Towards Focus Detection in Content Assessment

Ramon Ziai
Universität Tübingen
SFB 833, Project A4

Second TüBerlin Workshop on the Analysis of Learner Language
Tübingen, Dec 5, 2011
Outline

Introduction
  Project Background
  Empirical Basis

IS in Content Assessment
  Given/New
  Focus/Background

Annotation Experiment
  Data Selection
  Annotation Criteria
  Results

Conclusions

References
What is Content Assessment?

- The task of evaluating an answer to a question with respect to meaning given a concrete linguistic context.
- Context here means: Explicit question and reading text in Reading Comprehension data.
- Answers to RC questions are produced by language learners, language can be ill-formed and requires robust computational processing.
Content Assessment: A Simple Example
Meurers, Ziai, Ott & Kopp (2011b)

Q: Was sind die Kritikpunkte, die Leute über Hamburg äußern?
‘What are the objections people have about Hamburg?’

TA: Der Gestank von Fisch und Schiffsdiesel an den Kais.
The stink of fish and fuel on the quays.

<table>
<thead>
<tr>
<th>SemType</th>
<th>Spelling</th>
<th>Spelling</th>
<th>Ident</th>
<th>Ident</th>
<th>Similarity</th>
</tr>
</thead>
</table>

SA: Der Geruch von Fish und Schiffsdiesel beim Hafen.
The smell of fish and fuel at the port.
Together with Ohio State University and Kansas University we are collecting a corpus of
- reading texts,
- questions on the reading texts,
- student answers to the questions, and
- target answers pre-specified by teachers.

All student answers are rated by two annotators at OSU and KU with respect to meaning.
- Binary: correct, incorrect
- Detailed: correct answer, missing concept, extra concept, blend, non-answer
CREG: Numbers and Sizes

- Snapshot from September 6, 2011:
  - 118 Texts
  - 752 Questions
  - 1,059 Target Answers
  - 20,851 Student Answers
Today: IS of answers to RC questions

▶ So far, our approach has no notion of information required by the question, it only compares student and target answers

▶ Comparable systems such as C-rater (Leacock & Chodorow 2003) or Willow (Pérez Marin 2007) also don’t take the question into account

▶ However, questions naturally impose IS requirements on answers

▶ Research questions:
  ▶ What are the necessary IS notions? Given/New or Focus/Background?
  ▶ Can instances of them be identified reliably?
In order for a system to distinguish between potentially relevant and irrelevant content, a partitioning of the answers would be useful.

The system can only recognize previously mentioned lexical material so far (simple Givenness).

This simple case of Givenness is easy to detect, but does it suffice for the task?
Given/New: Example where it works

<table>
<thead>
<tr>
<th>Question:</th>
<th>Welche kulturelle Personlichkeit ist für die Salzburger am wichtigsten?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target Answer:</td>
<td>Mozart [ist für die Salzburger am wichtigsten]_{G}.</td>
</tr>
<tr>
<td>Student Answer:</td>
<td>Mozart [ist die kulturelle Persönlichkeit für die Salzburger an wichtigsten]_{G}.</td>
</tr>
</tbody>
</table>
Given/New: Why it works

- No extraneous, irrelevant new information
- \([Mozart]\) is the only new piece of information and also the correct answer
- When new information is identical to requested information, a Given/New distinction works as intended
- What happens in other cases?
Given/New: Example where it doesn’t work

<table>
<thead>
<tr>
<th>Question:</th>
<th>An was denken viele Menschen, wenn sie von Weißrussland hören?</th>
</tr>
</thead>
</table>
Given/New: Why it doesn’t work

- Lots of new and unrequested information in the student answer
- Relevant information [*die Tschernobyl-Katastrophe von 1986*] is included, but cannot be distinguished as such
- Newness is not an ideal category for requested answer content.
Given/New: Why it really doesn’t work

Extreme case: Alternative questions

<table>
<thead>
<tr>
<th>Question:</th>
<th>Ist die Wohnung in einem Neubau oder einem Altbau?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target Answer:</td>
<td>[Die Wohnung ist in einem Neubau]_G.</td>
</tr>
<tr>
<td>Student Answer:</td>
<td>[Die Wohnung ist in einem Neubau]_G</td>
</tr>
</tbody>
</table>

- Requested information is Given here, so Newness doesn’t help at all!

- What about Focus/Background?
What kind of focus?

- We need a notion of focus that selects the minimal acceptable answer out of the whole answer content
- Pragmatic notion of focus, no relation to prosody layer
- Needs to be able to cover phrases, because acceptable answers rarely consist of single words
  - Similar concept to “focus phrase” in ??
- Needs to correspond to the type of information requested by the question
**Focus/Background: Example**

<table>
<thead>
<tr>
<th>Question: An was denken viele Menschen, wenn sie von Weißrussland hören?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target Answer:</strong> Sie denken an [die Tschernobyl-Katastrophe von 1986]$_F$.</td>
</tr>
<tr>
<td><strong>Student Answer:</strong> Ausländer denken bei Weißrussland [weniger an Urlaub, sondern eher an die Tschernobyl-Katastrophe von 1986]$_F$. Damals explodierten in der Sowjetunion Teile eines Atomkraftwerks und wurden einige Regionen Weißrusslands von der radioaktiven Strahlung verseucht.</td>
</tr>
</tbody>
</table>
Focus/Background: Example 2

<table>
<thead>
<tr>
<th>Question:</th>
<th>Ist die Wohnung in einem Neubau oder einem Altbau?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target Answer:</td>
<td>Die Wohnung ist ([\text{in einem Neubau}]_F ).</td>
</tr>
<tr>
<td>Student Answer:</td>
<td>Die Wohnung ist ([\text{in einem Neubau}]_F ).</td>
</tr>
</tbody>
</table>
Focus Annotation Experiment

- In order to learn about Focus annotation and find out whether it is feasible, we annotated a selection of question/answer sets in our data.
- Given a certain level of consistency, this can also be used as training data for a computational approach later.
- Previous attempts at annotating focus (e.g. Ritz et al. 2008) were moderately successful.
- However, we believe that given the explicit questions in our corpus, our job is easier and thus can be done more consistently.
Data Selection

- Course level: intermediate and upwards
- Only correct answers with agreement between both raters in the binary decision task
- For each question, we chose the longest student answer ⇒ less chance of the minimal answer being identical to the whole answer
- Resulting data set consists of 82 questions, target answers and student answers
- Data set is tokenized but no other preprocessing is done
Build on Answer Types

- Since we have explicit questions and a correspondence between focus and requested content, we build our annotation scheme on so-called **Answer Types** as used in Question Answering literature (e.g. Li & Roth 2002)
  - An Answer Type is a label for the type of requested content, we use *Entity, Reason, Description, Place, Time, Degree* and *Polar*
  - Idea: Label the question part that defines the Answer Type, such as the *wh*-phrase in *wh*-questions
  - Depending on that, label the focused phrase in the target and student answer with the same type!
Annotation Criteria

Focus marking in the answer should

- cover the minimal part in the answer that can stand on its own and answer the question,
- ideally separate relevant from irrelevant content with respect to the question,
- obey syntactic borders, such as phrase boundaries,
- avoid Given material unless it is necessary to answer the question

The annotation was carried out in the EXMARaLDA tool (Schmidt 2004) independently by two annotators (Thanks to Philip Schulz for annotation and discussion).
Annotation Example

Q [tok] | Was tat Hitler im Gefängnis?
Q [foc] | Description

TA [tok] | Er schrieb sein Buch "Mein Kampf".
TA [foc] | Description

SA [tok] | Während Hitler in Gefängnis war, schrieb er sein Werk "Mein Kampf".
SA [foc] | Description
Quantitative Results: Measures

- In order to assess annotation consistency, we compared both annotators in different ways:
  - Exact match of focused part, only full matches count
  - Mean token overlap of focused parts to account for partial matches
  - Unlabeled percentage agreement per token
  - Labeled percentage agreement per token
  - Cohen’s Kappa on a token basis as a standard agreement measure
- Each of the above was calculated for questions, target answers and student answers separately.
Quantitative Results: Numbers

<table>
<thead>
<tr>
<th></th>
<th>Span-based</th>
<th></th>
<th>Token-based</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exact</td>
<td>Overlap</td>
<td>Unlabeled</td>
<td>Labeled</td>
</tr>
<tr>
<td>Q</td>
<td>74.4%</td>
<td>85.1%</td>
<td>95.7%</td>
<td>91.8%</td>
</tr>
<tr>
<td>T</td>
<td>47.6%</td>
<td>88.5%</td>
<td>89.9%</td>
<td>79.3%</td>
</tr>
<tr>
<td>S</td>
<td>29.3%</td>
<td>77.4%</td>
<td>86.8%</td>
<td>80.9%</td>
</tr>
</tbody>
</table>
Classifying and annotating questions is easier than consistently identifying the focus in answers.

Learner language is harder to annotate than the target answer material.

Agreement measures on a per-token basis are acceptable ⇒ task seems feasible.

Difference between exact match and overlap suggests that the boundaries of focused material need to be better defined.
Conclusions

- A Focus/Background distinction is a better fit for the task of Content Assessment than a Given/New one, because it allows minimal answer marking.
- Focus marking in Reading Comprehension data appears to be feasible.
- Domain of focus marking is still a problem, many smaller mismatches occur.
  - Constrain focus syntactically according to Answer Type. E.g. an Entity must be an NP.
- Set of Answer Types needs to be revised.
- Proper guidelines necessary with detailed description of conventions.
The End

Thank you!
References


