Detection of Semantic Relationships between Terms with a New Statistical Method

Nesrine Ksentini, Mohamed Tmar and Faïez Gargouri
MIRACL: Multimedia, InfoRmation Systems and Advanced Computing Laboratory
University of Sfax, Higher Institute of Computer Science and Multimedia of Sfax, Sfax, Tunisia
ksentini.nesrine@ieee.org, {mohamed.tmar, faiez.gargouri}@isimsf.rnu.tn

Keywords: Semantic Relatedness, Least Square Method, Information Retrieval, Query Expansion.

Abstract: Semantic relatedness between terms plays an important role in many applications, such as information retrieval, in order to disambiguate document content. This latter is generally studied among pairs of terms and is usually presented in a non-linear way. This paper presents a new statistical method for detecting relationships between terms called Least Square Mehod which defines these relations linear and between a set of terms. The evaluation of the proposed method has led to optimal results with low error rate and meaningful relationships. Experimental results show that the use of these relationships in query expansion process improves the retrieval results.

1 INTRODUCTION

With the increasing volume of textual data on the internet, effective access to semantic information becomes an important problem in information retrieval and other related domains such as natural language processing, Text Entailment and Information Extraction.

Measuring similarity and relatedness between terms in the corpus becomes decisive in order to improve search results (Agirre et al., 2009). Earlier approaches that have been investigating the latter idea can be classiffied into two main categories: those based on pre-available knowledge (ontology such as wordnet, thesauri, etc) (Agirre et al., 2010) and those inducing statistical methods (Sahami and Heilman, 2006),(Ruiz-Casado et al., 2005).

WordNet is a lexical database developed by linguists in the Cognitive Science Laboratory at Princeton University (Hearst, 1998). Its purpose is to identify, classify and relate in various ways the semantic and lexical content of the English language. WordNet versions for other languages exist, but the English version, however, is the most comprehensive to date.
Information in wordnet ;such as nouns, adjectives, verbs and adverbs; is grouped into synonyms sets called synsets. Each group expresses a distinct concept and it is interlinked with lexical and conceptual-semantic relations such as meronymy, hypernymy, causality, etc.

We represent WordNet as a graph $G = (V,E)$ as follows: graph nodes represent WordNet concepts (synsets) and dictionary words; undirected edges represent relations among synsets; and directed edges represent relations between dictionary words and synsets associated to them. Given a pair of words and a graph of related concepts, wordnet computes in the first time the personalized PageRank over WordNet for each word, giving a probability distribution over WordNet synsets. Then, it compares how similar these two probability distributions are by presenting them as vectors and computing the cosine between the vectors (Agirre et al., 2009).

For the second category, many previous studies used search engine collect co-occurrence between terms. In (Turney, 2001), author calculate the Pointwise Mutual Information (PMI) indicator of synonymy between terms by using the number of returned results by a web search engine.

In (Sahami and Heilman, 2006), the authors proposed a new method for calculating semantic similarity. They collected snippets from the returned results by a search engine and presented each of them as a vector. The semantic similarity is calculated as the inner product between the centroids of the vectors.

Another method to calculate the similarity of two words was presented by (Ruiz-Casado et al., 2005) it collected snippets containing the first word from a Web search engine, extracted a context around it, replaced it with the second word and checked if context

is modified in the Web.

However, all these methods measure relatedness between terms in pairs and cannot express them in a linear way. In this paper, we propose a new method which defines linear relations between a set of terms in a corpus based on their weights.

The paper is organized as follows, section 2 is devoted to detailing the proposed method followed by the evaluation in section 3. Finally, section 4 draws the conclusions and outlines future works.

2 PROPOSED METHOD

Our method is based on the extraction of relationships between terms $(t_1, t_2, \cdots, t_n)$ in a corpus of documents. Indeed, we try to find a linear relationship that may possibly exist between them with the following form:

$$t_i = f(t_1, t_2, \cdots, t_{i-1}, t_{i+1}, \cdots, t_n) \tag{1}$$

Least square method (Abdi., 2007), (Miller, 2006) is a frequently used method for solving this kind of problems in an approximate way. It requires some calculus and linear algebra.
In fact, this method seeks to highlight the connection being able to exist between an explained variable (y) and explanatory variables (x). It is a procedure to find the best fit line $(y = ax + b)$ to the data given that the pairs (x_i, y_i) are observed for $i \in 1, \cdots, n$.

The goal of this method is to find values of a and b that minimize the associated error (Err).

$$Err = \sum_{i=1}^{n} (y_i - (ax_i + b))^2 \tag{2}$$

Using a matrix form for the n pairs (x_i, y_i):

$$A = (X^T \times X)^{-1} \times X^T \times Y \tag{3}$$

where A represents vector of values $(a_1, a_2, \cdots, a_n)$ and X represents the coordinate matrix of n pairs.

In our case, let term (t_i) the explained variable and the remaining terms of the corpus $(t_1, t_2, \cdots, t_{i-1}, t_{i+1}, \cdots, t_n)$ the explanatory variables. We are interesting in the linear models; the relation between these variables is done by the following:

$$t_i \approx \alpha_1 t_1 + \alpha_2 t_2 + \cdots + \alpha_{i-1} t_{i-1} + \alpha_{i+1} t_{i+1}$$

$$+ \cdots + \alpha_n t_n + \varepsilon = \sum_{j=1}^{i-1} (\alpha_j t_j) + \sum_{j=i+1}^{n} (\alpha_j t_j) + \varepsilon \tag{4}$$

Where α are real coefficients of the model and present the weights of relationships between terms and ε represents the associated error of the relation.

We are looking for a model which enables us to obtain an exact solution for this problem.
Therefore, we proceed to calculate this relation for each document in the collection and define after that the final relationship between these terms in the whole collection. For that, m measurements are made for the explained and the explanatory variables to calculate the appropriate $\alpha_1, \alpha_2, \cdots, \alpha_n$ with m represent the number of documents in the collection.

$$\begin{cases} t_i^1 \approx \alpha_1 . t_1^1 + \alpha_2 . t_2^1 + \cdots + \alpha_n . t_n^1 \\ t_i^2 \approx \alpha_1 . t_1^2 + \alpha_2 . t_2^2 + \cdots + \alpha_n . t_n^2 \\ \vdots \\ t_i^m \approx \alpha_1 . t_1^m + \alpha_2 . t_2^m + \cdots + \alpha_n . t_n^m \end{cases} \tag{5}$$

Where t_i^j is the Tf-Idf weight of term i in document j. By using the matrix notations the system becomes:

$$\underbrace{\begin{pmatrix} t_i^1 \\ t_i^2 \\ \vdots \\ t_i^m \end{pmatrix}}_{t_i} \approx \underbrace{\begin{pmatrix} t_1^1 & t_2^1 & \cdots & t_n^1 \\ t_1^2 & t_2^2 & \cdots & t_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ t_1^m & t_2^m & \cdots & t_n^m \end{pmatrix}}_{X} \times \underbrace{\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix}}_{A} \tag{6}$$

where X is a TF-IDF (Term Frequency-Inverse Document Frequency) matrix whose rows represent the documents and columns represent the indexing terms (lemmas).
Thus, we seek $A = (\alpha_1, \cdots, \alpha_n)$ such as $X \times A$ is more near possible to t_i. Rather than solving this system of equations exactly, least square method tries to reduce the sum of the squares of the residuals. Indeed, it tries to obtain a low associated error (Err) for each relation.

We notice that the concept of distance appears. We expect that $d(X \times A, t_i)$is minimal , which is written:

$$min \mid\mid X \times A - t_i \mid\mid \tag{7}$$

To determine the vector A for each term in a corpus, we applied the least square method on the matrix X for each one.

$\forall i = 1, \ldots, n.$

$$A_i = (X^{iT} \times X^i)^{-1} \times X^T[i,.] \times t_i \tag{8}$$

Where X^i is obtained by removing the row of the $term_i$ in matrix X and n is the number of terms in a corpus.
$X^T[i,.]$ represents the transpose of the line weight of $term_i$ in all documents.

3 EXPERIMENTS

In this paper, we use our method to improve informa-

tion retrieval performance, mainly, by detecting relationships between terms in a corpus of documents.

We focus on the application of the least square method on a corpus of textual data in order to achieve expressive semantic relationships between terms.

In order to check the validity and the performance of our method, an experimental procedure was set up.

The evaluation is then based on a comparison of the list of documents retrieved by a developed information retrieval system and the documents deemed relevant.

To evaluate within a framework of real examples, we have resorted to a textual database, of 1400 documents, called Cranfield collection (Ahram, 2008)(Sanderson, 2010). This collection of tests includes a set of documents, a set of queries and the list of relevant documents in the collection for each query.

For each document of the collection, we proceed a handling and an analysis in order to lead it to lemmas which will be the index terms. Once the documents are each presented with a bag of words, we have reached by a set of 4300 terms in the whole collection. Hence, matrix X is sized $1400 * 4300$. After that, we applied on it the least square method for each term in order to determine the vector A for each one. The obtained values A_i indicate the relationship between $term_i$ and the remaining terms in the corpus. We obtain another square matrix T with 4300 lines expressing the semantic relationships between terms as follows:

$$\forall i \in 1,2,\ldots,4300, \forall j \in 1,2,\ldots,4300$$

$$term_i = \sum(T[i,j].term_j) \tag{9}$$

Example of obtained semantic relationships:

Term airborn = 0.279083 action + 0.222742 airforc + 0.221645 alon + 0.259213 analogu + 0.278371 assum + 0.275861 attempt + 0.210211 behaviour + 0.317462 cantilev + 0.215479 carrier + 0.277437 centr + 0.216453 chapman + 0.22567 character + 0.23094 conecylind + 0.347057 connect + 0.239277 contact + 0.225988 contrari + 0.217225 depth + 0.283544 drawn + 0.204302 eighth + 0.26399 ellipsoid + 0.312026 fact + 0.252312 ferri + 0.211903 glauert + 0.230067 grasshof + 0.223152 histori + 0.28336 hovercraft + 0.380206 inch + 0.238555 inelast + 0.205513 intermedi + 0.275635 interpret + 0.235573 interv + 0.216454 ioniz + 0.319457 meksyn + 0.200089 motion + 0.223062 movement + 0.233753 multicellular + 0.376881 multipli + 0.436183 nautic + 0.219787 orific + 0.414204 probabl + 0.214005 propos + 0.305503 question+ 0.204316 read + 0.222911 reciproc + 0.256728 reson + 0.237344 review + 0.202781 spanwis + 0.351152 telemet + 0.226465 termin + 0.212812 toroid+ 0.339988 tunnel + 0.25228 uniform + 0.233854 upper + 0.20262 vapor.

We notice that obtained relationships between terms are meaningful. Indeed, related terms in a relation talk about the same context, for example the relationship between the lemma airbon and the other lemmas (airborn, airforc, conecylind, action, tunnel ...) talks about the airborne aircraft carrier subject. To test these relationships, we calculate for each one the error rate (Err):

$$Err(term_i) = \sum_{j=1}^{1400}(X[j,i] - (\sum_{k=1}^{i-1}(X[j,k] \times T[i,k]) + \sum_{q=i+1}^{4300}(X[j,q] \times T[i,q])))^2 \tag{10}$$

The obtained values are all closed to zero, for example the error rate of the relationship between term (account) and the remaining of terms is $1.5 * 10^{-7}$ and for the term (capillari) is $5,23 * 10^{-11}$.

To check if obtained relations improve information retrieval results, we have implemented a vector space information retrieval system which test queries proposed by the Cranfield Collection.

The aim of this kind of system is to retrieve documents that are relevant to the user queries. To achieve this aim, the system attributes a value to each candidate document; then, it rank documents in the reverse order of this value. This value is called the Retrieval Status Value (RSV) (Imafouo and Tannier, 2005) and calculated with four measures (cosines, dice, jaccard and overlap).

Our system presents two kinds of evaluation; firstly, it calculates the similarity (RSV) of a document vector to a query vector. Then, it calculates the similarity of a document vector to an expanded query vector. The expansion is based on the relevant documents retrieved by the first model (Wasilewski, 2011) and the relationships obtained by least square method.

Indeed, if a term of a collection is very related with a term of query ($\alpha >= 0.5$) and appears in a the relevant returned documents, we add it to a query.
Mean Average Precision (MAP) is used to calculate precision of each evalution. Table1 shows the obtained results.

We notice from this evaluation, that relationships obtained by least square method are meaningful and can provide improvements in the information retrieval process. Indeed, the MAP values are increasing when these relations are used in information retrieval system. For example, our method improves information retrieval results using cosinus measure when $\alpha > 0.6$ with $MAP = 0.21826$ compared to the basic VSM model ($MAP = 0.20858$).

Compare our results with other works, we note

Table 1: Variation of MAP values.

	VSM	VSM with expanded query			
		$\alpha > 0.8$	$\alpha > 0.7$	$\alpha > 0.6$	$\alpha > 0.5$
Cosinus	0.20858	0.20654	0.21273	**0.21826**	0.21822
Dice	0.20943	0.20969	0.21529	0.21728	**0.22060**
Jaccard	0.20943	0.21043	**0.21455**	0.21341	0.20642
Overlap	0.12404	0.12073	0.12366	0.12311	0.12237

that this new statistical method (least square) improves search results. In (Ahram, 2008), experimental results from cranfield documents collection gave an average precision of 0.1384 which is less than that found in our work (0.21826 with cosinus measure, 0.22060 with dice measure).

4 SUMMARY AND FUTURE WORKS

We present in this paper a new method for detecting semantic relationships between terms. The proposed method (least square) defines meaningful relationships in a linear way and between a set of terms using weights of each one which represent the distribution of terms in the corpus.
These relationships give a low error rate. Indeed, they are used in the query expansion process for improving information retrieval results.
As future works, firstly, we will intend to participate in the competition TREC to evaluate our method on a large test collection. Secondly, we will look for how to use these relations in the process of weighting terms and the definition of terms-documents matrix to improve information retrieval results. Finally, we also will investigate these relations to induce the notion of context in the indexing process.

REFERENCES

Abdi., H. (2007). The method of least squares.

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 19–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

Agirre, E., Cuadros, M., Rigau, G., and Soroa, A. (2010). Exploring knowledge bases for similarity. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Ahram, T. Z. (2008). *Information retrieval performance enhancement using the average standard estimator and the multi-criteria decision weighted set of performance measures*. PhD thesis, University of Central Florida Orlando, Florida.

Hearst, M. (1998). WordNet: An electronic lexical database and some of its applications. In Fellbaum, C., editor, *Automated Discovery of WordNet Relations*. MIT Press.

Imafouo, A. and Tannier, X. (2005). Retrieval status values in information retrieval evaluation. In *String Processing and Information Retrieval*, pages 224–227. Springer.

Miller, S. J. (2006). The method of least squares.

Ruiz-Casado, M., Alfonseca, E., and Castells, P. (2005). Using context-window overlapping in synonym discovery and ontology extension. In *International Conference on Recent Advances in Natural Language Processing (RANLP 2005)*, Borovets, Bulgaria.

Sahami, M. and Heilman, T. D. (2006). A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 377–386, New York, NY, USA. ACM.

Sanderson, M. (2010). *Test collection based evaluation of information retrieval systems*. Now Publishers Inc.

Turney, P. D. (2001). Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proceedings of the 12th European Conference on Machine Learning*, EMCL '01, pages 491–502, London, UK, UK. Springer-Verlag.

Wasilewski, P. (2011). Query expansion by semantic modeling of information needs.