Lexical and Semantic Methods in Inner Text Topic Segmentation: A Comparison between C99 and Transeg

Alexandre Labadié and Violaine Prince

LIRMM
161 rue Ada
34392 Montpellier Cedex 5, France
{labadie,prince}@lirmm.fr
http://www.lirmm.fr

Abstract. This paper present a semantic and syntactic distance based method in topic text segmentation and compare it to a very well known text segmentation algorithm: c99. To do so we ran the two algorithms on a corpus of twenty two French political discourses and compared their results.

Key words: Text segmentation, topic change, c99

1 C99 and Transeg

In this paper we compare Transeg, a text segmentation method based on distances between text segments and a fairly deep syntactic and semantic analysis, to c99 [4]. Our goal is to assess the importance of syntactic and semantic information in text segmentation.

1.1 C99

Developed by Choi [4], c99 is text segmentation algorithm strongly based on the lexical cohesion principle. It is, at this time, one the best algorithms in the text segmentation domain.

1.2 Transeg: A distance based method

We have developed a distance based text segmentation specifically designated to find topic variations inside the text called Transeg.

The first step of our approach is to convert each text sentence into a semantic vector obtained using the French language parser SYGFRAN [2].

Using this sentence representation, we try to find transition zones inside the text using a protocol described in [5].

In our first implementation of this method, we used the angular distance to compute the transition score. In this paper we use an extended version of the concordance distance first proposed by [3]. This improvement has enhanced the discriminant capabilities of our method.

2 Experiment and result on French political discourses

In order to compare the two methods, we tried them on a set of twenty two French political discourses and we measured their scores in text topic boundaries detection.

We chose a set of French political discourses for two main reasons: (1) As they were identified by experts, internal boundaries looked less artificial than just beginnings of concatenated texts. (2) The topical structure of a political discourse

should be more visible than other more mundane texts.
From an original corpus of more than 300, 000 sentences of a questionable quality we extracted and cleaned 22 discourses totalizing 1, 895 sentences and 54, 551 words.
We set up a run of both Transeg and the LSA augmented c99 (Choi's algorithm) on each discourse separately. To be sure that there is not any implementation error, we used the 1.3 binary release that can be downloaded on Choi's personal Linguaware Internet page (`http://www.lingware.co.uk/homepage/freddy.choi/software/software.htm`).
To evaluate the results of both methods, we used the DEFT'06 **tolerant recall and precision** ([1]).
First of all, we see that results (table 1) are not spectacular. $FScore$ is a very

	Words	Sentences	Transeg			c99		
			Precision	Recall	FScore	Precision	Recall	FScore
Text 3	2767	92	42.86	85.71	**28.57**	20	14.29	8.33
Text 6	5348	212	8.7	18.18	5.88	20	18.18	**9.52**
Text 9	1789	53	75	100	**42.86**	25	16.67	10
Text 19	678	26	33.33	33.33	16.67	50	66.67	**28.57**
Text 22	1618	40	60	75	**33.33**	100	25	20

Table 1. c99 and Transeg topic segmentation results

strict measure, even when softened by using tolerant recall and precision.
Transeg has a better $FScore$ on 16 on the 22 documents composing the corpus. On these 16 texts, our recall is always better or equal to c99 and our $FScore$. Transeg has also the best $FScore$ of both runs with 42.86 on text 9. C99 has a better $FScore$ on 6 texts. Anyway, we should notice that c99 has comparatively good precision on most of the texts. Thus, when examining texts where c99 is better we see that they fit into two categories: (1) Texts with few boundaries. C99 seems to be very effective on short texts with just one inner topic boundary. (2) Enumerations. Text 6 for example, which is quite big, is a record of the government spokesman where he enumerates dealt subjects during the weekly minister reunion.

According to the experiment results, Transeg seems to be more effective at finding inner text segments than c99. It seems to be efficient on longer documents, with multiple and related topics. Whereas a lexically based method is efficient on either short texts with very few topics, or enumerations and/ or concatenation of unrelated topics or subjects.

References

1. J. Azé, T. Heitz, A. Mela, A. Mezaour, P. Peinl, and M. Roche. Présentation de deft'06 (defi fouille de textes). *Proceedings of DEFT'06*, 1:3–12, 2006.
2. J. Chauché. Un outil multidimensionnel de l'analyse du discours. *Proceedings of Coling'84*, 1:11–15, 1984.
3. J. Chauché, V. Prince, S. Jaillet, and M. Teisseire. Classification automatique de textes partir de leur analyse syntaxico-sémantique. *Proceedings of TALN'03*, pages 55–65, 2003.
4. F. Y. Y. Choi, P. Wiemer-Hastings, and J. Moore. Latent semantic analysis for text segmentation. *Proceedings of EMNLP*, pages 109–117, 2001.
5. V. Prince and A. Labadié. Text segmentation based on document understanding for information retrieval. *In Proceedings of NLDB'07*, pages 295–304, 2007.