Exploiting Multiple Semantic Resources for Answer Selection

Jeongwoo Ko, Laurie Hiyakumoto, Eric Nyberg

School of Computer Science, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh PA 15213, USA
{jko, hyaku, ehn}@cs.cmu.edu

Abstract

This paper describes the utility of semantic resources such as the Web, WordNet and gazetteers in the answer selection process for a question-answering system. In contrast with previous work using individual semantic resources to support answer selection, our work combines multiple resources to boost the confidence scores assigned to correct answers and evaluates different combination strategies based on unweighted sums, weighted linear combinations, and logistic regression. We apply our approach to select answers from candidates produced by three different extraction techniques of varying quality, focusing on TREC questions whose answers represent locations or proper-names. Our experimental results demonstrate that the combination of semantic resources is more effective than individual resources for all three extraction techniques, improving answer selection accuracy by as much as 32.35% for location questions and 72% for proper-name questions. Of the combination strategies tested, logistic regression models produced the best results for both location and proper-name questions.

1. Introduction

Question-answering (QA) systems aim to find precise answers to natural language questions within large document collections. Typical QA systems combine information retrieval with extraction techniques to identify a set of likely candidates, from which a final answer is selected. This selection process can be very challenging, as it often entails identifying the correct answer(s) amongst many incorrect ones.

To address this problem, several answer selection approaches have been developed that make use of external semantic resources. One of the most common relies on precompiled lists or ontologies such as WordNet, CYC, and gazetteers for answer validation or type checking. Answer candidates not found within the portion of the resource's hierarchy corresponding to the expected answer type of the question are either removed or discounted (Chu-Carroll et al., 2003; Nyberg et al., 2003; Xu et al., 2002). Moldovan et al. (2003) extract axioms from WordNet that are then used with a logic prover to verify the relationship between an answer candidate and the question. The Web has also been used in a data-driven approach in which answers are reranked according to search engine results produced by queries containing the answer candidate and question keywords (Magnini et al., 2002).

Although each of these approaches uses one or more semantic resources to independently support an answer, few have considered the potential benefits of combining resources together as evidence. Recently, Schlobach et al. (2004) combined geographical databases with WordNet in a type checker for location questions. However, in their experiments the combination actually hurt performance, a result they attribute to the increased semantic ambiguity that accompanied broader coverage of location names. We view this result as evidence that the combination method may matter as much as the choice of resources.

Our work addresses this issue further, evaluating three different strategies to combine semantic information from the Web, WordNet and gazetteers: an unweighted sum, a weighted linear combination, and logistic regression. We use the combined semantic evidence to rescale answer confidence scores for three extraction techniques of varying quality, focusing on questions whose answers represent locations (geopolitical and geographic entities) or proper-names (including person names and organization names). Experiments with questions from the TREC QA evaluations (Voorhees, 2003) demonstrate that the combination can be more effective than the individual resources, improving answer selection accuracy for all three extraction techniques by as much as 32.35% on location questions and 72% on proper-name questions.

The remainder of this paper is organized as follows. Section 2 briefly describes the QA system used as a testbed. Section 3 describes the semantic resources and their application in answer selection. Section 4 explains the three combination strategies evaluated. Section 5 summarizes our experimental setup and results. Finally, Section 6 presents the conclusions and briefly discusses future work.

2. JAVELIN QA System

JAVELIN is an open-domain QA system designed to support multi-strategy QA using modular components under the control of a planner (Nyberg et al., 2003). JAVELIN includes several different answer extraction techniques, each implemented as an interchangeable module. Among them, three answer extractors were used in our experiments:

- FST - an extractor based on finite state transducers that incorporate a set extraction patterns (both manually created and generalizations induced from examples), and learn their precision with respect to each answer type
- SVM - an extractor that uses Support Vector Machines to discriminate between correct and incorrect answers based on local semantic and syntactic context
- PROX - an extractor that selects candidates using a non-linear distance heuristic computed between the keywords and a candidate answer

Given a document set and question analysis produced by other components in JAVELIN, these extractors identify all possible answer candidates from the retrieved documents. Each candidate is assigned a normalized confidence score representing the likelihood it is correct.

An answer selection component then chooses the most probable answer(s) to the question from the extracted candidates. Input to the answer selector consists of a set of candidate answers produced by a single extraction technique, their corresponding confidence scores, and the expected answer type inferred for the question (i.e., a *location, proper-name,* or one of several predefined subtypes). The answer selection process starts with answer normalization to cluster redundant or complementary answer candidates. For example, "April 14th, 1912" and "14 April 1912" are normalized to "1912-04-14" and then clustered together. Each answer cluster is assigned a new confidence score representing the likelihood at least one candidate in the cluster is correct, given their individual confidence scores and the assumption that each candidate in the cluster is independent and equally weighted. After clustering, semantic resources are utilized to boost the confidence scores assigned to correct answers and lower the confidence scores assigned to incorrect answers. A more detailed description of how these resources are used is provided in the next section.

3. Semantic Resources

Our answer selection process incorporates three types of semantic resources: gazetteers, WordNet and the Web.

3.1. Gazetteers

Electronic gazetteers provide geographic information, such as a country's population, language, cities, continent and capital. As previously shown by Lita et al. (2004), gazetteers such as CIA World Factbook can answer specific types of TREC questions with high precision, but have limited coverage.

We used three gazetteer resources in our answer selection: the Tipster Gazetteer, the CIA World Factbook, and information about the US states provided by *www.50states.com*. These resources are used to assign an answer validity score between -1 and 1 to each candidate, following the algorithm in Figure 1. Effectively, a score of 0 means the gazetteers do not contribute to the answer selection process for that candidate.

1) If the answer candidate directly matches the gazetteer answer for the question, its gazetteer score is **1.0**. (e.g. Given the question "*What continent is Togo on?*", the candidate "*Africa*" receives a score of 1.0.)

2) If the answer candidate occurs in the gazetteer within the subcategory of the expected answer type, its score is **0.5**. (e.g., Given the question "*Which city in China has the largest number of foreign financial companies?*", the candidates "*Shanghai*" and "Boston" receive a score of 0.5 because they are both cities.)

3) If the answer candidate is not the correct semantic type, its score is **-1**. (e.g., Given the question "*Which city in China has the largest number of foreign financial companies?*", the candidate "*Taiwan*" receives a score of -1 because it is not a city.)

4) Otherwise, the score is **0.0**.

Figure 1. Algorithm to generate a score from gazetteers

3.2. WordNet

The WordNet lexical database includes English nouns, verbs, adjectives and adverbs organized in synonym sets, called synsets (Fellbaum, 1998). It has been used extensively for multiple QA tasks, including reasoning about answer correctness (Moldovan et al., 2003). Our answer selection process uses WordNet in a manner analogous to gazetteers: to produce an answer validity score between -1 and 1. This score is computed for each candidate using the algorithm in Figure 2. As with the gazetteer score, a score of 0 means that WordNet does not contribute to the answer selection process for a candidate.

1) If the answer candidate directly matches WordNet, its WordNet score is **1.0**. (e.g. Given the question "*What is the capital of Uruguay?*", the candidate "*Montevideo*" receives a score of 1.0.)

2) If the answer candidate's hypernyms include a subcategory of the expected answer type, its score is **0.5**. (e.g., Given the question "*Who wrote the book 'Song of Solomon'?*", the candidate "*Mark Twain*" receives a score of 0.5 because its hypernyms include "*writer*".)

3) If the answer candidate is not the correct semantic type, this candidate receives a score of **-1**. (e.g., Given the question "*What state is Niagara Falls located in?*", the candidate "*Toronto*" gets a score of -1 because it is not a state.)

4) Otherwise, the score is **0.0**.

Figure 2. Algorithm to generate a score from WordNet

3.3. World Wide Web

The Web has also been used for many different QA tasks, as a direct source of answers (Dumais et al., 2002), and to validate answer candidates based on the number of hits and text snippets (Magnini et al., 2002). Following Magnini et al. (2002), our answer selection process uses the Web to generate a numeric score for each candidate. A query consisting of an answer candidate and question keywords is sent to the Google search engine. The top 10 text snippets returned by Google are then analyzed using the algorithm in Figure 3 to calculate a Web score.

For each snippet *s*:
  Initialize the snippet co-occurrence score: $cs(s) = 1$
  For each question keyword *k* in *s*:
    1. Compute distance *d*, the minimum number of words between *k* and the answer candidate, excluding stopwords and other keywords
    2. Update the snippet co-occurrence score:
$$cs(s) = cs(s) \times 2^{(1+d)^{-1}}$$
  Add the snippet score to the web score
Normalize the web score by dividing by a constant *C*

Figure 3. Algorithm to generate a score from the Web

4. Resource Combination

We considered three different strategies for combining the semantic resource scores: an unweighted sum

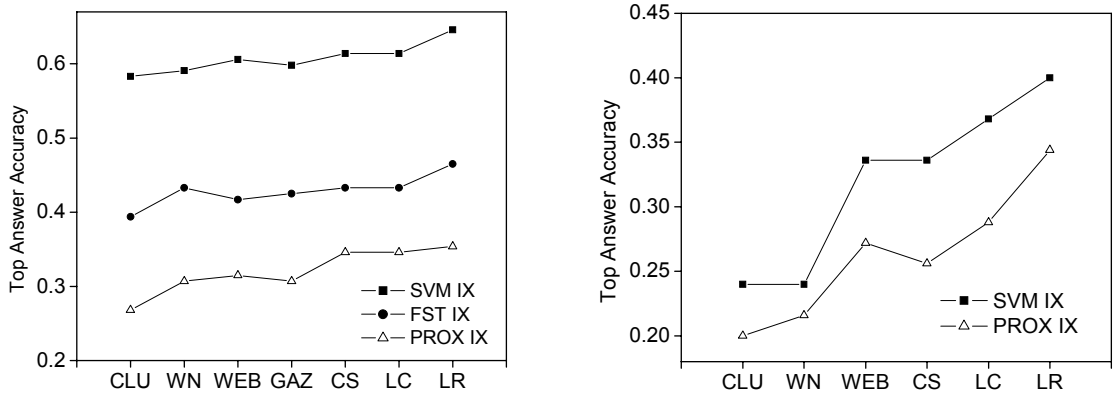

Figure 4. Top answer accuracy for location (left) and proper-name (right) questions (CLU: answer clustering baseline, GAZ: gazetteers, WN: WordNet, WEB: Web, CS: CombSUM, LC: linear combination, LR: logistic regression).

algorithm called CombSUM, a weighted linear combination and logistic regression. These methods have been previously applied to merge the ranked lists of documents returned from multiple search engines.

4.1. CombSUM

In the document retrieval domain, CombSUM (Fox and Shaw, 1994) has been used to rerank a document using the sum of the relevance scores that each search engine assigns to it. To adapt this to answer selection in the QA domain, we simply assign each candidate a score equal to the sum of scores from the answer cluster (s_{ac}), gazetteers (s_{gz}), WordNet (s_{wn}), and the Web (s_{web}):

$$s(a) = s_{ac} + s_{gz} + s_{wn} + s_{web} \quad (1)$$

4.2. Weighted Linear Combination

Weighted linear combinations have been used in metasearch by multiplying each relevance score by the weight assigned to the source that produced it (Vogt and Cottrell, 1999). We compute an analogous value for each answer candidate using the source's training set accuracy as its weight:

$$s(a) = w_{ac}s_{ac} + w_{gz}s_{gz} + w_{wn}s_{wn} + w_{web}s_{web} \quad (2)$$

4.3. Logistic Regression

Logistic regression is a statistical regression technique used to predict the probability of binary variables from a vector of discrete or continuous variables. It has been successfully employed to merge multilingual documents (Savoy and Berger, 2004). We adapt this approach to estimate the probability that an answer candidate is correct given scores from the answer cluster and three semantic resources:

$$s(a) = \frac{1}{1+\exp(-(w_0 + w_1 s_{ac} + w_2 s_{gz} + w_3 s_{wn} + w_4 s_{web}))} \quad (3)$$

The weights were directly estimated from the training set.

5. Experiments and Results

A total of 721 questions from the TREC8-12 QA evaluations (393 *location* and 328 *proper-name* questions) served as a dataset. Two-thirds of the questions were used for training, with the rest reserved for subsequent tests.

5.1. Experimental Setup

Clustering of complementary answers in the candidate set was used as a baseline process for answer selection. Additionally, to assess how well our resource combination strategies performed in comparison with selection methods using individual semantic resources, we implemented answer selectors that combine the baseline clustering with a single semantic resource by adding the cluster and resource scores together.

To better understand how the performance of the answer selection strategy varies for different extraction techniques, each approach was tested using three different extraction components provided by JAVELIN: the FST, PROX and SVM extractors. However, the FST extractor was excluded because it doesn't extract proper names.

Performance was measured as average accuracy: the number of correct top answers divided by the total number of questions.

5.2. Results and Analysis

Figure 4 compares the average accuracies for answer selection using the baseline, individual resources, and our three resource combination strategies. CLU is the average accuracy of the baseline approach. GAZ, WN and WEB represent the performance when adding gazetteers, WordNet and the Web individually to support answer selection. CS, LC and LR represent the average accuracies when merging the scores with the CombSUM, weighted linear combination and logistic regression methods, respectively. As can be seen, each resource tended to improve answer selection performance, with the Web providing significant gains for *proper-name* questions.

For *location* questions, the CS, LC and LR models improved answer selection performance an average of 14.94%, 14.94%, and 20.39%, respectively, over the baseline. The combination approaches also compared favorably with answer selection using a single resource

	LOCATION				PROPER-NAME			
Extractor	**Extractor Coverage**	**CS**	**LC**	**LR**	**Extractor Coverage**	**CS**	**LC**	**LR**
ML	0.836	0.709	0.709	0.745	0.652	0.375	0.411	0.446
PROX	0.706	0.647	0.647	0.662	0.518	0.291	0.327	0.391
FST	0.720	0.671	0.671	0.720	-	-	-	-
Average	0.754	0.676	0.676	0.709	0.585	0.333	0.369	0.419

Table 1. Answer selection coverage using combined semantic resources compared with extractor answer coverage.

for all three extractors. The biggest improvement was found with candidates produced by the PROX extractor using the LR combination, which provided a performance increase of 32.35% over the baseline, and a 12.5% improvement over selection using the Web alone.

On *proper-name* questions, the CS, LC and LR combinations improved answer selection performance an average of 34%, 48.67%, and 69.33%, respectively, over the baseline. However, on this question type, only the LC and LR strategies outperformed answer selection using a single resource for both extractors tested. Once again, selection with the PROX extractor candidates benefited most, with the LR combination producing a 72% improvement over the baseline and 26.47% over the Web.

5.3. Coverage of Answer Selection

As the performance of our approach is limited by the input quality, we also computed upper bounds for the *answer coverage* of the extractors (Table 1), which is the fraction of questions for which the extractor candidate set included at least one correct answer.

The difference between answer coverage and precision of the three combination methods represents the improvement possible without improving the extractors themselves.

6. CONCLUSION

This paper described an answer selection approach that combines semantic information from the Web, WordNet and gazetteers, and compares the performance of different strategies for combining these resources. Our empirical results on TREC questions show that the combination of semantic resources improves answer selection accuracy for all three extraction techniques tested, boosting performance by as much as 32.35% on location questions and 72% on proper-name questions.

Although these experiments focused specifically on questions related to locations and proper-names, we expect our approach to combine multiple semantic resources will provide comparable gains for other classes of questions such as dates and numeric-expressions. Currently we are adding encyclopedia and dictionaries. The integration of these new resources and extension to multilingual QA systems are the subject of ongoing work.

7. ACKNOWLEDGEMENTS

This work was supported in part by the Advanced Research and Development Activity (ARDA) under AQUAINT contract MDA904-01-C-0988.

8. REFERENCES

Chu-Carroll, J., Prager, J., Welty, C., Czuka, K., and Ferrucci, D. (2003). A Multi-Strategy and Multi-Source Approach to Question Answering. In *Proceedings of Text REtrieval Conference*.

Dumais, S., Banko, M., Brill, E., Lin, J., and Ng, A. (2002). Web question answering: Is more always better? In *Proceedings of SIGIR*.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. The MIT Press.

Fox, E. A. and Shaw, J. A. (1994). Combination of multiple searches. In *Proceedings of Text REtrieval Conference*.

Lita, L.V., Hunt, W., and Nyberg, E. (2004). Resource Analysis for Question Answering. In *Proceedings of ACL.*

Magnini, B., Negri, M., Pervete, R., and Tanev, H. (2002). Comparing statistical and content-based techniques for answer validation on the web. In *Proceedings of the VIII Convegno AI*IA*.

Moldovan, D., Clark, D., Harabagiu, S., and Maiorano, S. (2003). Cogex: A logic prover for question answering. In *Proceedings of HLT-NAACL*.

Nyberg, E., Mitamura, T., Callan, J., Carbonell, J., Frederking, R., Collins-Thompson, K., Hiyakumoto, L. , Huang, Y., Huttenhower, C., Judy, S., Ko, J., Kupsc, A., Lita, L.V., Pedro, V., Svoboda, D., and Van Durme, B. (2003). A multi-strategy approach with dynamic planning. In *Proceedings of Text REtrieval Conference*.

Savoy, J. and Berger, P-Y. (2004). Selection and Merging Strategies for Multilingual Information Retrieval. In *Proceedings of Cross Language Evaluation Forum*.

Schlobach, S., Olsthoorn, M., and de Rijke, M. (2004). Type Checking in Open-Domain Question Answering. In *Proceedings of European Conference on Artificial Intelligence.*

Vogt, C. C. and Cottrell, G. W. (1999). Fusion Via a Linear Combination of Scores. *Information Retrieval*. 1(3): 151-173. Oct. 1999.

Voorhees, E. (2003). Overview of the TREC 2003 question answering track. In *Proceedings of Text REtrieval Conference*.

Xu, J., Licuanan, A., May, J., Miller, S., and Weischedel, R. M. (2002). TREC 2002 QA at BBN: Answer Selection and Confidence Estimation. In *Proceedings of Text REtrieval Conference*.