USING TEXT SEGMENTATION ALGORITHMS FOR THE AUTOMATIC GENERATION OF E-LEARNING COURSES

Alexander Streicher, Can Beck, Andrea Zielinski
COLING 2015, Dublin
Agenda

- Introduction & Motivation

- Text Segmenters Application and Experimental Setup
  - Test Corpus
  - Segmentation Algorithms
  - Performance Measures

- Evaluation Results

- Conclusion
Vision & Research Question

*Reduce time consuming effort of e-learning course creation – generate courses automatically*

*When, where and how successful can text segmentation algorithms be applied?*
Motivation

Our 2-Level E-Learning Course Structure: Concept Containers (CC) and Knowledge Objects (KO)

- How to project texts on two-level course structure?
- How can we evaluate the usability of text segmentation algorithms for that task?
Setup Overview

Source  Test Corpus  Segmentation Algorithms  Performance Measures

while(noSuccess) {
    tryAgain();
    if(Dead) break;
}
Corpus

- Samples of unstructured text
- Sections from 530 featured Wikipedia articles, 6 categories
  - Ground truth on Macro and micro level to measure segmentation performance
- Macro level low coherence, micro level high coherence
- 1200 macro samples, 8231 micro samples
Segmentation Algorithms

- **Macro Level | Topics | ConceptContainers**
  - TopicTiling (Riedl & Biemann, 2012)

- **Micro Level | Units | KnowledgeObjects**
  - BayesSeg (Eisenstein & Barzily, 2008)
## Training & Testing - LDA based segmentation algorithms

### e.g. Choi Corpus

<table>
<thead>
<tr>
<th>k</th>
<th>Test Set Size (Macro)</th>
<th>Training Set Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>#samples = 120</td>
<td>139±7 featured Articles (26% of all articles)</td>
</tr>
<tr>
<td>20</td>
<td>60</td>
<td>267±8 featured Articles (51% of all articles)</td>
</tr>
<tr>
<td>30</td>
<td>40</td>
<td>338±7 featured Articles (64% of all articles)</td>
</tr>
</tbody>
</table>

### Consequences for the number of folds (k) used in cross validation:

- **10**: 10% of the macro dataset, 120 samples, 139±7 featured articles (26% of all articles).
- **20**: 5% of the macro dataset, 60 samples, 267±8 featured articles (51% of all articles).
- **30**: 3% of the macro dataset, 40 samples, 338±7 featured articles (64% of all articles).
Performance Measures

- Different metrics to measure segmenter performance
  - Penalty metrics
    - Probability Metric (Doug et al. 1998)
    - Window Diff (Pevzner & Hearst, 2002)
  - Rewarding metric
    - Boundary Similarity (Fournier & Inkpen 2013)

- Problem: What does 0.2 mean?
# Scalable Segmentation Performance – a new baseline

<table>
<thead>
<tr>
<th>Distance from true boundary</th>
<th>Standard deviation in % of avg. number of sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>very close</strong></td>
<td>$\sigma \in (0, 5]$</td>
</tr>
<tr>
<td><strong>close</strong></td>
<td>$\sigma \in (5, 15]$</td>
</tr>
<tr>
<td><strong>large</strong></td>
<td>$\sigma \in (15, 30]$</td>
</tr>
</tbody>
</table>

![Graph showing the distribution of distances from true boundaries and standard deviations.]
Results for TopicTiling on **Macro Dataset**

**Boundary Similarity**

With random segmenter as baseline:
Results for BayesSeg on Micro Dataset

Boundary Similarity

Subset

mean

0.0 0.2 0.4 0.6 0.8
3 4 5 6
Subset mean

worst                                                best
very close  close  large

segmenter

BS  σ=1%  σ=5%  σ=15%  σ=30%

© Fraunhofer IOSB        Alexander Streicher
Conclusion

- 2-level CC/KO block structure is extracted from unstructured text
- CC/KO structure forms basis for learning objects
- Good results on both levels in relation to own baseline
- Increased interpretability with presented baseline approach

Future Work:

- Comparison of other segmenters with RS on benchmark dataset, to gain unified overview
- Create full e-learning corpus based on real courses for further evaluation
Discussion

Contact:
Alexander.Streicher
@iosb.fraunhofer.de
Related Work

- Model-based; content-generation for e-learning courses based on existing course material  
  (Sathiyamurthy & Geetha, 2011)
  - hierarchical domain ontology
  - pedagogical ontology
  - LDA based segmentation

- Adaptation of existing courses to the learner or instructional plans
  - Particle swarm based content organization  (Lin et al., 2009)
  - Large-scale course generation  (Tan et al., 2010)
References

- Strassel, S., Graff, D., Martey, N. & Cieri, C., 2000. Quality Control in Large Annotation Projects Involving Multiple Judges: The Case of the TDT Corpora. s.l., s.n.