CoMeT: Integrating different levels of linguistic modeling for meaning assessment

Niels Ott Ramon Ziai Michael Hahn Detmar Meurers
Sonderforschungsbereich 833
Eberhard Karls Universität Tübingen
{nott,rziai,mhahn,dm}@sfs.uni-tuebingen.de

Abstract

This paper describes the CoMeT system, our contribution to the SemEval 2013 Task 7 challenge, focusing on the task of automatically assessing student answers to factual questions. CoMeT is based on a meta-classifier that uses the outputs of the sub-systems we developed: CoMiC, CoSeC, and three shallower bag approaches. We sketch the functionality of all sub-systems and evaluate their performance against the official test set of the challenge. CoMeT obtained the best result (73.1% accuracy) for the 3-way unseen answers in Beetle among all challenge participants. We also discuss possible improvements and directions for future research.

1 Introduction

Our contribution to the SemEval 2013 Task 7 challenge (Dzikovska et al., 2013) presented here is based on our research in the A4 project[1] of the SFB 833, which is dedicated to the question how meaning can be computationally compared in realistic situations. In realistic situations, utterances are not necessarily well-formed or complete, there may be individual differences in situative and world knowledge among the speakers. This can complicate or even preclude a complete linguistic analysis, leading us to the following research question: Which linguistic representations can be used effectively and robustly for comparing the meaning of sentences and text fragments computationally?

[1] http://purl.org/dm/projects/sfb833-a4

In order to work on effective and robust processing, we base our work on reading comprehension exercises for foreign language learners, of which we are also collecting a large corpus (Ott et al., 2012). Our first system, CoMiC, is an alignment-based approach which exists in English and German variants (Meurers et al., 2011a; Meurers et al., 2011b). CoMiC uses various levels of linguistic abstraction from surface tokens to dependency parses. Further work that we are starting to tackle includes the utilization of Information Structure (Krifka, 2007) in the system.

The second approach emerging from the research project is CoSeC (Hahn and Meurers, 2011; Hahn and Meurers, 2012), a semantics-based system for meaning comparison that was developed for German from the start and was ported to operate on English for this shared task. As a novel contribution in this paper, we present CoMeT (Comparing Meaning in Tübingen), a system that employs a meta-classifier for combining the output of CoMiC and CoSeC and three shallower bag approaches.

In terms of the general context of our work, short answer assessment essentially comes in the two flavors of meaning comparison and grading, the first trying to determine whether or not two utterances convey the same meaning, the latter aimed at grading the abilities of students (cf. Ziai et al., 2012). Short answer assessment is also closely related to the field of Recognizing Textual Entailment (RTE, Dagan et al., 2009), which this year is directly reflected by the fact that SemEval 2013 Task 7 is the Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge.

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 608–616, Atlanta, Georgia, June 14-15, 2013.

Turning to the organization of this paper, section 2 introduces the three types of sub-systems and the meta-classifier. In section 3, we report on the evaluation results of each sub-system both for our development set as well as for the official test set of the shared task. We then discuss possible causes and implications of the findings we made by participating in the shared task.

2 Systems

The CoMeT system that we describe in this paper is a combination of three types of sub-systems in one meta-classifier. CoSeC and CoMiC are systems that align linguistic units in the student answer to those in the reference answer. In contrast, the bag-based approaches employ a vocabulary of words, lemmas, and Soundex hashes constructed from all of the student answers in the training data. In the meta-classifier, we tried to combine the benefits of the named sub-systems into one large system that eventually computed our submission to the SemEval 2013 Task 7 challenge.

2.1 CoMiC

CoMiC (Comparing Meaning in Context) is an alignment-based system, i.e., it operates on a mapping of linguistic units found in a student answer to those given in a reference answer. CoMiC started off as a re-implementation of the Content Assessment Module (CAM) of Bailey and Meurers (2008). It exists in two flavors: CoMiC-DE for German, described in Meurers et al. (2011b), and CoMiC-EN for English, described in Meurers et al. (2011a). Both systems are positioned in the landscape of the short answer assessment field in Ziai et al. (2012). In this paper, we refer to CoMiC-EN simply as CoMiC.

Sketched briefly, CoMiC operates in three stages:

1. *Annotation* uses various NLP modules to equip student answers and reference answers with linguistic abstractions of several types.

2. *Alignment* creates links between these linguistic abstractions from the reference answer to the student answer.

3. *Classification* uses summary statistics of these alignment links in machine learning in order to assign labels to each student answer.

Automatic *annotation* and *alignment* are implemented in the Unstructured Information Management Architecture (UIMA, Ferrucci and Lally, 2004). Our UIMA modules mainly wrap around standard NLP tools of which we provide an overview in Table 1. We used the standard statistical models which are provided with the NLP tools.

Annotation Task	NLP Component
Sentence Detection	OpenNLP[2]
Tokenization	OpenNLP
Lemmatization	morpha (Minnen et al., 2001)
Spell Checking	Edit distance (Levenshtein, 1966), SCOWL word list[3]
Part-of-speech Tagging	TreeTagger (Schmid, 1994)
Noun Phrase Chunking	OpenNLP
Synonyms and Semantic Types	WordNet (Fellbaum, 1998)
Similarity Scores	PMI-IR (Turney, 2001) on UkWaC (Baroni et al., 2009)
Dependency Relations	MaltParser (Nivre et al., 2007)
Keyword extraction	Heads from dependency parse

Table 1: NLP tools used for CoMiC and Bag Approaches

Annotation ranges from very basic linguistic units such as sentences and tokens with POS and lemmas, over NP chunks, up to full dependency parses of the input. For distributional semantic similarity via PMI-IR (Turney, 2001), a local search engine based on Lucene (Gospodnetić and Hatcher, 2005) querying the UkWaC corpus (Baroni et al., 2009) was used, since all major search engines meanwhile have shut down their APIs.

After the annotation of linguistic units has taken place, *candidate alignment links* are created within UIMA. In a simple example case, a candidate alignment link is a pair of tokens that is token identical in the student answer and in the reference answer. The same token in the student answer may also be part of a candidate alignment link that maps to another token in the reference answer that, e.g., has the same lemma, or is a possible synonym, or again is token identical. Other possible links are based on spelling-corrected tokens, semantic types, or high values of the PMI-IR similarity measure.

Words that are present in the reading comprehension question and that are also found in the student answer are excluded from alignment, resulting in a very

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

[3] http://wordlist.sourceforge.net

basic implementation of an approach to *givenness* (cf. Halliday, 1967, p. 204 and many others since).

Subsequently, a *globally optimal alignment* of linguistic units in the reference answer and student answer is determined using the Traditional Marriage Algorithm (Gale and Shapley, 1962).

At this point, processing within UIMA comes to an end with an output module that generates the files containing the features for machine learning. These features basically are summary statistics of the types of alignment links. An overview of these numeric features used is given in Table 2.

Feature	Description
1. Keyword Overlap	Percent of keywords aligned (relative to target)
2./3. Token Overlap	Percent of aligned target/learner tokens
4./5. Chunk Overlap	Percent of aligned target/learner chunks
6./7. Triple Overlap	Percent of aligned target/learner triples
8. Token Match	Percent of token alignments that were token-identical
9. Similarity Match	Percent of token alignments that were similarity-resolved
10. Type Match	Percent of token alignments that were type-resolved
11. Lemma Match	Percent of token alignments that were lemma-resolved
12. Synonym Match	Percent of token alignments that were synonym-resolved
13. Variety of Match (0-5)	Number of kinds of token-level alignments

Table 2: Features used in CoMiC's classification phase

Current versions of CoMiC use the WEKA toolkit (Hall et al., 2009), allowing us to experiment with different machine learning strategies. In general, any type of classification can be trained in this machine learning phase, a binary *correct* vs. *incorrect* decision as in the 2-way task being the simplest case. The best results with CoMiC on our held-out development set were achieved using WEKA's J48 classifier, which is an implementation of decision tree based on Quinlan (1993).

In terms of linguistic abstractions, CoMiC leaves the choice of representations used to its alignment step. However, in the final machine learning step, no concrete information about linguistic units is present any more. The machine learning component only sees alignment configurations which are independent of concrete words, phrases, or any other linguistic information. This high level of abstraction suggests that CoMiC should perform better than other approaches on unseen topics and unseen questions, since it does not rely on concrete units as, e.g., a bag-of-words approach does.

2.2 CoSeC

CoSeC (Comparing Semantics in Context) performs meaning comparison on the basis of an underspecified semantic representation robustly derived from the learner and the reference answers. The system was developed for German (Hahn and Meurers, 2012), on the basis of which we created the English CoSeC-EN for the SemEval 2013 Task 7 challenge.

Using an explicit semantic formalism in principle makes it possible to precisely represent meaning differences. It also supports a direct representation of Information Structure as a structuring of semantics representations (Krifka, 2007).

CoSeC is based on Lexical Resource Semantics (LRS, Richter and Sailer, 2004). Being an underspecified semantic formalism, LRS avoids the costly computation of all readings and provides access to the building blocks of the semantic representation, while additional constraints provide the information about their composition.

As described in Hahn and Meurers (2011), LRS representations can be derived automatically using a two-step approach based on part-of-speech tags assigned by TreeTagger (Schmid, 1994) and dependency parses by MaltParser (Nivre et al., 2007). First, the dependency structure is transformed into a completely lexicalized syntax-semantics interface representation, which abstracts away from some form variation at the surface. These representations are then mapped to LRS representations. The approach is robust in that it always results in an LRS structure, even for ill-formed sentences.

CoSeC then aligns the LRS representations of the reference answer and the student answer to each other and also to the representation of the question. The alignment approach takes into account local criteria, namely the semantic similarity of pairs of elements that are linked by the alignment, as well as global criteria measuring the extent to which the alignment

preserves structure at the levels of variables and the subterm structure of the semantic formulas.

Local similarity of semantic expressions is estimated using WordNet (Fellbaum, 1998), FrameNet (Baker et al., 1998), PMI-IR (Turney, 2001) on the UkWaC (Baroni et al., 2009) as used in CoMiC, the Minimum Edit Distance (Levenshtein, 1966), and special parameters for comparing functional elements such as quantifiers and grammatical function labels.

Based on the alignments, the system marks elements which are not linked to elements in the question or which are linked to the semantic contribution of an alternative in an alternative question as "focused". This is intended as a first approximation of the concept of *focus* in the sense of Information Structure (von Heusinger, 1999; Kruijff-Korbayová and Steedman, 2003; Krifka, 2007), an active field of research in linguistics addressing the question how the information in sentences is packaged and integrated into discourse. Focus elements are expected to be particularly relevant for determining the correctness of an answer (Meurers et al., 2011b).

Overall meaning comparison is then done based on a set of numerical scores computed from the alignments and their quality. For each of these scores, a threshold is empirically determined, over which the student answer is considered to be correct. Among the scores discussed by Hahn and Meurers (2011), *weighted*-target *focus*, consistently scored best in the development set. This score measures the percentage of terms in the semantic representation of the reference answer which are linked to elements of the student answer in relation to the number of all elements in the representation of the reference answer. Only terms that were marked as focused in the preceding step are counted. Functional elements, i.e., quantifiers, predicates representing grammatical function labels, or the lambda operator, are weighted differently from other elements.

This threshold method can only be used to perform 2-way classification. Unlike the machine learning step in CoMiC, it does not generalize to 3-way or 5-way classification.

The alignment algorithm uses several numerical parameters, such as weights for the different components measuring semantic similarities, weights for the different overall local and global criteria, and the weight of the *weighted*-target *focus* score. These parameters are optimized using Powells algorithm combined with grid-based line optimization (Press et al., 2002). To avoid overfitting, the parameters and the threshold are determined on disjoint partitions of the training set.

In terms of linguistic abstractions, meaning assessment in CoSeC is based entirely on underspecified semantic representations. Surface forms are indirectly encoded by the structure of the representation and the predicate names, which are usually derived from the lemmas. As with CoMiC, parameter optimization and the determination of the thresholds for the numerical scores do not involve concrete information about linguistic objects. Again, the high level of abstraction suggests that CoSeC should perform better than other approaches on unseen topics and unseen questions.

2.3 The Bag Approaches

Inspired by the bag-of-words concept that emerged from information retrieval (Salton and McGill, 1983), we designed a system that uses bag representations of student answers. For each student answer, there are three bags, each containing one of the following representations: words, lemmas and Soundex hashes of that answer. The question ID corresponding to the answer is added to each bag as a pseudo-word, allowing the machine learner to adjust to question-specific properties. Based on the bag representations, the approach compares a given student answer to a model trained on all other known student answers. On the one hand, this method ignores the presence of reference answers (although they could be added to the training set as additional correct answers), on the other hand it makes use of information not taken into account by alignment-based systems such as CoMiC or CoSeC.

Concerning pre-processing, the linguistic analyses such as tokenization and lemmatization are identical to those of CoMiC, since the bag generator technically is just another output module of the UIMA-based pipeline used there. No stop-word list is used. The bags are fed into a support vector-based machine learner. We used WEKA's Sequential Minimal Optimization (SMO, Platt, 1998) implementation with the radial basis function (RBF) kernel, since it yielded good results on our development set and since it supports output of the estimated probabilities

for each class. The optimal gamma parameter and complexity constant were estimated via 10-fold grid search.

In terms of abstractions, all bag-based approaches simply disregard word order and in case of binary bags even word frequency. Still, a bit of the relation between words is essentially encoded in their morphology. This piece of information is discarded in the bags of lemmas, eventually, e.g., putting words like "bulb" and "bulbs" in the same vector slot. Further away from the surface are the Soundex hashes, a phonetic representation of English words patented by Russell (1918). The well-known algorithm transforms similar-sounding English words into the same representation of characters and numbers, thereby ironing out many spelling mistakes and common confusion cases of homophones such as "there" vs. "their". The MorphAdorner[4] implementation we used returns empty Soundex hashes for input tokens that do not start with a letter of the alphabet. However, we found in our experiments, that the presence of these empty hashes in the bags has a positive impact on performance. This is most likely due to the fact that it discriminates answers containing punctuation (not a letter of the alphabet) from those which do not.

Since the bag approaches use Soundex as phonetic equivalence classes, but no semantic equivalence classes, they should perform best on the unseen answers data in which most lexical material from the test set is likely to already be present in the training set.

2.4 CoMeT: A Meta-Classifier

As described in the previous sections, our subsystems perform short answer evaluation on different representations and at different levels of abstraction. The bag approaches are very surface-oriented, whereas CoSeC uses a semantic formalism to compare answers to each other. We expected each system to show its strengths in different test scenarios, so a way was needed to combine the predictions of different systems into the final result.

CoMeT (Comparing Meaning in Tübingen) is a meta-classifier which builds on the predictions of our individual systems (feature stacking, see Wolpert, 1992). The rationale is that if systems are complementary, their combination will perform better (or at least as good) than any individual system on its own. The design is as follows:

Each system produces predictions on the training set, using 10-fold cross-validation, and on the test set. In addition to the predicted class, each system was also made to output probabilities for each possible class (cf., e.g., Tetreault et al., 2012a). The class probabilities were then used as features in the meta classifier to train a model for the test data. In addition to the probabilities, we also used the question ID and module ID in the meta-classifier, in the hope that they would allow differentiation between scenarios. For example, an unseen question ID means that we are not testing on unseen answers and thus predictions from systems with more abstraction from the surface may be preferred.

The class probabilities come from different sources, depending on the system. In the case of CoMiC, they are extracted directly from the decision trees. For the bag approaches, we used WEKA's option to fit logistic models to the SVM output after classification in order to estimate probabilities. Finally, the CoSeC probabilities are derived directly from its final score. As mentioned in section 2.2, CoSeC only does binary classification, so those probabilities are used in the meta-classifier for all tasks.

Based on the results on our internal development set (see section 3.1), we chose different system combinations for different scenarios. For unseen topics and unseen questions, we used only CoMiC in combination with CoSeC, since the inclusion of the bag approaches had a negative impact on results. For unseen answers, we additionally included the bag models. All meta-classification was done using WEKA's Logistic Regression implementation. The results are discussed in section 3.

3 Evaluation

In this section, we present the results for each of the sub-systems, both on the custom-made split of the training data we used in our development, as well as on the official test data of the SemEval 2013 Task 7 challenge. Subsequently, we discuss possible causes for issues raised by our evaluation results.

[4] http://morphadorner.northwestern.edu

3.1 Development Set

In order to be as close as possible to the final test setting, we replicated the official test scenarios on the training set, resulting in a *train*/*dev*/*test* split for each of the corpora. For Beetle, we held out all answers to two random questions for each module to form the unseen questions scenario, and five random answers from each remaining question to form the unseen answers scenario. For SciEntsBank, we held out module *LF* for *dev* and module *VB* for *test* to form the unseen topics scenario, because they have an average number of questions (11). The *LF* module turned out to be far more skewed towards incorrect answers (76.8%) than the training set on average (57.5%). While this skewedness needs to be taken into account for the interpretation of the development results, it did not have a negative effect on our final test results. Furthermore, analogous to Beetle, we held out all answers to one random question for each remaining module for unseen-questions, and two random answers from each remaining question for unseen answers.

The *dev* set was used for tuning and design decisions concerning which individual systems to combine in the stacked classifier, while we envisaged the *test* set to be used as a final checkpoint before submission.

The accuracy results for all sub-systems on the development set are reported in detail in Table 3. The majority baseline reflects the accuracy a system would achieve by always labelling any student answer as "incorrect", hence it is equivalent to the percentage of incorrect answers in the data. The lexical baseline is the performance of the system provided by the challenge organizers.

	Beetle		SciEntsBank		
System	d-uA	d-uQ	d-uA	d-uQ	d-uT
Maj. Baseline	57.14%	59.28%	54.30%	60.70%	76.84%
Lex. Baseline	75.43%	71.10%	63.44%	66.05%	59.54%
CoMiC	76.57%	71.52%	67.20%	70.23%	64.63%
Bag of Words	85.14%	62.03%	80.65%	54.65%	73.79%
$\sim$ of Lemmas	85.71%	58.02%	80.11%	52.33%	74.55%
$\sim$ of Soundex	86.86%	60.76%	81.18%	53.95%	72.77%
CoSeC	76.00%	74.89%	64.52%	73.49%	68.96%
CoMeT	88.00%	75.95%	81.18%	66.74%	68.45%

Table 3: Development set: accuracy for 2-way task (uA: unseen answers, uQ: unseen questions, uT: unseen topics)

The systems presented in section 2 performed as expected: The Bag-of-Soundex system achieved its best scores on the unseen answers where overlap of vocabulary was most likely, outperforming CoMiC and CoSeC with accuracy values as high as 86.86%. For Beetle unseen answers, the meta-classifier operated as expected and improved the overall results to 88.86%. For SciEntsBank unseen answers, it remained stable at 81.18%.

As expected, CoMiC and CoSeC with their alignment not depending on vocabulary outperformed the bag approaches in the other scenarios, in which the question or even the domain were not known during training. However, both alignment-based systems failed on SciEntsBank's unseen topics in comparison to the rather high majority baseline.

3.2 Official Test Set

For our submission to the SemEval 2013 Task 7 challenge, we trained our sub-systems on the entire official training set. The overall performance of the CoMeT system on all sub-tasks is shown in Table 4.

		Beetle		SciEntsBank		
		uA	uQ	uA	uQ	uT
Lexical	2-way	79.7%	74.0%	66.1%	67.4%	67.6%
Overlap	3-way	59.5%	51.2%	55.6%	54.0%	57.7%
Baseline	5-way	51.9%	48.0%	43.7%	41.3%	41.5%
Best	2-way	84.5%	74.1%	77.6%	74.5%	71.1%
System	3-way	73.1%	59.6%	72.0%	66.3%	63.7%
	5-way	71.5%	62.1%	64.3%	53.2%	51.2%
CoMeT	2-way	83.8%	70.2%	77.4%	60.3%	67.6%
	3-way	**73.1**%	51.8%	71.3%	54.6%	57.9%
	5-way	68.8%	48.8%	60.0%	43.7%	42.1%

Table 4: Official test set: overall accuracy of CoMeT (uA: unseen answers, uQ: unseen questions, uT: unseen topics)

While CoMeT won the Beetle 3-way task in unseen answers, our main focus is on the 2-way task. The results for the 2-way task of our sub-systems on the official test set are shown in Table 5.

The first row of the table reports the results of the winning system of the challenge; the two baselines are computed as before. In general, the accuracy values of CoMeT exhibit a drop of around 5% from our development set to the official test set. The meta-classifier was unable to benefit from the different sub-systems except for the unseen answers in SciEntsBank that slightly outperformed the best bag approach.

System	Beetle		SciEntsBank		
	uA	uQ	uA	uQ	uT
Best	84.50%	74.10%	77.60%	74.50%	71.10%
Maj. Baseline	59.91%	58.00%	56.85%	58.94%	57.98%
Lex. Baseline	79.70%	74.00%	66.10%	67.40%	67.60%
CoMiC	76.08%	70.57%	67.96%	66.30%	67.97%
Bag of Words	83.14%	67.52%	75.93%	57.84%	59.84%
$\sim$ of Lemmas	83.60%	67.16%	76.67%	58.25%	58.81%
$\sim$ of Soundex	84.05%	68.38%	75.93%	57.57%	58.02%
CoSeC	62.19%	63.61%	67.22%	58.94%	62.36%
CoMeT	83.83%	70.21%	77.41%	60.30%	67.62%
CoSeC*	75.40%	70.82%	72.04%	64.94%	70.60%
CoMeT*	84.51%	71.43%	79.26%	65.35%	69.53%

Table 5: Official test set: accuracy for 2-way task (uA: unseen answers, uQ: unseen questions, uT: unseen topics)

Even though it does not live up to the standards of the bag approaches in their area of expertise (unseen answers), the CoMiC systems outperforms the bags on the unseen question and unseen topic sub-sets as expected. Note that on unseen topics, CoMiC still scores 10% above the majority baseline on the official test set, in contrast to the drop of more than 10% below the baseline for the corresponding (skewed) development set.

However, the results for CoSeC are around 10% lower on the unseen questions, and almost 7% lower on the unseen topics of the test data than on the development set, a drop that the overall meta-classifier (CoMeT) was unable to catch. Investigating this drop in comparison to our development set, we checked the correctness of the training script and discovered a bug in the CoSeC setup that led to the parameters and the thresholds being computed on the same partition of the training set, i.e., the system overfitted to this partition, while the remainder of the training set was not used for training. Correcting the bug resulted in CoSeC accuracy values broadly comparable to those of CoMiC, as was the case on the development set. This confirms that the reason for the drop in the submission was not a flaw in the CoSeC system as such, but a programming bug in a peripheral component.

With this bug fixed, CoSeC performs 5%–13% better on the test set, and the meta-classifier would have been able to benefit from the regularly performing CoSeC, improving in performance up to 5%. These two amended systems are listed as CoSeC* and CoMeT* in Table 5. For the two unseen answers scenarios, CoMeT* would outperform the best scoring systems of the challenge in the 2-way task.

3.3 Discussion

In this section, we try to identify some general tendencies from studying the results. Firstly, we can observe that due to the strong performance of the bag models, unseen answers scores are generally higher than their counterparts. It seems that if questions have been seen before, surface-oriented methods outperform more abstract approaches. However, the picture is different for unseen domains and unseen questions. We are generally puzzled by the fact that many systems in the shared task scored worse on unseen questions, where in-domain training data is available, than on unseen domains, where this is not the case. The CoMeT classifier suffered especially in unseen questions of SciEntsBank, scoring lower than our best system would have on its own (see Table 5); even after the CoSeC bug was fixed, CoMeT* still scored worse there than CoMiC on its own.

In general, we likely would have benefited from domain adaptation, as described in, e.g., Daume III (2007). Consider that the input for the meta-classifier always consists of the same set of features produced via standard cross-validation, regardless of the test scenario. Instead, the trained model should have different feature weights depending on what the model will be tested on.

4 Conclusion and Outlook

We presented our approach to Task 7 of SemEval 2013, consisting of a combination of surface-oriented bag models and the increasingly abstract alignment-based systems CoMiC and CoSeC. Predictions of all systems were combined using a meta classifier in order to produce the final result for CoMeT.

The results presented show that our approach performs competitively, especially in the unseen answers test scenarios, where we obtained the best result of all participants in the 3-way task with the Beetle corpus (73.1% accuracy). As expected, the unseen topics scenario proved to be more challenging, with results at 67.6% accuracy in the 2-way task for CoMeT. Surprisingly, CoMeT performed consistently worse in the unseen questions scenarios, which we attribute to rather low CoSeC results there and to the way the meta classifier is trained, which currently does not take into account the test scenario it is trained for and instead uses the module and question IDs as fea-

tures, which turned out not to be an effective domain adaptation approach.

In our future research, work on CoMiC will concentrate on integrating two aspects of the context: First, we are planning to develop an automatic approach to focus identification in order to pinpoint the essential parts of the student answers. Second, for data sets where a reading text is available, we will try to automatically determine the location of the relevant source information given the question, which can then be used as alternative or additional reference material for answer evaluation.

The CoMiC system currently also relies on the Traditional Marriage Algorithm to select the optimal global alignment between student answer and reference answer. We plan to replace this algorithm by a machine learning component that can handle this selection in a data-driven way.

For CoSeC, we plan to develop an extension that allows for *n-to-m* mappings, hence improving the alignment performance for multi-word units such as, e.g., phrasal verb constructions.

The bag approaches could be augmented by exploring additional levels of abstractions, e.g., semantic equivalence classes constructed via WordNet lookup.

In sum, while we will also plan to explore optimizations to the training setup of the meta-classifier (e.g., domain adaptation along the lines of Daume III, 2007), the main focus of our further research lies in improving the individual sub-systems, which then again are expected to push the overall performance of the CoMeT meta-classifier system.

Acknowledgements

We are thankful to Sowmya Vajjala and Serhiy Bykh for their valuable advice on meta-classifiers and other machine learning techniques. We also thank the reviewers for their comments; in consultation with the SemEval organizers we kept the length at 8 pages plus references, the page limit for papers describing multiple systems.

References

Stacey Bailey and Detmar Meurers. 2008. Diagnosing meaning errors in short answers to reading comprehension questions. In Joel Tetreault, Jill Burstein, and Rachele De Felice, editors, *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications (BEA-3) at ACL'08*, pages 107–115, Columbus, Ohio. `http://aclweb.org/anthology/W08-0913.pdf`.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, volume 1, pages 86–90, Montreal, Quebec, Canada.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Journal of Language Resources and Evaluation*, 3(43):209–226.

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.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(4):i–xvii, 10.

Hal Daume III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June. Association for Computational Linguistics.

Myroslava O. Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In **SEM 2013: The First Joint Conference on Lexical and Computational Semantics*, Atlanta, Georgia, USA, 13-14 June.

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

David Ferrucci and Adam Lally. 2004. UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3–4):327–348.

David Gale and Lloyd S. Shapley. 1962. College admissions and the stability of marriage. *American Mathematical Monthly*, 69:9–15.

Otis Gospodnetić and Erik Hatcher. 2005. *Lucene in Action*. Manning, Greenwich, CT.

Michael Hahn and Detmar Meurers. 2011. On deriving semantic representations from dependencies: A

practical approach for evaluating meaning in learner corpora. In *Proceedings of the Intern. Conference on Dependency Linguistics (DEPLING 2011)*, pages 94–103, Barcelona. http://purl.org/dm/papers/hahn-meurers-11.html.

Michael Hahn and Detmar Meurers. 2012. Evaluating the meaning of answers to reading comprehension questions: A semantics-based approach. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-7) at NAACL-HLT 2012*, pages 94–103, Montreal. http://aclweb.org/anthology/W12-2039.pdf.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. In *The SIGKDD Explorations*, volume 11, pages 10–18.

Michael Halliday. 1967. Notes on Transitivity and Theme in English. Part 1 and 2. *Journal of Linguistics*, 3:37–81, 199–244.

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)*. Universitätsverlag Potsdam, Potsdam.

Ivana Kruijff-Korbayová and Mark Steedman. 2003. Discourse and information structure. *Journal of Logic, Language and Information (Introduction to the Special Issue)*, 12(3):249–259.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Detmar Meurers, Ramon Ziai, Niels Ott, and Stacey Bailey. 2011a. 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. http://purl.org/dm/papers/meurers-ea-11.html.

Detmar Meurers, Ramon Ziai, Niels Ott, and Janina Kopp. 2011b. Evaluating answers to reading comprehension questions in context: Results for German and the role of information structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, pages 1–9, Edinburgh, Scotland, UK, July. http://aclweb.org/anthology/W11-2401.pdf.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–233.

Joakim Nivre, Jens Nilsson, Johan Hall, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(1):1–41.

Niels Ott, Ramon Ziai, and Detmar Meurers. 2012. Creation and analysis of a reading comprehension exercise corpus: Towards evaluating meaning in context. In Thomas Schmidt and Kai Wörner, editors, *Multilingual Corpora and Multilingual Corpus Analysis*, Hamburg Studies in Multilingualism (HSM), pages 47–69. Benjamins, Amsterdam. http://purl.org/dm/papers/ott-ziai-meurers-12.html.

John C. Platt. 1998. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical Report MSR-TR-98-14, Microsoft Research.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 2002. *Numerical Recipes in C++*. Cambridge University Press, Cambridge, UK.

J.R. Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

Frank Richter and Manfred Sailer. 2004. Basic concepts of lexical resource semantics. In Arnold Beckmann and Norbert Preining, editors, *European Summer School in Logic, Language and Information 2003. Course Material I*, volume 5 of *Collegium Logicum*, pages 87–143. Publication Series of the Kurt Gödel Society, Wien.

Robert C. Russell. 1918. US patent number 1.261.167, 4.

Gerard Salton and Michael J. McGill. 1983. *Introduction to modern information retrieval*. McGraw-Hill, New York.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Joel Tetreault, Daniel Blanchard, Aoife Cahill, and Martin Chodorow. 2012. Native tongues, lost and found: Resources and empirical evaluations in native language identification. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 2585–2602, Mumbai, India.

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.

Klaus von Heusinger. 1999. *Intonation and Information Structure. The Representation of Focus in Phonology and Semantics*. Habilitationsschrift, Universität Konstanz, Konstanz, Germany.

David H. Wolpert. 1992. Stacked generalization. *Neural Networks*, 5(2):241–259.

Ramon Ziai, Niels Ott, and Detmar Meurers. 2012. Short answer assessment: Establishing links between research strands. In Joel Tetreault, Jill Burstein, and Claudial Leacock, editors, *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-7) at NAACL-HLT 2012*, pages 190–200, Montreal, June. http://aclweb.org/anthology/W12-2022.pdf.