

Semantic Models for Question Answering

Piero Molino
Supervisor: Pasquale Lops

Dept. of Computer Science, University of Bari
Via Orabona – I-70125 Bari, Italy
`piero.molino@uniba.it`

Abstract. The research presented in this paper focuses on the adoption of semantic models for *Question Answering* (QA) systems. We propose a framework which exploits semantic technologies to analyze the question, retrieve and rank relevant passages. It exploits: (a) *Natural Language Processing* algorithms for the analysis of questions and candidate answers both in English and Italian; (b) *Information Retrieval* (IR) probabilistic models for retrieving candidate answers and (c) *Machine Learning* methods for question classification. The data source for the answers is an unstructured text document collection stored in search indices. The aim of the research is to improve the system performances by introducing semantic models in every step of the answering process.

1 Introduction

Question Answering (QA) is the task of answering users' questions with answers obtained from a collection of documents or from the Web.

Traditional search engines usually retrieve long lists of full-text documents that must be checked by the user in order to find the needed information. Instead QA systems exploit *Information Retrieval* (IR) and *Natural Language Processing* (NLP) [18, 9], to find the answer, or short passages of text containing it, to a natural language question. Open-domain QA systems search on the Web and exploit redundancy, textual pattern extraction and matching to solve the problem [14, 12].

QA emerged in the last decade as one of the most promising fields in *Artificial Intelligence* thanks to some competitions organized during international conferences [19, 16], but the first studies on the subject can be dated back to 1960s [3]. In the last years some enterprise applications, such as IBM's Watson/DeepQA [7], have shown the potential of the state-of-the-art technology.

This paper describes a study on the introduction of semantic models inside a QA framework in order to improve its performances in answering users' questions. Although Semantic Role Labelling and Word Sense Disambiguation have been already employed in the past [18], distributional and latent models for QA are a completely new approach to investigate for QA.

A framework for building real-time QA systems with focus on closed domains was built for this purpose. The generality of the framework allows that also its

application to open domains can be rather easy. It exploits NLP algorithms for both English and Italian and integrates a question categorization component based on Machine Learning techniques and linguistic rules written by human experts. Text document collections used as data sources are organized in indices for generic unstructured data storage with fast and reliable search functions exploiting state-of-the-art IR weighting schemes.

The paper is structured as follows. Section 2 provides a generic overview of the framework architecture, while Section 3 presents the different semantic models. In Section 4 a preliminary evaluation of the impact of the adoption of semantic models is provided. Final conclusions, then, close the paper.

2 Framework overview

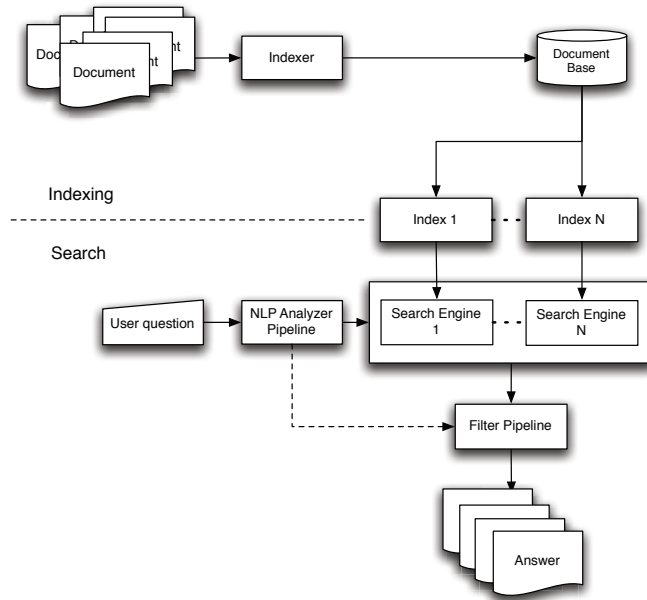


Fig. 1. QC1 Framework architecture overview

The architecture, shown in Figure 1, introduces some new aspects that make it general and easier to expand, such as the adoption of different indices, parallel search engines and different NLP and filtering pipelines, which can also run in parallel.

The first step is a linguistic analysis of the user's question. Question analysis is performed by a pipeline of NLP analyzers. NLP analyzers are provided both for

English and Italian and include a stemmer, a Part of Speech tagger, a lemmatizer, a Named Entity Recognizer and a chunker.

This step includes also a question classifier that uses an ensemble learning approach exploiting both hand-written rules and rules inferred by machine learning categorization techniques (*Support Vector Machines* are adopted), thus bringing together the hand-written rules' effectiveness and precision and the machine learning classifier's recall.

The question is then passed to the search engines, whose architecture is highly parallel and distributed. Moreover, each single engine has its own query generator, because the query's structure and syntax could change between different engines. For this purpose two different query expansion techniques are implemented: *Kullback-Liebler Divergence* [4] and *Divergence From Randomness* [1].

The filter pipeline is then responsible for the scoring and filtering of the passages retrieved by the search engines. Finally, a ranked list of passages is presented to the user.

3 Semantic Models

The word "semantics" describes the study of meaning. In NLP and IR it is used to refer to "lexical semantics", i.e. the meaning of words, and to "semantic role", the role of a phrase in a sentence. The aim of this research is to investigate whether semantic models can improve performances of QA systems and under which conditions improvements are achieved. A deep cost-benefit analysis analysis of semantic models will be also performed. New models will also be developed and tested.

From the point of view of QA, semantic models can be applied to different parts of the process.

Among the NLP Analyzers, *Word Sense Disambiguation* techniques should be applied to find the explicit meaning of every word taken from a semantic lexicon like *WordNet* [6]. The application of *Semantic Role Labelling* algorithms is also needed to extract the role of a phrase, thus helping the identification of the most important parts of the user's question [18]. *Word Sense Induction* [13] methods should be also applied to discriminate word meaning depending on the use of the word inside the collection, in an implicit way.

In the query expansion step, the use of synonyms taken from explicit repositories and similar words obtained from the similarity of contexts of use need both to be investigated.

During the search step, different approaches to semantic analysis should be applied, ranging from algebraic matrix approaches like *Latent Semantic Analysis* [5], *Non-negative Matrix Factorization* [11] and *Random Indexing* [10], to explicit representation approaches like *Explicit Semantic Analysis* [8]. As most of these techniques are expensive to be applied in real time for all documents, their adoption can be shifted to the filtering step, thus applying them only on a reduced and pre-filtered subset of all the candidate answers. To the best of my knowledge those semantic models have never been adopted in QA, in particular

for candidate answer filtering and scoring. Different search models like fuzzy and neural network based models for IR [2] will also be investigated.

In the filtering step, semantic distance and semantic correlation [15] measures can be applied to score the candidate answers, according to the adopted representation of meaning.

4 Preliminary Evaluation

A preliminary evaluation has been conducted on the *ResPubliQA 2010 Dataset* adopted into the *2010 CLEF QA Competition* [16]. This dataset contains about 10700 documents about European Union Legislation and European Parliament transcriptions, aligned in several languages including English and Italian, with 200 questions.

The adopted metric is the $c@1$ proposed for the competition:

$$c@1 = \frac{1}{N} \left(n_c + n_n \frac{n_c}{N} \right) \quad (1)$$

where N is the number of the questions, n_c is the number of the system's correct answers and n_n is the number of unanswered questions.

Several combinations of different parameters have been tested, but in the table below only the best one is shown. It employs a two BM25 based searchers [17] as search component, one using a keyword index, the other using a lemma index, and the filters pipeline is built with keyword matching, lemma matching, exact overlap, density and n-grams filters. The framework is named **qc1**. The adoption of a *Random Indexing* based semantic filter improves the system performance of 0.045 for English and 0.04 for Italian, as shown in Table 1.

System	Search	k1	b	RI	c@1 score
qc1 english	BM25	1.6	0.8	no	0.705
qc1+sf english	BM25	1.6	0.8	yes	0.75
best CLEF2010 en					0.73
qc1 italian	BM25	1.8	0.75	no	0.635
qc1+sf italian	BM25	1.8	0.75	yes	0.675
best CLEF2010 it					0.63

Table 1. Preliminary evaluation

5 Conclusions

In this paper, a research proposal about the adoption of semantic models for QA has been presented. A short overview of the adopted QA framework, alongside with a description of the different semantic models to adopt for the research purpose, has been provided. Finally, a preliminary evaluation on a standard dataset,

CLEF 2010 ResPubliQA, has been given, which shows an improvement in comparison to other state-of-the-art systems and demonstrates how promising the use of semantic models is for the field of QA.

References

1. Amati, G., Van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.* 20, 357–389 (October 2002)
2. Baeza-Yates, R., Ribeiro-Neto, B.: *Modern information retrieval*. Paperback (May 1999)
3. Bert F. Green, J., Wolf, A.K., Chomsky, C., Laughery, K.: Baseball: an automatic question-answerer. In: *Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference*. pp. 219–224. IRE-AIEE-ACM '61 (Western), ACM, New York, NY, USA (1961)
4. Carpineto, C., de Mori, R., Romano, G., Bigi, B.: An information-theoretic approach to automatic query expansion. *ACM Trans. Inf. Syst.* 19, 1–27 (January 2001)
5. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. *Journal of the American Society of Information Science* 41(6), 391–407 (1990)
6. Fellbaum, C.: *WordNet: an electronic lexical database*. Language, speech, and communication, MIT Press (1998)
7. Ferrucci, D.A.: *Ibm's watson/deepqa*. SIGARCH Computer Architecture News 39(3) (2011)
8. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. pp. 1606–1611 (2007)
9. Hovy, E.H., Gerber, L., Hermjakob, U., Junk, M., Lin, C.Y.: Question answering in webclopedia. In: *TREC* (2000)
10. Kanerva, P.: *Sparse distributed memory*. Bradford Books, MIT Press (1988)
11. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755), 788–791 (Oct 1999)
12. Lin, J.: An exploration of the principles underlying redundancy-based factoid question answering. *ACM Trans. Inf. Syst.* 25 (April 2007)
13. Navigli, R.: Word sense disambiguation: A survey. *ACM Comput. Surv.* 41(2), 10:1–10:69 (Feb 2009)
14. Paşca, M.: *Open-domain question answering from large text collections*. Studies in computational linguistics, CSLI Publications (2003)
15. Pedersen, T., Patwardhan, S., Michelizzi, J.: Wordnet::similarity: measuring the relatedness of concepts. In: *Demonstration Papers at HLT-NAACL 2004*. pp. 38–41. HLT-NAACL–Demonstrations '04, Association for Computational Linguistics, Stroudsburg, PA, USA (2004)
16. Penas, A., Forner, P., Rodrigo, A., Sutcliffe, R.F.E., Forascu, C., Mota, C.: Overview of ResPubliQA 2010: Question Answering Evaluation over European Legislation. In: *Braschler, M., Harman, D., Pianta, E. (eds.) Working notes of ResPubliQA 2010 Lab at CLEF 2010* (2010)
17. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.* 3, 333–389 (April 2009)