Alignment Weighting for Short Answer Assessment

Björn Rudzewitz\textsuperscript{1}
University of Tübingen

Presentation of B.A. Thesis

October 30, 2015

\textsuperscript{1}bjoern.rudzewitz@student.uni-tuebingen.de
Introduction

Data

System

Alignment Weighting
  General Linguistic Weighting
  Task-Specific Weighting
  Hybrid Approach

Experimental Testing

Discussion

Conclusion

Appendix
Reading Comprehension

Reading comprehension in foreign language learning context:

- text
- questions
- target answers
- student (language learner) answers
Reading Comprehension

Learners need to:

- ... understand the text and questions
- ... use L2 to formulate answers
Reading Comprehension

Learners need to ...

- ... understand the text and questions
  → **task** competence

- ... use L2 to formulate answers
  → **language** competence / performance
Reading Comprehension

Learners need to ...

▶ ... understand the text and questions  
  → **task** competence
▶ ... use L2 to formulate answers  
  → **language** competence / performance

**Goal of this work:** incorporate aspects of concrete task and general language in automatic SAA approach by alignment weighting
Data: CREG

Corpus of Reading Exercises in German [Meurers et al., 2010]

- longitudinal learner corpus collected at 2 German programs in USA (OSU, KU)
- structure:
  - texts
  - questions
  - target answers (TA)
  - student answers (SA)
  - meta data
  - links between elements
    (SA → TA, SA → Diagnosis, ...)
- significant variation / deviation of form and meaning in SAs
- binary (and detailed) gold diagnosis of *semantic* correctness of SAs
Data: CREG

Various subsets used for experiments

<table>
<thead>
<tr>
<th>data set</th>
<th># questions</th>
<th># SAs</th>
<th># TAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>CREG-1032-KU</td>
<td>117</td>
<td>610</td>
<td>180</td>
</tr>
<tr>
<td>CREG-1032-OSU</td>
<td>60</td>
<td>422</td>
<td>147</td>
</tr>
<tr>
<td>CREG-3620-KU</td>
<td>89</td>
<td>735</td>
<td>181</td>
</tr>
<tr>
<td>CREG-3620-OSU</td>
<td>585</td>
<td>2885</td>
<td>705</td>
</tr>
<tr>
<td>CREG-5K-KU</td>
<td>214</td>
<td>1814</td>
<td>382</td>
</tr>
<tr>
<td>CREG-5K-OSU</td>
<td>663</td>
<td>3324</td>
<td>875</td>
</tr>
</tbody>
</table>

Table: Data distribution of CREG subsets used in this study.
Baseline System

**CoMiC-DE** system [Meurers et al., 2011]

- **Comparing Meaning in Context**
- alignment-based short answer assessment system
- UIMA pipeline [Ferrucci and Lally, 2004]
- goal: diagnose form-independent meaning of SAs
CoMiC: System Architecture

3-step approach:

1. *Annotation*
   use NLP tools to generate linguistic multi-layer markup

2. *Alignment*
   use annotations to align similar elements between SA and TA

3. *Diagnosis*
   use features measuring quantity and quality of alignments for binary diagnosis
CoMiC: System Architecture

3-step approach:

1. **Annotation**
   use NLP tools to generate linguistic multi-layer markup

2. **Alignment**
   use annotations to align similar elements between SA and TA

3. **Diagnosis**
   use features measuring quantity and quality of alignments for binary diagnosis
## CoMiC Phase 1: Annotation

<table>
<thead>
<tr>
<th>Task</th>
<th>NLP Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence Detection</td>
<td>OpenNLP [Baldridge, 2005]</td>
</tr>
<tr>
<td>Tokenization</td>
<td>OpenNLP [Baldridge, 2005]</td>
</tr>
<tr>
<td>Lemmatization</td>
<td>TreeTagger [Schmid, 1994]</td>
</tr>
<tr>
<td>Spell Checking</td>
<td>Edit distance [Levenshtein, 1966], igerman98 word list</td>
</tr>
<tr>
<td>Part of Speech Tagging</td>
<td>TreeTagger [Schmid, 1994]</td>
</tr>
<tr>
<td>Noun Phrase Chunking</td>
<td>OpenNLP [Baldridge, 2005]</td>
</tr>
<tr>
<td>Lexical Relations</td>
<td>GermaNet [Hamp et al., 1997]</td>
</tr>
<tr>
<td>Similarity Score</td>
<td>PMI-IR [Turney, 2001]</td>
</tr>
<tr>
<td>Dependency Relations</td>
<td>MaltParser [Nivre et al., 2007]</td>
</tr>
</tbody>
</table>

**Table:** NLP tools used in the CoMiC-DE system.
CoMiC: System Architecture

3-step approach:

1. *Annotation*
   use NLP tools to generate linguistic multi-layer markup

2. *Alignment*
   use annotations to align similar elements between SA and TA

3. *Diagnosis*
   use features measuring quantity and quality of alignments for binary diagnosis
CoMiC Phase 2: Alignment

- align tokens, chunks, dependency triples
- elements given in question are excluded
- alignment candidates: words with overlaps on various linguistic levels
- use TMA [Gale and Shapley, 1962] for annotation matching
- alignment annotation contains alignment label
CoMiC Phase 2: Alignment

Figure: Alignment between target answer (top) and student answer (bottom) on different levels.
CoMiC: System Architecture

3-step approach:

1. **Annotation**
   use NLP tools to generate linguistic multi-layer markup

2. **Alignment**
   use annotations to align similar elements between SA and TA

3. **Diagnosis**
   use features measuring quantity and quality of alignments for binary diagnosis
CoMiC Phase 3: Diagnosis

- extract number and kinds of alignments for each SA → 13 ml features
- use TiMBL Daelemans et al. [2004] for LOO k-NN classification
- result: binary diagnosis for each SA
CoMiC Phase 3: Diagnosis

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Keyword Overlap</td>
<td>% keywords aligned</td>
</tr>
<tr>
<td>2. TA Token Overlap</td>
<td>% aligned TA tokens</td>
</tr>
<tr>
<td>3. Learner Token Overlap</td>
<td>% aligned SA tokens</td>
</tr>
<tr>
<td>4. TA Chunk Overlap</td>
<td>% aligned TA chunks</td>
</tr>
<tr>
<td>5. Learner Chunk Overlap</td>
<td>% aligned SA chunks</td>
</tr>
<tr>
<td>6. TA Triple Overlap</td>
<td>% aligned TA dependency triples</td>
</tr>
<tr>
<td>7. Learner Triple Overlap</td>
<td>% aligned SA dependency triples</td>
</tr>
<tr>
<td>8. Token Match</td>
<td>% token-identical token alignments</td>
</tr>
<tr>
<td>9. Similarity Match</td>
<td>% similarity-resolved token alignments</td>
</tr>
<tr>
<td>10. Type Match</td>
<td>% type-resolved token alignments</td>
</tr>
<tr>
<td>11. Lemma Match</td>
<td>% lemma-resolved token alignments</td>
</tr>
<tr>
<td>12. Synonym Match</td>
<td>% synonym-resolved token alignments</td>
</tr>
<tr>
<td>13. Variety</td>
<td>Number of kinds of token-level alignments (features 8-12)</td>
</tr>
</tbody>
</table>

**Table:** CoMiC baseline features.
Alignment Weighting: Motivation

Idea:

- aligned elements have different properties
- alignments between certain elements may be more important

→ weight existing alignments in new dimension of similarity
Alignment Weighting

2 conceptual weighting approaches → 3 implementations

1. General Linguistic Weighting
2. Task-Specific Weighting
3. Hybrid Approach

global vs. local weighting schemes
General Linguistic Weighting

- weighting of aligned elements by language-wide property in new dimension of similarity
- operationalization of abstract concept of general linguistic property: **part of speech tag classes**
- pos tags represent syntactic, semantic, morphological language-wide properties
General Linguistic Weighting

- problem: data sparsity
- solution: abstraction/generalization via equivalence classes of outcomes
  - pos tag classes

How to find equivalence classes:

- *top-down* approach:
  - using linguistic intuition to form classes of tags
- *bottom-up* approach:
  - induce classes of tags from sample data
### Option 1: top-down approach

<table>
<thead>
<tr>
<th>Group</th>
<th>STTS tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>nominal</td>
<td>NN, NE</td>
</tr>
<tr>
<td>verbal</td>
<td>VVFIN, VVIMP, VVINF, VVIZU, VVPP, VAFIN, VAIMP, VAINF, VAPP, VMFIN, VMINF, VMPP</td>
</tr>
<tr>
<td>adv</td>
<td>ADJA, ADJD, ADV</td>
</tr>
<tr>
<td>rest</td>
<td>APPR, APPRART, APPO, APZR, ART, CARD, FM, ITJ, KOUI, KOUS, KON, KOKOM, PDS, PDAT, PIS, PIAT, PIDAT, PPER, PPOSS, PPOSAT, PRELS, PRELAT, PRF, PWS, PWAT, PWAV, PAV, PTKZU, PTKNEG, PTKVZ, PTKANT, PTKA, TRUNC</td>
</tr>
</tbody>
</table>

**Table:** Coarse STTS subsets used for the general linguistic weighting, adapted from [Rudzewitz and Ziai, 2015].
Option 2: bottom-up approach

- choose a development set
- output single pos features for every tag for TA and SA
- perform hierarchical agglomerative clustering
- use clusters as equivalence classes for features
Option 2: bottom-up approach

Figure: Hierarchical Agglomerative Clustering of Part of Speech Tags over all instances of CREG-1032.
Option 2: bottom-up approach

Figure: Part of Hierarchical Agglomerative Clustering of Part of Speech Tags over all instances of CREG-1032.
Option 2: bottom-up approach

- observation: distinct clusters are representatives for 'main word' classes defined in STTS tag set [Schiller et al., 1995]
- hclust algorithm is given no assumptions about main word classes!

$\rightarrow$ use STTS main word classes as equivalence classes
Feature Variants

- problem with features: how to normalize?
- more concrete: given numeric quantities of aligned elements, how to account for effects of answer length?
- solution (in this work): explore and report results for all variants
Feature Variants

\[
A_h \in A(\text{"Answers"}), w_j \in W_{A_h} \subset W(\text{"Words"}), t_{w_j} \in T_i \subset T(\text{"tag from tag group"})
\]

\[
ol(A_h, T_i) = \frac{\sum_{t \in T_i} \sum_{w_j \in W_{A_h}} [w_j \text{ is aligned AND } t_{w_j} = t \text{ AND } w_j \text{ is new}]}{\sum_{t \in T_i} \sum_{w_j \in W_{A_h}} [\text{see Table !}]}
\]

<table>
<thead>
<tr>
<th>variant</th>
<th>(t_{w_j} = t)</th>
<th>(w_j) is new</th>
<th>(w_j) is aligned</th>
</tr>
</thead>
<tbody>
<tr>
<td>local</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>semi-global</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>global</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

**Table:** Denominator constraints for different feature variants. Logical conjunction AND between row values.
Feature Variant Interpretation

- **local**: Are many of the new tokens with this part of speech tag aligned?
- **semi-global**: Are many of the aligned tokens from a certain part of speech group?
- **global**: Do many of the new words have a tag from this part of speech group and are at the same time aligned?
Interpolated Features

\[ ol_{\text{ip}}(A_h, T_i) = ol_{\text{local}}(A_h, T_i) \times ol_{\text{sglobal}}(A_h, T_i) \times ol_{\text{global}}(A_h, T_i) \]

\[ ol_{\text{lip}}(A_h, T_i) = \frac{1}{3} \times (ol_{\text{local}}(A_h, T_i) + ol_{\text{sglobal}}(A_h, T_i) + ol_{\text{global}}(A_h, T_i)) \]

- combine the different feature variants
Task-Specific Weighting

- goal: include the specific (local) task context in SAA
- "task": complex concept, many aspects
- operationalization: implement question-type features
- binary indicator function for each question type
- gold standard from previous study [Meurers et al., 2011] as development set
Hybrid Weighting Approach

- *tf.idf* lemma-based weighting, adapted from Manning and Schütze [1999]
- generally applicable measure, but task-specific training
- document collection: all reading texts in CREG-5K
- for each aligned token, get *tf.idf* weight in reading text to which the SA refers

\[
ol_{tf.idf}(A_h) = \sum_{w_j \in W_{A_h}} \text{weight}_{tf.idf}(w_j, d_i)
\]

\[
\text{weight}_{tf.idf}(w_j, d_i) = \begin{cases} 
0 & \text{, if } (w_j \text{ NOT new}) \text{ OR } (w_j \text{ NOT aligned}) \text{ OR } (w_j \notin d_i) \\
(1 + \log(tf_j,i)) \times \log \frac{N}{df_j} & \text{, otherwise}
\end{cases}
\]
Experimental Testing

Significance Testing: McNemar’s test ($\alpha = 0.05$)

$H_0$: The binary classification performance of an alignment-based short answer assessment system does not change if it is augmented with part of speech or $tf.idf$ features.

$H_1$: The binary classification performance of an alignment-based short answer assessment system significantly improves if it is augmented with part of speech or $tf.idf$ features.
Experimental Testing: Coarse POS

<table>
<thead>
<tr>
<th>system</th>
<th>3620-KU</th>
<th>3620-OSU</th>
<th>1032-KU</th>
<th>1032-OSU</th>
<th>5K-KU</th>
<th>5K-OSU</th>
</tr>
</thead>
<tbody>
<tr>
<td>base</td>
<td>81.5</td>
<td>82.2</td>
<td>84.6</td>
<td>87.0</td>
<td>80.9</td>
<td>82.5</td>
</tr>
<tr>
<td>local</td>
<td>82.0</td>
<td>82.6</td>
<td>85.2</td>
<td>90.0*</td>
<td>82.0</td>
<td>82.8</td>
</tr>
<tr>
<td>semi-global</td>
<td>81.2</td>
<td>84.1*</td>
<td>85.4</td>
<td>87.2</td>
<td>81.3</td>
<td>84.0*</td>
</tr>
<tr>
<td>global</td>
<td>83.0</td>
<td>83.6*</td>
<td>84.8</td>
<td>85.8</td>
<td>81.6</td>
<td>83.6*</td>
</tr>
<tr>
<td>ip</td>
<td>80.5</td>
<td>84.1*</td>
<td>85.1</td>
<td>85.1</td>
<td>81.7</td>
<td>84.4*</td>
</tr>
<tr>
<td>lip</td>
<td>82.6</td>
<td>84.1*</td>
<td>84.4</td>
<td>87.0</td>
<td>81.4</td>
<td>84.1*</td>
</tr>
</tbody>
</table>

**Table:** System performance for the baseline system augmented with part of speech features in terms of accuracy. The symbol * denotes a statistically significant improvement over the baseline ($\alpha = 0.05$).
Experimental Results: Question Types and tf.idf

<table>
<thead>
<tr>
<th>system variant</th>
<th>3620-KU</th>
<th>3620-OSU</th>
<th>1032-KU</th>
<th>1032-OSU</th>
<th>5K-KU</th>
<th>5K-OSU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>81.5</td>
<td>82.2</td>
<td>84.6</td>
<td>87.0</td>
<td>80.9</td>
<td>82.5</td>
</tr>
<tr>
<td>q-types</td>
<td>80.8</td>
<td>\textbf{83.1} *</td>
<td>85.4</td>
<td>87.2</td>
<td>80.9</td>
<td>82.8</td>
</tr>
</tbody>
</table>

Table: System performance for the baseline system augmented with question type features in terms of accuracy. The symbol * denotes a statistically significant improvement over the baseline ($\alpha = 0.05$).

<table>
<thead>
<tr>
<th>system variant</th>
<th>3620-KU</th>
<th>3620-OSU</th>
<th>1032-KU</th>
<th>1032-OSU</th>
<th>5K-KU</th>
<th>5K-OSU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>81.5</td>
<td>82.2</td>
<td>84.6</td>
<td>87.0</td>
<td>80.9</td>
<td>82.5</td>
</tr>
<tr>
<td>tf.idf</td>
<td>\textbf{84.2} *</td>
<td>\textbf{84.1} *</td>
<td>86.1</td>
<td>88.4</td>
<td>\textbf{83.1} *</td>
<td>\textbf{84.3} *</td>
</tr>
</tbody>
</table>

Table: System performance for the baseline system augmented with tf.idf features in terms of accuracy. The symbol * denotes a statistically significant improvement over the baseline ($\alpha = 0.05$).
Experimental Testing: Combination

<table>
<thead>
<tr>
<th>system variant</th>
<th>3620-KU</th>
<th>3620-OSU</th>
<th>1032-KU</th>
<th>1032-OSU</th>
<th>5K-KU</th>
<th>5K-OSU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>81.5</td>
<td>82.2</td>
<td>84.6</td>
<td>87.0</td>
<td>80.9</td>
<td>82.5</td>
</tr>
<tr>
<td>q-types + stts local + tf.idf</td>
<td>83.8</td>
<td>84.7*</td>
<td>87.9*</td>
<td>86.5</td>
<td>82.4</td>
<td>84.9</td>
</tr>
<tr>
<td>q-types + stts semi-global + tf.idf</td>
<td>83.1</td>
<td>84.6*</td>
<td>85.4</td>
<td>88.2</td>
<td>82.1</td>
<td>84.9</td>
</tr>
<tr>
<td>q-types + stts global + tf.idf</td>
<td>84.2*</td>
<td>84.5*</td>
<td>87.9*</td>
<td>84.6</td>
<td>82.6*</td>
<td>84.6*</td>
</tr>
<tr>
<td>q-types + stts ip + tf.idf</td>
<td>83.3</td>
<td>84.7*</td>
<td>88.9*</td>
<td>84.1</td>
<td>82.8*</td>
<td>85.3</td>
</tr>
<tr>
<td>q-types + stts lip + tf.idf</td>
<td>84.5*</td>
<td>85.0*</td>
<td>88.0*</td>
<td>85.8</td>
<td>82.8*</td>
<td>84.9*</td>
</tr>
</tbody>
</table>

**Table:** System performance for the baseline system augmented with question type and STTS group part of speech features and \textit{tf.idf} weighting in terms of accuracy. The symbol * denotes a statistically significant improvement over the baseline ($\alpha = 0.05$).
Experimental Testing: Main results

- *many* more tables with accuracies and test statistics ...
- pos features alone result in highest accuracy on one data set (90%)
- *tf.idf* always yields improvement
- question-types alone not as effective
- best overall result for combination of all 3 weightings
- linguistically interpretable question-type specific pos alignment patterns (Appendix 1)
- question-type specific macro-averages show improvement from Meurers et al. [2011] (Appendix 2)
Discussion: Related work

- Ziai and Meurers [2014]: CoMiC + information structure
- Horbach et al. [2013]: CoMiC-reimplementation + pos-align criteria + use of reading text
- Hahn and Meurers [2012]: CoSeC
- many other SAA systems (see thesis)
Conclusion

- significant improvements with novel techniques
- results highly competitive to state-of-the-art systems
- no human annotation needed
- linguistically interesting insights from ml algorithms
- combination of all feature variants most effective
## Appendix 1: q-type pos align patterns

<table>
<thead>
<tr>
<th>q-type</th>
<th>#inst.</th>
<th>10 most informative Part of Speech tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternative</td>
<td>7</td>
<td>VVPP, PPOSAT, PPER, PPOS, VMFIN, PRELAT, PIS, PIDAT, PIAT, PDS</td>
</tr>
<tr>
<td>How</td>
<td>144</td>
<td>NN, CARD, VVFIN, ADJA, ART, VAFIN, NE, PIAT, PRELS, PTKNEG</td>
</tr>
<tr>
<td>What</td>
<td>276</td>
<td>NN, KON, ADJA, VVPP, VVINF, APPRART, PIS, CARD, PTKNEG, PWA</td>
</tr>
<tr>
<td>When</td>
<td>6</td>
<td>ADV, KOKOM, KOUS, NN, PIS, PWF, PIDAT, PWA, PPOSAT, VAFIN</td>
</tr>
<tr>
<td>Where</td>
<td>9</td>
<td>PIDAT, PPER, PPOSAT, PRELAT, PIS, VVPP, PRF, PIAT, PAVDAT</td>
</tr>
<tr>
<td>Which</td>
<td>170</td>
<td>NN, ADV, VVPP, PTKNEG, VAFIN, NE, VAINF, CARD, KON, PIS</td>
</tr>
<tr>
<td>Why</td>
<td>174</td>
<td>NN, VVFIN, ART, APPR, PIAT, VAFIN, KON, NE, ADJA, KOKOM</td>
</tr>
<tr>
<td>Who</td>
<td>41</td>
<td>NN, VVINF, ADJD, VMFIN, PPER, PRELAT, Prels, PPOS, PPOSAT, PTKANT</td>
</tr>
<tr>
<td>Yes/No</td>
<td>5</td>
<td>PTKANT, PPOSAT, PRELAT, PPOS, PIS, PPER, PIDAT, PRF, PIAT, PAV</td>
</tr>
<tr>
<td>Several</td>
<td>200</td>
<td>NN, NE, ADJA, PIAT, VMFIN, KON, PIS, VVPP, KON, PTKNEG</td>
</tr>
</tbody>
</table>

**Table:** Most informative part of speech alignments by question type.
Appendix 2: q-type macro-averages

<table>
<thead>
<tr>
<th>q-type</th>
<th># inst.</th>
<th>local</th>
<th>sglobal</th>
<th>global</th>
<th>ip</th>
<th>lip</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternative</td>
<td>7</td>
<td>0.57</td>
<td>0.57</td>
<td>0.57</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>How</td>
<td>144</td>
<td>0.88</td>
<td>0.89</td>
<td>0.91</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>What</td>
<td>276</td>
<td>0.87</td>
<td>0.88</td>
<td>0.87</td>
<td>0.85</td>
<td>0.88</td>
</tr>
<tr>
<td>When</td>
<td>6</td>
<td>1.00</td>
<td>0.83</td>
<td>1.00</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>Where</td>
<td>9</td>
<td>0.67</td>
<td>0.56</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>Which</td>
<td>170</td>
<td>0.91</td>
<td>0.92</td>
<td>0.93</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>Why</td>
<td>174</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.83</td>
<td>0.84</td>
</tr>
<tr>
<td>Who</td>
<td>41</td>
<td>0.88</td>
<td>0.90</td>
<td>0.85</td>
<td>0.88</td>
<td>0.85</td>
</tr>
<tr>
<td>Yes/No</td>
<td>5</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>Several</td>
<td>200</td>
<td>0.86</td>
<td>0.83</td>
<td>0.83</td>
<td>0.86</td>
<td>0.85</td>
</tr>
<tr>
<td>Micro</td>
<td>1032</td>
<td>86.7</td>
<td>86.8</td>
<td>87.0</td>
<td>86.5</td>
<td>87.3</td>
</tr>
</tbody>
</table>

Table: Macro-averages of the best system variant on CREG-1032 obtained by grouping results by question type. Boldface indicates an improvement upon the results by Meurers et al. [2011]


