Combining Term-Based and Event-Based Matching for Question Answering

Michael Wiegand
Spoken Language Systems
Saarland University
D-66125
Saarbrücken,Germany
Michael.Wiegand@lsv.uni-saarland.de

Jochen L. Leidner
Language Technology Group
School of Informatics
University of Edinburgh
Edinburgh EH8 9LW, UK
Jochen.Leidner@ed.ac.uk

Dietrich Klakow
Spoken Language Systems
Saarland University
D-66125 Saarbrücken,
Germany
Dietrich.Klakow@lsv.uni-saarland.de

ABSTRACT

In question answering, two main kinds of matching methods for finding answer sentences for a question are term-based approaches—which are simple, efficient, effective, and yield high recall—and event-based approaches that take syntactic and semantic information into account. The latter often sacrifice recall for increased precision, but actually capture the meaning of the events denoted by the textual units of a passage or sentence. We propose a robust, data-driven method that learns the mapping between questions and answers using logistic regression and show that combining term-based and event-based approaches significantly outperforms the individual methods.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic processing*; I.2.6 [**Artificial Intelligence**]: Learning—*Induction*

General Terms

Measurement, Design

Keywords

Question Answering, Machine Learning

1. INTRODUCTION

An obvious method in question answering (QA) for assessing the relevance of candidate answer sentences is by considering their underlying *event structures*, i.e. syntactic and semantic information. Unlike simple *term-based matching*, these approaches can be more precise since they reflect more accurately the meaning of these textual units. However, even with state-of-the-art NLP software, such linguistic processing is error-prone. Moreover, there are relevant answer sentences of questions which cannot be matched by event structures. In some of these cases, term-based approaches still work. We propose a robust, data-driven method that learns the mapping between questions and answers using logistic regression and show that combining term-based and event-based approaches significantly outperforms the individual methods.

2. RELATED WORK

One popular term-based approach is presented in [2] where *span-size ratio (SSR)* and *matching-term ratio (MTR)* are interpolated to a combined measure. Though this method already fails at simple paraphrases, such as *kill* and *murder*, one still achieves fairly reasonable results on QA data sets, such as TREC QA[1]. Most QA systems making use of event-based modelling consult lexical resources. [5], for example, proposes a successful method for expanding queries using WordNet[2]. Additionally, grammatical relations are important for sentence relevance detection or answer extraction, as [3] point out. With these information, sentences, such as question-answer pair (1)-(2), can be properly matched. The two events *assassinated* and *killed* can be identified as synonyms and their arguments properly matched despite the active-passive alternation due to the usage of grammatical functions.

(1) [Who]$_{SUBJ}$ assassinated [President Kennedy]$_{OBJ}$?
(2) [John F. Kennedy]$_{OBJ}$ was killed by [Lee H. Oswald]$_{SUBJ}$.

Most event-based QA systems suffer from lacking any simpler backing-off processing which should support matching of sentences when event-based processing fails. The causes for failure are diverse. The underlying event structures may be too complicated to match or the event processing erroneous. The following question-answer pair exemplifies a situation in which event structure cannot be used for matching since the full-verb *write*, which is the *event denoting expression (EDE)* of the question, is not reflected by any word in the answer sentence.

(3) Which famous book did Rachel Carson [write]$_{EDE}$?
(4) Rachel Carson's most famous book "Silent Spring" caused the banning of DDT.

The reflection of EDEs in answer sentences is essential since they are the linguistic units from which event structures are bootstrapped. It should be obvious that, in the current example, term-based matching works in order to establish the relevance of the answer sentence. In the next section, we show an event-based model that even supports matching of event structures which are bootstrapped by EDEs of different parts-of-speech. Thus a verb-based event structure can be mapped onto noun-based event structure[3], as in question-answer pair (1)-(2):

(1) [Who]$_{SUBJ}$ [won]$_{EDE}$ [the Super Bowl]$_{OBJ}$?

(2) The [Rams']$_{SUBJ}$ 23-16 [victory]$_{EDE}$ of [the Super Bowl]$_{OBJ}$ initiated the NFL's new epoch.

[1] `http://trec.nist.gov/data/qa.html`
[2] `http://wordnet.princeton.edu/`
[3] These nouns are either *nominalizations*, i.e. nouns which have been derived from full-verbs, e.g. *explanation* from *to explain*, or nominalization-like expressions, i.e. nouns which behave like nominalizations but are not lexically derived from a verb, such as *victory* or *home*.

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

3. METHOD

The algorithm we propose is based on three different kinds of mappings of type

$$map : qap \rightarrow [0; 1] \quad (1)$$

where qap represents a tuple comprising question and candidate answer sentence (1.0 means optimal match). We call the overall quality of matching a question and a candidate answer sentence $qaMap$. We are looking for the best matching formally denoted by:

$$\hat{qap} := \arg\max_{qap} (qaMap\,(qap)) \quad (2)$$

This measure combines matching the underlying event structures ($esMap$) and occurring terms ($tMap$):

$$qaMap\,(qap) := \alpha \cdot (esMap\,(qap)) + (1 - \alpha) \cdot (tMap\,(qap)) \quad (3)$$

Event-based matching is done by a linear binary classifier. We choose *logistic regression*:

$$esMap\,(qap) := \sigma\left(\vec{w}^T \vec{f} + b\right) \quad (4)$$

where σ is the logistic function (*S-curve*), $\vec{f}$ is a feature vector, $\vec{w}$ the corresponding weights and b is a bias. The features in $\vec{f}$ are similarity functions comprising information associated with event structure from various linguistic levels. The most prominent features[4] are:

- grammatical functions ($SUBJ, OBJ$, etc.);
- subcategorization information in order to distinguish complements from adjuncts;
- textual proximity of arguments to event descriptions[5];
- semantic comparison via WordNet.

Each similarity function is either binary or continuous, i.e. it is defined over $[0; 1]$. In order to be able to match event structures across different parts of speech, we use NOMLEX-Plus [1]. This enables us to match EDEs, such as *win* and *victory* in question-answer pair (1)-(2), and assign grammatical functions to arguments of EDEs being nouns. We assess term-based matching with the help of SSR and MTR:

$$tMap(qap) := SSR(qap)^{\alpha'} \cdot MTR(qap)^{\beta'} \quad (5)$$

The optimal weights are taken from [2], i.e. $\alpha' = 0.125$ and $\beta' = 1.0$. The other unknown parameters are estimated on the TREC QA 2005 data[6]. The parameters for $esMap$, i.e. $\vec{w}$ and b, are learned on a manually labelled subset of the corpus. The only unknown parameter for $qaMap$, i.e. α, is determined via iterative optimization using a separate subset of the same TREC collection. Note that this optimization does not require a separate training set.

4. EVALUATION

Figure 1 shows the plot of the performance on the complete parameter space of α for Equation 3. It illustrates that $tMap$ ($\alpha = 0.0$) has a high recall but a lower precision whereas $esMap$ ($\alpha = 1.0$) has a high precision but a lower recall. The fact that the optimal configuration is $\alpha = 0.4$ is plausible since the weighting trades recall against precision in the best possible way. This iterative optimization illustrates that the combination is successful since it outperforms the best individual method, i.e. $tMap$, with an observed absolute F-score increase from 0.68 to 0.74 by including the information offered by our *event-based* approach based on logistic regression.

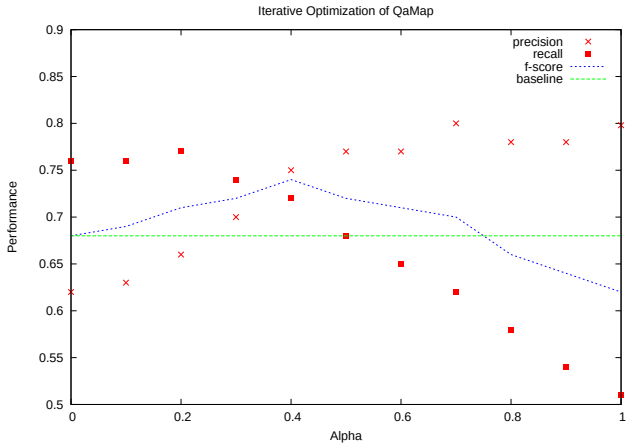

Figure 1: Optimum of *qaMap* at $\alpha = 0.4$ shows the best possible trade-off between precision and recall.

5. CONCLUSION

We proposed a data-driven algorithm for sentence relevance detection for QA which used event-based metrics to enhance term-based matching (i.e. *span-size ratio* and *matching-term ratio*). The resulting matching method achieved an increase in F-Score from 0.68 to 0.74.

6. ACKNOWLEDGMENTS

This research was partially funded by the BMBF project SmartWeb under Federal Ministry of Education and Research grant 01IM D01M.

7. REFERENCES

[1] A. Meyers, R. Reeves, C. Macleod, R. Szekeley, V. Zielinska, and B. Young. Proteus Project. Technical Report 04-005, New York University, New York, NY, USA, 2004.

[2] C. Monz. Minimal Span Weighting Retrieval for Question Answering. In R. Gaizauskas, M. Greenwood, and M. Hepple, editors, *Proceedings of the SIGIR Workshop on Information Retrieval for Question Answering*, pages 23–30, Sheffield, UK, 2004.

[3] B. van Durme, Y. Huang, A. Kupść, and E. Nyberg. Towards Light Semantic Processing for Question Answering. In *HLT/NAACL Workshop on Text Meaning*, Morristown, NJ, USA, 2003.

[4] M. Wiegand. Event-Based Modelling in Quesion Answering. Master's thesis, Saarland University, Saarbrücken,Germany, March 2007. `http://www.coli.uni-saarland.de/~miwieg/pub/dipl-thesis.pdf`.

[5] H. Yang, T. Chua, S. Wang, and C. Koh. Structured Use of External Knowledge for Event-Based Open Domain Question Answering. In *SIGIR*, pages 33–40, Toronto, Canada, 2003. ACM Press.

[4] A full description of the features is given in [4].

[5] This is a simple alternative event-based representation to grammatical functions which is still more informative than term-based representations.

[6] We can only use this corpus since the amount of event questions was too small in previous TREC collections.