Automatic Text Segmentation: TextTiling

Will Roberts
wroberts@coli.uni-sb.de

Tuesday, 20 November, 2007

1. Introduction
   - Motivation
   - Approaches
2. TextTiling
   - Theory
   - Method
3. Evaluation
4. Conclusion

Uses for Automatic Text Segmentation

- hypertext display
- information/passage retrieval
- text summarization
- automatic text generation
- measuring stylistic variation for genre detection
- aligning parallel multilingual corpora
- breaking up connected documents

Other Text Segmentation Approaches

- Clustering or similarity matrices based on word co-occurance
- Machine-learning or hand-crafted solutions for detection of cue words

TextTiling Theory

- Focus on multi-paragraph units in expository text
  - Topics not always contained in single paragraphs
- Identify subtopic shifts
  - Subtopic: piece of text “about” something
  - Identify topic shift, not topic
  - Linear segmentation
- Subtopic shifts associated with change in vocabulary
  - Linguistically simple: no prosody, discourse markers, pronoun reference resolution, . . .

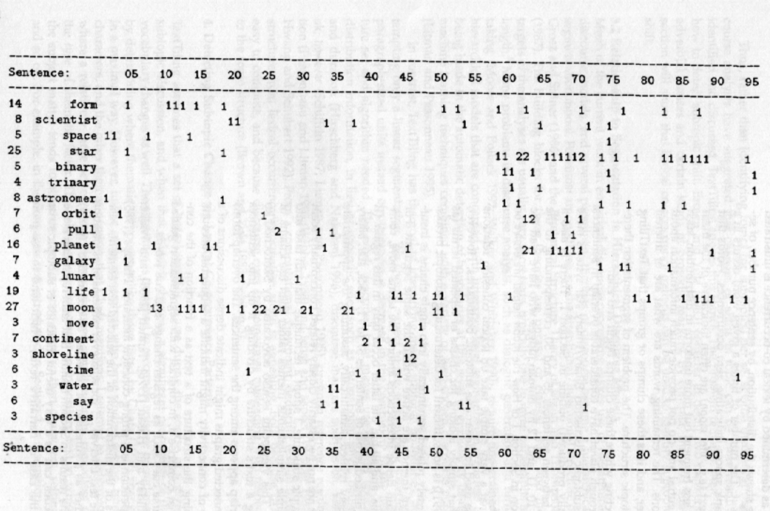
TextTiling Theory

- Focus on multi-paragraph units in expository text
  - Topics not always contained in single paragraphs
- Identify subtopic shifts
  - Subtopic: piece of text "about" something
  - Identify topic shift, not topic
  - Linear segmentation
- Subtopic shifts associated with change in vocabulary
  - Linguistically simple: no prosody, discourse markers, pronoun reference resolution, . . .

TextTiling Theory

- Focus on multi-paragraph units in expository text
    - Topics not always contained in single paragraphs
- Identify subtopic shifts
    - Subtopic: piece of text "about" something
    - Identify topic shift, not topic
    - Linear segmentation
- Subtopic shifts associated with change in vocabulary
    - Linguistically simple: no prosody, discourse markers, pronoun reference resolution, . . .

Word Occurance Counts



```
-------------------------------------------------------------------------------------------------------------
Sentence:        05   10   15   20   25   30   35   40   45   50   55   60   65   70   75   80   85   90   95
-------------------------------------------------------------------------------------------------------------
14       form    1     111 1   1                               1 1     1   1       1     1     1    1
 8  scientist                   11            1   1               1          1      1  1
 5      space 11    1     1
25       star    1             1                                       11 22  111112  1 1  1   11 1111     1
 5     binary                                                            11  1          1                    1
 4    trinary                                                             1   1         1                    1
 8 astronomer 1                1
 7      orbit   1                    1                                  1 1           1   1    1  1
 6       pull                          2     1 1                           12    1 1
16     planet   1    1       11               1           1                21  11111                  1     1
 7     galaxy    1                                                   1                1  11     1           1
 4      lunar            1  1     1       1
19       life 1  1  1                              1    11 1  11  1      1                 1 1    1 111  1 1
27       moon        13  1111   1 1 22 21  21    21           11 1
 3       move                                      1   1   1
 7  continent                                      2 1 1 2 1
 3  shoreline                                            12
 6       time                      1               1  1  1     1                                          1
 3      water                                 11            1
 6        say                                1 1        1        11                1
 3    species                                        1  1  1
-------------------------------------------------------------------------------------------------------------
Sentence:        05   10   15   20   25   30   35   40   45   50   55   60   65   70   75   80   85   90   95
-------------------------------------------------------------------------------------------------------------
```

TextTiling Algorithm

1. Tokenize
2. Calculate lexical similarity scores
3. Determine inter-sentence boundaries

Lexical Similarity

- Computed for each sentence gap in text
- Measure of the lexical similarity of the two sentences/blocks on either side
- Moving window of size k

Lexical Similarity

- Block comparison
  - Normalized inner product of two word vectors
- New vocabulary
  - New words introduced in a window centered around the sentence boundary
- Lexical chaining
  - Number of lexical chains active at the sentence boundary

Lexical Similarity

- Block comparison
  - Normalized inner product of two word vectors
- New vocabulary
  - New words introduced in a window centered around the sentence boundary
- Lexical chaining
  - Number of lexical chains active at the sentence boundary

Lexical Similarity

- Block comparison
  - Normalized inner product of two word vectors
- New vocabulary
  - New words introduced in a window centered around the sentence boundary
- Lexical chaining
  - Number of lexical chains active at the sentence boundary

Lexical Similarity

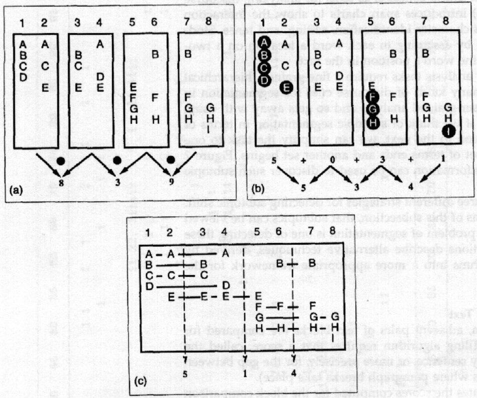


Figure 3
Illustration of three ways to compute the lexical score at gaps between sentences. Numbers indicate a numbered sequence of sentences, columns of letters signify the terms in the given sentence. (a) Blocks – dot product of vectors of word counts in the block on the left and the block on the right. (b) Vocabulary introduction – the number of words that occur for the first time within the interval centered at the sentence gap. (c) Chains – the number of active chains, or terms that repeat within threshold sentences and span the sentence gap.

Calculating Word Significance Values

Example

Saarland University (German: Universität des Saarlandes) is a university located in Saarbrücken, the capital of the German state of Saarland. It was founded in 1948 in co-operation with France and is organized in 8 faculties that cover all major fields of science. The university is particularly well known for research and education in Computer Science and Medicine.

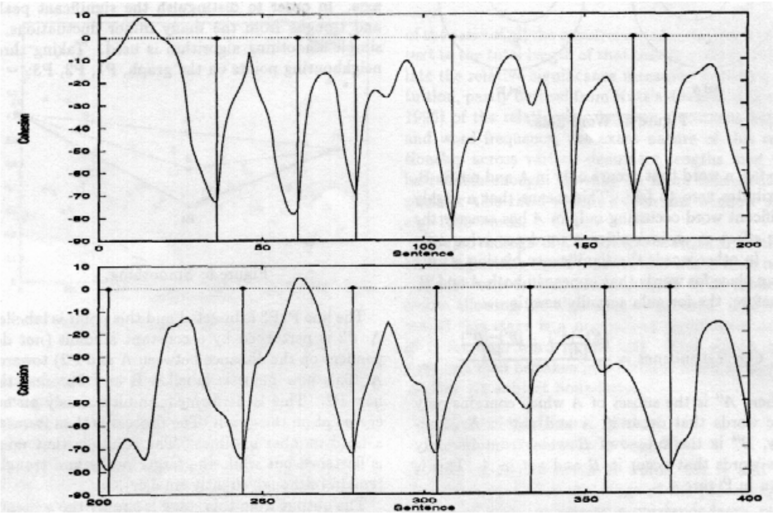
Saarland University, the first to be established after the Second World War, was founded in November 1948 with the support of the French Government and under the auspices of the University of Nancy.

At the time the Saarland found itself in the special situation of being partly autonomous and linked to France by economic . . .

Boundary Identification

- *Depth score* is computed at each sentence gap from the lexical scores
  - Score is the sum of the heights of the peaks on either side of the sentence gap
  - $D_i = (s_{i-1} - s_i) + (s_{i+1} - s_i) = s_{i-1} + s_{i+1} - 2s_i$
- Smooth depth scores and choose local minima

Boundary Identification



Evaluation Caveats

- Algorithm depends on parameters: window size, smoothing, number of boundaries
- Evaluation depends on desired application: precision, recall, and *near-misses*

Evaluation Methods

- Breaking consecutive documents
- Comparison with human judges
  - Humans rarely agree on correct segmentation
  - Consensus of human judges can be used as “gold standard”
- P_k metric

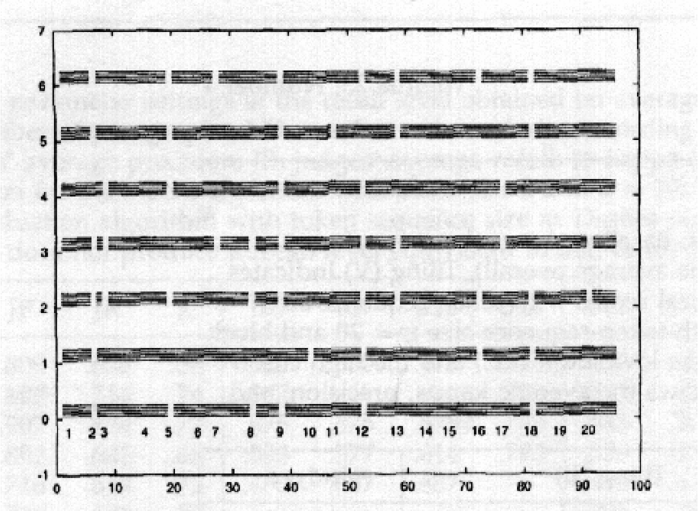
Evaluation Methods

- Breaking consecutive documents
- Comparison with human judges
    - Humans rarely agree on correct segmentation
    - Consensus of human judges can be used as “gold standard”
- P_k metric

Evaluation Methods

- Breaking consecutive documents
- Comparison with human judges
  - Humans rarely agree on correct segmentation
  - Consensus of human judges can be used as “gold standard”
- P_k metric

Human Judges



P_k Metric

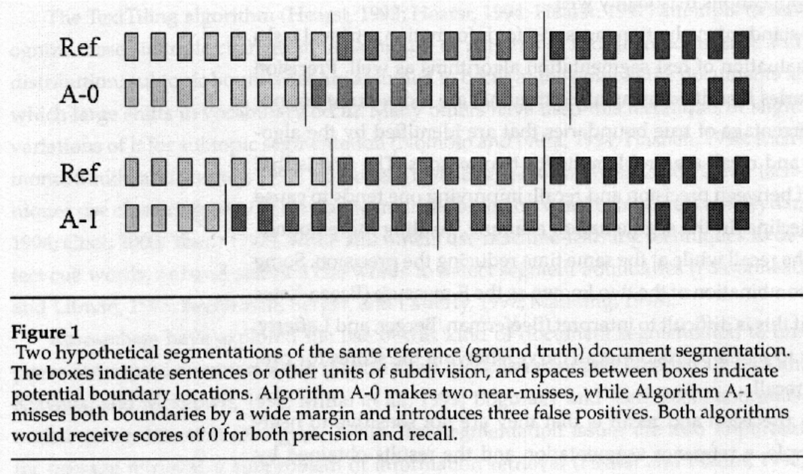


Figure 1
Two hypothetical segmentations of the same reference (ground truth) document segmentation. The boxes indicate sentences or other units of subdivision, and spaces between boxes indicate potential boundary locations. Algorithm A-0 makes two near-misses, while Algorithm A-1 misses both boundaries by a wide margin and introduces three false positives. Both algorithms would receive scores of 0 for both precision and recall.

P_k Metric

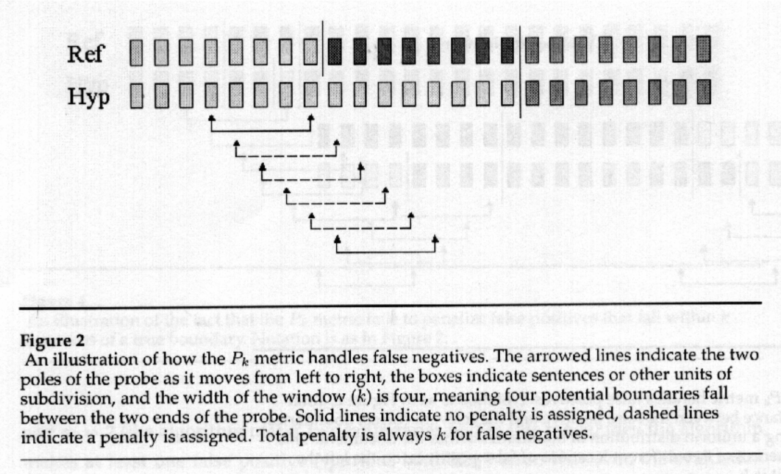


Figure 2
An illustration of how the P_k metric handles false negatives. The arrowed lines indicate the two poles of the probe as it moves from left to right, the boxes indicate sentences or other units of subdivision, and the width of the window (k) is four, meaning four potential boundaries fall between the two ends of the probe. Solid lines indicate no penalty is assigned, dashed lines indicate a penalty is assigned. Total penalty is always k for false negatives.

Conclusion

Strengths

- Linguistically and computationally simple
- Language independent

Weaknesses

- Designed for expository text; poor for narrative texts and discourse
- *Near-miss* errors might be unacceptable for some applications
- Cannot extract hierarchical text structure

Literature

Hearst, M.
Segmenting Text into Multi-paragraph Subtopic Passages
Computational Linguistics, 23(1), 1997.

Pevzner, L., and Hearst, M.
A Critique and Improvement of an Evaluation Metric for Text Segmentation
Computational Linguistics, pp. 9–16, 2001.

Richmond, K., Smith, A., and Amitay, E.
Detecting Subject Boundaries Within a Text: A Language Independent Statistical Approach
EMNLP, 1997.