Answer Credibility: A Language Modeling Approach to Answer Validation

Protima Banerjee Hyoil Han
College of Information Science and Technology
Drexel University
Philadelphia, PA 19104
pb66@drexel.edu, hyoil.han@acm.org

Abstract

Answer Validation is a topic of significant interest within the Question Answering community. In this paper, we propose the use of language modeling methodologies for Answer Validation, using corpus-based methods that do not require the use of external sources. Specifically, we propose a model for Answer Credibility which quantifies the reliability of a source document that contains a candidate answer and the Question's Context Model.

1 Introduction

In recent years, Answer Validation has become a topic of significant interest within the Question Answering community. In the general case, one can describe Answer Validation as the process that decides whether a Question is correctly answered by an Answer according to a given segment of supporting Text. Magnini et al. (Magnini, 2002) presents an approach to Answer Validation that uses redundant information sources on the Web; they propose that the number of Web documents in which the question and the answer co-occurred can serve as an indicator of answer validity. Other recent approaches to Answer Validation Exercise in the Cross-Language Evaluation Forum (CLEF) (Peters, 2008) make use of textual entailment methodologies for the purposes of Answer Validation.

In this paper, we propose the use of language modeling methodologies for Answer Validation, using corpus-based methods that do not require the use of external sources. Specifically, we propose the development of an Answer Credibility score which quantifies reliability of a source document that contains a candidate answer with respect to the Question's Context Model. Unlike many textual entailment methods, our methodology has the advantage of being applicable to question types for which hypothesis generation is not easily accomplished.

The remainder of this paper describes our work in progress, including our model for Answer Credibility, our experiments and results to date, and future work.

2 Answer Credibility

Credibility has been extensively studied in the field of information science (Metzger, 2002). Credibility in the computational sciences has been characterized as being synonymous with believability, and has been broken down into the dimensions of trustworthiness and expertise.

Our mathematical model of Answer Credibility attempts to quantify the reliability of a source using the semantic Question Context. The semantic Question Context is built using the Aspect-Based Relevance Language Model that was presented in (Banerjee, 2008) and (Banerjee, 2009). This model builds upon the Relevance Based Language Model (Lavrenko, 2001) and Probabilisitic Latent Semantic Analysis (PLSA) (Hofmann, 1999) to provide a mechanism for relating sense disambiguated Concept Terms (CT) to a query by their likelihood of relevance.

The Aspect-Based Relevance Language Model assumes that for every question there exists an un-

Proceedings of NAACL HLT 2009: Short Papers, pages 157–160,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

derlying relevance model R, which is assigned probabilities *P(z|R)* where *z* is a latent aspect of the information need, as defined by PLSA. Thus, we can obtain a distribution of aspects according to their likelihood of relevancy to the user's information need. By considering terms from the aspects that have the highest likelihood of relevance (eg. highest *P(z|R)* values), we can build a distribution that models a semantic Question Context.

We define Answer Credibility to be a similarity measure between the Question Context (QC) and the source document from which the answer was derived. We consider the Question Context to be a document, which has a corresponding document language model. We then use the well-known Kullback-Leibler divergence method (Lafferty, 2001) to compute the similarity between the Question Context document model and the document model for a document containing a candidate answer:

$$AnswerCredibility = \sum_{w \in CT} P\ (w \mid QC) \log \frac{P\ (w \mid QC)}{P(w \mid d)}$$

Here, *P(w|QC)* is the language model of the Question Context, *P(w|d)* is the language model o the document containing the candidate answer. To insert this model into the Answer Validation process, we propose an interpolation technique that modulates the answer score during the process using Answer Credibility.

3 Experimental Setup

The experimental methodology we used is shown as a block diagram in Figure 1. To validate our approach, we used the set of all factoid questions from the Text Retrieval Conference (TREC) 2006 Question Answering Track (Voorhees, 2006).

The OpenEphyra Question Answering testbed (Schlaefer, 2006) was then used as the framework for our Answer Credibility implementation. OpenEphyra uses a baseline Answer Validation mechanism which uses documents retrieved using Yahoo! search to support candidate answers found in retrieved passages. In our experiments, we constructed the Question Context according to the methodology described in (Banerjee, 2008). Our experiments used the Lemur Language Modeling toolkit (Strohman, 2005) and the Indri search engine (Ogilvie, 2001) to construct the Question Context and document language models.

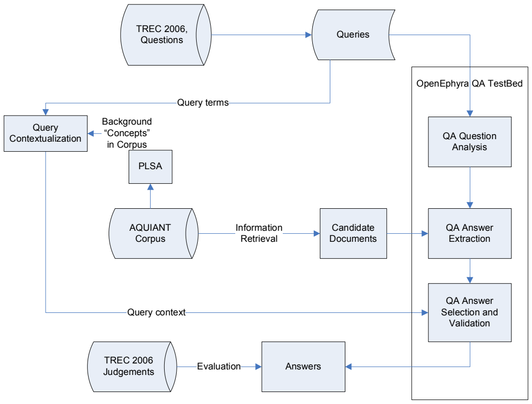

Figure 1: Experiment Methodology

We then inserted an Answer Credibility filter into the OpenEphyra processing pipeline which modulates the OpenEphyra answer score according to the following formula:

$$score' = (1 - \lambda) * score + \lambda * AnswerCredibility$$

Here score is the original OpenEphyra answer score and score' is the modulated answer score. In this model, λ is an interpolation constant which we set using the average of the *P(z|R)* values for those aspects that are included in the Question Context.

For the purposes of evaluating the effectiveness of our theoretical model, we use the accuracy and Mean Reciprocal Rank (MRR) metrics (Voorhees, 2005).

4 Results

We compare the results of the baseline OpenEphyra Answer Validation approach against the results after our Answer Credibility processing has been included as a part of the OpenEphyra pipeline. Our results are presented in Table 1 and Table 2.

To facilitate interpretation of our results, we subdivided the set of factoid questions into categories by their question words, following the example of (Murdock, 2006). The light grey shaded cells in both tables indicate categories for which improvements were observed after our Answer Credibility model was applied. The dark grey shaded cells in both tables indicate categories for which no change was observed. The paired Wilcoxon signed rank test was used to measure significance in improvements for MRR; the shaded cells in Table 2 indicate results for which the results were significant ($p<0.05$). Due to the binary results for accuracy at the question level (eg. a question is either correct or incorrect), the Wilcoxon test was found to be inappropriate for measuring statistical significance in accuracy.

Table 1: Average MRR of Baseline vs. Baseline Including Answer Credibility

Question Category	Question Count	Baseline MRR	Baseline + Answer Credibility MRR
How	20	0.33	0.28
how many	58	0.21	0.16
how much	6	0.08	0.02
in what	47	0.68	0.60
What	114	0.30	0.33
what is	28	0.26	0.26
When	29	0.30	0.19
Where	23	0.37	0.37
where is	6	0.40	0.40
Which	17	0.38	0.26
Who	17	0.51	0.63
who is	14	0.60	0.74
who was	24	0.43	0.55

Table 2: Average Accuracy of Baseline vs. Baseline Including Answer Credibility

Question Category	Question Count	Baseline Accuracy	Baseline + Answer Credibility Accuracy
How	20	0.25	0.20
how many	58	0.12	0.07
how much	6	0.00	0.00
in what	47	0.64	0.55
What	114	0.23	0.28
what is	28	0.18	0.18
When	29	0.21	0.10
Where	23	0.30	0.30
where is	6	0.33	0.33
Which	17	0.29	0.18
Who	17	0.47	0.59
who is	14	0.57	0.71
who was	24	0.38	0.50

Our results show the following:

- A 5% improvement in accuracy over the baseline for “what”-type questions.
- An overall improvement of 13% in accuracy for “who”-type questions, which include the “who,” “who is” and “who was” categories
- A 9% improvements in MRR for “what” type questions
- An overall improvement of 25% in MRR for “who”-type questions, which include the “who,” “who is” and “who was” categories
- Overall, 7 out of 13 categories (58%) performed at the same level or better than the baseline

5 Discussion

In this section, we examine some examples of questions that showed improvement to better understand and interpret our results.

First, we examine a “who” type question which was not correctly answered by the baseline system, but which was correctly answered after including Answer Credibility. For the question “Who is the host of the Daily Show?” the baseline system correctly determined the answer was “Jon Stewart” but incorrectly identified the document that this answer was derived from. For this question, the Question Context included the terms “stewart,” “comedy,” “television,” “news,” and “kilborn.” (Craig Kilborn was the host of Daily Show until 1999, which makes his name a logical candidate for inclusion in the Question Context since the AQUAINT corpus spans 1996-2000). In this case, the correct document that the answer was derived from was actually ranked third in the list. The Answer Credibility filter was able to correctly increase the answer score of that document so that it was ranked as the most reliable source for the answer and chosen as the correct final result.

Next, we consider a case where the correct answer was ranked at a lower position in the answer list in the baseline results and correctly raised higher, though not to the top rank, after the application of our Answer Credibility filter. For the question “What position did Janet Reno assume in 1993?” the correct answer (“attorney general”) was ranked 5 in the list in the baseline results. However, in this case the score associated with the answer was lower than the top-ranked answer by an order of magnitude. The Question Context for this question included the terms “miami,” “elian,” “gonzales,” “boy,” “attorney” and “justice.” After the application of our Answer Credibility filter, the score and rank of the correct answer did increase (which contributed to an increase in MRR), but the increase was not enough to overshoot the original top-ranked answer.

Categories for which the Answer Credibility had negative effect included “how much” and “how many” questions. For these question types, the correct answer or correct document was frequently not present in the answer list. In this case, the Answer Credibility filter had no opportunity to increase the rank of correct answers or correct documents in the answer list. This same reasoning also limits our applicability to questions that require a date in response.

Finally, it is important to note here that the very nature of news data makes our methodology applicable to some categories of questions more than others. Since our methodology relies on the ability to derive semantic relationships via a statistical examination of text, it performs best on those questions for which some amount of supporting information is available.

6 Conclusions and Future Work

In conclusion, we have presented a work in progress that uses statistical language modeling methods to create a novel measure called Answer Credibility for the purpose of Answer Validation. Our results show performance increases in both accuracy and MRR for “what” and “who” type questions when Answer Credibility is included as a part of the Answer Validation process. Our goals for the future include further development of the Answer Credibility model to include not only terms from a Question Context, but terms that can be deduced to be in an Answer Context.

References

Banerjee, P., Han, H. 2008. "Incorporation of Corpus-Specific Semantic Information into Question Answering Context," CIKM 2008 - Ontologies and Information Systems for the Semantic Web Workshop, Napa Valley, CA.

Banerjee, P., Han, H 2009. "Modeling Semantic Question Context for Question Answering," *To appear in FLAIRS 2009*.

Hofmann, T. 1999. "Probabilistic latent semantic indexing," Proceedings of the 22nd Annual International SIGIR.

Lafferty, J. and Zhai, C. 2001. "Document language models, query models, and risk minimization for information retrieval," in Proceedings of the 24th Annual International ACM SIGIR, New Orleans, Louisiana: pp. 111-119.

Lavrenko, V. and Croft, W. B. 2001. "Relevance based language models," Proceedings of the 24th annual international ACM SIGIR, pp. 120-127.

Magnini, B., Negri, M., Prevete, R. Tanev, H. 2002. "Is It the Right Answer? Exploiting Web Redundancy for Answer Validation," in Association for Computational Lingustistics (ACL) 2002, Philadelphia, PA, pp. 425-432.

Metzger, M. 2007. "Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research," Journal of the American Society of Information Science and Technology (JASIST), vol. 58, p. 2078.

Murdock, V. 2006. Exploring Sentence Retrieval. VDM Verlag.

Ogilvie, P. and Callan, J. P. 2001. "Experiments Using the Lemur Toolkit," in Online Proceedings of the 2001 Text Retrieval Conference (TREC).

Peters, C. 2008. “What happened in CLEF 2008: Introduction to the Working Notes.” http://www.clef-campaign.org/2008/working_notes.

Schlaefer, N., Gieselmann, P., Schaaf, T., & A., W. 2006. A Pattern Learning Approach to Question Answering within the Ephyra Framework, In Proceedings of the Ninth International Conference on Text, Speech and Dialogue (TSD).

Strohman, T., Metzler, D., Turtle, H., and Croft, W. B. 2005. "Indri: A language model-based search engine for complex queries," International Conference on Intelligence Analysis McLean, VA.

Voorhees, E. M. and Harman, D. K. 2005. TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing): The MIT Press.

Voorhees, E. M. 2006. "Overview of the TREC 2006 Question Answering Track," in Online Proceedings of 2006 Text Retrieval Conference (TREC).