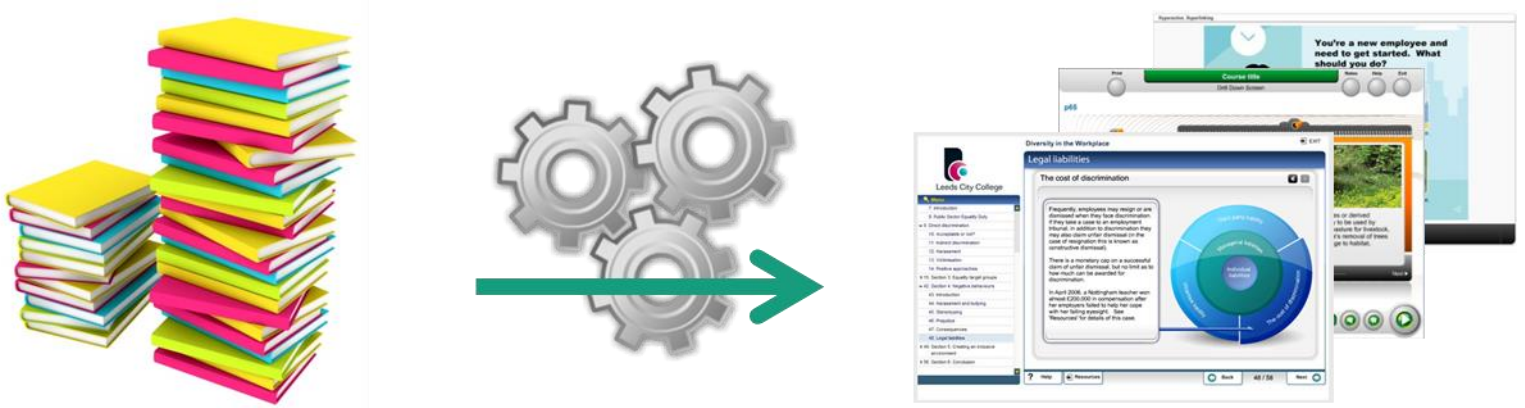
USING TEXT SEGMENTATION ALGORITHMS FOR THE AUTOMATIC GENERATION OF E-LEARNING COURSES

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

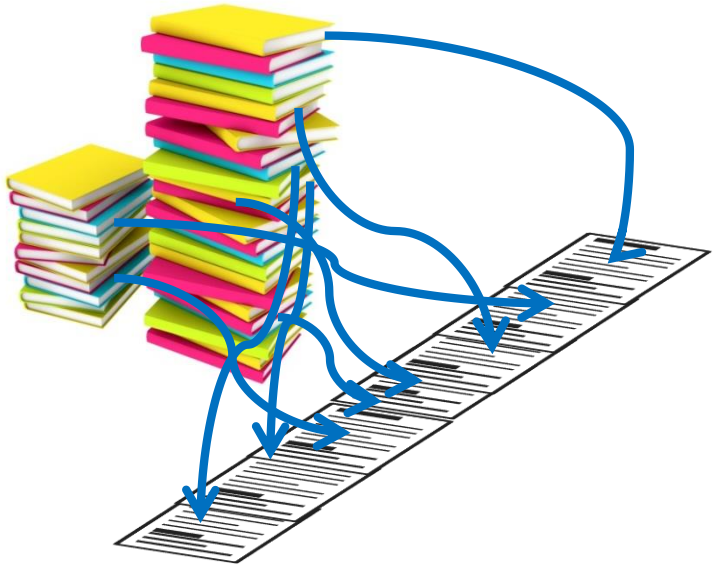
Vision & Research Question

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)



Course = {CC_1, CC_2, ..., CC_n}

Chapters

CC_1 = {KO_1,KO_2,... ,KO_n}

Pages

KO_1 =

- *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

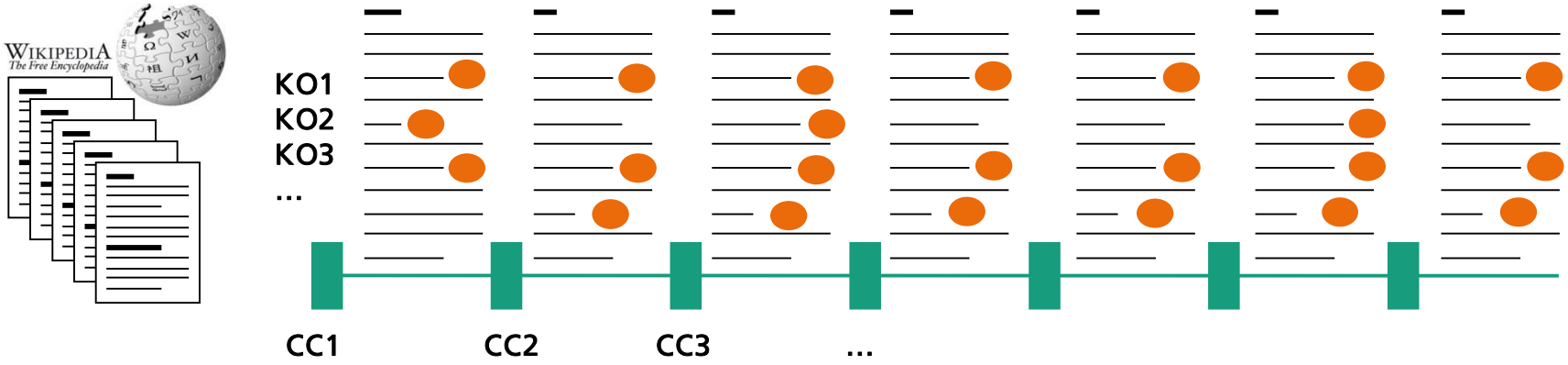
Test Corpus

Segmentation Algorithms

Performance Measures

Corpus

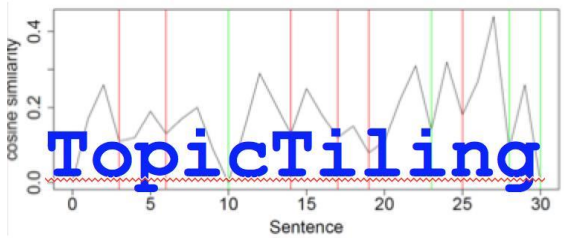
- Samples of unstructured text
- Sections from 530 featured Wikipedia articles, 6 categories
  - Ground truth on Macro and micro level to measure segmentation performance
- Macro level low coherence, micro level high coherence
- 1200 macro samples, 8231 micro samples



Segmentation Algorithms

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

- Macro Level | Topics | *ConceptContainers*

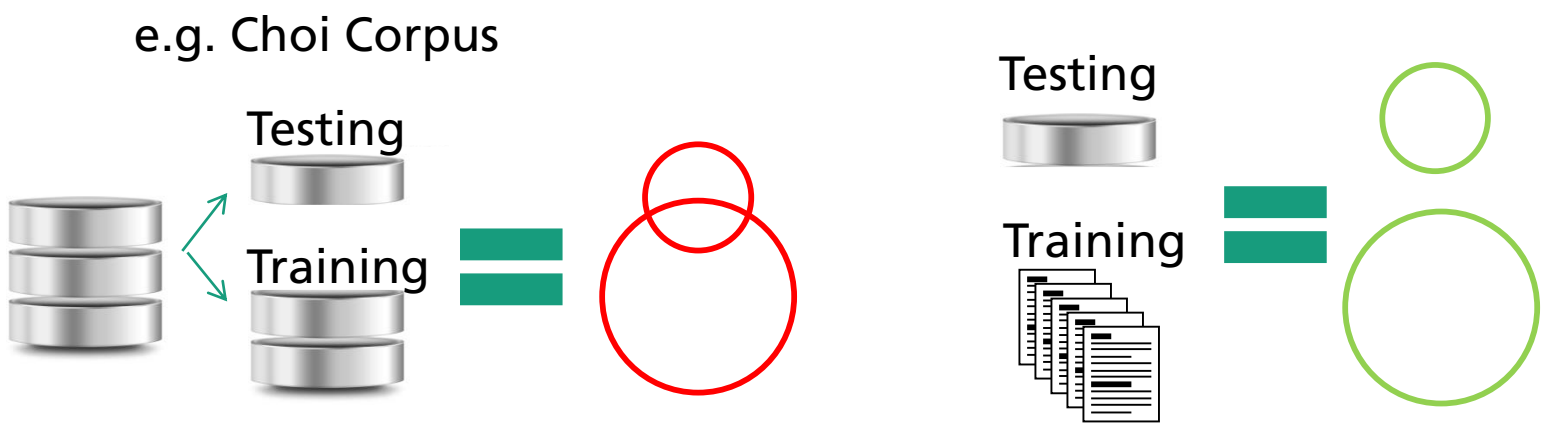
→ 

(Riedl & Biemann, 2012)

- Micro Level | Units | *KnowledgeObjects*

→BayesSeg (Eisenstein & Barzily, 2008)

Training & Testing - LDA based segmentation algorithms



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

k	Test Set Size (Macro)	Training Set Size
10	#samples = 120 10%	139±7 featured Articles (26% of all articles) ☹
20	60 5 %	267±8 featured Articles (51% of all articles)
30	40 3 %	338±7 featured Articles (64% of all articles) ☺

Performance Measures

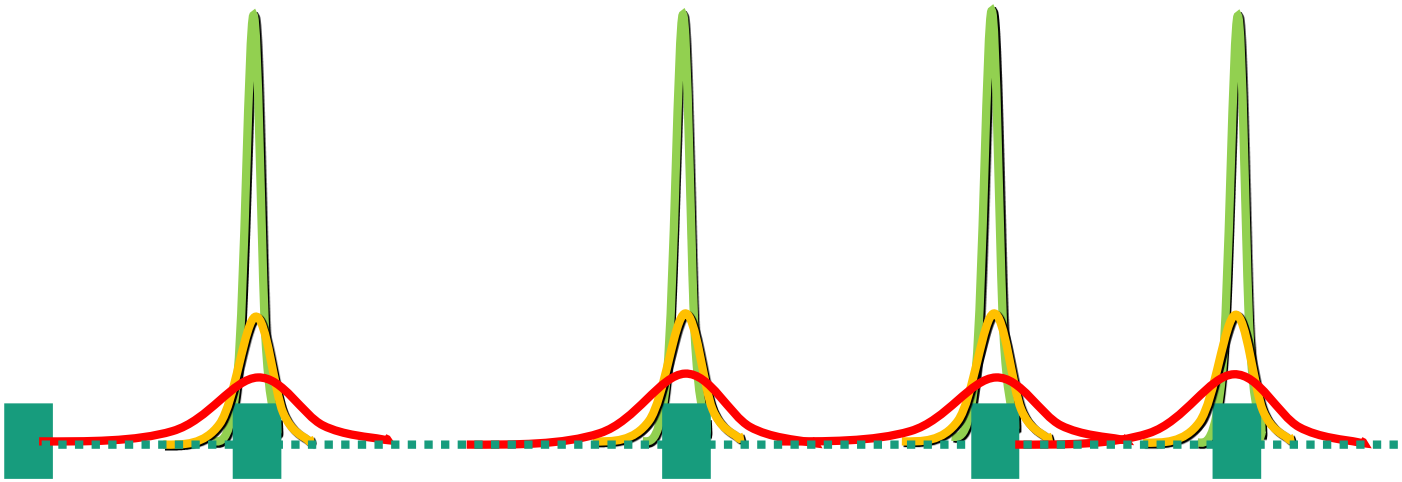
- Different metrics to measure segmenter performance
- Penalty metrics
  - Probability Metric (Doug et al. 1998)
  - Window Diff (Pevzner & Hearst, 2002)
- Rewarding metric
  - Boundary Similarity (Fournier & Inkpen 2013)

- Problem: What does 0.2 mean?

?

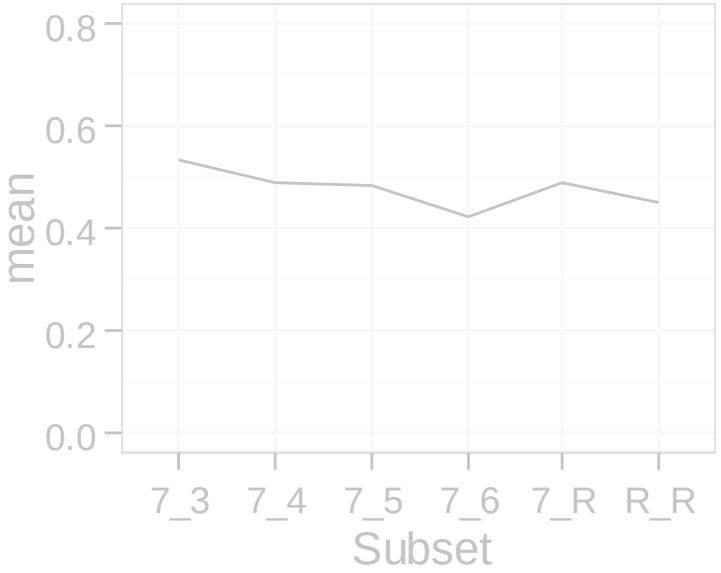
Scalable Segmentation Performance – a new baseline

Distance from true boundary	Standard deviation in % of avg. number of sentences
very close	$\sigma \in (0, 5]$
close	$\sigma \in (5, 15]$
large	$\sigma \in (15, 30]$

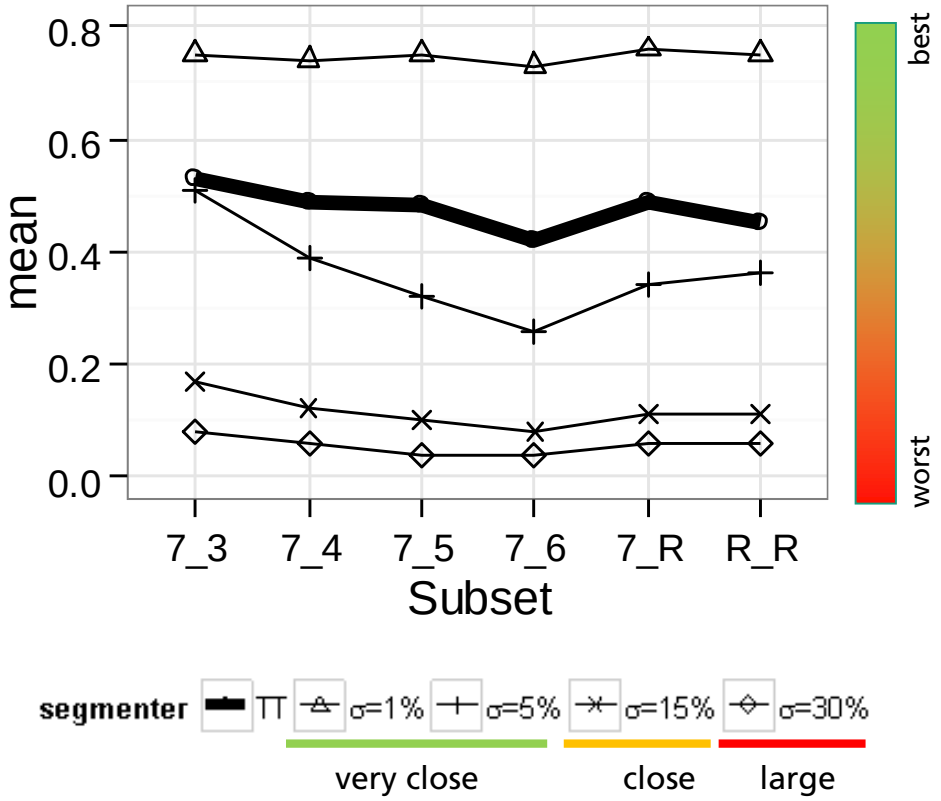


Results for TopicTiling on Macro Dataset

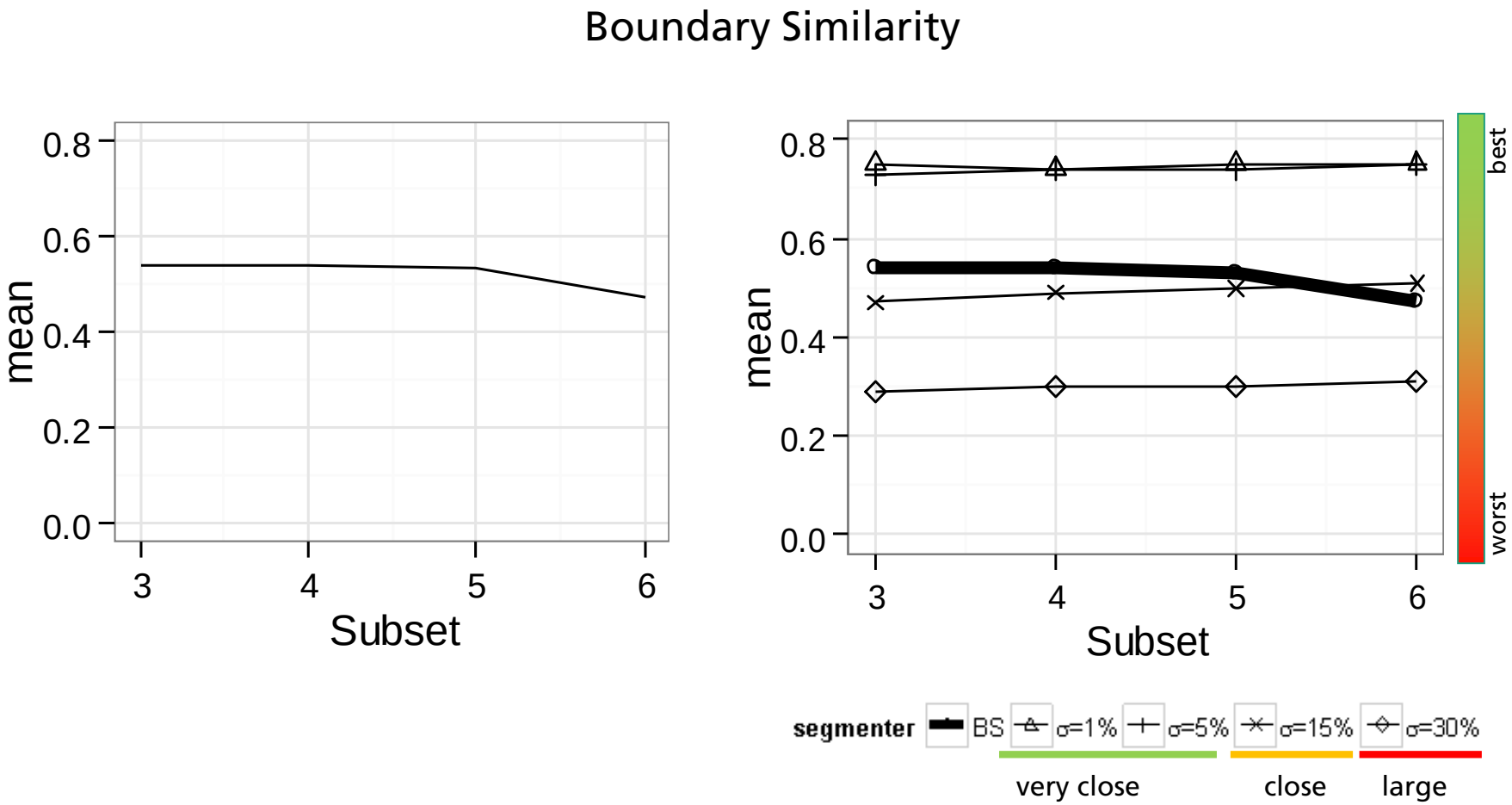
Boundary Similarity



With random segmenter as baseline:



Results for BayesSeg on Micro Dataset



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

- Beeferman, D., Berger, A. & Lafferty, J., 1999. Statistical Models for Text Segmentation. Mach. Learn., #feb#, 34(1-3), pp. 177-210.
- Bird, S., Klein, E. & Loper, E., 2009. Natural Language Processing with Python. s.l.:O'Reilly Media.
- Capuano, N. et al., 2009. LIA: an intelligent advisor for e-learning. Interactive Learning Environments, 17(3), pp. 221-239.
- Choi, F. Y. Y., 2000. Advances in Domain Independent Linear Text Segmentation. Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 26-33.
- Eisenstein, J. & Barzilay, R., 2008. Bayesian Unsupervised Topic Segmentation. Honolulu, Hawaii, Association for Computational Linguistics, pp. 334-343.
- Fournier, C., 2013. Evaluating Text Segmentation using Boundary Edit Distance. Stroudsburg, PA, USA, Association for Computational Linguistics, p. To appear.
- Fournier, C. & Inkpen, D., 2012. Segmentation Similarity and Agreement. Montreal, Canada, Association for Computational Linguistics, pp. 152-161.
- Galley, M., McKeown, K., Fosler-Lussier, E. & Jing, H., 2003. Discourse Segmentation of Multi-party Conversation. Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 562-569.
- Griffiths, T. L. & Steyvers, M., 2004. Finding scientific topics. Proceedings of the National Academy of Sciences, April, 101(Suppl. 1), pp. 5228-5235.
- Hearst, M. A., 1997. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. Comput. Linguist., #mar#, 23(1), pp. 33-64.
- Huang, X. et al., 2002. Applying Machine Learning to Text Segmentation for Information Retrieval. s.l.:s.n.
- Janin, A. et al., 2003. The ICSI Meeting Corpus. s.l., s.n., pp. I-364--I-367 vol.1.
- Kan, M.-Y., Klavans, J. L. & McKeown, K. R., 1998. Linear Segmentation and Segment Significance. s.l., s.n., pp. 197-205.
- Kim, S., Medelyan, O., Kan, M.-Y. & Baldwin, T., 2013. Automatic keyphrase extraction from scientific articles. Language Resources and Evaluation, 47(3), pp. 723-742.
- Lamprier, S., Amghar, T., Levrat, B. & Saubion, F., 2007. On Evaluation Methodologies for Text Segmentation Algorithms. s.l., s.n., pp. 19-26.
- Lin, Y.-T., Cheng, S.-C., Yang, J.-T. & Huang, Y.-M., 2009. An Automatic Course Generation System for Organizing Existent Learning Objects Using Particle Swarm Optimization. In: M. Chang, et al. Hrsg. Learning by Playing. Game-based Education System Design and Development. s.l.:Springer Berlin Heidelberg, pp. 565-570.
- Misra, H., Yvon, F., Jose, J. M. & Cappe, O., 2009. Text Segmentation via Topic Modeling: An Analytical Study. New York, NY, USA, ACM, pp. 1553-1556.
- Pevzner, L. & Hearst, M. A., 2002. A Critique and Improvement of an Evaluation Metric for Text Segmentation. Comput. Linguist., #mar#, 28(1), pp. 19-36.
- Riedl, M. & Biemann, C., 2012. TopicTiling: A Text Segmentation Algorithm Based on LDA. Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 37-42.
- 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.
- Sun, Q., Li, R., Luo, D. & Wu, X., 2008. Text Segmentation with LDA-based Fisher Kernel. Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 269-272.
- Swertz, C. et al., 2013. A Pedagogical Ontology as a Playground in Adaptive Elearning Environments.. s.l., GI, pp. 1955-1960.
- Tan, X., Ullrich, C., Wang, Y. & Shen, R., 2010. The Design and Application of an Automatic Course Generation System for Large-Scale Education. s.l., s.n., pp. 607-609.
- Utiyama, M. & Isahara, H., 2001. A Statistical Model for Domain-independent Text Segmentation. Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 499-506.
- Yaari, Y., 1997. Segmentation of Expository Texts by Hierarchical Agglomerative Clustering. s.l.s.n.