Paraphrase Detection for Short Answer Scoring

Nikolina Koleva, Andrea Horbach, Alexis Palmer, Simon Ostermann, Manfred Pinkal

Saarland University, Saarbrücken, Germany

`(nikkol|andrea|apalmer|simono|pinkal)@coli.uni-saarland.de`,

ABSTRACT

We describe a system that grades learner answers in reading comprehension tests in the context of foreign language learning. This task, also known as short answer scoring, essentially requires determining whether a semantic entailment relationship holds between an individual learner answer and a target answer; thus semantic information is a necessary part of any automatic short answer scoring system. At the same time the method must be robust to the particularities of learner language. We propose using paraphrase detection, a method that meets both requirements. The basis for our specific paraphrasing method is word alignment learned from parallel corpora which we create from the available data in the CREG corpus (Corpus for Reading Comprehension Exercises for German). We show the usefulness of this kind of information for the task of short answer scoring. Combining our results with existing approaches we obtain an improvement tendency.

KEYWORDS: paraphrase fragments, short answer scoring, reading comprehension.

Nikolina Koleva, Andrea Horbach, Alexis Palmer, Simon Ostermann and Manfred Pinkal 2014. Paraphrase detection for short answer scoring. *Proceedings of the third workshop on NLP for computer-assisted language learning.* NEALT Proceedings Series 22 / Linköping Electronic Conference Proceedings 107: 59–73.

1 Introduction

Reading comprehension exercises are a common means of assessment for language teaching: students read a text in the language they are learning and are then asked to answer questions about the text. Answers to such questions typically consist of one sentence, sometimes two or three. They are graded taking the semantic content into consideration, ignoring spelling or grammatical errors. Developing methods for the automatic scoring of answers (in short: **short answer scoring**) is a task of considerable practical relevance, in particular with regard to the increasing availability of online language courses. At the same time, it is an interesting challenge for computational semantics, and it calls for the use of methods from semantics-focused natural language processing. The short answer scoring (SAS) task stands in a close relationship to the task of recognizing textual entailment (RTE): A correct student answer should entail (ideally, be identical in content with) one of the *target answers*, i.e., the sample solutions created by a teacher. Moreover, the student answer should be entailed by the text.

Figure 1 shows an example of a passage of a reading text, a question about the text, the target answer and both a correct and incorrect learner answer. Note that the first learner answer is graded as correct because it is a paraphrase of the target answer, despite the errors it contains.

TEXT: (…) **Sent**$_i$: The Hessian government wants to prevent this reform, because “when it comes to apple wine all Hessians agree.” **Sent**$_{i+1}$: It’s easier for other apple wine nations like France or Spain. **Sent**$_{i+2}$: There, the beverage is called “Cidre” or “Sidra” and may keep that name, because the term “wine” is not part of the name. (…)	QUESTION: Do other European countries experience similar problems as Hesse. Why?
	TARGET ANSWER: No, [**in other apple wine nations like France or Spain the beverage is called “Cidre”**] or “Sidra” and may keep that name, because the term “wine” is not part of the name.
	LEARNER ANSWER (CORRECT): No. other countries, like [**France or Spain, have other name for apple drinking, like “Cidre”.**]
	LEARNER ANSWER (INCORRECT): Against - the Hessian government should this reform.

Figure 1: Example reading text, question, and answers (from CREG, translation by authors). The extracted paraphrase fragments between the target answer and the correct learner answer are in bold-print and square brackets.

In the standard RTE setting, the task is, given a text and a hypothesis sentence, to determine automatically whether the hypothesis is entailed by the text. Of course, there are substantial differences between SAS and the standard RTE setting. Most importantly, the linguistic quality of student answers may be very poor. Answers may be ungrammatical or contain many spelling errors, which makes deep entailment modeling difficult or even completely impossible, as can be seen in the learner answer of figure 1. Also, both the target answers and, in particular, the student answers, have a tendency to keep close to the text surface. Therefore, shallow approaches considering only surface information form a strong baseline. Existing approaches to automatic short answer scoring typically rely on alignments between learner and target answer, mostly using lexical and shallow syntactic information, plus possibly lexical-semantic resources such as WordNet (Fellbaum, 1998), in part with impressively good results (for an overview, see Ziai et al. (2012)).

In the present paper, we describe an approach to short answer scoring that uses semantic information which is easily obtained and robust to learner language and other requirements of the SAS setting. Central to our approach is a method that provides information about paraphrase relations between (parts of) student answer and target answer. We adopt the approach of Wang and Callison-Burch (2011) and Regneri and Wang (2012), who extract sub-sentential paraphrase candidates (“paraphrase fragments”) from monolingual parallel corpora, making essential use of GIZA++, a word alignment algorithm originally developed for aligning bilingual parallel texts in Machine Translation (Och and Ney, 2003). The alignment algorithm learns semantic information from the corpus in an unsupervised way, without any labeled training material. Once this semantic information is given, paraphrase fragments are predicted in a robust manner, using no or (in the chunk-based version of the algorithm) only very shallow additional linguistic information. An example for the fragments that are extracted from a learner answer and the corresponding target answer are the bold-print parts of the example in figure 1.

We create a parallel corpus using the Corpus of Reading Comprehension Exercises in German (CREG) (Ott et al., 2012) in a rather straightforward way by providing sentence pairs that consist of e.g. a learner answer and the corresponding target answer. We train a paraphrase fragment recognition system on this corpus following the approach by (Wang and Callison-Burch, 2011). The detected paraphrases are then used to assess the correctness of the learner answers in the CREG corpus. We do so by extracting features from the paraphrase fragments detected between a learner answer and the target answer and use these features as input to a linear regression learner. We consider features that are indicators for the strength of the semantic connection. The rationale is that a learner answer that shares no paraphrase fragment with the target answer is likely to be false, whereas a learner answer – target answer pair whose fragments are strongly linked is likely to involve a correct learner answer.

To our knowledge, we are the first to use automatic paraphrase fragment detection (and associated methods from machine translation) for the short answer scoring task. This method enables access to semantic knowledge in a robust and (almost) unsupervised way which is transferrable to other languages or domains with minimal additional effort. Evaluation on the CREG Corpus shows that information provided by paraphrase detection alone leads to quite good scoring results. More importantly, combining the system with shallow and deep semantic state of the art systems leads to consistent performance gains. A combination of all three systems results in an accuracy of 88.9 %, which surpasses the state of the art and seems to be appropriate for practical application.

The remainder of this paper is structured as follows: we discuss related approaches in section 2, and describe and evaluate paraphrase fragment detection on the CREG corpus in section 3. Section 4 describes and evaluates our use of paraphrases for short answer scoring, after which we conclude.

2 Related Work

Approaches to automatic short answer scoring usually target the grading task by comparing the learner answer to a target answer specified by a teacher. While early systems used handcrafted patterns (Pulman and Sukkarieh, 2005), most systems rely on alignments between learner and target answer, mostly using lexical and syntactic information (Leacock and Chodorow, 2003; Mohler et al., 2011; Meurers et al., 2011a,b), and sometimes explicitly using lexical paraphrase resources such as WordNet (Fellbaum, 1998).

Horbach et al. (2013) include the text as an additional source of information in grading learner answers, by comparing whether learner answer and target answer can be linked to the same text sentence. The restriction to sentence-sized units is one limitation addressed by our approach.

We compare our work to our reimplementation of the alignment-based approach by Meurers et al. (2011b). This model uses alignments on different linguistic levels (like words, lemmas, chunks and dependency triples) to align elements in the learner answer to elements in the target answer. Features (e.g. percentage of aligned tokens/chunks/triples in the learner answer and target answer, percentage of aligned words that are string-identical, lemma-identical, or synonyms, etc.) are then extracted for a machine learner that classifies an answer as correct or incorrect. They report an accuracy of 84.6% on the CREG corpus. Our reimplementation reaches an accuracy of 86.8% using a linear regression classifier.

The only deep semantic approach to short answer scoring known to us is described in Hahn and Meurers (2012). They provide an interesting solution to the robustness problem: as a semantic formalism they use Lexical Resource Semantics (LRS), which is a formalism enabling arbitrary degrees of underspecification, and a syntax-semantic interface using atomic dependency information. In effect, this guarantees that some kind of semantic representation is computed for any (grammatical or ungrammatical) input expression. The LRS representations for target and learner answer are aligned, and alignment features are extracted and used by a classifier. They reach state of the art accuracy of 86.3% on the CREG corpus, with a system that requires hand-coded language-specific semantic knowledge.

A widely used method for paraphrase detection is the extraction of equivalent sentences from either parallel or comparable monolingual corpora (Barzilay and McKeown, 2001; Barzilay and Elhadad, 2003; Quirk et al., 2004). However, for many NLP applications, sentences may turn out to be an impractical unit for paraphrasing, as the situation that two sentences convey exactly the same meaning is rather rare.

Recently, the research focus for paraphrase extraction has therefore been expanded to also consider sub-sentential paraphrase fragments as units of analysis that are not restricted to a particular category. This is done to account for partial semantic overlap between sentences that can be expressed using various types of categories, as e.g. *her preference* vs. *what she prefers*.

Recent approaches to paraphrase fragment extraction include Bannard and Callison-Burch (2005), Zhao et al. (2008) and Wang and Callison-Burch (2011). As pure word matching is not enough to achieve good results, most systems include syntactic information in the form of constituent or dependency structures (Callison-Burch, 2008; Regneri and Wang, 2012).

Gleize and Grau (2013) apply sentential paraphrase identification for scoring student answers. Their method is based on substitution by Basic English variants. They project the actual form of the answers onto a simple language and argue that in this way it is easier to draw inferences. However, by the mapping to the simplified representation not the entire semantic content is transferred. In addition, this method relies on available resources like dictionary and some hand-crafted rules, which is problematic when dealing with low resource languages.

3 Paraphrase Fragment Detection

This section describes our work on detecting paraphrase fragments in the context of reading comprehension exercises for learners of German as a foreign language. After describing the corpus (section 3.1) and method (section 3.2), we present an evaluation and analysis of the paraphrase fragments we detect (section 3.3 and section 3.4).

3.1 Data

We use the Corpus of Reading Comprehension Exercises in German (CREG) (Ott et al., 2012), first for paraphrase fragment detection and later (see section 4) as a testbed for using the extracted fragments in a short answer grading scenario. The corpus consists of *reading texts*, *questions* about the texts, *target answers* provided by teachers and *learner answers* given by German as a Foreign Language learners from two universities in the US; an example appears in figure 1. Each (paper-based) hand-written learner answer has been transcribed by two teachers, resulting in two potentially slightly different transcripts for each learner answer. The learner answers in CREG have also been scored as correct or incorrect by teachers. Following previous work, we use the balanced subset of 1032 learner answers, half correct and half incorrect.

Horbach et al. (2013) extend CREG with a set of annotations linking each target and learner answer to the sentence in the associated reading text that best matches the meaning of an answer and thus is its expected source. These are human annotations, providing a set of *gold text sentences* that we also use in our experiments. In the example in figure 1, both the target answer and the correct learner answer can be linked to sentence $i+1$, while the incorrect learner answer has a link to sentence i.

3.2 Method

Wang and Callison-Burch (2011) and Regneri and Wang (2012) describe a procedure for extracting paraphrase fragments which consists of the following steps: constructing a parallel/comparable corpus, estimating word alignments over this corpus, computing positive and negative lexical associations, refining the alignment and, finally, detecting paraphrases. We follow this general method, customizing some steps to suit the needs of our application context.

For paraphrase fragment detection, we present two versions of our system: *basic*, which uses only word alignments for the detection step, and *chunk-based*, which also makes use of shallow syntactic analysis.

Building a comparable corpus. The aim in building a comparable corpus is to collect pairs of sentences which are likely to contain paraphrase fragments. To build our collection of sentence pairs, we exploit properties of the short answer grading scenario (via the CREG corpus).

Target answers (TA) and (correct) learner answers (LA) are the first, most obvious candidate pairs, as they convey the same meaning. We also include TAs paired with incorrect LAs. Such pairs are sometimes completely unrelated, thus introducing noise to the data, but sometimes they overlap enough to share one or more paraphrase fragments. Our aim is specifically pairs of sentences. In cases where an answer consists of more than one sentence, we include all possible combinations of TA sentence and LA sentence. This expands the number of sentence pairs, but also introduces additional noise.

In order to provide a richer source of lexical variation, we extend the input with pairs consisting of a TA or LA and its corresponding sentence from the reading text: Horbach et al. (2013) describe both human annotations of the best fitting sentence from the reading text for an answer and a procedure for automatically identifying the most closely-linked text sentence. We use both in the experiments described below: the *goldlink condition* uses human annotations, and the *autolink condition* takes the sentence which has the highest alignment weight to the answer when the two sentences are aligned using the method described in (Meurers et al., 2011c).

We thus arrive at an input corpus, consisting of five sub-corpora: TA – *correct* LA, TA – *incorrect* LA, TA – *text sentence*, *correct* LA – *text sentence*, *incorrect* LA – *text sentence*.

We increase the training material available by boosting the corpus in several ways. First, to emphasize the importance of lexical identity for learning word alignments, we add trivially-identical pairs: each reading text sentence paired with itself, and each word in the CREG corpus vocabulary, also paired with itself. Additionally, we repeat non-identical sentence pairs, with the number of repetitions linked to the nature of the sub-corpus in which the pair appears. We have also begun experiments adding word pairs from GermaNet (Hamp and Feldweg, 1997), in order to learn lexical paraphrases, but the results reported here do not include GermaNet-based boosting.

For intrinsic evaluation of the detected paraphrase fragments (Section 3.3), we aim to reduce noise in the data and emphasize reliable sentence pairs. Accordingly, each pair involving correct LAs, as well as those with TAs and text sentences, is copied 10 times. Pairs involving incorrect LAs appear just one time. The trivially-identical pairs are entered 10 times for sentences and 20 times for word pairs.

Preprocessing. To prepare the data for word alignment, we apply a standard linguistic preprocessing toolchain, consisting of sentence segmentation using `OpenNLP`,[1] tokenization with the `Stanford Tokenizer`,[2] lemmatization and part-of-speech (POS) tagging, both using the `TreeTagger` (Schmid, 1995). We use the `Stanford Named Entity Recognizer`[3] to identify persons, organizations, locations and dates. For robustness against grammatical errors and to reduce vocabulary size, all tokens are replaced with their lemmatized forms. We replace all occurrences of NEs with the corresponding NE-tag (e.g. *PERSON*).

Learner answers frequently contain spelling errors. We treat them in the following way: we run all the learner answers through the German version of the spellchecker `aspell`[4] and check for non-words. For those non-words we first look up whether the word is nevertheless a correct word (like a proper name) from the connected material (target answer, question, text) that is for some reason not known to `aspell`. If that is not the case we look for a spelling alternative in the connected material, i.e. we check whether a token with a levenshtein distance up to a certain threshold occurs (in that order) either in the target answer, the question, or the text. If so, we replace the non-word learner answer token by this word.

Detecting paraphrase fragments. Following previous work (Wang and Callison-Burch, 2011; Regneri and Wang, 2012), we pass our input corpus to GIZA++ (Och and Ney, 2003) in order to: (a) estimate word alignments for input sentence pairs, and (b) obtain a lexical correspondence table with scores for individual word pairs.

Links between aligned words in the sentence pairs are then classified as positive or negative based on their scores, a technique which has previously been applied to extract paraphrase fragments from non-parallel bilingual corpora and has been shown to improve a state of the art machine translation system (Munteanu and Marcu, 2006). Word pairs containing punctuation or stop words are excluded from the alignment prior to scoring.[5]

Afterwards, the alignment is refined by removing all negatively-scored word pairs, such that only very strong alignments survive. We then smooth the alignment by recomputing scores for each word, averaging over a window of five words. In this way we often capture context words

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

[2] http://nlp.stanford.edu/software/tokenizer.shtml

[3] http://nlp.stanford.edu/software/CRF-NER.shtml

[4] http://aspell.net/

[5] http://www.ranks.nl/stopwords/german.html

that are left out of the alignment process (e.g. determiners, prepositions, or particles) but are nonetheless necessary for producing linguistically well-formed fragments.

For the *basic* version, a source-side fragment is detected by extracting sequences of adjacent words with positive scores after smoothing. The corresponding target-side fragment is induced using one of two methods: The *unidirectiona*l approach finds the target fragment by using the lexical scores for the source side plus alignment links to the target side. In the *bidirectional* approach, we also compute lexical scores for the target side and extract target-side fragments in that manner.

Despite the use of smoothing for producing more grammatical fragments, the basic approach often produces output of questionable readability, e.g. "hm, so" is a fragment that lacks context in order to understand the intended semantic content. Especially if these fragments might be used to give feedback to learners, it is important to produce readable output. This is the motivation for the second version of the system.

In the *chunk-based* version we reset the boundaries of the basic fragments in a post-processing step by taking syntactic chunk information into consideration. If a fragment has some overlap with a chunk, then the remainder of that chunk is also included in the fragment. We also apply some heuristics to account for aspects of the German language: e.g. prefixes of separable verbs and past participles often appear in sentence-final position and should be covered by the fragment.

The fragment extracted from the source sentence is the same for all configurations but the target fragment differs. Example (1) illustrates the difference of fragments extracted by the unidirectional vs. bidirectional method and example (2) the one of basic vs. chunk based.

(1) **source fragment**: in front of the PC or the TV
target fragments:
`uni`: with the PC or the TV
`bi`: all time with the the PC or the TV

(2) **source fragment**: in vegetable garden one has to chop and water
target fragments:
`basic`: in vegetable garden chop and waterz
`chunk`: one can chop and water in vegetable garden

An interesting observation is that the bidirectional method tends to be too greedy. Target fragments returned with it contain additional information that has no corresponding part on the source side. The chunk-based system is useful because it augments a fragment but also slightly modifies its semantic content.

3.3 Intrinsic Evaluation of Detected Paraphrases

To evaluate precision of the extracted paraphrases, we again follow Wang and Callison-Burch (2011) and Regneri and Wang (2012). For each of the two systems, 300 fragment pairs are randomly extracted, half with the unidirectional version and half with the bidirectional. These are evenly distributed across LA-TA pairs and answer-text sentence pairs. Each fragment is labeled by two annotators with one of four categories: paraphrase, related, unrelated, or invalid. The label *related* is assigned when there is overlap between the two fragments, but they are not

	unidirectional	bidirectional
basic	0.78	0.74
chunk-based	0.69	0.71

Table 2: Precision of paraphrase fragment detection

paraphrases, and *invalid* is assigned if one or both fragments are completely ungrammatical or not readable. Annotators were not told the type of the sentence pair, and they were instructed to ignore spelling and grammatical errors in evaluating paraphrases.
Table 1 shows the inter-annotator agreement in 2 conditions: if we consider all 4 labels separately, and if we instead merge *paraphrase* and *related* as well as *unrelated* and *invalid*. Results are along the lines of (Regneri and Wang, 2012) who report Kappa values of 0.55 for four-label annotation and 0.71 for a two-label condition. Our basic system shows worse agreement than the chunk-based. This is due to the fact that basic fragments are often linguistically not well-formed and are therefore harder to annotate. For the final gold-standard, all conflicts have been resolved by a third annotator.

	4 categories	2 categories
basic	0.22 (fair)	0.69 (good)
chunk-based	0.52 (moderate)	0.84 (very good)

Table 1: Inter-annotator-agreement

This gold-standard annotation is then used for evaluating the quality of the fragments. For measuring the precision of the extracted paraphrases, i.e. for measuring what percentage of the fragment pairs identified should be considered as paraphrases or related, we use the two-label condition. Results are presented in table 2. Precision on our dataset is in the same range as that reported by Wang and Callison-Burch (2011) (62 to 67%) on a monolingual comparable corpus. Note however that this evaluation covers only a very small dataset as compared to the overall parallel corpus. Overall the performance of the basic system is better than the chunk-based. This is an unexpected result because the chunk-based system was developed specifically to improve the quality of the basic fragments. However, missing tokens like prepositions that are added to a fragment by the chunk system can change its meaning and as a consequence the fragments are no longer related.

Between the unidirectional and bidirectional approaches there is no stastically significant difference, according to a chi-squared test (Pearson, 1900).

For the application of the extracted paraphrase fragments to short answer scoring, the unidirectional approach is used, because it gave us the best results for the generally better *basic* version of the system.

We expect variability across correct and incorrect answers, because in scoring a learner answer, strict paraphrases are not always necessary. For example a question in the corpus asking "*Wer war an der Tür*" (*Who was by the door?*) with the target answer "*Drei Soldaten (three soldiers) waren an der Tür*" the learner answer "*Drei Männer (three men) waren an der Tür*", although less specific, was also graded as correct by the teachers. To investigate this variability, we look at the distribution of the four categories across the various subcorpora.

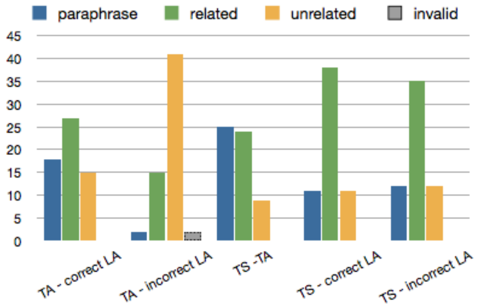

Figure 2: Distribution of annotation labels for the five subcorpora. TA stands for target answer, LA for learner answer and TS for the corresponding text sentence.

sub-corpus	**productivity in %**
`ta-corr la`	95
`ta-incorr la`	**78**
`textSent-ta`	95
`textSent-corr la`	94
`textSent-incorr la`	92
total	91

Table 3: Productivity by subcorpus

Figure 2 depicts the distribution of the labels – exemplarily for the chunk-based version – showing how often each annotation label occured within the five subcorpora TA – *correct* LA, TA – *incorrect* LA, TA – *text sentence*, *correct* LA – *text sentence*, *incorrect* LA – *text sentence*. We can see that correct learner answers lead to more paraphrases of the target answer (18) than do incorrect learner answers (2). Incorrect learner answers, however, have a much higher degree of unrelated fragments with the target answer (41 vs 15). Correctness has not much influence on the validity. In the subcorpora involving text sentences, both correct and incorrect learner answers have a similarly high degree of paraphrase and related cases. That is the case because both correct and incorrect learner answers are often paraphrases of some part of the text. In the case of an correct answer, the target answer is often a paraphrase of the same text sentences as the text sentence for the learner answer, in the case of an incorrect learner answer, the student often erroneously paraphrased a text sentence that has nothing to do with the correct answer.

3.4 Analysis of the Detected Fragments

This section presents continued analysis of the detected fragments from various subcorpora, covering productivity and variability of lexical material.

Productivity of the Detected Fragments

Table 3 shows productivity by subcorpus, measured by how often at least one fragment pair is detected per input sentence pair. As expected, productivity is lowest for incorrect LAs paired with TAs. Incorrect LAs paired with text sentences, however, show productivity similar to other

	source	target
1	Die Stadtverwaltung sagt nein	Die Stadtverwaltung ist dagegen
2	kein glücklisches Ende	ein schlechte Ende
3	die Broadway-Version erhielt sechs Tonys	Es hat sechs Tonys gewonnen
4	Damit lachen die anderen Kinder sie ja aus	die anderen Kinder lachen Julchen aus
5	darf nicht mehr verwendet werden	dann nicht mehr erlaubt
6	Die Leute wissen *nicht* ihre genauen monatlichen Ausgaben	die meisten Leute wissen wie eine Budgetplan zu machen
7	in einem *Neu*bau	in einem *Alt*bau
8	würde mit *Computer* arbeiten	würde mit *Wissenschaftlerin* arbeiten
9	[Nicht, sagten die Augen] der Frau, nicht lachen	[Er sollte nicht] lachen, weil das Kind [schlief]

Table 4: Fragments output with the `unidirectional` method for the *chunk-based* system

subcorpora. This is not surprising, as incorrect learner answers often stem from some part of the text (Horbach et al., 2013), although not necessarily the same as the target answer.

Lexical Variety of the Detected Paraphrases

In many cases, there are only minor differences between learner answers and target answers. Inspection of the data shows that our approach detects real paraphrase fragments, beyond the trivial case of identical spans of text in paired sentences.

To evaluate lexical variety, we measure the degree of lemma overlap between sentence pairs and fragment pairs. Figure 3 shows that there is a significantly higher overlap between paraphrase pairs than between sentences, but on the other hand, the overlap is not so extensive that it makes the paraphrase detection task trivial.

Table 4 shows example fragments detected by the chunk-based, unidirectional method. The qualitative analysis shows that non-identical material contained in the fragments often captures alternative expressions of the same semantic content. However, we can see that the method would benefit from handling of phenomena such as negation, antonymy, or relatedness between nouns or other content words.

Fragment pair 7 illustrates the difficulty faced in cases where antonymy is present. The compound words "Altbau" and "Neubau" both carry the main meaning of a building (*der Bau*) and are therefore related, but the modifying words "alt" and "neu" (*old* and *new*) are antonyms.

Fragment pair 8 highlights the problem that word alignments like *Computer-Wissenschaftlerin (scientist)* are learned, even though they are not valid paraphrases and the word *Wissenschaftlerin* only occurs in incorrect answers. This happens in cases when an input sentence pair shares many identical words, and one or more non-identical words that occur very infrequently (or even nowhere else) in the corpus. In such a case, GIZA++ learns strong alignments between the identical words and also between the two unrelated words, as there are no other options for linking those words.

The last fragment pair 9 shows an example of unrelated fragments, which are probably (mistakenly) classified as paraphrases because of the high token overlap.

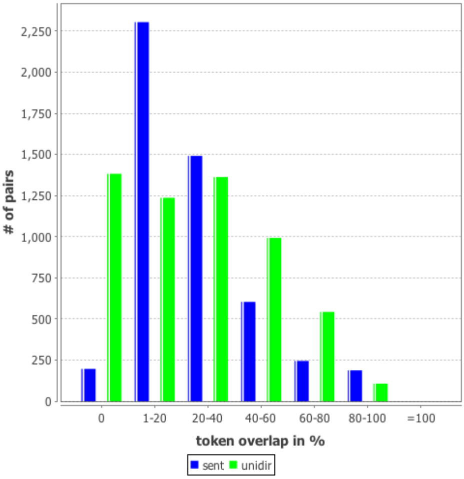

Figure 3: Percentage of identical tokens in sentence pairs (sent) and fragment pairs (unidir)

4 Short Answer Scoring

We use the results of the paraphrase fragment detection (section 3) as the basis for automatic short answer scoring. In this section we describe our method (section 4.1) and evaluate it on the CREG corpus (section 4.2).

4.1 Method

We base the assessment of the correctness of the learner answer on the paraphrase relation between learner answer and target answer. We take the case when no paraphrase is found to be strong evidence against correctness. If a paraphrase pair is detected, we want to make the scoring decision dependent on properties of the single paraphrases and their interrelation. Technically, we use a binary classifier, which bases its prediction on features extracted from the paraphrase fragments. Concretely, we employ the linear regression classifier from the Weka Toolkit (Witten and Frank, 2005).

As outlined in the introduction, we employ different modes to identify pairs of LA and TA paraphrases: In the direct mode, we directly determine a TA-LA paraphrase pair based on the alignment between LA and TA. In the indirect mode, we pair each of LA and TA with a text sentence (these may be identical or different sentences), independently derive paraphrase pairs for LA with text sentence and TA with text sentence, respectively, which in the success case gives us a TA-LA paraphrase pair obtained in an indirect way. We assume that the indirect mode provides additional information through the relatedness between LA, TA, and the text.

For each of the two comparison modes a set of features is extracted, which provide information about the relation between the paraphrase fragments f_1 and f_2, which are extracted from a sentence pair s_1 and s_2 or about single fragments.

The following features are considered:

1. token overlap: jaccard coefficient $J(tokens(f_1), tokens(f_2)) = \frac{|tokens(f_1) \cap tokens(f_2)|}{|tokens(f_1) \cup tokens(f_2)|}$, 0 if there are no fragments
2. difference in fragment length $|f_1 - f_2|$, -1 if there are no fragments
3. percentage of tokens in the s_1 covered by f_1
4. percentage of tokens in the s_2 covered by the f_2
5. average of lexical scores for the target answer (resulting from word alignment)

Because we use the unidirectional alignment version and take the text sentence to be the source sentence, only lexical scores for the text sentences are computed in the indirect case. Therefore the fifth feature is not available in the indirect mode.

4.2 Evaluation

We compare our approach to both the alignment model (as in (Meurers et al., 2011b; Horbach et al., 2013)) and the deep semantic model by (Hahn and Meurers, 2012). We re-implement the alignment model using features for token and chunk alignment reaching an accuracy of 86.8% on the CREG corpus (compared to 84.6% in the (Meurers et al., 2011b) model). The deep semantic model reaches an accuracy of 86.3%, also on the CREG data. We make direct comparison against these two scores; a random baseline for this balanced data set is 50%.

We evaluate using tenfold cross-validation, running the complete paraphrase fragment detection method (Section 3) on nine folds for training. For the test corpus, of course, we don't know ahead of time whether answers are correct or not. Thus we build our input corpus without taking advantage of this information. In this setting, each pair involving a LA or TA is included 10 times, regardless of the answer's correctness.

We evaluate our model alone and using additional features from the other two models, as is shown in table 5: In order to see the contribution of the direct and indirect feature sets, we evaluate those sets individually (*paraphrases direct* and *paraphrases indirect*) and together (*paraphrases combined*). For combining with the other models, we always use the combined set of paraphrase features.

To evaluate our model in combination with the alignment model (*paraphrases + aligment system*), we add the features from our reimplementation. We also combine our model with both of the other two models (*paraphrases + aligment model + deep semantics*), using the semantic scores obtained by Hahn and Meurers (2012) as an additional feature.

Evaluation Corpus	paraphrases direct	paraphrases indirect	paraphrases combined	paraphrases + alignment	paraphrases + deepSemScore	paraphrases + deepSemScore + Alignment	alignment + deepSemScore
autolink - basic	76.9	70.6	78.3	86.5	86.9	87.7	87.5
autolink - chunk	76.8	70.1	77.1	86.4	86.7	88.1	87.5
goldlink - basic	77.5	72.8	77.6	86.5	87.0	88.1	87.5
goldlink - chunk	76.6	72.1	77.4	86.7	87.1	**88.9**	87.5

Table 5: Accuracy on CREG balanced corpus with various model combinations

Table 5 summarizes our results: We can see that our system alone, while being far from reaching the state of the art, can reasonably differentiate between correct and incorrect answers. The direct comparison of learner answer and target answer (*paraphrases direct*) works better than just the indirect comparison via fragments obtained from alignment with the text. In combination, the indirect features still contribute to the performance *paraphrases combined*, although not in a statistically significant way.

When combining the paraphrase features with the features from the alignment system, we don't get an improvement over the alignment system (86.8%). When additionally adding the semantic score to both feature sets, we reach our best result with an accuracy of 88.9% which is not significantly better (α=0.25 according to a McNemar test) than the comparison figure of 86.8%.

When comparing the *goldlink* to the *autolink* condition, we see an advantage of having the optimal information about the best matching sentence in the indirect feature set.

There is no clear trend as to whether the *basic* or the *chunk-based* system performs better. The paraphrase fragments model on its own is not good enough to beat the other methods. However, combining the three systems gives an improvement of 2.1%, which is an indication of complementary information provided by the different feature sets.

5 Conclusion

In this paper we have presented the first approach which uses paraphrase information for automatically scoring short answers. We successfully adapt a paraphrase fragment extraction method to the new domain of reading comprehension data for learning German as a foreign language. In this way we frame the short answer scoring task with respect to semantic information that is robust to noise in the input. Because of this robustness, and because of its (nearly) unsupervised nature, the approach is readily adaptable for other languages or domains. We obtain good scoring results using detected paraphrases, and when we combine our method with shallow and deep semantic systems, we surpass the state of the art on the CREG corpus.

We see three obvious extensions for future research. First, paraphrase fragments detected between target and learner answers, or between learner answers and the reading text, could be very useful in practical applications, such as providing direct feedback to language learners. This could be done by highlighting for a learner the paraphrased regions of his answer and, more importantly, those which do not stand in such a semantic relationship to the target answer or the text. Second, we are interested in investigating the influence of information structure on scoring; fragments which cover information from the question should receive less weight than fragments which offer new information, and our fragment detection method is one way of making such distinctions. Finally, our method can be adapted to handle online input, computing alignments based on previously-existing lexical correspondence tables and in this way providing immediate output for new learner answers.

References

Bannard, C. and Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the ACL'05*, pages 597–604.

Barzilay, R. and Elhadad, N. (2003). Sentence alignment for monolingual comparable corpora. In Collins, M. and Steedman, M., editors, *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 25–32.

Barzilay, R. and McKeown, K. R. (2001). Extracting paraphrases from a parallel corpus. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 50–57, Toulouse, France. Association for Computational Linguistics.

Callison-Burch, C. (2008). Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 196–205, Honolulu, Hawaii. Association for Computational Linguistics.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.

Gleize, M. and Grau, B. (2013). Limsiiles: Basic english substitution for student answer assessment at semeval 2013. In **SEM, Volume 2: Proceedings of SemEval 2013*, pages 598–602, Atlanta, Georgia, USA. Association for Computational Linguistics.

Hahn, M. and Meurers, D. (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 (BEA7)*, pages 326–336, Montreal, Canada. Association for Computational Linguistics.

Hamp, B. and Feldweg, H. (1997). Germanet - a lexical-semantic net for German. In *In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

Horbach, A., Palmer, A., and Pinkal, M. (2013). Using the text to evaluate short answers for reading comprehension exercises. In **SEM, Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 286–295, Atlanta, Georgia, USA. Association for Computational Linguistics.

Leacock, C. and Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37(4):389–405.

Meurers, D., Ziai, R., Ott, N., and Bailey, S. (2011a). Integrating parallel analysis modules to evaluate the meaning of answers to reading comprehension questions. *Special Issue on Free-text Automatic Evaluation. International Journal of Continuing Engineering Education and Life-Long Learning (IJCEELL)*, 21(4):355–369.

Meurers, D., Ziai, R., Ott, N., and Kopp, J. (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.

Meurers, D., Ziai, R., Ott, N., and Kopp, J. (2011c). 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, Scottland, UK. Association for Computational Linguistics.

Mohler, M., Bunescu, R. C., and Mihalcea, R. (2011). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In Lin, D., Matsumoto, Y., and Mihalcea, R., editors, *ACL*, pages 752–762.

Munteanu, D. S. and Marcu, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 81–88, Stroudsburg, PA, USA. ACL.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51.

Ott, N., Ziai, R., and Meurers, D. (2012). Creation and analysis of a reading comprehension exercise corpus: Towards evaluating meaning in context. In Schmidt, T. and Wörner, K., editors, *Multilingual Corpora and Multilingual Corpus Analysis*, Hamburg Studies in Multilingualism (HSM), pages 47–69. Benjamins, Amsterdam.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175.

Pulman, S. G. and Sukkarieh, J. Z. (2005). Automatic short answer marking. In *Proceedings of the second workshop on Building Educational Applications Using NLP*, EdAppsNLP 05, pages 9–16.

Quirk, C., Brockett, C., and Dolan, W. (2004). Monolingual machine translation for paraphrase generation. In Lin, D. and Wu, D., editors, *Proceedings of EMNLP 2004*, pages 142–149, Barcelona, Spain. Association for Computational Linguistics.

Regneri, M. and Wang, R. (2012). Using discourse information for paraphrase extraction. In *Proceedings of EMNLP-CoNNL 2012*, Jeju, Korea.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

Wang, R. and Callison-Burch, C. (2011). Paraphrase fragment extraction from monolingual comparable corpora. In *Proceedings of the 4th Workshop on Building and Comparable Corpora: Comparable Corpora and the Web*, pages 52–60, Portland, Oregon. ACL.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition.

Zhao, S., Wang, H., Liu, T., and Li, S. (2008). Pivot approach for extracting paraphrase patterns from bilingual corpora. In *Proceedings of ACL-08: HLT*, pages 780–788, Columbus, Ohio. Association for Computational Linguistics.

Ziai, R., Ott, N., and Meurers, D. (2012). Short answer assessment: Establishing links between research strands. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7)*, Montreal, Canada.