Integrating Named Entities in a Semantic Search Engine[*]

Annalina Caputo
University of Bari
Dept. of Computer Science
via E. Orabona, 4
Bari, Italy
acaputo@di.uniba.it

Pierpaolo Basile
University of Bari
Dept. of Computer Science
via E. Orabona, 4
Bari, Italy
basilepp@di.uniba.it

Giovanni Semeraro
University of Bari
Dept. of Computer Science
via E. Orabona, 4
Bari, Italy
semeraro@di.uniba.it

ABSTRACT

Traditional Information Retrieval (IR) systems are based on bag-of-words representation. This approach retrieves relevant documents by lexical matching between query and document terms. Due to synonymy and polysemy, lexical methods produce imprecise or incomplete results. In this paper we present how named entities are integrated in SENSE (SEmantic N-levels Search Engine). SENSE is an IR system that 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, and named entities. Our aim is to prove that named entities are useful to improve retrieval performance.

1. BACKGROUND AND MOTIVATION

In recent years a lot of attention has been invested on Named Entities (NE), and their informative and discriminative power within documents. Due to the importance of research on NE, several sub-areas arose, such as entity detection and extraction, entity disambiguation and entity ranking. The typical information extraction task involving NE is Named Entity Recognition (NER). This task has been defined for the first time during the Message Understanding Conference (MUC) [4], and requires the identification and categorization of NE as entity names (for people and organization), place names, temporal expressions and numerical expressions. Named Entities play also a key role in the Information Retrieval context. Indeed, a very common task in that research area is the entity ranking, whose aim is to retrieve entities (rather than documents) that satisfy the user query. Most documents we deal on everyday contain a lot of references to persons, dates, monetary values and places. Moreover, named entity terms are among the most frequently searched terms on the Web. Statistics on Yahoo's top 10 search terms in 2008[1] showed that all the ten search terms consist of named entity terms: six persons, one sport organization, one role-playing game, one fictional character and one TV show.

In this paper we propose a new way of exploiting named entities in Information Retrieval. Named entities mentioned in a document constitute an important part of its semantics. However, when named entities are considered alone they may fail to capture the semantics expressed in a document or in a user query. For that reason we adopt an IR model, called *N-levels* [2], able to capture semantic information in a text by exploiting *word meanings*, described in a reference dictionary (e.g. WordNet), and named entities. Thus, we propose an IR system, called *SENSE* (SEmantic N-levels Search Engine), which manages documents indexed at multiple separate levels: keywords, senses (word meanings) and entities (named entities). The system is able to combine keyword search with semantic information provided by the two other indexing levels. Finally, we present the development of the full-fledged entity level based on a novel model called *Semantic Vectors*.

2. NAMED ENTITY LEVEL

Named entities are phrases that contain the names of persons, organizations, locations and, more generally, entities that can be identified by proper names. In order to identify named entities in a text, several methods can be applied such as Rule-based, Dictionary-based or Statistical ones. We adopted a statistical method exploiting YamCha[2], a generic open source text chunker useful for a lot of NLP tasks. YamCha adopts a state-of-the-art machine learning algorithm called Support Vector Machines (SVMs), introduced by Vapnik in 1995. We trained YamCha using the dataset provided by CoNLL-2003 organization during the Shared-Task 2003 [5]. The dataset contains entities extracted from Reuters dataset. In particular three types of entities are extracted: PERSON, LOCATION, ORGANIZATION and MISC, which contains entities that do not belong to the previous three categories. We extract entities from the CLEF 2008 collection [1]. The results of the entity recognition task are exported into a Lucene index. In detail, each document is split in two fields: HEADLINE and TEXT, in compliance with the document structure in CLEF. Each field contains the set of the recognized entities and, for each entity, the number of occurrences.

Building the entity level requires three steps:

1. **pre-processing and entity extraction:** XML files

[*] The full version appears in [3]

[1] http://buzz.yahoo.com/yearinreview2008/top10/

[2] http://chasen.org/ taku/software/YamCha/

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

provided by CLEF 2008 organizers are processed in order to extract entities. Named entities are stored in IOB2 format. In IOB2, words outside the Named Entity are tagged with O, while the first word in the entity is tagged with B-k (to begin class k), and further words receive the I-k tag, indicating that these words are inside the entity;

2. **entity indexing:** entities extracted in the previous step are stored into an index using Lucene. The entity extraction procedure allows to obtain an entity-based vector space representation, called bag-of-entities (BoE). In this model an entity vector, rather than a word vector, corresponds to a document.

3. **Semantic Vector building:** in this step semantic vectors are built by exploiting the Lucene index. The main idea behind models based on Semantic Vectors [6] is that words and concepts are represented by points in a mathematical space, and this representation is learned from text in such a way that concepts with similar or related meanings are near to one another in that space. The SemanticVectors package offers tools for indexing a collection of documents and their retrieval. It relies on Apache Lucene to create a basic term-document matrix. Then the Lucene API is exploited to create a Wordspace model from the term-document matrix, by using Random Projection to perform *on-the-fly* dimensionality reduction. This is a relevant point because it allows us to use the same entity index produced in step 2 to induce semantic vectors. A detailed discussion on Semantic Vectors can be found in [6], whilst a thorough explanation about the entity index can be found in [3].

3. EXPERIMENTAL SESSION

For the evaluation of the system effectiveness, we used the CLEF Ad Hoc WSD-Robust dataset derived from the English CLEF data, which comprises corpora from "Los Angeles Times" and "Glasgow Herald", amounting to 166, 726 documents and 160 topics in English and Spanish. The relevance judgments were taken from CLEF. The goal of the evaluation was to prove that the combination of three indexing levels outperforms a single level. In particular, that adding the entity level increases the effectiveness of the search with respect to the keyword and meaning levels. To evaluate system effectiveness, different runs were performed by exploiting a single level at a time, or a combination of two or more levels. Each experiment is identified by the names of the used levels. To measure retrieval performance, we adopted Mean-Average-Precision (MAP) and Geometric-Mean-Average-Precision (GMAP) calculated by *trec_eval 0.8.1*, a simple program supplied by the Text REtrieval Conference organizers[3], on the basis of 1,000 retrieved items per request. Table 1 shows the results for each run, with an overview on the exploited features.

Table 1: Results of the performed experiments

Run	***MAP***	***GMAP***
Keyword (K)	0.192	0.041
Meaning (M)	0.188	0.035
K+M	0.220	0.057
Entity (E)	0.134	0.006
K+E	0.220	0.048
M+E	0.228	0.054
K+M+E	**0.252**	**0.076**

The results confirm our hypothesis: named entity recognition, in conjunction with an IR model capable of expressing semantics, can greatly improve the retrieval performance. If evaluated individually, the entity level does not yield to satisfactory results. This result is due to the presence of topics in which no entity was recognized. Conversely, when search is performed by making use of multiple levels, the entity level is able to improve performance even on those (difficult) topics for which few relevant documents are returned. This result suggests that named entities play a key role in increasing the number of retrieved relevant results previously ignored. Specifically, considering the experiment *K+M+E* where we used all three levels, an improvement of 14.5% in the MAP and 33.3% in the GMAP was observed. Generally speaking, we noted an overall improvement in all the experiments that used the entity level, compared to the equivalent experiments in which that level was not exploited.

[3] http://trec.nist.gov/trec_eval/

4. REFERENCES

[1] E. Agirre, G. M. Di Nunzio, N. Ferro, T. Mandl, and C. Peters. CLEF 2008: Ad Hoc Track Overview. In *Working notes for the CLEF 2008 Workshop*, 2008.

[2] 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.

[3] A. Caputo, P. Basile, and G. Semeraro. Boosting a semantic search engine by named entities. In J. Rauch, Z. W. Ras, P. Berka, and T. Elomaa, editors, *ISMIS - Foundations of Intelligent Systems, 18th International Symposium, ISMIS 2009, Prague, Czech Republic, September 14-17, 2009. Proceedings*, volume 5722 of *Lecture Notes in Computer Science*, pages 241–250. Springer, 2009.

[4] R. Grishman and B. Sundheim. Message understanding conference-6: A brief history. In *COLING*, pages 466–471, 1996.

[5] E. F. Tjong Kim Sang and F. De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In W. Daelemans and M. Osborne, editors, *Proceedings of CoNLL-2003*, pages 142–147. Edmonton, Canada, 2003.

[6] 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.