Semantic Vectors: an Information Retrieval scenario

Pierpaolo Basile
Dept. of Computer Science
University of Bari
Via E. Orabona, 4
70125 Bari (ITALY)
basilepp@di.uniba.it

Annalina Caputo
Dept. of Computer Science
University of Bari
Via E. Orabona, 4
70125 Bari (ITALY)
acaputo@di.uniba.it

Giovanni Semeraro
Dept. of Computer Science
University of Bari
Via E. Orabona, 4
70125 Bari (ITALY)
semeraro@di.uniba.it

ABSTRACT

In this paper we exploit Semantic Vectors to develop an IR system. The idea is to use semantic spaces built on terms and documents to overcome the problem of word ambiguity. Word ambiguity is a key issue for those systems which have access to textual information. Semantic Vectors are able to dividing the usages of a word into different meanings, discriminating among word meanings based on information found in unannotated corpora. We provide an *in vivo* evaluation in an Information Retrieval scenario and we compare the proposed method with another one which exploits Word Sense Disambiguation (WSD). Contrary to sense discrimination, which is the task of discriminating among different meanings (not necessarily known a priori), WSD is the task of selecting a sense for a word from a set of predefined possibilities. The goal of the evaluation is to establish how Semantic Vectors affect the retrieval performance.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Indexing methods, Linguistic processing; H.3.3 [**Information Search and Retrieval**]: Retrieval models, Search process

Keywords

Semantic Vectors, Information Retrieval, Word Sense Discrimination

1. BACKGROUND AND MOTIVATIONS

Ranked keyword search has been quite successful in the past, in spite of its obvious limits basically due to polysemy, the presence of multiple meanings for one word, and synonymy, multiple words having the same meaning. The result is that, due to synonymy, relevant documents can be missed if they do not contain the exact query keywords, while, due to polysemy, wrong documents could be deemed as relevant. These problems call for alternative methods that work not only at the lexical level of the documents, but also at the meaning level.

In the field of computational linguistics, a number of important research problems still remain unresolved. A specific challenge for computational linguistics is ambiguity. Ambiguity means that a word can be interpreted in more than one way, since it has more than one meaning. Ambiguity usually is not a problem for humans therefore it is not perceived as such. Conversely, for a computer ambiguity is one of the main problems encountered in the analysis and generation of natural languages. Two main strategies have been proposed to cope with ambiguity:

1. **Word Sense Disambiguation:** the task of selecting a sense for a word from a set of predefined possibilities; usually the so called *sense inventory*[1] comes from a dictionary or thesaurus.

2. **Word Sense Discrimination:** the task of dividing the usages of a word into different meanings, ignoring any particular existing *sense inventory*. The goal is to discriminate among word meanings based on information found in unannotated corpora.

The main difference between the two strategies is that disambiguation relies on a sense inventory, while discrimination exploits unannotated corpora.

In the past years, several attempts were proposed to include sense disambiguation and discrimination techniques in IR systems. This is possible because discrimination and disambiguation are not an end in themselves, but rather "intermediate tasks" which contribute to more complex tasks such as information retrieval. This opens the possibility of an *in vivo* evaluation, where, rather then being evaluated in isolation, results are evaluated in terms of their contribution to the overall performance of a system designed for a particular application (e.g. Information Retrieval).

The goal of this paper is to present an IR system which exploits semantic spaces built on words and documents to overcome the problem of word ambiguity. Then we compare this system with another one which uses a Word Sense Disambiguation strategy. We evaluated the proposed system into the context of CLEF 2009 Ad-Hoc Robust WSD task [2].

The paper is organized as follows: Sections 2 presents the IR model involved into the evaluation, which embodies semantic vectors strategies. The evaluation and the results are reported in Section 3, while a brief discussion about the main works related to our research are in Section 4. Conclusions and future work close the paper.

[1]A sense inventory provides for each word a list of all possible meanings.

Appears in the Proceedings of the 1st Italian Information Retrieval Workshop (IIR'10), January 27–28, 2010, Padova, Italy.
http://ims.dei.unipd.it/websites/iir10/index.html

2. AN IR SYSTEM BASED ON SEMANTIC VECTORS

Semantic Vectors are based on WordSpace model [15]. This model is based on a vector space in which points are used to represent semantic concepts, such as words and documents. Using this strategy it is possible to build a vector space on both words and documents. These vector spaces can be exploited to develop an IR model as described in the following.

The main idea behind Semantic Vectors is that words are represented by points in a mathematical space, and words or documents with similar or related meanings are represented close in that space. This provide us an approach to perform sense discrimination. We adopt the Semantic Vectors package [18] which relies on a technique called Random Indexing (RI) introduced by Kanerva in [13]. This allows to build semantic vectors with no need for the factorization of document-term or term-term matrix , because vectors are inferred using an incremental strategy. This method allows to solve efficiently the problem of reducing dimensions, which is one of the key features used to uncover the "latent semantic dimensions" of a word distribution.

RI is based on the concept of Random Projection: the idea is that high dimensional vectors chosen randomly are "nearly orthogonal". This yields a result that is comparable to orthogonalization methods, such as Singular Value Decomposition, but saving computational resources. Specifically, RI creates semantic vectors in three steps:

1. a context vector is assigned to each document. This vector is sparse, high-dimensional and ternary, which means that its elements can take values in {-1, 0, 1}. The index vector contains a small number of randomly distributed non-zero elements, and the structure of this vector follows the hypothesis behind the concept of Random Projection;

2. context vectors are accumulated by analyzing terms and documents in which terms occur. In particular the semantic vector of each term is the sum of the context vectors of the documents which contain the term;

3. in the same way a semantic vector for a document is the sum of the semantic vectors of the terms (created in step 2) which occur in the document.

The two spaces built on terms and documents have the same dimension. We can use vectors built on word-space as query vectors and vectors built on document-space as search vectors. Then, we can compute the similarity between word-space vectors and document-space vectors by means of the classical cosine similarity measure. In this way we implement an information retrieval model based on semantic vectors.

Figure 1 shows a word-space with two only dimensions. If those two dimensions refer respectively to `LEGAL` and `SPORT` contexts, we can note that the vector of the word *soccer* is closer to the `SPORT` context than the `LEGAL` context, vice versa the word *law* is closer to the `LEGAL` context. The angle between *soccer* and *law* represents the similarity degree between the two words. It is important to emphasize that contexts in WordSpace have no tag, thus we know that each dimension is a context, but we cannot know the kind of the context. If we consider document-space rather than word-space, document semantically related will be represented closer in that space.

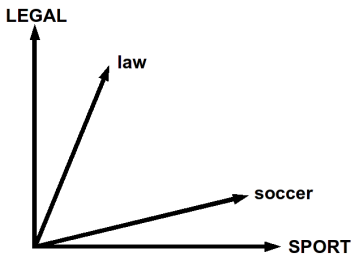

Figure 1: Word vectors in word-space

The Semantic Vectors package supplies tools for indexing a collection of documents and their retrieval adopting the Random Indexing strategy. This package relies on Apache Lucene[2] to create a basic term-document matrix, then it uses the Lucene API to create both a word-space and a document-space from the term-document matrix, using Random Projection to perform dimensionality reduction without matrix factorization. In order to evaluate Semantic Vectors model we must modify the standard Semantic Vectors package by adding some ad-hoc features to support our evaluation. In particular, documents are split in two fields, *headline* and *title*, and are not tokenized using the standard text analyzer in Lucene.

An important factor to take into account in semantic-space model is the number of contexts, that sets the dimensions of the context vector. We evaluated Semantic Vectors using several values of reduced dimensions. Results of the evaluation are reported in Section 3.

3. EVALUATION

The goal of the evaluation was to establish how Semantic Vectors influence the retrieval performance. The system is evaluated into the context of an Information Retrieval (IR) task. We adopted the dataset used for CLEF 2009 Ad-Hoc Robust WSD task [2]. Task organizers make available document collections (from the news domain) and topics which have been automatically tagged with word senses (synsets) from WordNet using several state-of-the-art disambiguation systems. Considering our goal, we exploit only the monolingual part of the task.

In particular, the Ad-Hoc WSD Robust task used existing CLEF news collections, but with WSD added. The dataset comprises corpora from "Los Angeles Times" and "Glasgow Herald", amounting to 169,477 documents, 160 test topics and 150 training topics. The WSD data were automatically added by systems from two leading research laboratories, UBC [1] and NUS [9]. Both systems returned word senses from the English WordNet, version 1.6. We used only the senses provided by NUS. Each term in the document is annotated by its senses with their respective scores, as assigned by the automatic WSD system. This kind of dataset supplies WordNet synsets that are useful for the development of search engines that rely on disambiguation.

In order to compare the IR system based on Semantic Vectors to other systems which cope with word ambiguity

[2]http://lucene.apache.org/

by means of methods based on Word Sense Disambiguation, we provide a baseline based on SENSE. SENSE: SEmantic N-levels Search Engine is an IR system which relies on Word Sense Disambiguation. SENSE is based on the N-Levels model [5]. This model tries to overcome the limitations of the ranked keyword approach by introducing *semantic levels*, which integrate (and not simply replace) the lexical level represented by keywords. Semantic levels provide information about word meanings, as described in a reference dictionary or other semantic resources. SENSE is able to manage documents indexed at separate levels (keywords, word meanings, and so on) as well as to combine keyword search with semantic information provided by the other indexing levels. In particular, for each level:

1. a *local scoring function* is used in order to weigh elements belonging to that level according to their informative power;
2. a *local similarity function* is used in order to compute document relevance by exploiting the above-mentioned scores.

Finally, a *global ranking function* is defined in order to combine document relevance computed at each level. The SENSE search engine is described in [4], while the setup of SENSE into the context of CLEF 2009 is thoroughly described in [7]

In CLEF, queries are represented by topics, which are structured statements representing information needs. Each topic typically consists of three parts: a brief TITLE statement, a one-sentence DESCRIPTION, and a more complex "narrative" specifying the criteria for assessing relevance. All topics are available with and without WSD. Topics in English are disambiguated by both UBC and NUS systems, yielding word senses from WordNet version 1.6.

We adopted as baseline the system which exploits only keywords during the indexing, identified by *KEYWORD*. Regarding disambiguation we used the *SENSE* system adopting two strategies: the former, called *MEANING*, exploits only word meanings, the latter, called *SENSE*, uses two levels of document representation: keywords and word meanings combined.

The query for the *KEYWORD* system is built using word stems in TITLE and DESCRIPTION fields of the topics. All query terms are joined adopting the OR boolean clause. Regarding the *MEANING* system each word in TITLE and DESCRIPTION fields is expanded using the synsets in WordNet provided by the WSD algorithm. More details regarding the evaluation of SENSE in CLEF 2009 are in [7].

The query for the *SENSE* system is built combining the strategies adopted for the *KEYWORD* and the *MEANING* systems. For all the runs we remove the stop words from both the index and the topics. In particular, we build a different stop words list for topics in order to remove non informative words such as *find*, *reports*, *describe*, that occur with high frequency in topics and are poorly discriminating.

In order to make results comparable we use the same index built for the *KEYWORD* system to infer semantic vectors using the Semantic Vectors package, as described in Section 2. We need to tune two parameters in Semantic Vectors: the number of dimensions (the number of contexts) and the frequency[3] threshold (T_f). The last value is used to discard terms that have a frequency below T_f. After a tuning step, we set the dimension to 2000 and T_f to 10. Tuning is performed using training topics provided by the CLEF organizers.

[3]In this instance word frequency refers to word occurrences.

Topic fields	**MAP**
TITLE	0.0892
TITLE+DESCRIPTION	**0.2141**
TITLE+DESCRIPTION+NARRATIVE	0.2041

Table 1: Semantic Vectors: Results of the performed experiments

System	**MAP**	**Imp.**
KEYWORD	0.3962	-
MEANING	0.2930	-26.04%
SENSE	0.4222	+6.56%
SV_{best}	0.2141	-45.96%

Table 2: Results of the performed experiments

Queries for the Semantic Vectors model are built using several combinations of topic fields. Table 1 reports the results of the experiments using Semantic Vectors and different combinations of topic fields.

To compare the systems we use a single measure of performance: the Mean Average Precision (MAP), due to its good stability and discrimination capabilities. Given the Average Precision [8], that is the mean of the precision scores obtained after retrieving each relevant document, the MAP is computed as the sample mean of the Average Precision scores over all topics. Zero precision is assigned to unretrieved relevant documents.

Table 2 reports the results of each system involved into the experiment. The column *Imp.* shows the improvement with respect to the baseline *KEYWORD*. The system SV_{best} refers to the best result obtained by Semantic Vectors reported in boldface in Table 1.

The main result of the evaluation is that *MEANING* works better than SV_{best}; in other words disambiguation wins over discrimination. Another important observation is that the combination of keywords and word meanings, the *SENSE* system, obtains the best result. It is important to note that SV_{best} obtains a performance below the *KEYWORD* system, about the 46% under the baseline. It is important to underline that the keyword level implemented in SENSE uses a modified version of Apache Lucene which implements Okapi BM25 model [14].

In the previous experiments we compared the performance of the Semantic Vectors-based IR system to SENSE. In the following, we describe a new kind of experiment in which we integrate the Semantic Vector as a new level in SENSE. The idea is to combine the results produced by Semantic Vectors with the results which come out from both the keyword level and the word meaning level. Table 3 shows that the combination of the keyword level with Semantic Vectors outperforms the keyword level alone.

Moreover, the combination of Semantic Vectors with word meaning level achieves an interesting result: the combination is able to outperform the word meaning level alone. Finally, the combination of Semantic Vectors with *SENSE* (keyword level+word meaning level) obtains the best MAP with an increase of about the 6% with respect to *KEY-*

System	MAP	Imp.
SV+KEYWORD	0.4150	+4.74%
SV+MEANING	0.3238	-18.27%
SV+SENSE	0.4216	+6.41%

Table 3: Results of the experiments: combination of Semantic Vectors with other levels

WORD. However, *SV* does not contribute to improve the effectiveness of *SENSE*, in fact *SENSE* without *SV* (see Table 2) outperforms *SV+SENSE*.

Analyzing results query by query, we discovered that for some queries the Semantic Vectors-based IR system achieves an high improvement wrt keyword search. This happen mainly when few relevant documents exist for a query. For example, query "10.2452/155-AH" has only three relevant documents. Both keyword and Semantic Vectors are able to retrieve all relevant documents for that query, but keyword achieves 0,1484 MAP, while for Semantic Vectors MAP grows to 0,7051. This means that Semantic Vectors are more accurate than keyword when few relevant documents exist for a query.

4. RELATED WORKS

The main motivation for focusing our attention on the evaluation of disambiguation or discrimination systems is the idea that ambiguity resolution can improve the performance of IR systems.

Many strategies have been used to incorporate semantic information coming from electronic dictionaries into search paradigms.

Query expansion with WordNet has shown to potentially improve recall, as it allows matching relevant documents even if they do not contain the exact keywords in the query [17]. On the other hand, semantic similarity measures have the potential to redefine the similarity between a document and a user query [10]. The semantic similarity between concepts is useful to understand how similar are the meanings of the concepts. However, computing the degree of relevance of a document with respect to a query means computing the similarity among all the synsets of the document and all the synsets of the user query, thus the matching process could have very high computational costs.

In [12] the authors performed a shift of representation from a lexical space, where each dimension is represented by a term, towards a semantic space, where each dimension is represented by a concept expressed using WordNet synsets. Then, they applied the Vector Space Model to WordNet synsets. The realization of the semantic tf-idf model was rather simple, because it was sufficient to index the documents or the user-query by using strings representing synsets. The retrieval phase is similar to the classic tf-idf model, with the only difference that matching is carried out between synsets.

Concerning the discrimination methods, in [11] some experiments in IR context adopting LSI technique are reported. In particular this method performs better than canonical vector space when queries and relevant documents do not share many words. In this case LSI takes advantage of the implicit higher-order structure in the association of terms with documents ("semantic structure") in order to improve the detection of relevant documents on the basis of terms found in queries.

In order to show that WordSpace model is an approach to ambiguity resolution that is beneficial in information retrieval, we summarize the experiment presented in [16]. This experiment evaluates sense-based retrieval, a modification of the standard vector-space model in information retrieval. In word-based retrieval, documents and queries are represented as vectors in a multidimensional space in which each dimension corresponds to a word. In sense-based retrieval, documents and queries are also represented in a multidimensional space, but its dimensions are senses, not words. The evaluation shows that sense-based retrieval improved average precision by 7.4% when compared to word-based retrieval.

Regarding the evaluation of word sense disambiguation systems in the context of IR it is important to cite SemEval-2007 task 1 [3]. This task is an application-driven one, where the application is a given cross-lingual information retrieval system. Participants disambiguate text by assigning WordNet synsets, then the system has to do the expansion to other languages, the indexing of the expanded documents and the retrieval for all the languages in batch. The retrieval results are taken as a measure for the effectiveness of the disambiguation. CLEF 2009 Ad-hoc Robust WSD [2] is inspired to SemEval-2007 task 1.

Finally, this work is strongly related to [6], in which a first attempt to integrate Semantic Vectors in an IR system was performed.

5. CONCLUSIONS AND FUTURE WORK

We have evaluated Semantic Vectors exploiting an information retrieval scenario. The IR system which we propose relies on semantic vectors to induce a WordSpace model exploited during the retrieval process. Moreover we compare the proposed IR system with another one which exploits word sense disambiguation. The main outcome of this comparison is that disambiguation works better than discrimination. This is a counterintuitive result: indeed it should be obvious that discrimination is better than disambiguation. Since, the former is able to infer the usages of a word directly from documents, while disambiguation works on a fixed distinction of word meanings encoded into the sense inventory such as WordNet.

It is important to note that the dataset used for the evaluation depends on the method adopted to compute document relevance, in this case the pooling techniques. This means that the results submitted by the groups participating in the previous ad hoc tasks are used to form a pool of documents for each topic by collecting the highly ranked documents. What we want to underline here is that generally the systems taken into account rely on keywords. This can produce relevance judgements that do not take into account evidence provided by other features, such as word meanings or context vectors. Moreover, distributional semantics methods, such as Semantic Vectors, do not provide a formal description of why two terms or documents are similar. The semantic associations derived by Semantic Vectors are similar to how human estimates similarity between terms or documents. It is not clear if current evaluation methods are able to detect these cognitive aspects typical of human thinking. More investigation on the strategy adopted for the evaluation is needed. As future work we intend to exploit several discrimination methods, such as Latent Semantic Indexing and Hyperspace Analogue to Language.

6. REFERENCES

[1] E. Agirre and O. L. de Lacalle. BC-ALM: Combining k-NN with SVD for WSD. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007), Prague, Czech Republic*, pages 341–325, 2007.

[2] E. Agirre, G. M. Di Nunzio, T. Mandl, and A. Otegi. CLEF 2009 Ad Hoc Track Overview: Robust - WSD Task. In *Working notes for the CLEF 2009 Workshop*, 2009. http://clef-campaign.org/2009/working_notes/agirre-robustWSDtask-paperCLEF2009.pdf.

[3] E. Agirre, B. Magnini, O. L. de Lacalle, A. Otegi, G. Rigau, and P. Vossen. SemEval-2007 Task 1: Evaluating WSD on Cross-Language Information Retrieval. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007), Prague, Czech Republic*, pages 7–12. ACL, 2007.

[4] P. Basile, A. Caputo, M. de Gemmis, A. L. Gentile, P. Lops, and G. Semeraro. Improving Ranked Keyword Search with SENSE: SEmantic N-levels Search Engine. *Communications of SIWN (formerly: System and Information Sciences Notes), special issue on DART 2008*, 5:39–45, August 2008. SIWN: The Systemics and Informatics World Network.

[5] P. Basile, A. Caputo, A. L. Gentile, M. Degemmis, P. Lops, and G. Semeraro. Enhancing Semantic Search using N-Levels Document Representation. In S. Bloehdorn, M. Grobelnik, P. Mika, and D. T. Tran, editors, *Proceedings of the Workshop on Semantic Search (SemSearch 2008) at the 5th European Semantic Web Conference (ESWC 2008), Tenerife, Spain, June 2nd, 2008*, volume 334 of *CEUR Workshop Proceedings*, pages 29–43. CEUR-WS.org, 2008.

[6] P. Basile, A. Caputo, and G. Semeraro. Exploiting Disambiguation and Discrimination in Information Retrieval Systems. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and International Conference on Intelligent Agent Technology - Workshops, Milan, Italy, 15-18 September 2009*, pages 539–542. IEEE, 2009.

[7] P. Basile, A. Caputo, and G. Semeraro. UNIBA-SENSE @ CLEF 2009: Robust WSD task. In *Working notes for the CLEF 2009 Workshop*, 2009. http://clef-campaign.org/2009/working_notes/basile-paperCLEF2009.pdf.

[8] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 33–40, New York, NY, USA, 2000. ACM.

[9] Y. S. Chan, H. T. Ng, and Z. Zhong. NUS-PT: Exploiting Parallel Texts for Word Sense Disambiguation in the English All-Words Tasks. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007), Prague, Czech Republic*, pages 253–256, 2007.

[10] C. Corley and R. Mihalcea. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13–18, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[11] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[12] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran. Indexing with WordNet synsets can improve text retrieval. In *Proceedings of the COLING/ACL*, pages 38–44, 1998.

[13] P. Kanerva. *Sparse Distributed Memory*. MIT Press, 1988.

[14] S. Robertson, H. Zaragoza, and M. Taylor. Simple bm25 extension to multiple weighted fields. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 42–49, New York, NY, USA, 2004. ACM.

[15] M. Sahlgren. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, Stockholm: Stockholm University, Faculty of Humanities, Department of Linguistics, 2006.

[16] H. Schütze and J. O. Pedersen. Information retrieval based on word senses. In *In Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1995.

[17] E. M. Voorhees. *WordNet: An Electronic Lexical Database*, chapter Using WordNet for text retrieval, pages 285–304. Cambridge (Mass.): The MIT Press, 1998.

[18] D. Widdows and K. Ferraro. Semantic Vectors: A Scalable Open Source Package and Online Technology Management Application. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2008.