CoMiC: Adapting a Short Answer Assessment System for Answer Selection

Björn Rudzewitz **Ramon Ziai**
Sonderforschungsbereich 833
Eberhard Karls Universität Tübingen
Nauklerstraße 35
72070 Tübingen, Germany
{brzdwtz,rziai}@sfs.uni-tuebingen.de

Abstract

Open forum threads exhibit a great variability in the quality and quantity of the answers they attract, making it difficult to manually moderate and separate relevant from irrelevant content. The goal of SemEval 2015 Task 3 (Subtask A, English) is to build systems that automatically distinguish between relevant and irrelevant content in forum threads.

We extend a short answer assessment system to build relations between forum questions and answers with respect to similarity, question type, and answer content. The features are used in a sequence classifier to account for the conversation character of threads. The performance of this approach is modest in comparison to the other task participants and also to the performance the system usually reaches in short answer assessment. However, the new features implemented for this task are a first step in developing more fine-grained question-answer features and identifying relevant answers.

1 Introduction

In this paper, we discuss the adaptation of our Short Answer Assessment (SAA) system CoMiC (Meurers et al., 2011) to Task 3, Subtask A (English) of SemEval 2015, *Answer Selection in Community Question Answering*. The aim in the task was to distinguish helpful from unhelpful answers in a community forum given a question.

We enter the QA landscape from the perspective of evaluating student answers to reading comprehension questions with respect to whether they contain the targeted content. In such settings, one generally has a reference answer to which a candidate answer can be compared, making alignment-based systems a natural solution. This is not the case for QA, where a system has to select or rank candidate answers with regard to a question posed. However, the present task is still interesting to us because it shares a central characteristic with SAA: one needs to identify the relevant part of an answer, given a question. In theoretical linguistics, that relevant part is usually called *focus* (cf., e.g., Krifka (2007)), and several research groups have made efforts to annotate it in corpus data (Hajičová and Sgall, 2001; Ritz et al., 2008; Calhoun et al., 2010; Ziai and Meurers, 2014).

Automatic approaches to identifying focus have however yet to be proposed, so for the current task, we adapted and used our SAA system to align candidate answers with the forum question, identifying whether and how question material was picked up, which in turn should indicate whether answers are on-topic. We then used a number of features to characterize the unaligned answer material, from POS classes to temporal expressions. We also encoded which question words were present in the question in the hope that the resulting classifier would pick up connections between individual question words and the different answer features in an approximation to identifying the focus of the answer.

The paper is organized as follows: Section 2 briefly discusses the data of the task before section 3 presents the details of our system architecture and the features we used. Section 4 then shows the results of our efforts and a short error analysis, and finally section 5 concludes and discusses directions for further efforts.

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 247–251,
Denver, Colorado, June 4-5, 2015.

2 Data

The English dataset used in the task is a collection of web-crawled forum[1] texts where each item consists of a question and responses to the question. Each response has one of the six labels *Good*, *Bad*, *Potential*, *Dialogue*, *non-English*, or *Other*, describing its potential for answering the corresponding question. The correct label for every response had to be predicted by the systems at test time. The dataset is not balanced since it contains more *Good* labelled answers than answers with another label. The language used in the questions and responses exhibits strong deviations from standard English. For a detailed description, refer to (Màrquez et al., 2015).

3 System Details

In this section, we describe the CoMiC system and its extensions for Task 3 of SemEval 2015. We begin by going briefly over the baseline system and its features and continue by describing in detail the new features introduced for this task.

The baseline CoMiC system is an alignment-based short answer assessment system. Alignments between a student and a target answer are computed on different linguistic levels. The quantities of alignments of a certain quality are used as features and given to a classifier that predicts a binary correctness label for the student answer. A detailed description can be found in (Meurers et al., 2011).

For this task, we adapt the system by making it establish alignments between forum questions and the corresponding answers. Thus it is used primarily as a text similarity system extended by features to differentiate between given and new material.

3.1 Features

The system uses the standard features from the CoMiC system and a range of new features. Although the new features described here were used in the context of Question Answering, we are planning to explore to what extent the usage of these features will improve the CoMiC system in the context of short answer assessment. The following sections will start with an overview about the standard CoMiC features and will continue with a detailed description of the new features.

[1] http://www.qatarliving.com/forum

3.1.1 CoMiC

As mentioned in the introduction, the CoMiC system is designed to judge the contents of a short answer to a reading comprehension question based on alignment with a target answer (Meurers et al., 2011). The features it uses express the linguistic unit and nature of the successful alignments found between candidate and target answer. In the present setting, we used the standard CoMiC features to determine the degree of similarity between the candidate answer and the forum question, in order to find out whether the answer does indeed pick up on question topic material. These features are summarized in Table 1.

Feature	Description
1. Keyword Overlap	Percent of dependency heads aligned (relative to question)
2./3. Token Overlap	Percent of aligned question/candidate tokens
4./5. Chunk Overlap	Percent of aligned question/candidate chunks (as identified by OpenNLP[2])
6./7. Triple Overlap	Percent of aligned question/candidate dependency triples
8. Token Match	Percent of token alignments that were token-identical
9. Similarity Match	Percent of token alignments resolved using PMI-IR (Turney, 2001)
10. Type Match	Percent of token alignments resolved using WordNet hierarchy (Fellbaum, 1998)
11. Lemma Match	Percent of token alignments that were lemma-resolved
12. Synonym Match	Percent of token alignments sharing same WordNet synset
13. Variety of Match (0-5)	Number of kinds of token-level alignments (features 8–12)

Table 1: Standard features in the CoMiC system

3.1.2 POS-Specific Weighting

The system uses four features that measure how much of the material not given in the question belongs to a group of syntactically related categories. The idea is to weight new material by estimating a distribution of general syntactic classes over it. After

[2] http://opennlp.apache.org/

the alignment process, the distribution of groups of POS categories of non-aligned tokens is computed with respect to all non-aligned tokens. As a basis, the Penn Treebank POS tags from prior annotation are used. Four groups are distinguished which are composed in the following way:

- *nouns*: subsumes all nominal categories
- *verbs*: subsumes full verbs, auxiliaries, modals, and participles
- *adj/v*: subsumes all adjectival and adverbial categories
- *rest*: subsumes all categories not listed above

For every of the four groups, the frequency of each POS tag in this group in the non-aligned material is computed, normalized against the frequency of all POS tags in the non-aligned material, and summed up to get the overall proportion of this group in the non-aligned material. Previous experiments suggested to prefer this approach with coarse groups over an approach with more fine-grained POS classes due to its overall robustness needed in this context.

3.1.3 Question Words

In an approximation to identifying question types, we encoded the presence or absence of the *wh*-words *who*, *how*, *why*, *when*, *where*, *which*, *whom*, *whose* and *what* with a binary feature for each. We also encode the presence of modal and auxiliary verbs in the first three tokens of a sentence in order to detect questions such as "Can anyone help me?".

The idea behind these features was to enable associations between them and the features characterizing the new material in the answer.

3.1.4 Named Entity Recognition

We used the Stanford Named Entity Recognizer (Finkel et al., 2005) to detect named entities in new answer material. For each of the three standard NE classes PERSON, ORGANIZATION and LOCATION, we encode its presence or absence in a binary feature. Additionally, we encode the total number of syntactic chunks found in the answer, of which the named entities constitute a subset.

By detecting NEs, we wanted to enable the resulting classifier to pick up connections between the previously mentioned *wh*-features and the named entities.

3.1.5 Temporal Expressions

The system uses a binary feature indicating the presence or absence of one or more temporal expressions in every answer. In combination with the question word features, the system can build relations between questions asking for temporal content and the presence of temporal expressions in the answer. The system therefore makes use of an adapted version of the HeidelTime temporal tagger (Strötgen and Gertz, 2013) due to its ability to parse web content with a high accuracy. No distinction is made between different kinds of temporal expressions recognized by the HeidelTime module.

3.2 Adaptation to Social Media Language

Since the CoMiC system is designed for the assessment of short answers of language learners, several adaptations were needed in order for the system to be able to deal with the noisiness of social media language. These adaptations consist of multiple steps that will be described in this section.

The first step towards normalizing the language consists of the removal of HTML markup present in several answers. For this purpose, the CoMiC system was extended by adding an additional module that parses the raw input and recursively extracts the text content while removing any HTML markup. The jsoup module[3] was used to accomplish this task.

The second step in the normalization process is driven by the idea to exclude certain tokens from further processing if they are recognized as being of a category unlikely to contribute usefully in deeper analysis by the system, such as emoticons, e-mail addresses, hashtags, abbreviations, symbols, punctuation sequences, etc. Therefore we use an adapted version of the ark-tweet-nlp module (Gimpel et al., 2011) in the tokenization step which allows parallel tokenization and POS tagging with a tagset tailored to cover the specifics of social media language. The exclusion of noisy material is done after sentence segmentation, allowing to preserve sentences including all tokens from the text, at the same time excluding unwanted material from further analysis and alignment.

[3] `http://jsoup.org/`

3.3 Model

We trained two different models based on separate classification methods. We first experimented with memory-based learning using TiMBL (Daelemans et al., 2007), using the cosine as distance metric and $k = 5$ nearest neighbors that each instance was compared to. In order to take advantage of the fact that a forum thread is in fact a conversation and the usefulness of a given forum answer may depend on previous answers, we also employed a CRF tagger (MALLET, McCallum (2002)) to classify a sequence of forum posts instead of a single instance. We used one Markov order for the CRF. To our knowledge, this is the only model in the competition that attempted to classify answer sequences.

The CRF performed slightly better than the memory-based approach on the development set, which we attribute to its ability to take an answer's context into account. We submitted it as our primary run and the memory-based one as the contrastive run.

4 Results

Evaluation was done using two scenarios: fine-grained (*Good*, *Potential*, *Dialogue*, *Bad*) and coarse-grained (*Good*, *Potential*, *Bad*), with missing classes always collapsed into *Bad*. Table 2 shows the coarse-grained accuracies and Macro F1 scores of our system variants on development and test set for the English Subtask A. The CRF approach used in the primary system outperforms the contrastive memory-based approach on both data sets in terms of accuracy. In case of the primary system, the model seems to transfer well since the accuracy on the test set is even higher than on the development set. In case of the contrastive system, the accuracy drops when the model is applied to the test set. The table also shows the accuracy for the best-performing system, JAIST-contrastive, and the majority baseline.

These accuracies are rather modest, both in comparison to accuracy values of the CoMiC system when used for the task of short answer assessment for which the system is intended and designed, and also in comparison to other task participants.

An error analysis showed several problems that influenced the performance of the system. The noisiness of the input text on the syntactic and morphological level caused the POS tagger to assign incorrect POS tags. This led to problems for modules that make use of POS information. The noisiness is reflected also in the fact that not all lemmas are identified correctly. Another problem is that the spelling correction component struggled with certain forms and did not always find the spelling-corrected form. The main problem was that too few tokens and hardly any chunks could be aligned to the question, severely influencing the alignment-based features. The system also got mislead in cases where the person who posed the question reformulated the question for others, since the classifier failed to use the high similarity between the question and the answer as a clear indicator for an unhelpful answer.

System	Dev. Set		Test Set	
	Acc.	F1	Acc.	F1
Best system	–	–	73.76	57.29
CoMiC-prim.	54.89	28.41	54.20	30.63
CoMiC-contr.	53.37	24.36	50.56	23.36
Maj. baseline	53.19	23.15	50.46	22.36

Table 2: Coarse-grained accuracy and Macro F1 of systems on development and test set for Subtask A, English

5 Conclusion

We applied the short answer system CoMiC to the task of question selection. The standard CoMiC system was used to determine the similarity between a question and an answer. We added new features to the CoMiC system to enable the classifier to build relations between the question type and certain answer features. Extensions to the system were necessary in order to deal with the noisiness of web texts. We applied a CRF classifier that takes into account the context of answers in the forum and found a positive effect on performance. The results of the task show that our system performs rather moderately when used for this task it is not designed or intended for. However, the new features implemented for this task are a first step in developing more fine-grained question-answer features which eventually could be useful for identifying the relevant part of an answer.

Acknowledgments

We would like to thank two anonymous reviewers for their detailed and helpful comments.

References

Sasha Calhoun, Jean Carletta, Jason Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The NXT-format switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44:387–419.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch, 2007. *TiMBL: Tilburg Memory-Based Learner Reference Guide, ILK Technical Report ILK 07-03*. Induction of Linguistic Knowledge Research Group Department of Communication and Information Sciences, Tilburg University, Tilburg, The Netherlands, July 11. Version 6.0.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics.

Eva Hajičová and Petr Sgall. 2001. Topic-focus and salience. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 276–281, Toulouse, France. Association for Computational Linguistics.

Manfred Krifka. 2007. Basic notions of information structure. In Caroline Fery, Gisbert Fanselow, and Manfred Krifka, editors, *The notions of information structure*, volume 6 of *Interdisciplinary Studies on Information Structure (ISIS)*, pages 13–55. Universitätsverlag Potsdam, Potsdam.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Detmar Meurers, Ramon Ziai, Niels Ott, and Stacey Bailey. 2011. Integrating parallel analysis modules to evaluate the meaning of answers to reading comprehension questions. *IJCEELL. Special Issue on Automatic Free-text Evaluation*, 21(4):355–369.

Lluís Màrquez, James Glass, Walid Magdy, Alessandro Moschitti, Preslav Nakov, and Bilal Randeree. 2015. Semeval-2015 task 3: Answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics.

Julia Ritz, Stefanie Dipper, and Michael Götze. 2008. Annotation of information structure: An evaluation across different types of texts. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 2137–2142, Marrakech, Morocco.

Jannik Strötgen and Michael Gertz. 2013. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298.

Peter Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, pages 491–502, Freiburg, Germany.

Ramon Ziai and Detmar Meurers. 2014. Focus annotation in reading comprehension data. In *Proceedings of the 8th Linguistic Annotation Workshop (LAW VIII, 2014)*, pages 159–168, Dublin, Ireland. COLING, Association for Computational Linguistics.