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

Michael Hahn    Detmar Meurers

Department of Linguistics, SFB 833
University of Tübingen

BEA 7 Workshop, NAACL-HLT
Montreal, 7. June 2012
Introduction

- A range of approaches have been proposed for short answer meaning assessment (Ziai, Ott & Meurers 2012).
Introduction

▶ A range of approaches have been proposed for short answer meaning assessment (Ziai, Ott & Meurers 2012).
▶ Meaning comparison generally relies on a combination of surface-based and deeper linguistic representations,
  ▶ but essentially no use is made of semantic formalisms created by theoretical linguists to represent meaning.
Introduction

- A range of approaches have been proposed for short answer meaning assessment (Ziai, Ott & Meurers 2012).
- Meaning comparison generally relies on a combination of surface-based and deeper linguistic representations,
  - but essentially no use is made of semantic formalisms created by theoretical linguists to represent meaning.
    - deep linguistic analysis of formal semantics often lacks coverage and robustness
    - semantic structures are complex to derive and compare
Introduction

- A range of approaches have been proposed for short answer meaning assessment (Ziai, Ott & Meurers 2012).
- Meaning comparison generally relies on a combination of surface-based and deeper linguistic representations, but essentially no use is made of semantic formalisms created by theoretical linguists to represent meaning.
  - deep linguistic analysis of formal semantics often lacks coverage and robustness
  - semantic structures are complex to derive and compare
  + semantic representations abstract away from lexical and syntactic variation in the realization of the same meaning
  + they precisely expose meaning distinctions and support linking meaning to discourse
A range of approaches have been proposed for short answer meaning assessment (Ziai, Ott & Meurers 2012).

Meaning comparison generally relies on a combination of surface-based and deeper linguistic representations, but essentially no use is made of semantic formalisms created by theoretical linguists to represent meaning.

- deep linguistic analysis of formal semantics often lacks coverage and robustness
- semantic structures are complex to derive and compare
+ semantic representations abstract away from lexical and syntactic variation in the realization of the same meaning
+ they precisely expose meaning distinctions and support linking meaning to discourse

We present a short answer assessment approach based on underspecified formal semantic representations.
Outline

General setup
  Corpus of Reading Comprehension Exercises in German
  Lexical Resource Semantics representations
  Our general approach

Aligning Meaning Representations

From Alignment to Meaning Comparison

Experiments

Conclusion
Empirical challenge: CREG

- Empirical basis: Corpus of Reading Comprehension Exercises in German (CREG; Ott, Ziai & Meurers 2012)
  - CREG consists of texts, questions, target answers, and student answers written by learners of German.
Empirical challenge: CREG

- Empirical basis: Corpus of Reading Comprehension Exercises in German (CREG; Ott, Ziai & Meurers 2012)
  - CREG consists of texts, questions, target answers, and student answers written by learners of German.
- CREG data was collected and assessed in two large German programs in the US: KU and OSU
  - For each student answer, two independent annotators evaluated whether it correctly answers the question.
    - Answers were only assessed with respect to meaning, not orthography or grammaticality.
Empirical challenge: CREG

- Empirical basis: Corpus of Reading Comprehension Exercises in German (CREG; Ott, Ziai & Meurers 2012)
  - CREG consists of texts, questions, target answers, and student answers written by learners of German.

- CREG data was collected and assessed in two large German programs in the US: KU and OSU
  - For each student answer, two independent annotators evaluated whether it correctly answers the question.
    - Answers were only assessed with respect to meaning, not orthography or grammaticality.

- Data freely available, and reference results available for CoMiC-DE system (Meurers, Ziai, Ott & Kopp 2011),
  - a system not using formal semantic representations
Representations: Lexical Resource Semantics

- LRS (Richter & Sailer 2003) is an underspecified semantic formalism:
  - standard model-theoretic semantics
  - semantic representations are not completely specified but subsume a set of possible resolved expressions
Representations: Lexical Resource Semantics

- LRS (Richter & Sailer 2003) is an underspecified semantic formalism:
  - standard model-theoretic semantics
  - semantic representations are not completely specified but subsume a set of possible resolved expressions

- Advantage of an underspecified semantic formalism for content assessment:
  - provides access to fine-grained semantic distinctions
    - all parts of the semantic representation are accessible in a flat representation
  - how the parts are combined is separately encoded (variable bindings, dominance)
    - avoids costly computation of all readings
    - similar parts can be compared independent of where they appear in the overall semantics
Example for LRS representation

(1) *Alle Zimmer haben nicht eine Dusche.*

all rooms have not a shower

‘Not every room has a shower.’ vs. ‘No room has a shower.’
Evaluating Meaning of RC Answers: A Semantics-Based Approach

Michael Hahn, Detmar Meurers

General setup
CREG as empirical challenge

LRS representations
Our general approach

Aligning meaning representations
Maximization Criterion
Alignment links
Unifiers
Consistency with dominance constraints
Finding the best alignment

From alignment to meaning comparison
Basic measures
Functional elements
Information structure

Experiments
Setup
Results

Conclusion

Example for LRS representation

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

- INTERNAL CONTENT: core semantic contribution of head
- EXTERNAL CONTENT: semantic representation of sentence
- PARTS: all subterms of the representation

\[
\begin{bmatrix}
\text{INCONT} & \text{have(e)} \\
\text{EXCONT} & A \\
\text{PARTS} & A, \text{have(e)}, \forall x_1 (B \rightarrow C), \exists x_2 (D \land E), \neg F, \\
& \text{room}(x_1), \text{shower}(x_2), \text{subj(e,x}_1), \text{obj(e,x}_2) \\
& \exists e(\text{have(e) \land subj(e,x}_1 \land \text{obj(e,x}_2))
\end{bmatrix}
\]
The readings of the sentence are obtained by identifying the meta-variables A, . . . , F with the subformulas.
Example for LRS representation (cont.)

► The readings of the sentence are obtained by identifying the meta-variables A, . . . , F with the subformulas.
► LRS representations include dominance constraints, which restrict possible identifications, e.g.:

\[
\forall x_1 (B \rightarrow C) \quad \exists x_2 (D \land E) \quad \neg F
\]

room(x1) shower(x2) \exists e (\text{have}(e) \land \text{subj}(e, x1) \land \text{obj}(e, x2))
Our general approach

1. automatically derive LRS representations for the student answer, the target answer, and the question
   ▶ method described in Hahn & Meurers (2011)
   ▶ based on statistical dependency parsing
   ▶ always results in an LRS structure, also for ill-formed input
Our general approach

1. automatically derive LRS representations for the student answer, the target answer, and the question
   ▶ method described in Hahn & Meurers (2011)
   ▶ based on statistical dependency parsing
   ▶ always results in an LRS structure, also for ill-formed input

2. align LRS representations of target and student answers
   ▶ *local* measures of semantic similarity
   ▶ *global* measures of extent to which alignment preserves semantic structure (variable bindings, dominance)
Our general approach

1. automatically derive LRS representations for the student answer, the target answer, and the question
   ▶ method described in Hahn & Meurers (2011)
   ▶ based on statistical dependency parsing
   ▶ always results in an LRS structure, also for ill-formed input

2. align LRS representations of target and student answers
   ▶ local measures of semantic similarity
   ▶ global measures of extent to which alignment preserves semantic structure (variable bindings, dominance)

alignments also computed between answers and question
Our general approach

1. automatically derive LRS representations for the student answer, the target answer, and the question
   - method described in Hahn & Meurers (2011)
   - based on statistical dependency parsing
   - always results in an LRS structure, also for ill-formed input

2. align LRS representations of target and student answers
   - local measures of semantic similarity
   - global measures of extent to which alignment preserves semantic structure (variable bindings, dominance)
   - alignments also computed between answers and question

3. perform overall meaning comparison based on numerical scores representing quality of alignment
Aligning Meaning Representations

▶ An alignment between two LRS representations is a bijective partial mapping between PARTS lists $p_1^n$ and $q_m^m$
  ▶ Every element of one representation can be aligned to at most one element of the other representation.
Aligning Meaning Representations

An alignment between two LRS representations is a bijective partial mapping between PARTS lists $p_1^n$ and $q_1^m$.

Every element of one representation can be aligned to at most one element of the other representation.

A simple example: “John left.” vs. “Jon vanished.”

\[
\begin{align*}
\exists x & \ A \ B \ \exists y & \ E \ F \\
john(x) & \ \exists e & \ C \ D \ \exists f & \ G \ H \\
subj(e,x) & \ \exists leave(e) & \\
subj(f,y) & \ \exists vanish(f)
\end{align*}
\]
Aligning Meaning Representations

- An alignment between two LRS representations is a bijective partial mapping between PARTS lists $p_i^n$ and $q_j^m$.
  - Every element of one representation can be aligned to at most one element of the other representation.

- A simple example: “John left.” vs. “Jon vanished.”
  
  $\exists x \ A \ B \ \exists y \ E \ F$
  
  john(x) $
  \exists e \ C \ D \ \exists f \ G \ H$

  subj(e, x) $
  \text{leave(e)} \ 
  \text{subj(f,y)} \ 
  \text{vanish(f)}$

- Best alignment is determined automatically using a maximization criterion.
Automatically Deriving Alignments
Maximization criterion

- combines three measures of alignment quality:
  - **LinkScore**: similarity of the alignment links
  - **VariableScore**: consistency of alignments with respect to the induced variable bindings $\theta$
  - **DominanceScore**: consistency with respect to dominance constraints

- $Q(a, \theta|S, T) = LinkScore(a|S, T) \cdot VariableScore(\theta) \cdot DominanceScore(a|S, T)$

- The alignment maximizing the criterion is found efficiently using the A* algorithm.
LinkScore: Similarity for Alignment Links

- Base cases:
  - Variables can be matched with any variable of same type.
  - For other semantic terms, compute the maximum score of:
    - Levenshtein distance, to account for spelling errors
    - Synonyms: score 1 if in GermaNet (Hamp & Feldweg 1997)
    - Dissimilar elements of same category: constant costs, empirically determined for pairs of
      - grammatical function terms
      - special terms (affirmative or negative natural language expressions and logical negation)
  - ...

"SFB 833"
LinkScore: Similarity for Alignment Links

- Base cases:
  - Variables can be matched with any variable of same type.
  - For other semantic terms, compute the maximum score of:
    - Levenshtein distance, to account for spelling errors
    - Synonyms: score 1 if in GermaNet (Hamp & Feldweg 1997)
    - Dissimilar elements of same category: constant costs, empirically determined for pairs of
      - grammatical function terms
      - special terms (affirmative or negative natural language expressions and logical negation)
  - ...

- Complex expressions are compared recursively.
LinkScore: Similarity for Alignment Links

- Base cases:
  - Variables can be matched with any variable of same type.
  - For other semantic terms, compute the maximum score of:
    - Levenshtein distance, to account for spelling errors
    - Synonyms: score 1 if in GermaNet (Hamp & Feldweg 1997)
    - Dissimilar elements of same category: constant costs, empirically determined for pairs of:
      - grammatical function terms
      - special terms (affirmative or negative natural language expressions and logical negation)
    - ...

- Complex expressions are compared recursively.

⇒ Overall LinkScore = sum of similarity of all alignment links

- unaligned elements: constant cost $\mu_{NULL}$ (may be smaller than costly alignment link in another overall alignment)
VariableScore

▶ Every alignment induces a unifier, which unifies all variables which are matched by the alignment.
Every alignment induces a unifier, which unifies all variables which are matched by the alignment.

“A woman sees a man.” vs. “A man sees a woman.”

- line links require unifying $x_1$ with $y_1$ and $x_2$ with $y_2$
- adding the dotted links would result in unifying all variables
VariableScore

- Every alignment induces a unifier, which unifies all variables which are matched by the alignment.

- “A woman sees a man.” vs. “A man sees a woman.”

- $\exists x_1 A B$
  - $\text{woman}(x_1)$
  - $\text{obj}(e, x_2)$
  - $\text{subj}(e, x_1)$
  - $\text{see}(e)$
  - $\exists x_2 C D$
  - $\text{man}(x_2)$

- $\exists y_1 A B$
  - $\text{man}(y_1)$
  - $\text{obj}(e, y_2)$
  - $\text{subj}(e, y_1)$
  - $\text{see}(e)$
  - $\exists y_2 C D$
  - $\text{woman}(y_2)$

- line links require unifying $x_1$ with $y_1$ and $x_2$ with $y_2$
- adding the dotted links would result in unifying all variables

- An alignment which preserves the structure will not unify two distinct variables from the same LRS representation.

$\Rightarrow$ **VariableScore** = information loss resulting from unification
Mismatches in the structure of the linked semantic representations need to be taken into account.

For example:

(2) a. *Peter will come but Hans will not come.*

b. *Peter will not come but Hans will come.*

\[ \text{DominanceScore} = \text{extent to which an alignment defines an isomorphism} \]
From Alignment to Meaning Comparison

Basic measures

- Based on the best overall alignment identified using A*, we compute several measures for meaning comparison.

- **ALIGN** measure, based on alignment quality Q:

  
  \[
  \text{ALIGN} = \frac{\text{alignment quality(\text{student answer}, \text{target answer})}}{\# \text{ of elements in shorter parts list}}
  \]
From Alignment to Meaning Comparison

Basic measures

- Based on the best overall alignment identified using A*, we compute several measures for meaning comparison.

  - **ALIGN** measure, based on alignment quality Q:
    \[
    \text{ALIGN} = \frac{\text{alignment quality}(\text{student answer, target answer})}{\text{# of elements in shorter parts list}}
    \]

  - **EQUAL** measure, based on number of alignment links:
    \[
    \begin{align*}
    \text{Student} &= \frac{\text{# of alignment links}(\text{student answer, target answer})}{\text{# of elements on parts list of student answer}} \\
    \text{Target} &= \frac{\text{# of alignment links}(\text{student answer, target answer})}{\text{# of elements of parts list of target answer}} \\
    \text{Average} &= \text{average of Student and Target measures}
    \end{align*}
    \]
From Alignment to Meaning Comparison
Studying Impact of Functional Elements

- **EQUAL** measures treat all semantic elements the same
From Alignment to Meaning Comparison
Studying Impact of Functional Elements

- **EQUAL** measures treat all semantic elements the same
- Define measures to help identify the impact of functional elements (quantifiers, lambda operator, subj, obj, ...):
From Alignment to Meaning Comparison
Studying Impact of Functional Elements

- **EQUAL** measures treat all semantic elements the same.
- Define measures to help identify the impact of functional elements (quantifiers, lambda operator, *subj*, *obj*, ...):
  - **IGNORE** measures: ignore all functional elements.
  - **WEIGHTED** measures: weight elements so that functional and non-functional ones differ in impact.
    - weights are empirically determined using grid search on a development set.

General setup:
- CREG as empirical challenge
- LRS representations
- Our general approach

Aligning meaning representations:
- Maximization Criterion
- Alignment links
- Unifiers
- Consistency with dominance constraints
- Finding the best alignment

From alignment to meaning comparison:
- Basic measures
- Functional elements
- Information structure

Experiments:
- Setup
- Results

Conclusion
From Alignment to Meaning Comparison
Studying Impact of Information Structure

- Information Structure (Krifka 2008): structuring of the meaning of a response in relation to the discourse
  - *given* (vs. *new*): part of meaning known from question
  - *focus* (vs. *background*): part of meaning selecting between the set of alternatives that the question raises

- Basing meaning comparison on semantic representation allows us to directly represent Information Structure.
- Some previous approaches exclude *given* material from alignment (Bailey & Meurers 2008; Mohler et al. 2011):
  - greatly improves classification accuracy
- Meurers, Ziai, Ott & Kopp (2011) show that the relevant linguistic aspect here is not *givenness* but *focus*. 
From Alignment to Meaning Comparison
Studying Impact of Information Structure

- Information Structure (Krifka 2008): structuring of the meaning of a response in relation to the discourse
  - given (vs. new): part of meaning known from question
  - focus (vs. background): part of meaning selecting between the set of alternatives that the question raises

- Basing meaning comparison on semantic representation allows us to directly represent Information Structure.
Information Structure (Krifka 2008): structuring of the meaning of a response in relation to the discourse

- *given* (vs. *new*): part of meaning known from question
- *focus* (vs. *background*): part of meaning selecting between the set of alternatives that the question raises

Basing meaning comparison on semantic representation allows us to directly represent Information Structure.

Some previous approaches exclude *given* material from alignment (Bailey & Meurers 2008; Mohler et al. 2011):

- greatly improves classification accuracy
Information Structure (Krifka 2008): structuring of the meaning of a response in relation to the discourse
  ▶ *given* (vs. *new*): part of meaning known from question
  ▶ *focus* (vs. *background*): part of meaning selecting between the set of alternatives that the question raises

Basing meaning comparison on semantic representation allows us to directly represent Information Structure.

Some previous approaches exclude *given* material from alignment (Bailey & Meurers 2008; Mohler et al. 2011):
  ▶ greatly improves classification accuracy

Meurers, Ziai, Ott & Kopp (2011) show that the relevant linguistic aspect here is not *givenness* but *focus*. 
From Alignment to Meaning Comparison
How we integrate information structure

- Needed: A component which automatically identifies the focus of an answer in a question-answer pair.
  - First approximation: an element on the parts lists of an answer is marked as focused if it is not aligned to the question, except for alignment with explicit alternatives.
From Alignment to Meaning Comparison
How we integrate information structure

- Needed: A component which automatically identifies the focus of an answer in a question-answer pair.
  - First approximation: an element on the parts lists of an answer is marked as focused if it is not aligned to the question, except for alignment with explicit alternatives.

- **FOCUS** measures: BASIC measures counting only those elements which are recognized as focused

- **GIVEN** measures: BASIC measures counting only elements not aligned to the question
Experiments

Setup

▶ Corpus

▶ 1032 answers from the CREG corpus, used for evaluating the CoMiC-DE system (Meurers, Ziai, Ott & Kopp 2011)
▶ balanced: same number of correct and incorrect answers
Experiments

Setup

- Corpus
  - 1032 answers from the CREG corpus, used for evaluating the CoMiC-DE system (Meurers, Ziai, Ott & Kopp 2011)
  - balanced: same number of correct and incorrect answers

- Preparation
  - optimized numerical parameters using grid search on a separate development set of 379 answers from CREG
Experiments

Setup

- Corpus
  - 1032 answers from the CREG corpus, used for evaluating the CoMiC-DE system (Meurers, Ziai, Ott & Kopp 2011)
  - balanced: same number of correct and incorrect answers

- Preparation
  - optimized numerical parameters using grid search on a separate development set of 379 answers from CREG

- Experiment
  - explored all measures for meaning assessment
  - binary classification is based on a threshold
    - arithmetic mean of the average result of correct and the average result of incorrect answers
  - training and testing performed using the leave-one-out scheme Weiss & Kulikowski (1991)
Results

<table>
<thead>
<tr>
<th></th>
<th>BASIC</th>
<th>GIVEN</th>
<th>FOCUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>WEIGHTED Average</td>
<td>80.9</td>
<td>86.1</td>
<td><strong>86.3</strong></td>
</tr>
<tr>
<td>IGNORE Average</td>
<td>79.8</td>
<td>84.7</td>
<td>84.9</td>
</tr>
<tr>
<td>EQUAL Average</td>
<td>76.6</td>
<td>80.8</td>
<td>80.7</td>
</tr>
<tr>
<td>ALIGN</td>
<td>77.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CoMiC-DE</td>
<td></td>
<td></td>
<td><strong>84.6</strong></td>
</tr>
</tbody>
</table>

- Best accuracy with WEIGHTED Average FOCUS measure
### Results

<table>
<thead>
<tr>
<th></th>
<th>BASIC</th>
<th>GIVEN</th>
<th>FOCUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>WEIGHTED Average</td>
<td>80.9</td>
<td>86.1</td>
<td><strong>86.3</strong></td>
</tr>
<tr>
<td>IGNORE Average</td>
<td>79.8</td>
<td>84.7</td>
<td><strong>84.9</strong></td>
</tr>
<tr>
<td>EQUAL Average</td>
<td>76.6</td>
<td>80.8</td>
<td><strong>80.7</strong></td>
</tr>
<tr>
<td>ALIGN</td>
<td>77.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CoMiC-DE</td>
<td></td>
<td>84.6</td>
<td></td>
</tr>
</tbody>
</table>

- Best accuracy with WEIGHTED Average FOCUS measure
- Including functional elements improves accuracy (+1.4%)
Results

<table>
<thead>
<tr>
<th></th>
<th>BASIC</th>
<th>GIVEN</th>
<th>FOCUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>WEIGHTED Average</td>
<td>80.9</td>
<td>86.1</td>
<td><strong>86.3</strong></td>
</tr>
<tr>
<td>IGNORE Average</td>
<td>79.8</td>
<td>84.7</td>
<td>84.9</td>
</tr>
<tr>
<td>EQUAL Average</td>
<td>76.6</td>
<td>80.8</td>
<td><strong>80.7</strong></td>
</tr>
<tr>
<td>ALIGN</td>
<td>77.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CoMiC-DE</td>
<td></td>
<td></td>
<td><strong>84.6</strong></td>
</tr>
</tbody>
</table>

- Best accuracy with WEIGHTED Average FOCUS measure
- Including functional elements improves accuracy (+1.4%)
  - weight should differ from content elements (+5.6%)
Results

<table>
<thead>
<tr>
<th></th>
<th>BASIC</th>
<th>GIVEN</th>
<th>FOCUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>WEIGHTED Average</td>
<td>80.9</td>
<td>86.1</td>
<td>86.3</td>
</tr>
<tr>
<td>IGNORE Average</td>
<td>79.8</td>
<td>84.7</td>
<td>84.9</td>
</tr>
<tr>
<td>EQUAL Average</td>
<td>76.6</td>
<td>80.8</td>
<td>80.7</td>
</tr>
<tr>
<td>ALIGN</td>
<td>77.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CoMiC-DE</td>
<td></td>
<td>84.6</td>
<td></td>
</tr>
</tbody>
</table>

- Best accuracy with WEIGHTED Average FOCUS measure
- Including functional elements improves accuracy (+1.4%)
  - weight should differ from content elements (+5.6%)
- Information Structure
  - Focus helps target relevant part of answer (+5.4%)
Results

<table>
<thead>
<tr>
<th></th>
<th>BASIC</th>
<th>GIVEN</th>
<th>FOCUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>WEIGHTED Average</td>
<td>80.9</td>
<td>86.1</td>
<td>86.3</td>
</tr>
<tr>
<td>IGNORE Average</td>
<td>79.8</td>
<td>84.7</td>
<td>84.9</td>
</tr>
<tr>
<td>EQUAL Average</td>
<td>76.6</td>
<td>80.8</td>
<td>80.7</td>
</tr>
<tr>
<td>ALIGN</td>
<td>77.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CoMiC-DE</td>
<td></td>
<td>84.6</td>
<td></td>
</tr>
</tbody>
</table>

- Best accuracy with WEIGHTED Average FOCUS measure
- Including functional elements improves accuracy (+1.4%)
  - weight should differ from content elements (+5.6%)
- Information Structure
  - Focus helps target relevant part of answer (+5.4%)
- Outperforms CoMiC-DE, also integrating givenness
  - supports usefulness of semantic representations

SFB 833

▶ Best accuracy with WEIGHTED Average FOCUS measure
▶ Including functional elements improves accuracy (+1.4%)
  ▶ weight should differ from content elements (+5.6%)
▶ Information Structure
  ▶ Focus helps target relevant part of answer (+5.4%)
▶ Outperforms CoMiC-DE, also integrating givenness
  ▶ supports usefulness of semantic representations
Results

Experiment testing impact of grammaticality

- We manually annotated 220 student answers for grammatical well-formedness.
  - 66% were ungrammatical
- Accuracy on this sample:
  - 83% for ungrammatical answers
  - 81% for grammatical answers

⇒ semantics-based approaches can be robust, not directly linked to grammaticality
Conclusion

- We presented a system for evaluating the content of answers to reading comprehension questions.
- Unlike previous content assessment systems, it is based on comparing formal semantic representations.
  - integrates a novel approach for comparing underspecified semantic representations
- Formal semantic representations readily support the integration of information structural differences.
  - connects content-assessment to information structure research in formal semantics and pragmatics
- The system presented outperforms our shallower CoMiC-DE system on the same CREG data set.
  - formal semantic representations can be competitive for content assessment in real-world contexts
Evaluating Meaning of RC Answers: A Semantics-Based Approach

Michael Hahn, Detmar Meurers

General setup
CREG as empirical challenge
LRS representations
Our general approach
Aligning meaning representations
Maximization Criterion
Alignment links
Unifiers
Consistency with dominance constraints
Finding the best alignment

From alignment to meaning comparison
Basic measures
Functional elements
Information structure

Experiments
Setup
Results

Conclusion

References


### Full Results

<table>
<thead>
<tr>
<th>Measure</th>
<th>BASIC</th>
<th>GIVEN</th>
<th>FOCUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALIGN</td>
<td>77.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EQUAL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>69.8</td>
<td>75.3</td>
<td>75.2</td>
</tr>
<tr>
<td>Target</td>
<td>70.0</td>
<td>75.5</td>
<td>75.2</td>
</tr>
<tr>
<td>Average</td>
<td>76.6</td>
<td>80.8</td>
<td>80.7</td>
</tr>
<tr>
<td>IGNORE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>75.8</td>
<td>80.1</td>
<td>80.3</td>
</tr>
<tr>
<td>Target</td>
<td>77.2</td>
<td>82.2</td>
<td>82.3</td>
</tr>
<tr>
<td>Average</td>
<td>79.8</td>
<td>84.7</td>
<td>84.9</td>
</tr>
<tr>
<td>WEIGHTED</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>75.0</td>
<td>80.6</td>
<td>80.7</td>
</tr>
<tr>
<td>Target</td>
<td>76.1</td>
<td>83.3</td>
<td>83.3</td>
</tr>
<tr>
<td>Average</td>
<td>80.9</td>
<td>86.1</td>
<td><strong>86.3</strong></td>
</tr>
<tr>
<td>CoMiC-DE</td>
<td>84.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
From Alignment to Meaning Comparison
Information Structure: Example for Focus vs. Given

► Alternative questions: focused information determining whether answer is correct is explicitly given in question.

(3) *Ist die Wohnung in einem Altbau oder Neubau?*
    is the flat in a old house or new house

    the flat is in a old house

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

► All words in answers mentioned in the question, but some are **focused**, shown in boldface.