Semantic Models for Re-ranking in Question Answering

Piero Molino
Department of Computer Science
University of Bari Aldo Moro
Via Orabona 4, I-70125
Bari, Italy
piero.molino@uniba.it

This paper describes a research aimed at unveiling the role of Semantic Models in Question Answering. In these systems questions and answers are often expressed in quite different languages, so our objective is to bridge this "lexical chasm" adopting semantic representations. The aim of the research is to find out if Semantic Models are useful for this task and if they can improve the answer re-ranking performance. We have carried out an initial evaluation of a subset of the semantic models on the CLEF2010 QA dataset, showing their effectiveness. We also did a first attempt in combining them by means of Learning to Rank algorithms.

Question Answering, Semantics, Learning to Rank

1. INTRODUCTION

Question Answering (QA) emerged in the last decade as one of the most promising fields in Artificial Intelligence. By exploiting techniques borrowed from IR and Natural Language Processing (NLP), QA systems are able to answer users' questions expressed in natural language with either the exact answer or with short passages of text containing the exact answer.

One of the pitfalls of QA systems is the difference between the language in which users ask the questions and the language of the answers, the so called "lexical chasm" (Berger et al. (2000)). We believe that a semantic representation of both question and answers would "bridge the lexical chasm", so we want to adopt Semantic Models (SM) in order to re-rank the candidate answers in a more sensible way. Based on the usefulness that those models show in other tasks, we think that SMs can have a significant role in improving current state-of-the-art systems' performance in answer re-ranking.

The general aim of this research is to find out if SMs are useful for answer re-ranking. The questions it wants to answer are:

- Are additional semantic features useful for answer re-ranking? Does their adoption improve systems' performance?
- Which of them is more effective and under which circumstances?
- Do semantic features bring information that is not present in the bag-of-words and syntactic features?
- Is there any Learning-to-Rank (MLR) algorithm that exploits semantic features more than others (has more relative or absolute improvement by their adoption) and why?

The paper is structured as follows. A brief overview of related work is reported in Section 2, while the proposed methodology is described in Section 3. Some preliminary experiments are shown in Section 4 and the description of the future work closes the paper.

2. RELATED WORK

Much of the work in QA has been done on factoid questions, where answers are short excerpts of text, usually named entities, dates or quantities. Factoid QA systems rely heavily on information extraction techniques in order to obtain the specific answer, including the adoption of linguistic patterns.

In the last few years non-factoid QA has received more attention. It focuses on causation, manner and reason questions, where the expected answer has the form of a passage of text. Depending on the structure of the corpus, the passages can be single sentences, groups of sentences, paragraphs or short texts.

The presence of annotated corpora from Text REtrieval Conference (TREC) and Cross Language Evaluation Forum (CLEF) makes it possible to use machine learning techniques to tackle the problem of ranking the passages for further extraction in factoid QA (Agarwal et al. (2012)). In non-factoid QA the training data adopted is of different types, like hand annotated answers from Wikipedia used by Verberne et al. (2011), small hand built corpora adopted by Higashinaka and Isozaki (2008), Frequently Asked Questions lists employed in Soricut and Brill (2006) and Yahoo! Answers Extracted corpus used by Surdeanu et al. (2011).

These corpora enable the adoption of MLR algorithms in order to output a sensible ranking of the candidate answers. MLR algorithms apply Machine Learning techniques to the problem of ordering a set of items with respect to queries. In the case of QA the items are answers and the queries are the questions. Usually, the features for the learning task are different similarity measures between the query and the item, in the Information Retrieval tasks TF-IDF, BM25 and Language Modeling based features are often used. Verberne et al. (2008) adopted linguistically motivated features showing their effectiveness for the QA task, while Verberne et al. (2011) compare different MLR algorithms over the same set of features.

Bilotti et al. (2010) showed the importance of semantic role labeling features, while Verberne et al. (2011) use WordNet synsets for expansion and comparison. Surdeanu et al. (2011) carried out a comprehensive large scale evaluation, alongside with the introduction of new features based on translation models and web correlation.

As mentioned, just few experiments adopted semantic features for answer re-ranking, but their use showed significant performance improvement. There are still several different possible semantic features that have not been taken into account so far and we want to find out if their use could be helpful. For example features coming from SM like Distributional Semantic Models (DSMs) (Turney and Pantel (2010)), Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch (2007)) and Latent Dirichlet Allocation (LDA) (Blei et al. (2003)) have never been applied to the task so far.

3. METHODOLOGY

We are going to test if these insights are correct starting from the design and implementation of a QA framework that helps us to set up several systems with different settings.

We have already built the cornerstone: Question-Cube (Molino and Basile (2012)) is a multilingual QA framework created using Natural Language Processing and Information Retrieval techniques.

Question analysis is carried out by a full-featured NLP pipeline. The document search step is carried out by Lucene, a standard off-the-shelf retrieval framework that allows TF-IDF, Language Modeling and BM25 weighting. All the passages in the top documents from Lucene are considered as candidate answers. The answer re-ranking component is designed as a pipeline of different scoring criteria. We derive a global re-ranking function by combining the scores with CombSum (Shaw et al. (1994)). The CombSum function can be replaced by MLR algorithms if available training data is available. More details on the framework and a description of the main scorers is reported by Molino and Basile (2012).

The next step is the implementation of different MLR algorithms in order to combine the features obtained by the scoring criteria with linear and non-linear models and replace the CombSum function. QA datasets are usually really skewed. For each question, only one correct answer is usually given and thus Pairwise MLR algorithms could be more effective than Pointwise and Listwise approaches. Verberne et al. (2011) showed that this still has to be proved, so we implemented a whole collection of different MLR algorithms inside the framework including Pointwise, Pairwise and Listwise algorithms. This will allow us to compare their performance on the non-factoid QA task and to find out if they exploit the additional information given by the semantic features in different ways.

The implemented algorithms are: Linear Regression (used as a baseline), Logistic Regression (Agarwal et al. (2012) reported to be very effective in IBM's DeepQA / Watson), RankNet (Burges et al. (2005)), RankBoost (Freund et al. (2003)) and LambdaMART (Wu et al. (2010)). More, in particular SVM based ones, will be added in the near future.

As a proof of concept we implemented some scoring criteria based on DSMs in order to realize if their adoption as unique rankers or combined with simple similarity and density criteria would improve ranking over the one obtained with classic Information Retrieval weighting schemes.

3.1. Distributional Semantic Models

Distributional Semantic Models (DSMs) represent word meanings through linguistic contexts. The meaning of a word can be inferred by the linguistic contexts in which the word occurs.

The idea behind DSMs can be summarized as follows: if two words share the same linguistic context they are somehow similar in meaning. For example, in analyzing the sentences “drink a glass of wine” and “drink a glass of beer”, we can assume that the words “wine” and “beer” have a similar meaning.

Using that assumption, the meaning of a word can be expressed by the geometrical representation in a *semantic space*. In this space a word is represented by a vector whose dimensions correspond to linguistic contexts surrounding the word. The word vector is built by analyzing (e.g. counting) the contexts in which the term occurs across a corpus. Some definitions of context may be the set of co-occurring words in a document, in a sentence or in a window of surrounding terms.

The earliest and simplest formulation of such a space stems from the use of the Vector Space Model in IR.

Semantic space scalability and independence from external resources resulted in their practical use in many different tasks. For example they have been applied in several linguistic and cognitive tasks, such as synonyms detection by Landauer and Dumais (1997), semantic priming by Jones and Mewhort (2007), automatic construction of thesauri by Schütze and Pedersen (1995) and word sense induction by Schütze (1998).

Our DSMs are constructed over a co-occurrence matrix. The linguistic context taken into account is a window w of co-occurring terms. Given a reference corpus, the collection of documents indexed by the QA system, and its vocabulary V , a $n \times n$ co-occurrence matrix is defined as the matrix $\mathbf{M} = (m_{ij})$ whose coefficients $m_{ij} \in \mathbb{R}$ are the number of co-occurrences of the words t_i and t_j within a predetermined distance w .

The *term $\times$ term* matrix $\mathbf{M}$, based on simple word co-occurrences, represents the simplest semantic space, called Term-Term co-occurrence Matrix (TTM).

In literature, several methods to approximate the original matrix by rank reduction have been proposed. The aim of these methods varies from discovering high-order relations between entries

to improving efficiency by reducing its noise and dimensionality. We exploit three methods for building our semantic spaces: Latent Semantic Analysis (*LSA*) (Deerwester et al. (1990)), Random Indexing (*RI*) (Kanerva (1988)) and LSA over RI (*LSARI*) (Sellberg and Jönsson (2008)). *LSARI* applies the SVD factorization to the reduced approximation of $\mathbf{M}$ obtained through RI. All these methods produce a new matrix $\hat{\mathbf{M}}$, which is a $n \times k$ approximation of the co-occurrence matrix $\mathbf{M}$ with n row vectors corresponding to vocabulary terms, while k is the number of reduced dimensions. Molino et al. (2012) give more details.

We integrated the DSMs into the framework creating a new scorer, the **Distributional Scorer**, that represents both question and passage by applying the addition operator to the vector representation of terms they are composed of. The vectors of the single words came from the different $\hat{\mathbf{M}}$, so changing the method to construct it we obtain one different scorer for each semantic space. This way it is possible to compute the similarity between question and passage by exploiting the cosine similarity between the summed vectors.

The simple scorers employed alongside with the ones based on DSMs in the evaluation are: the **Overlap Scorer**, a scorer that counts the term overlap between the question and the candidate answer; the **Exact Sequence Scorer**, a scorer that counts the number of consecutive overlapping terms between the question and the answer, and the **Density Scorer**, a scorer that assigns a score to a passage based on the distance of the question terms inside it as proposed by Monz (2004). All the scorers can use linguistic annotations like stems, Part-of-Speech Tags, lemmas, Named Entities and combinations of annotations as features instead of simple words.

4. EVALUATION

The goal of the evaluation is twofold: (1) to give empirical support for the thesis that DSMs could be effective into our QA system and (2) to provide a comparison between the different DSMs.

The evaluation has been performed on the *ReS-PubliQA 2010 Dataset* adopted in the *2010 CLEF QA Competition* (Penas et al. (2010)). The dataset contains about 10,700 documents of the European Union legislation and European Parliament transcriptions, aligned in several languages including English and Italian, with 200 questions.

The first metric adopted in the evaluation is the accuracy $a@n$ (known in literature as *success@n*),

calculated considering only the first n answers. If the correct answer occurs in the top n retrieved answers, the question is marked as correctly answered. In particular, we take into account several values of $n = 1, 5, 10$ and 30 . Moreover, we adopt the Mean Reciprocal Rank (MRR) as well, that considers the rank of the correct answer.

The framework setup used for the evaluation adopts Lucene as document searcher, and uses a NLP Pipeline made of a lemmatizer, a Part-of-Speech tagger and a named entity recognizer.

The different DSMs and the classic TTM have been used both as scorers alone, which means no other scorers are adopted, and combined with a standard scorer pipeline. The composition of the standard pipeline includes: the Terms Overlap (TO) scorer, the Lemma+Part-Of-Speech Overlap (LPO) scorer, the Lemma+Part-Of-Speech Density (LPD) scorer, the Exact Term Sequence (ET) scorer.

Moreover, we empirically chose the parameters for the DSMs: the window w of terms considered for computing the co-occurrence matrix is 4, while the number of reduced dimensions considered in LSA, RI and LSARI is equal to 1000.

The performance of the standard pipeline, without the distributional scorer, is shown as a baseline. The experiments have been carried out both for English and Italian. Results are shown in Table 1 for English and in Table 2 for Italian.

The results in the rows marked as "alone" refer to DSMs used as unique rankers, while the results reported in the "combined" part of tables refer to the CombSum of TO, LPO, LPD, ET and the specified DSM scorers.

Both tables report the accuracy $a@n$ computed considering a different number of answers, the MRR and the significance of the results with respect to both the baseline ([†]) and the distributional model based on TTM ([‡]). The significance is computed using the non-parametric Randomization test as suggested by Smucker et al. (2007). The best results are reported in bold.

Considering each distributional scorer on its own, the results show that all the proposed DSMs are better than the TTM, and the improvement is always significant. The best improvement for the MRR in English is obtained by LSA (+180%), while in Italian by LSARI (+161%).

As for the distributional scorers combined with the standard scorer pipeline, the results show that all the combinations are able to overcome the baseline. For

Table 1: Evaluation Results for English

		English				
	Run	a@1	a@5	a@10	a@30	MRR
alone	TTM	0.060	0.145	0.215	0.345	0.107
	RI	0.180	0.370	0.425	0.535	0.267 [‡]
	LSA	0.205	0.415	0.490	0.600	0.300 [‡]
	LSARI	0.190	0.405	0.490	0.620	0.295 [‡]
combined	baseline	0.445	0.635	0.690	0.780	0.549
	TTM	0.535	0.715	0.775	0.810	0.614
	RI	0.550	0.730	0.785	0.870	0.637 ^{†‡}
	LSA	0.560	0.725	0.790	0.855	0.637 [†]
	LSARI	0.555	0.730	0.790	0.870	0.634 [†]

Table 2: Evaluation Results for Italian

		Italian				
	Run	a@1	a@5	a@10	a@30	MRR
alone	TTM	0.060	0.140	0.175	0.280	0.097
	RI	0.175	0.305	0.385	0.465	0.241 [‡]
	LSA	0.155	0.315	0.390	0.480	0.229 [‡]
	LSARI	0.180	0.335	0.400	0.500	0.254 [‡]
combined	baseline	0.445	0.635	0.690	0.780	0.549
	TTM	0.405	0.565	0.645	0.740	0.539 [†]
	RI	0.465	0.645	0.720	0.785	0.555 [†]
	LSA	0.470	0.645	0.690	0.785	0.551 [†]
	LSARI	0.480	0.635	0.690	0.785	0.557 ^{†‡}

English we have obtained an improvement in MRR of about 16% compared to the baseline and the result obtained by the TTM is significant. For Italian, we have achieved an even higher improvement in MRR of 26% compared to the baseline using LSARI.

The slight difference in performance between LSA and LSARI suggests that LSA applied to the matrix obtained by RI produces the same result of LSA applied to TTM, but requiring less computation time, as the matrix obtained by RI contains less dimensions than the TTM matrix.

Those results give a preliminary answer to some of the research questions: they suggest that the semantic features can be useful for answer re-ranking because they improve the performances significantly and that LSA and LSARI are the most effective among the adopted ones.

4.1. Preliminary MLR experiment

A preliminary experiment with MLR algorithms has been carried out separately from the main evaluation. The features we employed are the outputs of the scorers adopted in the previous experiment, a really small number, but the aim of the experiment was to find out if a better combination of

the same scorers of the main evaluation could lead to better results.

The experiment was carried out using the RankNet (Borges et al. (2005)) MLR algorithm, performing a 10-fold Cross Validation on the same dataset of the main evaluation. We did four runs with the 4 fixed features coming from the standard scorers described in Section 4 and changing the fifth feature among the four different DSMs scorers. The best average score of MRR on the 10 different folds is 0.68 for English and 0.605 for Italian obtained with the LSARI DSM. Far from being significant, this little MLR experiment still encourages us to follow the path of semantic features combined with MLR algorithms.

5. FUTURE PLANS

There are several future steps to follow in order to answer the research questions. Carrying out the reported evaluation, we discovered that some of the semantic features obtained from DSMs can be useful for answer re-ranking both alone and combined with other features.

What we still don't know is how effective they can be inside a MLR setting and we still don't know if this can be generalized to other datasets.

To this purpose, the following activities will be carried out:

- To add more MLR algorithms for re-ranking. More MLR algorithms are fundamental in order to carry out a comprehensive comparative analysis.
- To experiment further the usefulness of other semantic features, coming from ESA, LDA and even more semantic models. This could help in catching different aspects of the semantics of questions and answers that the DSMs alone do not cover.
- To incorporate other state-of-the-art linguistic features in order to realize if the information they convey overlaps with the information from the semantic features. Good candidates are the ones proposed by Verberne et al. (2008), that cover lexical and syntactic information, and the ones proposed by Surdeanu et al. (2011). In particular, the translation based features are really effective and can help to "bridge the lexical chasm" (Berger et al. (2000)).
- Other operations for combining vectors coming from the applied DSMs will also be investigated, in order to tackle more deeply the semantic compositionality problem. For doing so Mitchell and Lapata (2010) used operators

like product, tensor product and circular convolution, and their adoption for our task can be helpful.

- Once all the features are ready, a MLR algorithm comparison will be carried out, in order to find out which algorithms take more advantage from the semantic features. An ablation test will be used to understand how much of the improvement is obtained thanks to the semantic features.

Alongside with those steps, different datasets will be collected, focusing mainly on non-factoid QA. The Yahoo! Answers Manner Questions datasets are a good starting point, but also non-factoid questions from Webclopedia (Hovy et al. (2000)) can be helpful. The aim is to compare directly to state-of-the-art systems in order to find out if the semantic features can lead to better results in a general setting.

Another dataset will be collected with aligned English and Italian non-factoid questions with answers taken from Wikipedia. The answers will be posted by the users of Wikied¹ and the dataset will contain textual answers in the form of paragraphs from Wikipedia pages, their relevance judgment (the number of votes from the users) and a feature list that will contain the output of the different scorers.

REFERENCES

- Arvind Agarwal, Hema Raghavan, Karthik Subbian, Prem Melville, Richard D. Lawrence, David Gondek, and James Fan. Learning to rank for robust question answering. In *CIKM 2012*, pages 833–842, 2012.
- Adam L. Berger, Rich Caruana, David Cohn, Dayne Freitag, and Vibhu O. Mittal. Bridging the lexical chasm: statistical approaches to answer-finding. In *SIGIR*, pages 192–199, 2000.
- Matthew W. Bilotti, Jonathan L. Elsas, Jaime G. Carbonell, and Eric Nyberg. Rank learning for factoid question answering with linguistic and semantic constraints. In *CIKM*, pages 459–468, 2010.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3: 993–1022, March 2003. ISSN 1532-4435.
- Christopher J. C. Borges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N. Hullender. Learning to rank using gradient descent. In *ICML*, pages 89–96, 2005.

¹A QA system over Wikipedia articles, www.wikiedi.it

- Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.
- Yoav Freund, Raj D. Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003.
- Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, 2007.
- Ryuichiro Higashinaka and Hideki Isozaki. Corpus-based question answering for why-questions. In *Proceedings of IJCNLP*, pages 418–425, 2008.
- Eduard H. Hovy, Laurie Gerber, Ulf Hermjakob, Michael Junk, and Chin-Yew Lin. Question answering in webclopedia. In *TREC*, 2000.
- Michael N. Jones and Douglas J. K. Mewhort. Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114(1):1–37, 2007.
- Pentti Kanerva. *Sparse Distributed Memory*. MIT Press, 1988.
- Thomas K. Landauer and Susan T. Dumais. A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240, 1997.
- Jeff Mitchell and Mirella Lapata. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429, 2010. ISSN 1551-6709.
- Piero Molino and Pierpaolo Basile. Questioncube: a framework for question answering. In Giambattista Amati, Claudio Carpineto, and Giovanni Semeraro, editors, *IIR 2012*, volume 835 of *CEUR Workshop Proceedings*, pages 167–178, 2012.
- Piero Molino, Pierpaolo Basile, Annalina Caputo, Pasquale Lops, and Giovanni Semeraro. Exploiting distributional semantic models in question answering. In *ICSC*, pages 146–153, 2012.
- Christof Monz. Minimal span weighting retrieval for question answering. In Rob Gaizauskas, Mark Greenwood, and Mark Hepple, editors, *Proceedings of the SIGIR Workshop on Information Retrieval for Question Answering*, pages 23–30, 2004.
- A. Penas, P. Forner, A. Rodrigo, R. F. E. Sutcliffe, C. Forascu, and C. Mota. Overview of republiqa 2010: Question answering evaluation over european legislation. In Martin Braschler, Donna Harman, and Emanuele Pianta, editors, *Working notes of ResPubliQA 2010 Lab at CLEF 2010*, 2010.
- Hinrich Schütze. Automatic word sense discrimination. *Comput. Linguist.*, 24:97–123, March 1998. ISSN 0891-2017.
- Hinrich Schütze and Jan O. Pedersen. Information retrieval based on word senses. In *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1995.
- Linus Sellberg and Arne Jönsson. Using random indexing to improve singular value decomposition for latent semantic analysis. In *LREC*, 2008.
- Joseph A. Shaw, Edward A. Fox, Joseph A. Shaw, and Edward A. Fox. Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2)*, pages 243–252, 1994.
- Mark D. Smucker, James Allan, and Ben Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *CIKM*, pages 623–632, 2007.
- Radu Soricut and Eric Brill. Automatic question answering using the web: Beyond the factoid. *Inf. Retr.*, 9(2):191–206, March 2006. ISSN 1386-4564.
- Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. Learning to rank answers to non-factoid questions from web collections. *Computational Linguistics*, 37(2):351–383, 2011.
- Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res. (JAIR)*, 37:141–188, 2010.
- Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. Using syntactic information for improving why-question answering. In *COLING*, pages 953–960, 2008.
- Suzan Verberne, Hans van Halteren, Daphne Theijssen, Stephan Raaijmakers, and Lou Boves. Learning to rank for why-question answering. *Inf. Retr.*, 14(2):107–132, 2011.
- Qiang Wu, Christopher J. C. Burges, Krysta Marie Svore, and Jianfeng Gao. Adapting boosting for information retrieval measures. *Inf. Retr.*, 13(3): 254–270, 2010.