and r are drawn from the same distribution.

Notice that $A(q||r)$ is non-negative ($A(q||r) \geq 0$ with equality holding if and only if $\forall y \in Y\ q(y)=r(y)$), but it is not a metric, because it does not verify the triangle inequality.

3. Experiments

To validate our DS model, we have conducted several experiments on a subset of the reference data from the AMARYLLIS evaluation campaign for Information Retrieval systems for French. Two corpora have been considered:

Corpus	Source	Nb docs	Nb queries	total Nb of relevant docs
LRSA	books on Melanesia	502	15	423
OFIL	newspaper articles	11016	26	587

The documents and queries were first analysed by a syntactic parser (the Sylex software, from Ingenia-LN), in order to find the part-of-speech tags and lemmas of the words. A set of 62895 terms (lemmas of nouns, verbs and adjectives) were extracted from the documents and queries. The co-occurrences were computed on sentences, on a basis of a set of 2382 indexing features with document frequencies in the range of [450,1500].

This model has been tested for the two interpretations described above: the geometric one, in terms of directions (denoted DIR) and the stochastic one, in terms of distributions (denoted DIST), with their associated similarity measure. 250 documents were retrieved for each query, and ordered according to their similarity to the query.

Figures 1 and 2 present the results obtained for the two corpora in terms of *precision/recall* graphs. Precision is the proportion of retrieved documents that are relevant and recall is the proportion of relevant documents retrieved (Salton Mc Gill 1983). Results obtained with the standard VS model are also provided for comparison.

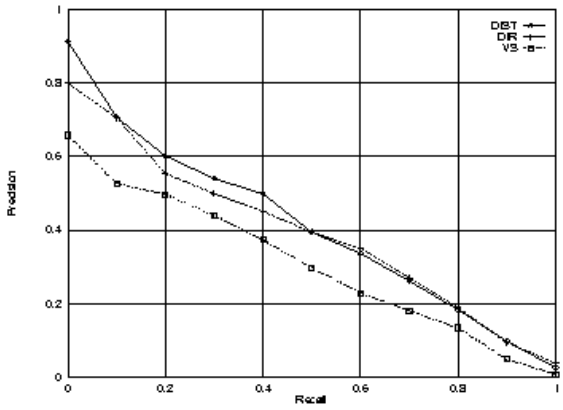

Fig 1: Results for the LRSA corpus

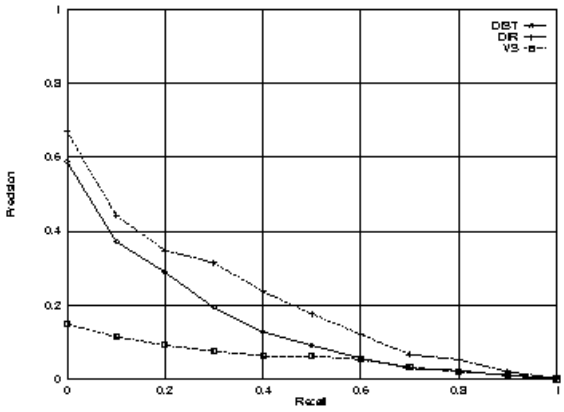

Fig 2: Results for the OFIL corpus

These results show, for both interpretations, a significant improvement of the performance when using a distributional model over the standard VS model. Notice however that they do not provide significant conclusions for the comparison of the two interpretations.

4. Conclusion

This paper presents a textual similarity model based on a distributional representation of the semantics of a term. Two interpretations can be given to this model: a geometric interpretation in terms of directions in a semantic vector space, and a stochastic interpretation in terms of distributions over the indexing feature set. For these two interpretations (and the associated similarity metrics), the model has been tested on

reference data from the AMARYLLIS evaluation campaign. The results show a significant improvement when using a distributional model instead of the standard VS model.

Further validation of the DS model will be conducted on additional tasks, such as 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 sense definitions 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). These experiments should also provide better insight for a more detailed analysis of the differences between the two interpretations (geometric and stochastic) of the DS model.

References

Cover, T.M. and Thomas, J.A (1991). *Elements of Information Theory*, Wiley Series in Telecommunications, Wiley Sons.

Dagan, I., Lee, L. and Pereira, F. (1997). Similarity-based Methods for Word Sense Desambiguation, in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pp 56--63.

Rajman, M. and Bonnet A (1992). Corpora-base linguistics: new tools for natural language processing, in *1st Annual Conference of the Association for Global Strategic Information*.

Rajman, M. and Lebart L (1998). Similarités pour données textuelles, *in 4th$ International Conference on Statistical Analysis of Textual Data (JADT'98)*.

Rungsawang, A. (1997). *Recherche Documentaire à base de sémantique distributionnelle*, PhD thesis, École Nationale Supérieure des Télécommunications, Paris.

Salton, G. and Buckley, C. (1988). Term weighting approaches in automatic text retrieval, *Information Processing and Management*, 24:513--523.

Salton, G. and McGill, M. (1988). *Introduction to Modern Information Retrieval*, McGraw Hill.

Salton, G., Yang C.S., and Yu C.T. (1975). A theory of term importance in automatic text analysis, *Journal of the American Society for Information Science*.

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