A Maximum Entropy Framework that Integrates Word Dependencies and Grammatical Relations for Reading Comprehension

Kui Xu[1,2] **and Helen Meng**[1]
[1]Human-Computer Communications Laboratory
Dept. of Systems Engineering and
Engineering Management
The Chinese University of Hong Kong
Hong Kong SAR, China
{kxu, hmmeng}@se.cuhk.edu.hk

Fuliang Weng[2]
[2]Research and Technology Center
Robert Bosch Corp.
Palo Alto, CA 94304, USA
Fuliang.weng@rtc.bosch.com

Abstract

Automatic reading comprehension (RC) systems can analyze a given passage and generate/extract answers in response to questions about the passage. The RC passages are often constrained in their lengths and the target answer sentence usually occurs very few times. In order to generate/extract a specific precise answer, this paper proposes the integration of two types of "deep" linguistic features, namely word dependencies and grammatical relations, in a maximum entropy (ME) framework to handle the RC task. The proposed approach achieves 44.7% and 73.2% HumSent accuracy on the Remedia and ChungHwa corpora respectively. This result is competitive with other results reported thus far.

1 Introduction

Automatic reading comprehension (RC) systems can analyze a given passage and generate/extract answers in response to questions about the passage. The RC passages are often constrained in their lengths and the target answer sentence usually occurs only once (or very few times). This differentiates the RC task from other tasks such as open-domain question answering (QA) in the Text Retrieval Conference (Light et al., 2001). In order to generate/extract a specific precise answer to a given question from a short passage, "deep" linguistic analysis of sentences in a passage is needed.

Previous efforts in RC often use the bag-of-words (BOW) approach as the baseline, which is further augmented with techniques such as shallow syntactic analysis, the use of named entities (NE) and pronoun references. For example, Hirschman et al. (1999) have augmented the BOW approach with stemming, NE recognition, NE filtering, semantic class identification and pronoun resolution to achieve 36% HumSent[1] accuracy in the Remedia test set. Based on these technologies, Riloff and Thelen (2000) improved the HumSent accuracy to 40% by applying a set of heuristic rules that assign handcrafted weights to matching words and NE. Charniak et al. (2000) used additional strategies for different question types to achieve 41%. An example strategy for *why* questions is that if the first word of the matching sentence is "this," "that," "these" or "those," the system should select the previous sentence as an answer. Light et al. (2001) also introduced an approach to estimate the performance upper bound of the BOW approach. When we apply the same approach to the Remedia test set, we obtained the upper bound of 48.3% HumSent accuracy. The state-of-art performance reached 42% with answer patterns derived from web (Du et al., 2005).

This paper investigates the possibility of enhancing RC performance by applying "deep" linguistic analysis for every sentence in the passage. We refer to the use of two types of features, namely word dependencies and grammatical relations, that

[1]If the system's answer sentence is identical to the corresponding human marked answer sentence, the question scores one point. Otherwise, the question scores no point. HumSent accuracy is the average score across all questions.

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 185–188,
New York, June 2006.

are integrated in a maximum entropy framework. Word dependencies refer to the headword dependencies in lexicalized syntactic parse trees, together with part-of-speech (POS) information. Grammatical relations (GR) refer to linkages such as subject, object, modifier, etc. The ME framework has shown its effectiveness in solving QA tasks (Ittycheriah et al., 1994). In comparison with previous approaches mentioned earlier, the current approach involves richer syntactic information that cover longer-distance relationships.

2 Corpora

We used the Remedia corpus (Hirschman et al., 1999) and ChungHwa corpus (Xu and Meng, 2005) in our experiments. The Remedia corpus contains 55 training stories and 60 testing stories (about 20K words). Each story contains 20 sentences on average and is accompanied by five types of questions: *who, what, when, where* and *why*. The ChungHwa corpus contains 50 training stories and 50 test stories (about 18K words). Each story contains 9 sentences and is accompanied by four questions on average. Both the Remedia and ChungHwa corpora contain the annotation of NE, anaphor referents and answer sentences.

3 The Maximum Entropy Framework

Suppose a story S contains n sentences, $C_0, \ldots, C_n$, the objective of an RC system can be described as:

$$A = \arg\max_{C_i \in S} P(C_i \textit{ answers } Q|Q). \quad (1)$$

Let "x" be the question (Q) and "y" be the answer sentence C_i that answers "x". Equation 1 can be computed by the ME method (Zhou et al., 2003):

$$p(y|x) = \frac{1}{Z(x)} \exp^{\sum_j \lambda_j f_j(x,y)}, \quad (2)$$

where $Z(x) = \sum_y \exp^{\sum_j \lambda_j f_j(x,y)}$ is a normalization factor, $f_j(x, y)$ is the indicator function for feature f_j; f_j occurs in the context x, λ_j is the weight of f_j. For a given question Q, the C_i with the highest probability is selected. If multiple sentences have the maximum probability, the one that occurs the earliest in the passage is returned. We used the selective gain computation (SGC) algorithm (Zhou et al., 2003) to select features and estimate parameters for its fast performance.

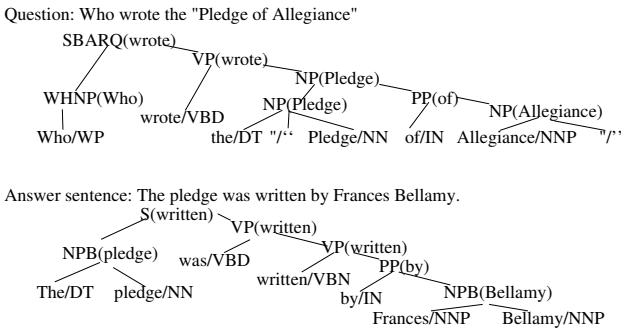

Figure 1. The lexicalized syntactic parse trees of a question and a candidate answer sentence.

4 Features Used in the "Deep" Linguistic Analysis

A feature in the ME approach typically has binary values: $f_j(x, y) = 1$ if the feature j occurs; otherwise $f_j(x, y) = 0$. This section describes two types of "deep" linguistic features to be integrated in the ME framework in two subsections.

4.1 POS Tags of Matching Words and Dependencies

Consider the following question Q and sentence C,

Q: *Who wrote the "Pledge of Allegiance"*

C: *The pledge was written by Frances Bellamy*.

The set of words and POS tags[2] are:

Q: {*write/VB, pledge/NN, allegiance/NNP*}

C: {*write/VB, pledge/NN, by/IN, Frances/NNP, Bellamy/NNP*}.

Two matching words between Q and C (i.e. "*write*" and "*pledge*") activate two POS tag features:

$f_{VB}(x, y)$=1 and $f_{NN}(x, y)$=1.

We extracted dependencies from lexicalized syntactic parse trees, which can be obtained according to the head-rules in (Collins, 1999) (e.g. see Figure 1). In a lexicalized syntactic parse tree, a dependency can be defined as:

$$< hc \rightarrow hp > \text{ or } < hr \rightarrow TOP >,$$

where hc is the headword of the child node, hp is the headword of the parent node ($hc \neq hp$), hr is the headword of the root node. Sample

[2]We used the MXPOST toolkit downloaded from *ftp://ftp.cis.upenn.edu/pub/adwait/jmx/* to generate POS tags. Stop words including *who, what, when, where, why, be, the, a, an*, and *of* are removed in all questions and story sentences. All plural noun POS tags are replaced by their single forms (e.g. NNS→NN); all verb POS tags are replaced by their base forms (e.g. VBN→VB) due to stemming.

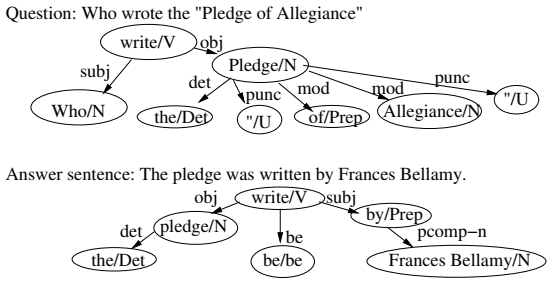

Figure 2. The dependency trees produced by MINIPAR for a question and a candidate answer sentence.

dependencies in C (see Figure 1) are:

<write→TOP> and <pledge→write>.

The dependency features are represented by the combined POS tags of the modifiers and headwords of (identical) matching dependencies[3]. A matching dependency between Q and C, *<pledge→write>* activates a dependency feature: $f_{NN-VB}(x, y)$=1. In total, we obtained 169 and 180 word dependency features from the Remedia and ChungHwa training sets respectively.

4.2 Matching Grammatical Relationships (GR)

We extracted grammatical relationships from the dependency trees produced by MINIPAR (Lin, 1998), which covers 79% of the dependency relationships in the SUSANNE corpus with 89% precision[4]. IN a MINIPAR dependency relationship:

$(word1$ *CATE1:RELATION:CATE2* $word2)$,

CATE1 and CATE2 represent such grammatical categories as nouns, verbs, adjectives, etc.; RELATION represents the grammatical relationships such as subject, objects, modifiers, etc.[5] Figure 2 shows dependency trees of Q and C produced by MINIPAR. Sample grammatical relationships in C are *pledge N:det:Det the,* and *write V:by-subj:Prep by.* GR features are extracted from identical matching relationships between questions and candidate sentences. The only identical matching relationship between Q and C, "*write V:obj:N pledge*" activates a grammatical relationship feature: $f_{obj}(x, y)$=1. In total, we extracted 44 and 45 GR features from the Remedia and ChungHwa training sets respectively.

[3]We extracted dependencies from parse trees generated by Collins' parser (Collins, 1999).

[4]MINIPAR outputs GR directly, while Collins' parser gives better result for dependencies.

[5]Refer to the *readme* file of MINIPAR downloaded from *http://www.cs.ualberta.ca/ lindek/minipar.htm*

5 Experimental Results

We selected the features used in Quarc (Riloff and Thelen, 2000) to establish the reference performance level. In our experiments, the 24 rules in Quarc are transferred[6] to ME features:

"If contains(Q,{*start, begin*}) and contains(S,{*start, begin, since, year*}) Then Score(S)+=20" $\rightarrow$ $f_j(x, y) = 1$ (0< j <25) if Q is a *when* question that contains *"start"* or *"begin"* and C contains *"start," "begin," "since"* or *"year"*; $f_j(x, y) = 0$ otherwise.

In addition to the Quarc features, we resolved five pronouns (*he, him, his, she* and *her*) in the stories based on the annotation in the corpora. The result of using Quarc features in the ME framework is 38.3% HumSent accuracy on the Remedia test set. This is lower than the result (40%) obtained by our re-implementation of Quarc that uses handcrafted scores. A possible explanation is that handcrafted scores are more reliable than ME, since humans can generalize the score even for sparse data. Therefore, we refined our reference performance level by combining the ME models (MEM) and handcrafted models (HCM). Suppose the score of a question-answer pair is $score(Q, C_i)$, the conditional probability that C_i answers Q in HCM is:

$HCM(Q, C_i) = P(C_i \textit{ answers } Q|Q) = \frac{score(Q,C_i)}{\Sigma_{j \leq n} score(Q,C_j)}$.

We combined the probabilities from MEM and HCM in the following manner:

$score'(Q, C_i) = \alpha MEM(Q, C_i) + (1 - \alpha)HCM(Q, C_i)$.

To obtain the optimal α, we partitioned the training set into four bins. The ME models are trained on three different bins; the optimal α is determined on the other bins. By trying different bins combinations and different α such that $0 < \alpha < 1$ with interval 0.1, we obtained the average optimal $\alpha = 0.15$ and 0.9 from the Remedia and ChungHwa training sets respectively[7]. Our baseline used the combined ME models and handcrafted models to achieve 40.3% and 70.6% HumSent accuracy in the Remedia and ChungHwa test sets respectively.

We set up our experiments such that the linguistic features are applied incrementally - (i) First , we use only POS tags of matching words among questions

[6]The features in (Charniak et al., 2000) and (Du et al., 2005) could have been included similarly if they were available.

[7]HCM are tuned by hand on Remedia, thus a bigger weight, 0.85 represents their reliability. For ChungHwa, a weight, 0.1 means that HCM are less reliable.

and candidate answer sentences. (ii) Then we add POS tags of the matching dependencies. (iii) We apply only GR features from MINIPAR. (iv) All features are used. These four feature sets are denoted as "+wp," "+wp+dp," "+mini" and "+wp+dp+mini" respectively. The results are shown in Figure 3 for the Remedia and ChungHwa test sets.

With the significance level 0.05, the pairwise t-test (for every question) to the statistical significance of the improvements shows that the p-value is 0.009 and 0.025 for the Remedia and ChungHwa test sets respectively. The "deep" syntactic features significantly improve the performance over the baseline system on the Remedia and ChungHwa test sets[8].

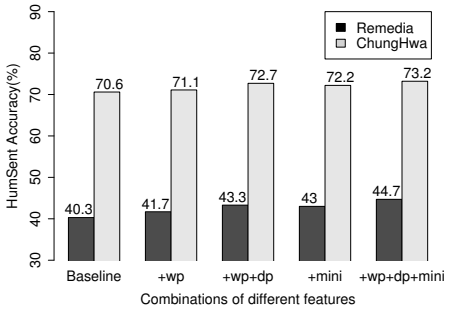

Figure 3. Baseline and proposed feature results on the Remedia and ChungHwa test sets.

6 Conclusions

This paper proposes the integration of two types of "deep" linguistic features, namely word dependencies and grammatical relations, in a ME framework to handle the RC task. Our system leverages linguistic information such as POS, word dependencies and grammatical relationships in order to extract the appropriate answer sentence for a given question from all available sentences in the passage. Our system achieves 44.7% and 73.2% HumSent accuracy on the Remedia and ChungHwa test sets respectively. This shows a statistically significant improvement over the reference performance levels, 40.3% and 70.6% on the same test sets.

Acknowledgements

This work is done during the first author's internship at RTC Bosch Corp. The work is also affiliated with the CUHK Shun Hing Institute of Advanced Engineering and partially supported by CUHK4237/03E from RGC of HKSAR Government.

[8] Our previous work about developing the ChungHwa corpus (Xu and Meng, 2005) shows that most errors can only be solved by reasoning with domain ontologies and world knowledge.

References

Dekang Lin. 1998. *Dependency-based Evaluation of MINIPAR*. Workshop on the Evaluation of Parsing Systems 1998.

Ellen Riloff and Michael Thelen. 2000. *A Rule-based Question Answering System for Reading Comprehension Test*. ANLP/NAACL-2000 Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems.

Eugene Charniak, Yasemin Altun, Rofrigo D. Braz, Benjamin Garrett, Margaret Kosmala, Tomer Moscovich, Lixin Pang, Changhee Pyo, Ye Sun, Wei Wy, Zhongfa Yang, Shawn Zeller, and Lisa Zorn. 2000. *Reading Comprehension Programs In a Statistical-Language-Processing Class*. ANLP-NAACL 2000 Workshop: Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems.

Kui Xu and Helen Meng. 2005. *Design and Development of a Bilingual Reading Comprehension Corpus*. International Journal of Computational Linguistics & Chinese Language Processing, Vol. 10, No. 2.

Lynette Hirschman, Marc Light, Eric Breck, and John D. Burger. 1999. *Deep Read: A Reading Comprehension System*. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics.

Marc Light, Gideon S. Mann, Ellen Riloff, and Eric Breck. 2001. *Analyses for Elucidating Current Question Answering Technology*. Journal of Natural Language Engineering, No. 4 Vol. 7.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania.

Abraham Ittycheriah, Martin Franz, Wei-Jing Zhu and Adwait Ratnaparkhi. 2001. *Question Answering Using Maximum-Entropy Components*. Proceedings of NAACL 2001.

Yaqian Zhou, Fuliang Weng, Lide Wu, Hauke Schmidt. 2003. *A Fast Algorithm for Feature Selection in Conditional Maximum Entropy Modeling*. Proceedings of EMNLP 2003.

Yongping Du, Helen Meng, Xuanjing Huang, Lide Wu. 2005. *The Use of Metadata, Web-derived Answer Patterns and Passage Context to Improve Reading Comprehension Performance*. Proceedings of HLT/EMNLP 2005.