NAIST at the CLEF 2013 QA4MRE Pilot Task

Philip Arthur, Graham Neubig, Sakriani Sakti,
Tomoki Toda, and Satoshi Nakamura

Nara Institute of Science and Technology,
8916-5, Takayama-cho, Ikoma-shi, Nara 630-0192 JAPAN
{philip-a,neubig,ssakti,tomoki,s-nakamura}@is.naist.jp

Abstract. This paper describes the Nara Institute of Science and Technology's system for the entrance exam pilot task of CLEF 2013 QA4MRE. The core of the system is a similar to the system for the main task of CLEF 2013 QA4MRE. We use minimum error rate training (MERT) to train the weights of the model and also propose a novel method for MERT with the addition of a threshold that defines the certainty with which we must answer questions. The system received a score of 22% c@1.

Keywords: discriminative learning, minimum error rate training, linear feature model, question answering, machine reading, inter-sentence features

1 Introduction

While years of research on Question Answering (QA) have greatly improved the state-of-the-art, we know that this problem is far from solved. Question answering campaigns such as CLEF [7] and TREC [4] have resulted in a large number of distinct proposals about how to build robust systems that can provide correct answers in the general domain.

One of the features of QA that is widely accepted is that "two heads are better than one" [3]. By combining different information sources, we gain the ability to cover up the disadvantages of one system with another information source, which results in more effective QA on the whole. One way to combine multiple systems is to weight each system's score with some value and choose the maximum value from a linear combination [8]. Another important aspect of QA is that it is sometimes good not to answer the question [6]. Many systems currently return No Answer (NoA) if they are not confident because a wrong answer is often worse than no answer [2]. Our system for the CLEF Question Answering for Machine Reading Evaluation (QA4MRE) this year is based on these two principles, devising a number of features that provide useful information to identify the correct answer, and combining them together with a learning framework that is also able to learn when not to answer questions.

We introduce several new features that span multiple sentences in addition to more traditional features such as cosine similarity. These features are combined

in a framework that learns both *how* and *when* to answer questions in a single weighted linear model. In particular, we find *how* to answer questions by learning appropriate weigths for each feature, with final score of an answer being their weighted linear combination. We define *when* not to answer by not returning candidates for which scores are less than a set threshold t from other candidates. Finally, we propose a method to intelligently weight the features and threshold using minimum error rate training.

As results, our system received a score of 22% on the Entrance Exam pilot task according to the c@1 evaluation metric.

2 System Description

The core of our system relies on a log linear scoring model that is fully described in [1]. Before we score the answer, our system use several basic preprocessing methods such as *tokenization*, *named entity recognition*, *anaphora resolution*, *lowercasing*, *stop word deletion*, and *stemming* to process the text before hand. Our model is based on bags-of-n-grams vector space model that takes the union from higher and lower order of n-grams. We weight the features of the model based on tf-idf term weighting and also use this criterion to measure the similarity between vectors. Next, we score each candidate answer for each question with features that are based on traditional intra-sentence features and some proposed inter-sentence features multiplied by their trained weight. The candidate answer with the best score that exceeds a defined threshold will be chosen as system's answer, or the system will return no answer if the score is below the threshold.

To train the model, we used a new training method based on minimum error rate training (MERT, [5]) for question answering. The training method takes a set of questions, candidate answers and their particular features score and train it accordingly. Furthermore, we define a threshold t, and the system will only answer if the highest scoring candidate exceeds the second candidate by more than the threshold. This MERT plus its threshold is a new training method called TMERT that is described in [1].

3 Evaluation

3.1 Evaluation Measures

To evaluate the system's performance, we used "c@1," which is used for the QA4MRE evaluation metric [6],

$$c@1 = \frac{1}{n}\left(\mathrm{ca} + \frac{\mathrm{ca} * \mathrm{ca}}{n}\right) \tag{1}$$

where "ca", "na", and "n" correspond to "correct answer", "no answer", and number of questions.

3.2 Experimental Setup

The system used only the English test set document and did not reference the background collection. The "Entrance Exams" task aims to evaluate systems under the same conditions under which humans are evaluated for entering university. This new task consists of 9 test sets containing 10 questions with 4 candidate answers each. To train the parameters of the model, we use both test set documents from past CLEF 2011 and 2012 QA4MRE campaigns [7] and receive a c@1 score of 42% on the training data [1].

3.3 Entrance Exam Task Results

Reading ID	Correct Answer	Wrong Answer	No Answer	c@1 Score
1	1	4	0	20%
2	0	5	1	0%
3	2	3	0	40%
4	1	4	0	20%
5	0	5	0	0%
6	1	4	0	20%
7	1	4	0	20%
8	3	2	0	60%
9	1	4	0	20%

Table 1. Result of Participation in Pilot Task

First we show the result in Table 1. For the Japanese entrance exam pilot task, we only submitted 1 run which achieved 10 correct answers, 35 wrong answers and 1 unanswered question resulting a c@1 score of 22.22%, which is lower than a random baseline (25%). We take a look at the results carefully and spot some mistakes the system made. This sample question is taken from the r_id=1 and q_id=1.

> When I was a child, our dining room had two kinds of chairs - two large ones with arm rests and four small ones without. The larger ones stood at the ends of the table, the small ones on the sides. Mom and Dad sat in the big chairs, except when one of us was away; then Mom would sit in one of the smaller chairs. I always remained in the same place, at my father's right. He always sat at the end, at the "head" of the table.
>
> Question: Where did the author's mother sit when one of her children was away?
>
> 1. She didn't change her chair.
> 2. She moved her own chair next to Dad's.
> 3. She moved to an empty chair on the side.
> 4. She sat opposite to Dad.

The system return 4 as its answer because the keyword "sat" occurs in it. Normally, to answer this question, we need deep comprehension of the reading document. While all of the sentences are constructively describing the scene, we

know that the answer must be 2 or 3 because candidate answers number 1 and 4 are contradicting the evidence. Further, because there is not enough evidence to answer candidate answer number 2, the most probable answer is candidate answer number 3. However, our system is incapable of constructing this kind of proof. Currently, our features are only based on statistical analysis of keywords that occured in the passage, question, and candidate answer so this type of logical inferences can't be solved. This problem shows that our system needs further refinement in terms of processing, inference, and more knowledge to answer these type of questions.

4 Conclusion

As part of our participation in QA4MRE Pilot Task@CLEF 2013, we have developed QA-system that is simple but lacks in terms of answering more complex question types found in the pilot task. For future work, we believe that it is necessary to use external knowledge such as background knowledge so the system can provide further analysis in classifying questions and determining certain type of strategies to answer the questions. Further work will be focussed on integrating external knowledge derived from sources such as Wikipedia and the background collections by adding more features.

References

1. Arthur, P., Neubig, G., Sakti, S., Toda, T., Nakamura, S.: Inter Sentence Features and Thresholded Minimum Error Rate Training: NAIST at CLEF 2013 QAMRE. In: CLEF (2013)
2. Brill, E., Dumais, S., Banko, M.: An Analysis of the AskMSR Question-Answering System. In: In Proceedings of EMNLP. pp. 257–264 (2002)
3. Chu-Carroll, J., Czuba, K., Prager, J., Ittycheriah, A.: In Question Answering, Two Heads Are Better Than One. In: In HLT-NAACL. pp. 24–31 (2003)
4. Dang, H.T., Lin, J.J., Kelly, D.: Overview of the TREC 2006 Question Answering Track 99. In: TREC (2006)
5. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proceedings of ACL (2003)
6. Peñas, A., Rodrigo, A.: A Simple Measure to Assess Non-response. In: Proceedings of ACL. pp. 1415–1424. Association for Computational Linguistics, Portland, Oregon, USA (June 2011)
7. Peñas, A., Hovy, E.H., Forner, P., Rodrigo, Á., Sutcliffe, R.F.E., Forascu, C., Sporleder, C.: Overview of QA4MRE at CLEF 2011: Question Answering for Machine Reading Evaluation. In: Petras, V., Forner, P., Clough, P.D. (eds.) CLEF (Notebook Papers/Labs/Workshop) (2011)
8. Tufis, D.: Natural Language Question Answering in Open Domains. The Computer Science Journal of Moldova 19(2), 146–164 (2011)