Semantic Models for Answer Re-ranking in Question Answering

Piero Molino
2^{nd} year PhD. student
Department of Computer Science
University of Bari Aldo Moro
Via Orabona, I-70125
Bari, Italy
piero.molino@uniba.it

ABSTRACT

This paper describes a research aimed at unveiling the role of Semantic Models into Question Answering systems. The objective is to use Semantic Models for answer re-ranking in order to improve the passage retrieval performance and the overall performance of the system.

Semantic Models use concepts rather than simple words to represent texts, expressing them in explicit or implicit ways. The different representation allows to compute relatedness between users' questions and candidate answers to provide better answer re-ranking. This is done by exploiting different properties, like explicit relations between concepts or latent similarities between words expressed as the similarity of the contexts in which they appear.

We want to find out if the combination of different semantic relatedness measures by means of Learning to Rank algorithms will show a significant improvement over the state-of-the-art. We have carried out an initial evaluation of a subset of the semantic models on the CLEF2010 QA dataset, proving their effectiveness.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.7 [**Artificial Intelligence**]: Natural Language Processing; I.2.6 [**Artificial Intelligence**]: Learning

Keywords

Question Answering, Learning to Rank, Semantics

1. INTRODUCTION

The task of Question Answering (QA) is to find correct answers to users' questions expressed in natural language. 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 heavly 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 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 passage retrieval step is, anyway, fundamental in both factoid and non-factoid QA as in the former the answers are extracted from the obtained passages, while in the latter the passage corresponds to the candidate answer itself, even if the length of the passage for non-factoid QA is much larger, as shown in [28].

The presence of annotated corpora from Text REtrieval Conference (TREC) and Cross Language Evaluation Forum (CLEF) allows to use machine learning techniques to tackle the problem of ranking the passages for further extraction in factoid QA [1]. In non-factoid QA the training data adopted is of different types, like hand annotated answers from Wikipedia [29], small hand built corpora [10], Frequently Asked Questions lists [2, 25] and Yahoo! Answers Extracted corpus [26].

This allows the adoption of Learning to Rank (MLR) algorithms in order to output a sensible ranking of the candidate answers. MLR algorithms applies Machine Learning techniques to the problem of ordering a set of items with respect to queries. In the QA case 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. In [28] the adoption of linguistically motivated features is shown to be effective for the QA task, while in [30] different MLR algorithms were compared over the same set of features.

The importance of semantic features, in the form of semantic role labelling features, was shown in [4], while in [30] WordNet synsets are used for expansion and comparison. A comprehensive large scale evaluation, alongside with the introduction of new features based on translation models and web correlation, was carried out in [26].

As mentioned, just few experiments adopted semantic features for answer re-ranking, but their use showed significant

SIGIR 2013 Dublin, Ireland

performance improvement. There are still several different possible semantic features that have not been taken into account so far and our goal is to find out if their use could lead to performance improvement.

For example features coming from Semantic Models (SM) like Distributional Semantic Models (DSMs) [27], Explicit Semantic Analysis (ESA) [9], Latent Dirichlet Allocation (LDA) [5] induced topics have never been applied to the task so far.

Based on the usefulness that those models show in other tasks, we think that SM can have a significant role in improving current state-of-the-art systems' performance in answer re-ranking.

The questions this research wants to answer are:

- Do semantic features bring information that is not present in the bag-of-words and syntactic features?
- Do they bring different information or does it overlap with that of other features?
- 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?
- Is there any MLR algorithm that exploits semantic features more than others (has more relative or absolute improvement by their adoption) and why?

2. 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: QuestionCube is a multilingual QA framework created using Natural Language Processing and Information Retrieval techniques. The overall architecture of the framework is shown in Figure 1.

Question analysis is carried out by a full-featured NLP pipeline. The passage search step is carried out by Lucene, a standard off-the-shelf retrieval framework that allows TF-IDF, Language Modeling and BM25 weighting. The question re-ranking component is designed as a pipeline of different scoring criteria. We derive a global re-ranking function combining the scores with CombSum [23]. 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 in [16].

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. This still has to be proved [30] so we implemented a whole collection of different MLR algorithms inside the framework. This will allow us to compare their performance on the non-factoid QA task and to find out if they exploit the additional information diven by the semantic features in different ways.

The implemented algorithms are: Linear Regression (used as a baseline), Logistic Regression (reported to be very effective in IBM's DeepQA / Watson [1]), RankNet [6], RankBoost [8] and LambdaMART [31]. More, in particular SVM based ones, will be added in the near future.

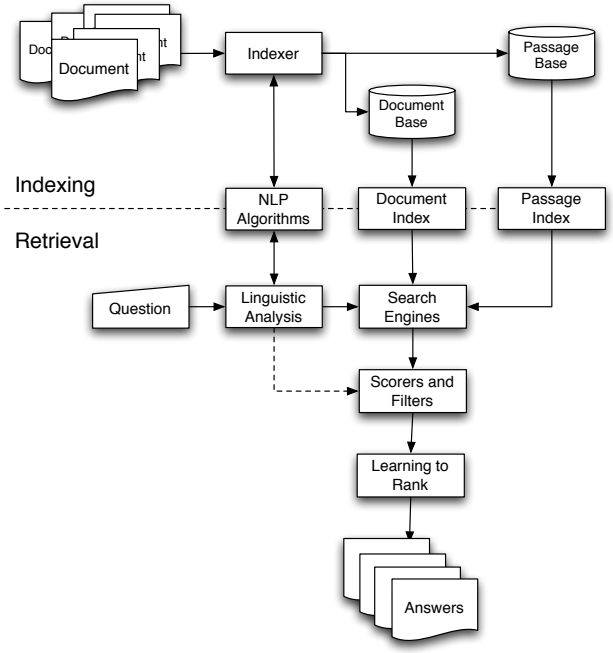

Figure 1: QuestionCube architecture overview

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.

2.1 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 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 [19].

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 choice [14], seman-

tic priming [12, 14], automatic construction of thesauri [21] and word sense induction [20].

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) [7], Random Indexing (RI) [13] and LSA over RI ($LSARI$) [22]. $LSARI$ 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. More details can be found in [17].

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. Furthermore, it is possible to compute the similarity between question and passage by exploiting the cosine similarity between vectors using the different matrices.

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. 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.

3. EVALUATION

The goal of the evaluation is twofold: (1) proving the effectiveness of DSMs into our question answering system and (2) providing a comparison between the different DSMs.

The evaluation has been performed on the *ResPubliQA 2010 Dataset* adopted in the *2010 CLEF QA Competition* [18]. 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 stemmer, 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 in [24]. The best results are reported in bold.

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†

Considering each distributional scorer on its own, the results prove 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 prove that all the combinations are able to overcome the baseline. For English we have

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	$0.241^{\ddagger}$
	LSA	0.155	0.315	0.390	0.480	$0.229^{\ddagger}$
	LSARI	**0.180**	**0.335**	**0.400**	**0.500**	$\mathbf{0.254}^{\ddagger}$
combined	*baseline*	*0.445*	*0.635*	*0.690*	*0.780*	*0.549*
	TTM	0.405	0.565	0.645	0.740	$0.539^{\dagger}$
	RI	0.465	**0.645**	**0.720**	**0.785**	$0.555^{\dagger}$
	LSA	0.470	**0.645**	0.690	**0.785**	$0.551^{\dagger}$
	LSARI	**0.480**	0.635	0.690	**0.785**	$\mathbf{0.557}^{\dagger\ddagger}$

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

The slight difference in performance between LSA and LSARI proves 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.

Finally, the improvement obtained considering each distributional scorers on its own is higher than their combination with the standard scorer pipeline.

3.1 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 experimentation, 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 the RankNet [6] 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 3 and changing the fifth feature among the four diffferent 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 for being significative, this little MLR experiment still encurages us to follow the path of semantic features combined with MLR algorithms.

4. 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 we want to adopt can be useful. In particular, 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 that use features from the different scorers. This will also probably lead to even better performances than the ones already achieved, as suggested from the preliminary experiment discussed in Section 3.1. More MLR algorithms are fundamental in order to carry out a comprehensive comparative analysis.
- To experiment further the usefulness of other semantic features, such as 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.
- Incorporating other state-of-the-art linguistic features will also be of fundamental importance in order to realize if the conveyed information of the semantic features overlaps with the information from other linguistic features. Candidate features are the ones proposed in [28], that cover lexical and syntactic information, and the ones proposed in [26]. In particular, the translation based features are really effective and can help to "bridge the lexical chasm" [3].
- Other operations for combining vectors coming from the applied DSMs will also be investigated, in order to tackle more deeply the semantic compositionality problem. In [15] operators like product, tensor product and circular convolution are used, and their adoption for our task can be helpful.
- Once all the features are ready, 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 useful 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 [11] can be helpful. The reason for that is the possibility to compare our evaluation directly to the state-of-the-art ones in order to realize if the semantic features can lead to better results also on those datasets.

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 Wikiedi[1] 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 implementing different similarity criteria between the question and the answer.

5. REFERENCES

[1] A. Agarwal, H. Raghavan, K. Subbian, P. Melville, R. D. Lawrence, D. Gondek, and J. Fan. Learning to rank for robust question answering. In *CIKM*, pages 833–842, 2012.

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

[2] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 183–194, New York, NY, USA, 2008. ACM.

[3] A. L. Berger, R. Caruana, D. Cohn, D. Freitag, and V. O. Mittal. Bridging the lexical chasm: statistical approaches to answer-finding. In *SIGIR*, pages 192–199, 2000.

[4] M. W. Bilotti, J. L. Elsas, J. G. Carbonell, and E. Nyberg. Rank learning for factoid question answering with linguistic and semantic constraints. In *CIKM*, pages 459–468, 2010.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[6] C. J. C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. N. Hullender. Learning to rank using gradient descent. In *ICML*, pages 89–96, 2005.

[7] 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(6):391–407, 1990.

[8] Y. Freund, R. D. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003.

[9] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *In Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, 2007.

[10] R. Higashinaka and H. Isozaki. Corpus-based question answering for why-questions. In *In Proceedings of IJCNLP*, pages 418–425, 2008.

[11] E. H. Hovy, L. Gerber, U. Hermjakob, M. Junk, and C.-Y. Lin. Question answering in webclopedia. In *TREC*, 2000.

[12] M. N. Jones and D. J. K. Mewhort. Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114(1):1–37, 2007.

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

[14] T. K. Landauer and S. 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.

[15] J. Mitchell and M. Lapata. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429, 2010.

[16] P. Molino and P. Basile. QuestionCube: a Framework for Question Answering. In G. Amati, C. Carpineto, and G. Semeraro, editors, *IIR*, volume 835 of *CEUR Workshop Proceedings*, pages 167–178. CEUR-WS.org, 2012.

[17] P. Molino, P. Basile, A. Caputo, P. Lops, and G. Semeraro. Exploiting distributional semantic models in question answering. In *ICSC*, pages 146–153, 2012.

[18] A. Penas, P. Forner, A. Rodrigo, R. F. E. Sutcliffe, C. Forascu, and C. Mota. Overview of ResPubliQA 2010: Question Answering Evaluation over European Legislation. In M. Braschler, D. Harman, and E. Pianta, editors, *Working notes of ResPubliQA 2010 Lab at CLEF 2010*, 2010.

[19] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18:613–620, November 1975.

[20] H. Schütze. Automatic word sense discrimination. *Comput. Linguist.*, 24:97–123, March 1998.

[21] H. Schütze and J. 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.

[22] L. Sellberg and A. Jönsson. Using random indexing to improve singular value decomposition for latent semantic analysis. In *LREC*, 2008.

[23] J. A. Shaw, E. A. Fox, J. A. Shaw, and E. A. Fox. Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2*, pages 243–252, 1994.

[24] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 623–632, New York, NY, USA, 2007. ACM.

[25] R. Soricut and E. Brill. Automatic question answering using the web: Beyond the factoid. *Inf. Retr.*, 9(2):191–206, Mar. 2006.

[26] M. Surdeanu, M. Ciaramita, and H. Zaragoza. Learning to rank answers to non-factoid questions from web collections. *Computational Linguistics*, 37(2):351–383, 2011.

[27] P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res. (JAIR)*, 37:141–188, 2010.

[28] S. Verberne, L. Boves, N. Oostdijk, and P.-A. Coppen. Using syntactic information for improving why-question answering. In *COLING*, pages 953–960, 2008.

[29] S. Verberne, L. Boves, N. Oostdijk, and P.-A. Coppen. What is not in the bag of words for *why*-qa? *Computational Linguistics*, 36(2):229–245, 2010.

[30] S. Verberne, H. van Halteren, D. Theijssen, S. Raaijmakers, and L. Boves. Learning to rank for *why*-question answering. *Inf. Retr.*, 14(2):107–132, 2011.

[31] Q. Wu, C. J. C. Burges, K. M. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Inf. Retr.*, 13(3):254–270, 2010.

APPENDIX

A. STUDENT STATEMENT

The main reason why I would like to participate to the SIGIR Doctoral Consortium is that I am just starting the 2^{nd} year of my PhD and my research proposal is defined enough to be presented, but at the same time I still have enough time to straighten out the direction I am taking with the feedbacks I would obtain at SIGIR. Participating to the next edition would not leave me enough time to improve my research on the basis of the feedback.

I am also absolutely confident that there would be no better opportunity than SIGIR Doctoral Consortium to show my work and get useful feedback because the topics I am working on fit really well with the topics of the conference and because SIGIR is so prestigious that most of the researchers in my areas of interest will probably be present. This would help me to get the best feedback possible, helping me to discover new direction I would probably have never imagined and guiding me toward a significant improvement in the quality of my research.

I am also certain that the interaction with other students disserting about their research could be helpful to learn some new insights in cutting edge research and would help me in building relationships that can turn out to be important for my career and could also help me in creating new fruitful collaborations.

B. ADVISOR STATEMENT

I have known Piero Molino since 2008, soon before his Bachelor graduation. I have had the chance to closely monitor his progress since he started working on his thesis about Question Answering (QA). At the moment, Piero is expanding his knowledge and preparation through a PhD in Computer Science under my supervision, and he is taking part in research projects in the QA area. He is accomplishing all this while working actively for QuestionCube, a startup company he founded, operating in the field of Semantic Search Engines and Question Answering.

My opinion is that the SIGIR Doctoral Consortium will be the perfect venue to understand whether the specific direction chosen to improve QA systems is worth to be investigated and can allow the development of advanced QA systems which can be applied in the context of real-world applications.

This connotation of the work towards real industrial scenarios is an expertise that is not available at our institution, in particular in the field of QA.

Piero has just started the second year of the PhD course and already identified the main topic of the work. The Italian PhD program is 3-years long and the defense of the thesis is expected on June 2015. This means that after the precious feedback by the members of the SIGIR Doctoral Consortium Program Committee, there would be enough time to identify problems and related solutions.