Candidate Evaluation Strategies for Improved Difficulty Prediction of Language Proficiency Tests

BEA Workshop 2015
Lisa Beinborn, Torsten Zesch, Iryna Gurevych
UKP Lab
Department of Computer Science
Technische Universität Darmstadt
Follow-up work

TACL paper 2014: “Predicting the Difficulty of Language Proficiency Tests”

What is difficult for language learners? (And why?)
How can we predict difficulty?

Data: Language Proficiency Tests
Task: Predict difficulty of the items
Reduced Redundancy Principle

- Spolsky (1969): “Natural language is redundant”
- The ability to deal with reduced redundancy distinguishes beginners from advanced language learners

Thanks to the redundancy of language,

yxx cxn xndxrstxnd whxt x xm wrxtxng xvxn xf x rxplxcx xll thx vxwxls wxth xn ‘x‘.

The roots of humanity can be traced back to millions of years ago. The primary evidence comes from fossils - skulls, skeletons, and bone fragments. Scientists have made tools that allow them to extract subtle information from ancient bones about their environment settings. Modern forensic work in the field allows laboratories to provide a rich understanding of how our ancestors lived.
The roots of humanity can be traced back to millions of years ago. The primary evidence comes from fossils - skulls, skeletons, and bone fragments. Scientists have made tools that allow them to extract subtle information from ancient bones about their environments. Modern forensic work in this field allows laboratorians to provide a rich understanding of how our ancestors lived.
The roots of humanity can be traced back to millions of years ago. **T** primary evidence comes from fossils - skulls, skeletons, and bone fragments. Scientists have made tools that allow them to extract subtle information from ancient bones and their environments. Modern forensic work in this field in laboratories can now provide a rich understanding of how our ancestors lived.

### C-Test Difficulty

- **Micro-level processing**
  - Solution Difficulty
  - Candidate Ambiguity
- **Macro-level processing**
  - Inter-Gap Dependency
  - Paragraph Difficulty
Outlook of TACL paper

1) Adapt to other languages and to other test variants
2) Improve the dimensions candidate ambiguity and inter-gap dependency
TODAY: Part 1

1) Adapt to **German and French** and to **two test variants**

2) Improve the dimensions candidate ambiguity and inter-gap dependency
TODAY, Part 2

1) Adapt to German and French and to two test variants
2) Improve the dimensions candidate ambiguity and inter-gap dependency

---

13. His characteristic talk, with its keen ____ of detail and subtle power of inference held me amused and enthralled.

- instincts
- presumption
- observance
- expiation
- implements
Different Languages

Is it a language-independent approach?  
Can it be adapted to other languages?

Markus Koljonen: iki.fi/markus.koljonen
Le Noël des familles
Noël semble immuable, et pourtant il change ! Au co____ du te____ on
s’aper____, discrètement ma____ sûrement, q____ la fê____ de l____
Nativité pr____ de nouv____ allures e____ alliant l____ traditions e____
la mode____. Traditionnellement, l____ sapin d____ Noël e____ surtout
déc____ de bou____, de guirl____ lumineuses e____ de boules
multicolores. Toutefois, une certaine influence germanique se fait de plus
en plus sentir, avec la présence de figurines en bois ou de dessins aux
fenêtres.

- articles
- accents
- richer morphology
Le Noël des familles
Noël semble immuable, et pourtant il change ! Au co____ du te____ on s’aper____, discrètement ma____ sûrement, q____ la fê____ de l____
Nativité pr____ de nouv____ allures e____ alliant l____ traditions e____
la mode____. Traditionnellement, l____ sapin d____ Noël e____ surtout
déc____ de bou____, de guir____ lumineuses e____ de boules
multicolores. Toutefois, une certaine influence germanique se fait de plus en plus sentir, avec la présence de figurines en bois ou de dessins aux fenêtres.
Auf einer Weltkarte kann man sehen, dass Asien und Nordamerika im Norden nur durch einen schmalen Meeresstreifen voneinander getrennt sind: durch die Beringstraße. Während d______ Eiszeit herrs_____ auf d_____ ganzen Er_____ niedrigere Temper_____ , und d_____ Beringstraße w_____ zugefroren. Üb_____ das E_____ gelangten Volksg_____ aus As_____ auf d_____ amerikanischen Kont_____ . Manche bli_____ in Norda_____ und bild_____ mehr a_____ tausend versch_____ Stämme - jew_____ mit ei_____ eigenen Sprache, die anderen zogen weiter bis nach Südamerika.

- articles, cases
- Umlaute, case sensitivity
- old/new orthography
- compounds
Auf einer Weltkarte kann man sehen, dass Asien und Nordamerika im Norden nur durch einen schmalen Meeresstreifen voneinander getrennt sind: durch die Beringstraße. Während d____ Eiszeit herrs_____ auf d_____ ganzen Er_____ niedrigere Temper_____ , und d_____ Beringstraße w_____ zugefroren. Üb_____ das E_____ gelangten Volksg_____ aus As_____ auf d_____ amerikanischen Kont_____ . Manche bli_____ in Norda_____ und bild_____ mehr a_____ tausend versch_____ Stämme - jew_____ mit ei_____ eigenen Sprache, die anderen zogen weiter bis nach Südamerika.
## Different Languages...

<table>
<thead>
<tr>
<th>Language</th>
<th>Words</th>
<th>Mean word length</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>99,171</td>
<td>8.5 ± 2.6</td>
</tr>
<tr>
<td>French</td>
<td>139,719</td>
<td>9.6 ± 2.6</td>
</tr>
<tr>
<td>German</td>
<td>332,263</td>
<td>12.0 ± 3.5</td>
</tr>
</tbody>
</table>
Different Candidate Space

- German C-test
- French C-test
- English C-test

Mean number of candidates vs. Word length
Different Test Types

The reduced redundancy principle comprises more than just the C-test.
Prefix Deletion Test

Was ist Kreativität?
Die meisten halten Kreativität für eine seltene Gabe, über die nur eine exklusive Minderheit verfügt. Dabei ____t jeder ____sch auf _____ne Weise _____erisch. Wir _____nen nicht _____ht kreativ _____in. Die _____idende Frage _____t, ob _____r diese _____liche Fähigkeit _____iv pflegen _____er verkümmern _____sen. Denn _____vität bezieht _____ch nicht _____f ein _____mmtes Themengebiet, _____ern ist _____all möglich. _____ativ sein _____utet, sich _____as anderes _____ellen zu _____nen als das, was man gerade sieht. Kreativität wird häufig mit Innovation verwechselt.

- prefix vs postfix
- multiple solutions
Was ist Kreativität?
Die meisten halten Kreativität für eine seltene Gabe, über die nur eine exklusive Minderheit verfügt. Dabei ______t jeder ______sch auf ______ne Weise ______erisch. Wir ______nen nicht ______cht kreativ ______in. Die ______idende Frage ______t, ob ______r diese ______liche Fähigkeit ______iv pflegen ______er verkümmern ______sen. Denn ______vität bezieht ______ch nicht ______f ein ______mmtes Themengebiet, ______ern ist ______all möglich. ______ativ sein ______utet, sich ______as anderes ______ellen zu ______nen als das, was man gerade sieht. Kreativität wird häufig mit Innovation verwechselt.

Data: University of Duisburg-Essen
German proficiency test for prospective teachers
13. His characteristic talk, with its keen _____ of detail and subtle power of inference held me amused and entralled.

- instincts
- presumption
- observance
- expiation
- implements

- Closed format
- Distractors
13. His characteristic talk, with its keen ____ of detail and subtle power of inference held me amused and enthralled.

- instincts
- presumption
- observance
- expiation
- implements

Questions:
Microsoft Research Zweig & Burges (2012)
Error Rates: own study
Candidate Space

![Graph showing the mean number of candidates for different word lengths. The graph compares three methods: Prefix, C-test, and Cloze. The x-axis represents word length, and the y-axis represents the mean number of candidates. The graph shows peaks and troughs for each method, indicating variations in candidate space across different word lengths.](image-url)
## Data

<table>
<thead>
<tr>
<th>Format</th>
<th>Test Type</th>
<th>Texts</th>
<th>Gaps</th>
<th>Participants</th>
<th>Av. Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open</td>
<td>C-test en</td>
<td>39</td>
<td>775</td>
<td>210</td>
<td>.35 ± .25</td>
</tr>
<tr>
<td></td>
<td>C-test fr</td>
<td>40</td>
<td>799</td>
<td>24</td>
<td>.52 ± .28</td>
</tr>
<tr>
<td></td>
<td>C-test de</td>
<td>82</td>
<td>1,640</td>
<td>251</td>
<td>.55 ± .26</td>
</tr>
<tr>
<td></td>
<td>Prefix de</td>
<td>14</td>
<td>348</td>
<td>225</td>
<td>.36 ± .23</td>
</tr>
<tr>
<td>Closed</td>
<td>Cloze en</td>
<td>100</td>
<td>100</td>
<td>22</td>
<td>.27 ± .22</td>
</tr>
</tbody>
</table>
Generalization

- Adapt format and types
- Adapt linguistic pre-processing (DKPro)
- Adapt resources
- Adapt features (Reduction from 87 to 70)
Prediction Task

- SMO regression, Leave-one-out cross-validation on texts
## Prediction Results: Languages

<table>
<thead>
<tr>
<th>Test Type</th>
<th>Pearson’s r</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-test en</td>
<td>.47</td>
</tr>
<tr>
<td>C-test fr</td>
<td>.67</td>
</tr>
<tr>
<td>C-test de</td>
<td>.61</td>
</tr>
</tbody>
</table>
## Prediction Results

<table>
<thead>
<tr>
<th>Test Type</th>
<th>Pearson’s r</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-test en</td>
<td>0.47</td>
</tr>
<tr>
<td>C-test fr</td>
<td>0.67</td>
</tr>
<tr>
<td>C-test de</td>
<td>0.61</td>
</tr>
<tr>
<td>Prefix de</td>
<td>0.27</td>
</tr>
<tr>
<td>Cloze en</td>
<td>0.20</td>
</tr>
</tbody>
</table>
## Prediction Results

<table>
<thead>
<tr>
<th>Test Type</th>
<th>Pearson’s r</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-test en</td>
<td>.47</td>
</tr>
<tr>
<td>C-test fr</td>
<td>.67</td>
</tr>
<tr>
<td>C-test de</td>
<td>.61</td>
</tr>
<tr>
<td>Prefix de</td>
<td>.27</td>
</tr>
<tr>
<td>Cloze en</td>
<td>.20</td>
</tr>
</tbody>
</table>

![Scatter plot showing prediction results](image-url)
Prefix Deletion: Outlier
Open vs closed format

- Many of the features developed for difficulty prediction were targeted at production problems (e.g. spelling).
- Closed cloze tests only require recognition skills.

<table>
<thead>
<tr>
<th>Test Type</th>
<th>Pearson’s r</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-test en</td>
<td>.47</td>
</tr>
<tr>
<td>C-test fr</td>
<td>.67</td>
</tr>
<tr>
<td>C-test de</td>
<td>.61</td>
</tr>
<tr>
<td>Prefix de</td>
<td>.27</td>
</tr>
<tr>
<td>Cloze en</td>
<td>.20</td>
</tr>
</tbody>
</table>
A closer look at cloze

49. The stage lost a fine ____ , even as science lost an acute reasoner, when he became a specialist in crime.
   o linguist
   o hunter
   o actor
   o estate
   o horseman

21. When his body had been carried from the cellar we found ourselves still confronted with a problem which was almost as ____ as that with which we had started.
   o tall
   o loud
   o invisible
   o quick
   o formidable

Error Rate: 0.0
Error Rate: 0.6
Inspiration from automated solving

How would we proceed for nlp-based solving of language tests?

Rank the candidates

1) According to language model probability
2) According to semantic relatedness between candidate and context
Differences

- **Automatic solving:**
  train on domain-specific data (Holmes novels)

- **Difficulty prediction:**
  train on general data to model learner knowledge

Language model: trained with Berkeley LM on Leipzig corpora
- 1 million sentences, news domain

Explicit Semantic Analysis Index: trained on Wikipedia
- 931,559 concepts
### LM ranker

- **Candidate fitness**: log probability of sentence in language model

The stage lost a fine _______ , even as science lost an acute reasoner, when he became a specialist in crime.

<table>
<thead>
<tr>
<th>Role</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>actor</td>
<td>-358.83</td>
</tr>
<tr>
<td>estate</td>
<td>-361.22</td>
</tr>
<tr>
<td>hunter</td>
<td>-361.96</td>
</tr>
<tr>
<td>linguist</td>
<td>-362.71</td>
</tr>
<tr>
<td>horseman</td>
<td>-362.93</td>
</tr>
</tbody>
</table>

**Rank of solution**: **1**
LM ranker

- Candidate fitness: log probability of sentence in language model

When his body had been carried from the cellar we found ourselves still confronted with a problem which was almost as _____ as that with which we had started.

- tall: -175.17
- quick: -176.61
- loud: -178.60
- formidable: **-179.52**
- invisible: -179.52

Rank of solution: **4**
ESA ranker

- Candidate fitness: sum over the cosine similarity between the candidate and every content word in the sentence (similar to Zweig et al. 2012)

The stage lost a fine _______ , even as science lost an acute reasoner, when he became a specialist in crime.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>actor</td>
<td>1.06</td>
</tr>
<tr>
<td>estate</td>
<td>0.70</td>
</tr>
<tr>
<td>hunter</td>
<td>0.58</td>
</tr>
<tr>
<td>horseman</td>
<td>0.29</td>
</tr>
<tr>
<td>linguist</td>
<td>0.29</td>
</tr>
</tbody>
</table>

Rank of solution: 1
## ESA ranker

- Candidate fitness: sum over the cosine similarity between the candidate and every content word in the sentence (similar to Zweig et al. 2012)

When his body had been carried from the cellar we found ourselves still confronted with a problem which was almost as _____ as that with which we had started.

<table>
<thead>
<tr>
<th>Word</th>
<th>Cosine Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>quick</td>
<td>1.03</td>
</tr>
<tr>
<td>tall</td>
<td>0.80</td>
</tr>
<tr>
<td>invisible</td>
<td>0.63</td>
</tr>
<tr>
<td>loud</td>
<td>0.61</td>
</tr>
<tr>
<td>formidable</td>
<td>0.56</td>
</tr>
</tbody>
</table>

Rank of solution: 5
Improved Cloze Results

Reduced features

Reduced + Ranking
Improved Cloze Results

![Graph showing Improved Cloze Results for LM, ESA, and LM + ESA]
Summary and Outlook

- Successfully adapted the prediction framework to new languages and test types.
- We ❤️ French.
- Candidate evaluation strategies from automatic solving can be applied to simulate learner difficulties.

Plans:
- Currently collecting more error rates for cloze tests.
- Use candidate evaluation strategies for distractor generation and difficulty manipulation.
- Strategies for inter-item dependencies.
Thank you very much for your ________.

- attention
- presence
- patience
- enthusiasm

Do you have any_______?

- questions
- comments
- praises
- doubts