Evaluating the Meaning of Answers to Reading Comprehension Questions A Semantics-Based Approach

Michael Hahn **Detmar Meurers**
SFB 833 / Seminar für Sprachwissenschaft
Universität Tübingen
{mhahn,dm}@sfs.uni-tuebingen.de

Abstract

There is a rise in interest in the evaluation of meaning in real-life applications, e.g., for assessing the content of short answers. The approaches typically use a combination of shallow and deep representations, but little use is made of the semantic formalisms created by theoretical linguists to represent meaning.

In this paper, we explore the use of the underspecified semantic formalism LRS, which combines the capability of precisely representing semantic distinctions with the robustness and modularity needed to represent meaning in real-life applications.

We show that a content-assessment approach built on LRS outperforms a previous approach on the CREG data set, a freely available corpus of answers to reading comprehension exercises by learners of German. The use of such a formalism also readily supports the integration of notions building on semantic distinctions, such as the information structuring in discourse, which we show to be useful for content assessment.

1 Introduction

There is range of systems for the evaluation of short answers. While the task is essentially about evaluating sentences based on their meaning, the approaches typically use a combination of shallow and deep representations, but little use is made of the semantic formalisms created by theoretical linguists to represent meaning. One of the reasons for this is that semantic structures are difficult to derive because of the complex compositionality of natural language. Another difficulty is that form errors in the input create problems for deep processing, which is required for extracting semantic representations.

On the other hand, semantic representations have the significant advantage that they on the one hand abstract away from variation in the syntactic realization of the same meaning and on the other hand clearly expose those distinctions which do make a difference in meaning. For example, the difference between *dog bites man* and *man bites dog* is still present in deeper syntactic or semantic representations, while semantic representations abstract way from meaning-preserving form variation, such as the active-passive alternation (*dog bites man – man was bitten by dog*). This suggests that sufficiently robust approaches using appropriate semantic formalisms can be useful for the evaluation of short answers.

In this paper, we explore the use of Lexical Resource Semantics (Richter and Sailer, 2003), one of the underspecified semantic formalisms combining the capability of precisely representing semantic distinctions with the robustness and modularity needed to represent meaning in real-life applications. Specifically, we address the task of evaluating the meaning of answers to reading comprehension exercises.

We will base our experiments on the freely available data set used for the evaluation of the CoMiC-DE system (Meurers et al., 2011), which does not use semantic representations. The data consists of answers to reading comprehension exercise written by learners of German together with questions and corresponding target answers.

2 Related Work

There are several systems which assess the content of short answers. Mitchell et al. (2002) use handcrafted patterns which indicate correct answers to a question. Similarly, Nielsen et al. (2009) use manually annotated word-word relations or "facets". Pulman and Sukkarieh (2005) use machine learning to automatically find such patterns. Other systems evaluate the correctness of answers by comparing them to one or more manually annotated target answers. C-Rater (Leacock and Chodorow, 2003) and the system of Mohler et al. (2011) compare the syntactic parse to the parse of target answers. A comparison of a range of content assessment approaches can be found in Ziai et al. (2012).

The work in this paper is most similar to a line of work started by Bailey and Meurers (2008), who present a system for automatically assessing answers to reading comprehension questions written by learners of English. The basic idea is to align the student answers to a target answer using a parallel approach with several levels on which words or chunks can be matched to each other. Classification is done by a machine learning component. The CoMiC-DE system for German is also based on this approach (Meurers et al., 2011).

In terms of broader context, the task is related to the research on Recognizing Textual Entailment (RTE) (Dagan et al., 2006). In particular, alignment (e.g., MacCartney et al., 2008, Sammons et al., 2009) and graph matching approaches (Haghighi et al., 2005, Rus et al., 2007) are broadly similar to our approach.

3 General Setup

3.1 Empirical challenge: CREG

Our experiments are based on the freely available Corpus of Reading comprehension Exercises in German (CREG, Ott et al., 2012) . It consists of texts, questions, target answers, and corresponding student answers written by learners of German. For each student answer, two independent annotators evaluated whether it correctly answers the question. Answers were only assessed with respect to meaning; the assessment is in principle intended to be independent of grammaticality and orthography. The task of our system is to decide which answers correctly answer the given question and which do not.

3.2 Formal basis: Lexical Resource Semantics

Lexical Resource Semantics (LRS) (Richter and Sailer, 2003) is an underspecified semantic formalism which embeds model-theoretic semantic languages like IL or Ty2 into constraint-based typed feature structure formalisms as used in HPSG. It is formalized in the *Relational Speciate Reentrancy Language* (RSRL) (Richter, 2000).

While classical formal semantics uses fully explicit logical formulae, the idea of underspecified formalisms such as LRS is to derive semantic representations which are not completely specified and subsume a set of possible resolved expressions, thus abstracting away from ambiguities, in particular, but not exclusively, scope ambiguities.

As an example for the representations, consider the ambiguous example (1) from the CREG corpus.

(1) Alle Zimmer haben nicht eine Dusche.
all rooms have not a shower
'Not every room has a shower.'
'No room has a shower.'

The LRS representation of (1) is shown in Figure 1, where INCONT (INTERNAL CONTENT) encodes the core semantic contribution of the head, EXCONT (EXTERNAL CONTENT) the semantic representation of the sentence, and PARTS is a list containing the subterms of the representation.

$$\begin{bmatrix} \text{INCONT} & \textit{haben(e)} \\ \text{EXCONT} & \text{A} \\ \text{PARTS} & \left\langle \begin{array}{l} \text{A, haben(e)}, \forall x_1(\text{B} \rightarrow \text{C}), \\ \text{zimmer}(x_1),\ \exists x_2\ (\text{D} \wedge \text{E}),\ \neg\ \text{F}, \\ \text{dusche}(x_2),\ \text{subj(e,}x_1\text{)},\ \text{obj(e,}x_2\text{)} \\ \exists \text{e(haben(e)} \wedge \text{subj(e,}x_1\text{)} \wedge \text{obj(e,}x_2\text{)} \end{array} \right\rangle \end{bmatrix}$$

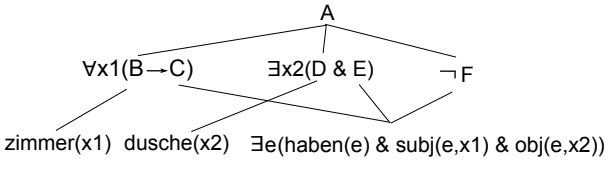

Figure 1: LRS and dominance graph for (1)

The representation also includes a set of subterm constraints, visualized as a dominance graph at the

bottom of the figure. The example (1) has several readings, which is reflected in the fact that the relative scope of the two quantifiers and the negation is not specified. The different readings of the sentence can be obtained by identifying each of the meta-variables A, ..., F with one of the subformulas. Meta-variables are labels that indicate where a formula can be plugged in; they are only part of the underspecified representation and do not occur in the resolved representation.

This illustrates the main strengths of an underspecified semantic formalism such as LRS for practical applications. All elements of the semantic representation are explicitly available on the PARTS list, with dominance constraints and variable bindings providing separate control over the structure of the representation. The underspecified nature of LRS also supports partial analyses for severely ill-formed input or fragments, which is problematic for classical approaches to semantic compositionality such as Montague semantics (Montague, 1973). Another advantage of LRS as an underspecified formalism is that it abstracts away from the computationally costly combinatorial explosion of possible readings of ambiguous sentences, yet it also is able to represent fine-grained semantic distinctions which are difficult for shallow semantic methods to capture.

3.3 Our general approach

In a first step, LRS representations for the student answer, the target answer, and the question are automatically derived on the basis of the part-of-speech tags assigned by TreeTagger (Schmid, 1994) and the dependency parses by MaltParser (Nivre and Hall, 2005) in the way discussed in Hahn and Meurers (2011). In this approach, LRS structures are derived in two steps. First, surface representations are mapped to syntax-semantics-interface representations, which abstract away from some form variation at the surface. In the second step, rules map these interface representations to LRS representations. The approach is robust in that it always results in an LRS structure, even for ill-formed sentences.

Our system then aligns the LRS representations of the target answer and the student answer to each other and also to the representation of the question. Alignment takes into account both local criteria, in particular semantic similarity, and global criteria, which measure the extent to which the alignment preserves structure on the level of variables and dominance constraints.

The alignments between answers and the question are used to determine which elements of the semantic representations are *focused* in the sense of Information Structure (von Heusinger, 1999; Kruijff-Korbayová and Steedman, 2003; Krifka, 2008), an active field of research in linguistics addressing the question how the information in sentences is packaged and integrated into discourse.

Overall meaning comparison in our approach is then done based on a set of numerical scores computed from potential alignments and their quality. Given its LRS basis, we will call the system CoSeC-DE (Comparing Semantics in Context).

4 Aligning Meaning Representations

The alignment is done on the level of the PARTS lists, on which all elements of the semantic representation are available:

Definition 1. An alignment a between two LRS representations S and T with PARTS lists p_1^n and q_1^m is an injective partial function from $\{1,...,n\}$ to $\{1,...,m\}$.

Requiring a to be injective ensures that every element of one representation can be aligned to at most one element of the other representation. Note that this definition is symmetrical in the sense that the direction can be inverted simply by inverting the injective alignment function.

To automatically derive alignments, we define a maximization criterion which combines three factors measuring different aspects of alignment quality. In addition to i) the similarity of the alignment links, the quality Q of the alignment a takes into account the structural correspondence between aligned elements by evaluating the consistency of alignments ii) with respect to the induced variable bindings θ and, and iii) with respect to dominance constraints:

$$\begin{aligned} Q(a,\theta|S,T) = & \; linksScore(a|S,T) \\ & \cdot variableScore(\theta) \\ & \cdot dominanceScore(a|S,T) \end{aligned} \quad (1)$$

The approach thus uses a deep representation abstracting away from the surface, but the meaning

comparison approach on this deep level is flat, yet at the same time is able to take into account structural criteria. In consequence, the approach is modular because it uses the minimal building blocks of semantic representations, but is able to make use of the full expressive power of the semantic formalism.

4.1 Evaluating the Quality of Alignment Links

The quality of an alignment link between two expressions is evaluated by recursively evaluating the similarity of their components. In the base case, variables can be matched with any variable of the same semantic type:

$sim(x_\tau, y_\tau) = 1$

Meta-variables can be matched with any meta-variable of the same semantic type:

$sim(\mathrm{A}_\tau, \mathrm{B}_\tau) = 1$

For predicates with arguments, both the predicate name and the arguments are compared:

$$sim(P_1(a_1^k), P_2(b_1^k)) = sim(P_1, P_2) \cdot \prod_{i=1}^{k} sim(a_i, b_i) \quad (2)$$

If the predicates have different numbers of arguments, similarity is zero. Linguistically well-known phenomena where the number of arguments of semantically similar predicates differ do not cause a problem for this definition, because semantic roles are linked to the verbal predicate via grammatical function terms such as *subj* and *obj* predicating over a Davidsonian event variable, as in Figure 1.[1]

For formulas with generalized quantifiers, the quantifiers, the variables, the scopes and the restrictors are compared:

$$\begin{aligned} sim(Q_1 x_1(\phi \circ \psi), Q_2 x_2(\sigma \circ \tau)) = \\ sim(Q_1, Q_2) \cdot sim(x_1, x_2) \\ \cdot sim(\phi, \sigma) \cdot sim(\psi, \tau) \end{aligned} \quad (3)$$

Lambda abstraction is dealt with analogously. The similarity $sim(P_1, P_2)$ of names of predicates and generalized quantifiers takes into account several sources of evidence and is estimated as the maximum of the following quantities:

As a basic similarity, the Levenshtein distance normalized to the interval [0,1] (with 1 denoting identity and 0 total dissimilarity) is used. This accounts for the high frequency of spelling errors in learner language.

Synonyms in GermaNet (Hamp and Feldweg, 1997) receive the score 1.

For numbers, the (normalized) difference $\frac{|n_1 - n_2|}{max(n_1, n_2)}$ is used.

For certain pairs of dissimilar elements which belong to the same category, constant costs are defined. This encourages the system to align these elements, unless the structural factors, i.e., the quality of the unifier and the consistency with dominance constraints, discourage this. Such constants are defined for pairs of grammatical function terms. Other constants are defined for pairs of numerical terms and for pairs of terms encoding affirmative and negative natural language expressions and logical negation.

Having defined how to compute the quality for single alignment links, we still need to define how to compute the combined score of the alignment links, which we define to be the sum of the qualities of the links:

$$linksScore(a|p_1^n, q_1^m) = \sum_{k=1}^{n} \begin{cases} sim(p_k, q_{a(k)}) & \text{if } a(k) \text{ is defined,} \\ \mu_{NULL} & \text{else.} \end{cases} \quad (4)$$

The quality of a given overall alignment thus is determined by the quality of the alignment links of the PARTS elements which are aligned. For those PARTS elements not aligned, a constant cost μ_{NULL} must be paid, which, however, may be smaller than a costly alignment link in another overall alignment.

4.2 Evaluating Unifiers

Alignments between structurally corresponding semantic elements should be preferred. For situations in which they structurally do not correspond, this may have the effect of dispreferring the pairing of elements which in terms of the words on the surface are identical or very similar. Consider the sentence pair in (2), where *Frau* in (2a) syntactically corresponds to *Mann* in (2b).

[1] In this paper, we simply use grammatical function names in place of semantic role labels in the formulas. A more sophisticated, real mapping from syntactic functions to semantic roles could usefully be incorporated.

(2) a. Eine Frau sieht einen Mann
a woman sees a man
'A woman sees a man.'

b. Ein Mann sieht eine Frau
a man sees a woman
'A man sees a woman.'

On the level of the semantic representation, this is reflected in the correspondence between the variables x_1 and y_1, both of which occur as arguments of *subj*, as shown in Figure 2.

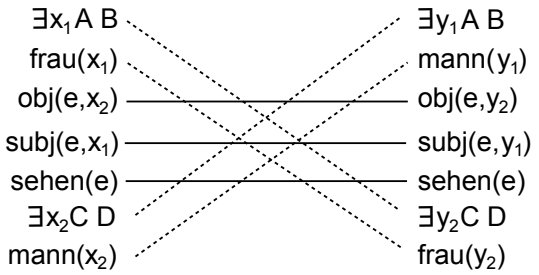

Figure 2: An excerpt of an alignment between the PARTS lists of (2a) on the left and (2b) on the right. Dotted alignment links are the ones only plausible on the surface.

Our solution to capture this distinction is to use the concept of a unifier, well-known from logic programming. A *unifier* for terms ϕ, ψ is a substitution θ such that $\phi\theta = \psi\theta$. Every alignment induces a unifier, which unifies all variables which are matched by the alignment.

The alignment in Figure 2 (without the dotted links) induces the unifier

$\theta_1 = [(x_1, y_1) \mapsto z_1; (x_2, y_2) \mapsto z_2]$.

If links between the matching predicates *mann* and *frau*, respectively, are added, one also has to unify x_1 with y_2 and x_2 with y_1 and thus obtains the unifier

$\theta_2 = [(x_1, x_2, y_1, y_2) \mapsto z]$.

Intuitively, a good unifier unifies only variables which correspond to the same places in the semantic structures to be aligned. In the case of Figure 2, choosing an alignment including the dotted links results in the unifier θ_2 which unifies x_1 and x_2 – yet they are structurally different, with one belonging to the subject and the other one to the object.

In general, it can be expected that an alignment which preserves the structure will not unify two distinct variables from the same LRS representation, since they are known to be structurally distinct. So we want to capture the information loss resulting from unification. This intuition is captured by (5), which answers the following question: Given some variable z in a unified expression, how many additional bits do we need on average[2] to encode the original pair of variables x, y in the PARTS lists p and q, respectively?

$$H(\theta) = \frac{1}{Z_{p,q}} \sum_{z \in Ran(\theta)} W_\theta(z)\, log(W_\theta(z)) \quad (5)$$

$$\text{where} \quad W_\theta(z) = |\{x \in Var(p) | x\theta = z\}| \cdot |\{y \in Var(q) | y\theta = z\}| \quad (6)$$

$$Z_{p,q} = |Var(p)| \cdot |Var(q)| \quad (7)$$

The value of a unifier θ is then defined as follows:

$$variableScore(\theta) = \left(1 - \frac{H(\theta)}{\hat{H}}\right)^k \quad (8)$$

where k is a numerical parameter with $0 \leq k \leq 1$ and $\hat{H}$ is a (tight) upper bound on $H(\theta)$ obtained by evaluating the worst unifier, i.e., the unifier that unifies all variables $\hat{H} = log(Z_{p,q})$.

4.3 Evaluating consistency with dominance constraints

While evaluating unifiers ensures that alignments preserve the structure on the level of variables, it is also important to evaluate their consistency with the dominance structure of the underspecified semantic representations, such as the one we saw in Figure 1. Consider the following pair:

(3) a. Peter kommt und Hans kommt nicht.
Peter comes and Hans comes not
'Peter comes and Hans does not come.'

b. Peter kommt nicht und Hans kommt.
Peter comes not and Hans comes
'Peter does not come and Hans comes.'

While the words and also the PARTS lists of the sentences are identical, they clearly differ in meaning. Figure 3 on the next page shows the LRS dominance graphs for the two sentences together with an

[2]For simplicity, it is assumed that every combination in $Var(p) \times Var(q)$ occurs the same number of times.

alignment between them. The semantic difference between the two sentences is reflected in the position of the negation in the dominance graph: while it dominates *kommen(e_2)* $\wedge$ *subj(e_2,hans)* in (3a), it dominates *kommen(f_1)* $\wedge$ *subj(f_1,peter)* in (3b).

To account for this issue, we evaluate the consistency of the alignment with respect to dominance constraints. An alignment a is optimally consistent with respect to dominance structure if it defines an isomorphism between its range and its domain with respect to the relation $\triangleleft$ 'is dominated by'.

Figure 3 shows an alignment which aligns all matching elements in (3b) and (3a). The link between the negations violates the isomorphism requirement: the negation dominates *kommen(e_2)* $\wedge$ *subj(e_2,hans)* in (3a), while it does not dominate the corresponding elements in (3b). An optimally consistent alignment will thus leave the negations unaligned. Unaligned negations can later be used in the overall meaning comparison as strong evidence that the sentences do not mean the same.

dominanceScore measures how "close" a is to defining an isomorphism. We use the following simple score, which is equal to 1 if and only if a defines an isomorphism:

$$dominanceScore(a|S,T) = \frac{1}{1+\sum_{i,j\in Dom(a)} \kappa \begin{pmatrix} p_i \triangleleft p_j, \\ p_i \triangleright p_j, \\ q_{a(i)} \triangleleft q_{a(j)}, \\ q_{a(i)} \triangleright q_{a(j)} \end{pmatrix}} \quad (9)$$

where κ is a function taking four truth values as its arguments. It measures the extent to which the isomorphism requirement is violated by an alignment. $\kappa(t_1,t_2,t_1,t_2)$ is defined as 0 because there is no violation if the dominance relation between p_i and p_j is equal to that between the elements they are aligned with, $q_{a(i)}$ and $q_{a(j)}$. For other combinations of truth values, κ should be set to values greater than zero, empirically determined on a development set.

4.4 Finding the best alignment

Because of the use of non-local criteria in the maximization criterion $Q(a,\theta|S,T)$ defined in equation (1), an efficient method is needed to find the alignment maximizing the criterion. We exploit the structure inherent in the set of possible alignments to apply the A* algorithm (Russel and Norvig, 2010). We first generalize the notion of an alignment.

Definition 2. A **partial alignment** of order i is an index i together with an alignment which does not have alignment links for any p_j with $j > i$.

A partial alignment can be interpreted as a class of alignments which agree on the first i elements.

Definition 3. The **refinements** $\rho(a)$ of the partial alignment a (of order i) are the partial alignments b such that (1) b is of order $i+1$, and (2) a and b agree on $\{1,...,i\}$.

Intuitively, refinements of an alignment of order i are obtained by deciding how to align element $i+1$. ρ induces a tree over the set of partial alignments, whose leaves are exactly the complete alignments.

A simple optimistic estimate for the value of all complete descendants of an alignment a of order i is given by the following expression:

$$\begin{aligned} optimistic(a,\theta|S,T) = variableScore(\theta) \\ \cdot dominanceScore(a,S,T) \\ \cdot (linksScore_i(a,\theta|p,q) + \\ \sum_{k=i+1}^{n} heuristic(k,a,p_1^n,q_1^m)) \end{aligned} \quad (10)$$

where $linksScore_i$ is the sum in (4) restricted to $1 \leq k \leq i$, and $heuristic(k,a,p_1^n,q_1^m)$ is 0 if p_k is aligned and a simple, optimistic estimate for the quality of the best possible alignment link containing p_k if p_k is unaligned. It is estimated as the maximum of μ_{NULL} and $max\{sim(p_k,q_j) \mid q_j \ unaligned\}$.

The estimate in (10) is optimistic in the sense that it provides an upper bound on the values of all complete alignments below a. It defines a monotone heuristic and thus allows complete and optimal search using the A* algorithm. To obtain an efficient implementation, additional issues such as the order of elements in the PARTS lists were taken care of. As they do not play a role for the conceptualization of our approach, they are not discussed here.

The crucial part at this point of the discussion is that the A* search can determine the best alignment between two PARTS lists. As mentioned in the overview in section 3.3, we compute three such

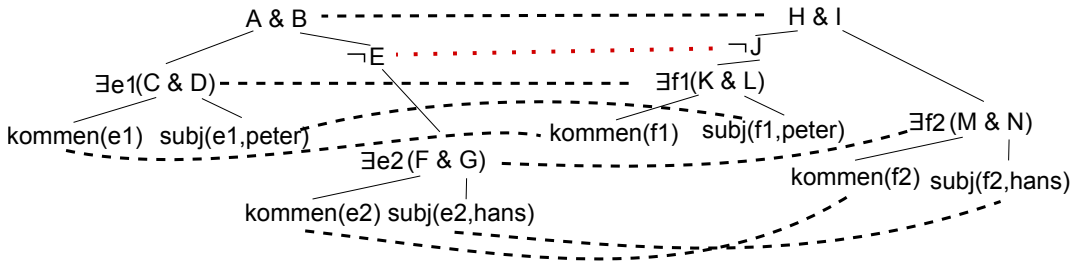

Figure 3: Alignment between the dominance graphs of (3a) and (3b). The red dotted link violates isomorphism.

alignments: between the student and the target answer, between the question and the student answer, and between the question and the target answer.

5 From Alignment to Meaning Comparison

Based on the three alignments computed using the just discussed algorithm, we now explore different options for computing whether the student answer is correct or not. We discuss several alternatives, all involving the computation of a numerical score based on the alignments. For each of these scores, a threshold is empirically determined, over which the student answer is considered to be correct.

Basic Scores The simplest score, ALIGN, is computed by dividing the alignment quality Q between the student answer and the target answer as defined in equation (1) by the number of elements in the smaller PARTS list. Two other scores are computed based on the number of alignment links between student and target answer, which for the EQUAL-Student score is divided by the number of elements of the PARTS list of the *student* answer, and for the EQUAL-Target score by those of the *target* answer.

For dealing with functional elements, i.e., predicates like *subj, obj*, quantifiers and the lambda operator, we tried out three options. The straight case is the one mentioned above, treating all elements on the PARTS list equally (EQUAL). As a second option, to see how important the semantic relations between words are, and how much is just the effect of the elements themselves, we defined a score which ignores functional elements (IGNORE). A third possibility is to weight elements so that functional and non-functional ones differ in impact (WEIGHTED).

Each of the three scores (EQUAL, IGNORE, WEIGHTED) is either divided by the number of elements of the PARTS list of the *student* answer or the *target*, resulting in six scores. In addition, three more scores result from computing the average of the student and target answer scores.

Information Structure Scores Basing meaning comparison on actual semantic representation also allows us to directly take into account Information Structure as a structuring of the meaning of a sentence in relation to the discourse. Bailey and Meurers (2008), Meurers et al. (2011), and Mohler et al. (2011) showed that excluding those parts of the answer which are mentioned (*given*) in the question greatly improves classification accuracy. Meurers et al. (2011) argue that the relevant linguistic aspect is not whether the material was mentioned in the question, but the distinction between *focus* and *background* in Information Structure (Krifka, 2008). The focus essentially is the information in the answer which selects between the set of alternatives that the question raises.

This issue becomes relevant, e.g., in the case of '*or*' questions, where the focused information determining whether the answer is correct is explicitly given in the question. This is illustrated by the question in (4) with target answer (5a) and student answer (5b), from the CREG corpus. While all words in the answers are mentioned in the question, the part of the answers which actually answer the question are the *focused* elements shown in boldface.

(4) Ist die Wohnung in einem Altbau oder
is the flat in a old building or
Neubau?
new building

(5) a. Die Wohnung ist in einem **Altbau**.
the flat is in a old.building

b. Die Wohnung ist in einem **Neubau**.
the flat is in a new.building

To realize a focus-based approach, one naturally needs a component which automatically identifies the focus of an answer in a question-answer pair. As a first approximation, this currently is implemented by a module which marks the elements of the PARTS lists of the answers for information structure. Elements which are not aligned to the question are marked as focused. Furthermore, in answers to '*or*' questions, it marks as focused all elements which are aligned to the semantic contribution of a word belonging to one of the alternatives. '*Or*' questions are recognized by the presence of *oder* ('or') and the absence of a *wh*-word.

While previous systems simply ignored all words given in the question during classification, our system aligns all elements and recognizes givenness based on the alignments. Therefore, givenness is still recognized if the surface realization is different. Furthermore, material which incidentally is also found in the question, but which is structurally different, is not assumed to be given.

Scores using information structure were obtained in the way of the BASIC scores but counting only those elements which are recognized as focused (FOCUS). For comparison, we also used the same scores with givenness detection instead of focus detection, i.e., in these scores, all elements aligned to the question were excluded (GIVEN).

Annotating semantic rather than surface representations for information structure has the advantage that the approach can be extended to cover focusing of relations in addition to focusing of entities. The general comparison approach also is compatible with more sophisticated focus detection techniques capable of integrating a range of cues, including syntactic cues and specialized constructions such as clefts, or prosodic information for spoken language answers – an avenue we intend to pursue in future research.

Dissimilar score We also explored one specialized score paying particular attention to dissimilar aligned elements, as mentioned in section 4.1. Where a focused number is aligned to a different number, or a focused polarity expression is aligned to the opposite polarity, or a logical negation is not aligned, then 0 is returned as score, i.e., the student answer is false. In all other cases, the DISSIMILAR score is identical to the WEIGHTED-Average FOCUS score, i.e., the score based on the average of the student and target scores with weighting and focus detection.

6 Experiments

6.1 Corpus

We base the experiments on the 1032 answers from the CREG corpus which are used in the evaluation of the CoMiC-DE system reported by Meurers et al. (2011). The corpus is balanced, i.e., the numbers of correct and of incorrect answers are the same. It contains only answers where the two human annotators agreed on the binary label.

6.2 Setup

The alignment algorithm contains a set of numerical parameters which need to be determined empirically, such as μ_{NULL} and the function κ. In a first step, we optimized these parameters and the weights used in the WEIGHTED scores using grid search on a development set of 379 answers. These answers are from CREG, but do not belong to the 1032 answers used for testing. We used the accuracy of the DISSIMILAR score as performance metric.

In our experiment, we explored each score separately to predict which answers are correct and which not. For each score, classification is based on a threshold which is estimated as the arithmetic mean of the average score of correct and the average score of incorrect answers. Training and testing was performed using the leave-one-out scheme (Weiss and Kulikowski, 1991). When testing on a particular answer, student answers answering the same question were excluded from training.

6.3 Results

Figure 4 shows the accuracy results obtained in our experiments together with the result of CoMiC-DE on the same dataset. With an accuracy of up to 86.3%, the WEIGHTED-Average FOCUS score outperform the 84.6% reported for CoMiC-DE (Meurers et al., 2011) on the same dataset. This is remarkable given that CoMiC-DE uses several (but comparably shallow) levels of linguistic abstraction for finding alignment links, whereas our approach is exclusively based on the semantic representations.

Score	BASIC	GIVEN	FOCUS
ALIGN	77.1		
EQUAL			
Student	69.8	75.3	75.2
Target	70.0	75.5	75.2
Average	76.6	80.8	80.7
IGNORE			
Student	75.8	80.1	80.3
Target	77.2	82.2	82.3
Average	79.8	84.7	84.9
WEIGHTED			
Student	75.0	80.6	80.7
Target	76.1	83.3	83.3
Average	80.9	86.1	**86.3**
DISSIMILAR	85.9		
CoMiC-DE	84.6		

Figure 4: Classification accuracy of CoSeC-DE

The fact that WEIGHTED-Average outperforms the IGNORE-Average scores shows that the inclusion of functional element (i.e., predicates like *subj, obj*), which are not available to approaches based on aligning surface strings, improves the accuracy.[3] On the other hand, the lower performance of EQUAL shows that functional elements should be treated differently from content-bearing elements.

Of the 13.7% answers misclassified by WEIGHTED-Average FOCUS, 53.5% are false negatives and 46.5% are false positives.

We also investigated the impact of grammaticality on the result by manually annotating a sample of 220 student answers for grammatical well-formedness, 66% of which were ungrammatical. On this sample, grammatical and ungrammatical student answers were evaluated with essentially the same accuracy (83% for ungrammatical answers, 81% for grammatical answers).

The decrease in accuracy of the COMBINED score over the best score can be traced to some yes-no-questions which have an unaligned negation but are correct. On the other hand, testing only on answers with focused numbers results in an accuracy of 97%.

The performance of GIVEN and FOCUS scores compared to BASIC confirms that information structuring helps in targeting the relevant parts of the answers. Since CoMiC-DE also demotes given material, the better GIVEN results of our approach must result from other aspects than the information structure awareness. Unlike previous approaches, the FOCUS scores support reference to the material focused in the answers. However, since currently the FOCUS scores only differs from the GIVEN scores for alternative questions, and the test corpus only contains seven answers to such '*or*' questions, we see no serious quantitative difference in accuracy between the FOCUS and GIVENNESS results.

While the somewhat lower accuracy of the score ALIGN shows that the alignment scores are not sufficient for classification, the best-performing scores do not require much additional computation and do not need any information that is not in the alignments or the automatic focus annotation.

7 Future Work

The alert reader will have noticed that our approach currently does not support many-to-many alignments. As is known, e.g., from phrase-based machine translation, this is an interesting avenue for dealing with non-compositional expressions, which we intend to explore in future work. The alignment approach can be adapted to such alignments by adding a factor measuring the quality of many-to-many links to $linkScore$ (4) and $optimistic$ (10).

8 Conclusion

We presented the CoSeC-DE system for evaluating the content of answers to reading comprehension questions. Unlike previous content assessment systems, it is based on formal semantics, using a novel approach for aligning underspecified semantic representations. The approach readily supports the integration of important information structural differences in a way that is closely related to the information structure research in formal semantics and pragmatics. Our experiments showed the system to outperform our shallower multi-level system CoMiC-DE on the same CREG-1032 data set, suggesting that formal semantic representations can indeed be useful for content assessment in real-world contexts.

[3] We also evaluated IGNORE scores using parameter values optimized for these scores, but their performance was still below those of the corresponding WEIGHTED-Average scores.

Acknowledgements

We are grateful to the three anonymous BEA reviewers for their very encouraging and helpful comments.

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.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. In J. Quionero-Candela, I. Dagan, B. Magnini, and F. d'Alch Buc, editors, *Machine Learning Challenges*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

Aria D. Haghighi, Andrew Y. Ng, and Christopher D. Manning. 2005. Robust textual inference via graph matching. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 387–394. Association for Computational Linguistics.

Michael Hahn and Detmar Meurers. 2011. On deriving semantic representations from dependencies: A practical approach for evaluating meaning in learner corpora. In Kim Gerdes, Eva Hajicov, and Leo Wanner, editors, *Depling 2011 Proceedings*, pages 94–103, Barcelona.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

Manfred Krifka. 2008. Basic notions of information structure. *Acta Linguistica Hungarica*, 55(3):243–276.

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

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

Bill MacCartney, Michel Galley, and Christopher D. Manning. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 802–811. Association for Computational Linguistics.

Detmar Meurers, Ramon Ziai, Niels Ott, and Janina Kopp. 2011. 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. Association for Computational Linguistics.

Tom Mitchell, Terry Russell, Peter Broomhead, and Nicola Aldridge. 2002. Towards robust computerised marking of free-text responses. In *Proceedings of the 6th International Computer Assisted Assessment (CAA) Conference*.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th Annual Meeting of the Association for Comnputational Linguistics*, pages 752–762.

Richard Montague. 1973. The Proper Treatment of Quntification in Ordinary English. In Jaakko Hintikka, Julius Moravcsik, and Patrick Suppes, editors, *Approaches to Natural Language*, pages 221–242. Reidel, Dordrecht.

Rodney D. Nielsen, Wayne Ward, and James H. Martin. 2009. Recognizing entailment in intelligent tutoring systems. *Natural Language Engineering*, 15(4):479–501.

Joakim Nivre and Johan Hall. 2005. Maltparser: A language-independent system for data-driven dependency parsing. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories*, pages 13–95.

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 Wrner, editors, *Multilingual Corpora and Multilingual Corpus Analysis*, Hamburg Studies in Multilingualism (HSM). Benjamins, Amsterdam.

Stephen G. Pulman and Jana Z. Sukkarieh. 2005. Automatic short answer marking. In *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP*, pages 9–16.

Frank Richter and Manfred Sailer. 2003. Basic Concepts of Lexical Resource Semantics. In Arnold Beckmann and Norbert Preining, editors, *ESSLLI 2003 – Course Material I*, volume 5 of *Collegium Logicum*, pages 87–143, Wien. Kurt Gödel Society.

Frank Richter. 2000. *A Mathematical Formalism for Linguistic Theories with an Application in Head-Driven Phrase Structure Grammar*. Phil. dissertation, Eberhard-Karls-Universität Tüingen.

Vasile Rus, Arthur Graesser, and Kirtan Desai. 2007. Lexico-syntactic subsumption for textual entailment.

Recent Advances in Natural Language Processing IV: Selected Papers frp, RANLP 2005, pages 187–196.

Stuart Russel and Peter Norvig. 2010. *Artificial Intelligence. A Modern Approach*. Pearson, 2nd edition.

Mark Sammons, V.G.Vinod Vydiswaran, Tim Vieira, Nikhil Johri, Ming-Wei Chang, Dan Goldwasser, Vivek Srikumar, Gourab Kundu, Yuancheng Tu, Kevin Small, Joshua Rule, Quang Do, and Dan Roth. 2009. Relation Alignment for Textual Entailment Recognition. In *Text Analysis Conference (TAC)*.

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.

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

Sholom M. Weiss and Casimir A. Kulikowski. 1991. *Computer systems that learn*. Morgan Kaufmann, San Mateo, CA.

Ramon Ziai, Niels Ott, and Detmar Meurers. 2012. Short answer assessment: Establishing links between research strands. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-7) at NAACL-HLT 2012*, Montreal.