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

Alexandre Labadié, Violaine Prince

1

1

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. Our two conclusions are that the two approaches are complementary and that evaluation methods in this domain should be revised.

Key words: Text segmentation, topic change, c99

Introduction

There are many distinct tasks labeled as 'text segmentation'. For instance, identifying and extracting text from multimedia support where it is mixed with pictures or videos is called as such [9]. The task of grouping words into morphemes or bigger linguistic units is sometimes also referred as text segmentation (e.g. in written Asiatic languages where words boundaries are not easy to assess[18], [19]). In this paper, we concentrate on '**topic based text segmentation**'. This type of process tries to find the topical structure [8] of a text and thus provide a possible thematic decomposition of a given document [14]. Most texts do not talk about only one topic. The bigger the documents, the more topics they include. The goal of topic based text segmentation is to find where a topic begins and where it ends, within a given text. For practical purposes, we will use the label 'text segmentation' to refer to topic based text segmentation.

Basically, the goal of text segmentation is to divide a text into multiple segments which are thematically coherent and distinct. Each of these text segments should ideally bear one topic, but topics could be complex units from a rhetorical point of view, needing explanations, examples or argumentations.

This brings out the question of defining the concept of a topic. Browsing literature shows that there are several definitions and a large body of works in (topical) text segmentation. Generally speaking, a topic is: *The subject matter of a conversation or discussion.* In linguistics, it is defined as: *The part of the*

proposition that is being talked about (predicated). Thus one may admit that the topic of a text segment is *what talking is about.* So, the goal of an automatized text segmentation could be simplified into dividing a text in segments, each sentence of which 'talks about' the same subject.

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, a well known lexical text segmentation algorithm. Our goal is to assess the importance of syntactic and semantic information in text segmentation. In a first part we will present the two methods compared in this paper. In a second part we will compare them on a corpus of twenty two French political discourses and examine the results. In the conclusion we will discus the validity of automatic evaluation methods of such approaches and present possible evolutions of Transeg.

1   C99 and Transeg

As said in introduction, literature is abundant on the subject, and mostly methods divide into two main categories: Supervised ones, more or less data dependent, and unsupervised methods, trying to avoid the liabilities of learning. In this paper we concentrate on unsupervised methods, since they can be evaluated on corpora as broadly distinct as possible, which is a better case for evaluation.

1.1   C99

Developed by Choi [6], c99 is text segmentation algorithm strongly based on the lexical cohesion principle [13]. It is, at this time, one the best algorithms (if not the best) in the text segmentation domain (as defined in the introduction) [2].

Quick description C99 uses similarity matrix of the text sentences. First projected in a word vectorial space representation, sentences are then compared using the cosine similarity measure (by the way, the most used measure). Similarity values are used to build the similarity matrix. More recently, Choi improved c99 by using the Latent Semantic Analysis (LSA) achievements to reduce the size of the word vectorial space [7].

But the author does not work on the similarity matrix. Instead, he builds a second matrix known as the rank matrix. The latter is computed by giving to each case of the similarity a rank equal to the number of cases around the examined one (in a layer) which have a lesser similarity score. This rank is normalized by the number of cases that were really inside the layer to avoid side effects.

C99 then finds topic boundaries by recursively seeking the optimum density of matrices along the rank matrix diagonal. The algorithm stops when the optimal boundaries returned are the end of the current matrix or, if the user gave this parameter to the algorithm, when the maximum number of text segments is reached.

Most of lexically based algorithms use few, if not at all, syntactic and semantic

information. C99 is no exception. Even if LSA adds some semantic information to the method, it is also a lexical cohesion based method and thus only partially compensates this absence. **Lexical Cohesion** is the fundamental concept upon which word-based algorithms. The idea is that sets of close words create a topic, and another topic emerges when another set is detected. It has been properly described in [13].

Limits As an almost exclusively lexical cohesion based method, c99 only looks for similar and/or different words to find text segments or boundaries. In natural language, the word/constituent syntactic function also bears information. If a noun is the subject of a verb, it could mean something totally different from what it would have meant if it were its object. This information is not taken into account in word-based methods.

Another limitation of such methods, pointed out by [17], is the intensive use of synonyms as a stylistic effect. In many languages, and particularly in French, the language on which we experiment, repeating several times the same word in a paragraph or even a short text is considered unsightly. This massive use of synonyms makes these approaches quite inefficient as they are based on the exact repetition of words. It is possible to use some semantic resources like WordNet to counterbalance this, but languages requiring such a use of synonyms have also great polysemy issues. So, doing so only changes the issue into another.

Thus, the following question becomes relevant: What would a method involving syntactic infomation and sentence semantics, bring to text segmentation? We have tried to provide a first answer in the next paragraphs.

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.

Textual representation The first step of our approach is to convert each text sentence into a semantic vector obtained using the French language parser SYGFRAN [3] (It could be transposed to any language with a good parser producing constituents and dependencies). These vectors are Roget-like semantic vectors ([16]), but using the Larousse thesaurus ([11]) as a reference (Roget being the reference for English). Sentence vectors are recursively computed by linearly combining sentence constituents, which are themselves computed by a linear combination of word vectors. The weight of each word vector is proportional to its function, determined through a constituent and dependency decomposition of the sentence performed by the parser (The formula is given in [4]). So, these vectors bear both the semantic and the syntactic information of the sentence.

Text segmentation Using this sentence representation, we try to find transition zones inside the text. The notion of **transition zone** comes from the idea

that topic change boundaries inside a text are not isolated sentences, but small groups of sentences, acting as a transition between two topics. To find them, we slide a window along the text, considering each half of the window as a potential segment. Each potential text segment is then represented by one vector, which is a weighted barycenter of its sentence vectors. We take into account stylistic information by giving a better weight to first sentences, relying on the fact that introductions bear the important information ([10],[12]). Then we compute a distance (we call it **thematic distance**) between the two barycenters, and consider it as the window central sentence transition score.

Practically, transition zones are successive sentences with a transition score
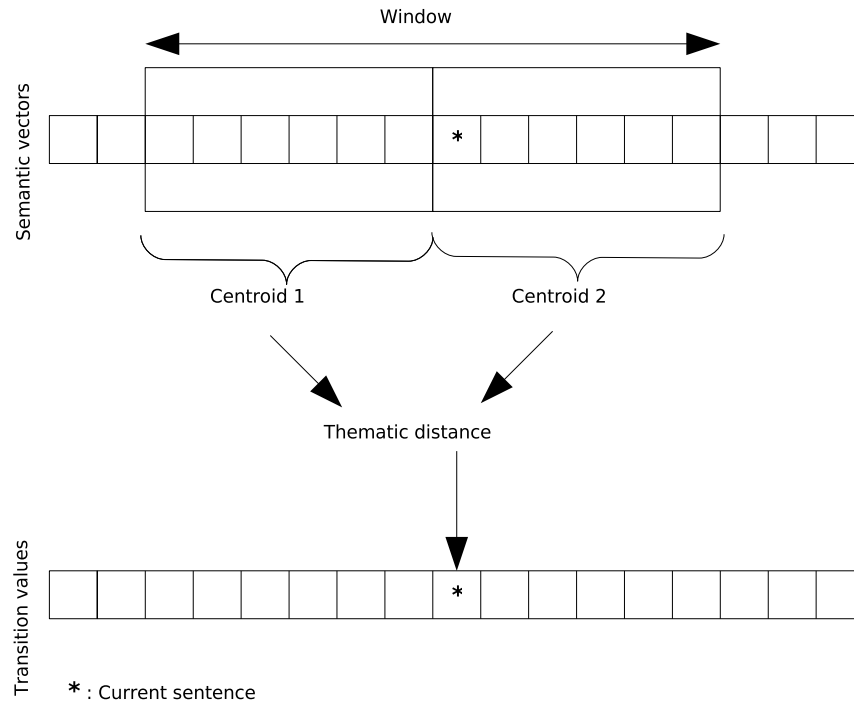
Fig. 1. Giving a transition score to each sentences

greater than a threshold. This threshold is the result of a detailed observation of the DEFT'06 political corpus (DEFT is a challenge conference close to the TREC Novelty Task, but focusing on French as a main language. The DEFT06 [1] topic was about text segmentation, and several corpora, among which political discourses, were provided to competitors. A more detailed information about data is given in next section). We computed distances on many discourses and

their topic segments (the total number of their sentences was around $100,000$) and obtained an average distance of 0.45 and a σ of 0.08. Boundary sentences are selected in the transition zones. A more detailed description of this approach can be found in [15] where a successful application to relevant segment retrieval has been described.

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 [5]. This improvement has enhanced the discriminant capabilities of our method.

Concordance distance Semantic vectors resulting from parsing have 873 components and most of them are not even activated. With so many null values in the vector, the angular distance is not discriminant enough. The goal of the concordance distance is to be a sharper tool. It does not only consider the vectors components values, but their ranks too.

Let two vectors $\overrightarrow{A}$ and $\overrightarrow{B}$ be; We sort their values from the most activated (highest) to the less activated (lowest value) and choose to keep only the first values of the new vectors ($\frac{1}{3}$ of the original vector). $\overrightarrow{A_{sr}}$ and $\overrightarrow{B_{sr}}$ are respectively the sorted and reduced versions of $\overrightarrow{A}$ and $\overrightarrow{B}$. Obviously $\overrightarrow{A_{sr}}$ and $\overrightarrow{B_{sr}}$ could have no common strong component (so the distance will be 1), but if they have some, we can compute two differences :

The rank difference: if i is the rank of C_t a component of $\overrightarrow{A_{sr}}$ and $\rho(i)$ the rank of the same component in $\overrightarrow{B_{sr}}$, we have :

$$E_{i,\rho(i)} = \frac{(i - \rho(i))^2}{Nb^2 + (1 + \frac{i}{2})} \tag{1}$$

Where Nb is the number of values kept.

The intensity difference: We also have to compare the intensity of common strong components. If a_i is the intensity of i rank component from $\overrightarrow{A_{sr}}$ and $b_{\rho(i)}$ the intensity of the same component in $\overrightarrow{B_{sr}}$ (its rank is $\rho(i)$), we have:

$$I_{i,\rho(i)} = \frac{\left\| a_i - b_{\rho(i)} \right\|}{Nb^2 + (\frac{1+i}{2})} \tag{2}$$

These two differences allow us to compute an intermediate value P:

$$P(\overrightarrow{A_{sr}}, \overrightarrow{B_{sr}}) = (\frac{\sum_{i=0}^{Nb-1} \frac{1}{1 + E_{i,\rho(i)} * I_{i,\rho(i)}}}{Nb})^2 \tag{3}$$

As P concentrate on components intensities and ranks, we introduce the overall components direction by mixing P with the angular distance. If $\delta(\overrightarrow{A}, \overrightarrow{B})$ is the angular distance between $\overrightarrow{A}$ and $\overrightarrow{B}$, then we have:

$$\Delta(\overrightarrow{A_{sr}}, \overrightarrow{B_{sr}}) = \frac{P(\overrightarrow{A_{sr}}, \overrightarrow{B_{sr}}) * \delta(\overrightarrow{A}, \overrightarrow{B})}{\beta * P(\boldsymbol{A_{sr}}, \overrightarrow{B_{sr}}) + (1 - \beta) * \delta(\overrightarrow{A}, \overrightarrow{B})} \tag{4}$$

Where β is a coefficient used to give more (or less) weight to P. It is easy to prove that neither P nor $\Delta(\overrightarrow{A_{sr}}, \overrightarrow{B_{sr}})$ are symmetric.

But $\Delta(\overrightarrow{A_{sr}}, \overrightarrow{B_{sr}})$ was designed in a context of text classification, to compare text vectors to class vectors. As only the likelihood of a text to the class center had to be measured, $\Delta(\overrightarrow{A_{sr}}, \overrightarrow{B_{sr}})$ did not need to be symmetric. But in our context of text segmentation we needed a symmetric value. even if $\overrightarrow{A}$ come before $\overrightarrow{B}$ in a text, $\overrightarrow{A}$ is not more important than $\overrightarrow{B}$. So the final concordance distance $D(\overrightarrow{A}, \overrightarrow{B})$ we use, is:

$$D(\overrightarrow{A}, \overrightarrow{B}) = \frac{\Delta(\overrightarrow{A_{sr}}, \overrightarrow{B_{sr}}) + \Delta(\overrightarrow{B_{sr}}, \overrightarrow{A_{sr}})}{2} \tag{5}$$

2 Experiment and result on French political discourses

We have set up an experiment comparing c99 and our method. Both are unsupervised, therefore not data sensitive (they don't learn, don't adapt to data specificities, therefore a given corpus could be used several times with no effect on results). The first is recognized as one the best in the text segmentation field, the second has been tested in the DEFT'06 challenge and used to improve information retrieval in [15]. So, in order to compare methods, we tried them on a set of twenty two French political discourses and we measured their scores in text topic boundaries detection. The following subsections describe data, experiments and results.

2.1 Data: a corpus of French political discourse

We chose a set of French political discourses, among several other corpora, for two main reasons:

– As they were identified by experts, internal boundaries looked less artificial than just beginnings of concatenated texts.
– As an argumentative text, the topical structure of a political discourse should be more visible than other more mundane texts.

As previously said the 2006 DEFT edition was about finding topic boundaries in three different corpora in politics (the one we chose) law and science, but we discarded the other domains because of several biases introduced by artificial devices (e.g., words such as 'article' in European law texts, or paragraph line breaks that were questionably considered as a topic frontiers in the science corpora by the organizers).

There was a lot of noise inside the political corpus. Some discourses were exclusively in capital letters, which is quite annoying when processing a language like French, that discriminates words according to accents on vowels. And some of the 'discourses' were, in fact, interviews. So, we manually selected, separated and cleaned twenty two discourses from this corpus.

From an original corpus of more than $300,000$ sentences of a questionable quality we extracted 22 discourses totalizing $1,895$ sentences and $54,551$ words. No information on the discourses were at our disposal, except the beginning of topic segments (which could have been beginnings of texts or real topic boundaries), so this manual cleaning of the corpus took a lot of time and significantly reduced the amount data. But it was a necessity to have a workable data set.

The original corpus was fairly corrupted: Beside the already cited entire sentences in capital letters, empty sentences, punctuation repetition, and other liabilities degraded the available data. That aspect, in our opinion, brings some discredit to DEFT'06 results. However, noise is a common problem in natural language processing and as it should be done with, it should not invalidate the DEFT'06 experiment as such. In our case, as we tried to compare two different approaches on a very specific task, we needed the cleanest data set possible.

2.2 Experiments

We set up a run of both Transeg and the LSA augmented c99 (Choi's algorithm) on each discourse separately. We chose to use the latest version of c99 because it is commonly recognized as one of the best text segmentation methods (if not the best at all). 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]). These 'fuzzy' values of recall and precision count as relevant, potential boundary sentences which are in a window around the boundary sentence identified by experts. This evaluation gives a better idea of algorithms efficiency on the task of finding inner texts topic boundaries and does not have a significant influence on the task of finding concatenated texts boundaries. The team of DEFT'06 noticed in [1] that the use of either strict or tolerant measures had no effect on the ranking of the submissions they had to evaluate.

We computed the corresponding tolerant $FScore$ using the well known formula:

$$FScore = \frac{(\beta^2 + 1) * recall * precision}{\beta^2 * precision + recall} \tag{6}$$

With $\beta = 1$.

We have to note that both methods consider first sentences of texts as a boundaries and that every first sentence of each text is considered as a boundary when computing recall, precision and $FScore$ (so both methods have always at least one good answer).

2.3 Results

First of all, we see that results (table 1) are not spectacular. $FScore$ is a very strict measure, even when softened by using tolerant recall and precision. The

best $FScore$, obtained by Transeg in text 9, is only of 42.86 (all results were multiplied by 100 for legibility purpose and to be read as percentages) for a precision of 75 and a recall of 100. That gives us a good view of the quality of current text segmentation methods and of the progress we still have to aim at in this domain.

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 $FScore$ ranges from 20% (text 1) up to 329% (text 9) better than c99 corresponding result. Transeg has also the best $FScore$ of both runs with 42.86 on text 9. C99 has a better $FScore$ on 6 texts, but it is at best twice Transeg $FScore$ on the same text. 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:

– Texts with few boundaries. C99 seems to be very effective on short texts with just one inner topic boundary. With few boundaries identified, and first sentences always identified as boundaries, mathematically, c99 has a very good precision on such short texts (text 10 for example).
– 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. So it is basically an enumeration of different subjects with different vocabularies and no real transition between the different segments.

3   Conclusion

According to the experiment results, Transeg seems to be more effective at finding inner text segments than c99. So, the introduction of semantic and syntactic does have an appreciable effect on text segmentation. 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. This brings up an interesting highlighting of the different skills developed through the difference of language granularity.

By comparing methods that are not data sensitive (with no learning), and experimenting both algorithms on the same set of clean data, we have limited as strictly as possible the introduction of biases in evaluation. This is not the case in most evaluation tracks with which we have been confronted. In a great majority text or information retrieval challenges, learning based algorithms tune to their learning corpora. If the test data is not utterly different, then they are automatically favored in any competition. Unsupervised algorithms such as those we compared also perform according to their innerbuilt capabilities. If competitions concatenate distinct texts and offer them as a text segmentation task, then lexical based methods are also intrinsically favored. Whereas if the real situation for text segmentation, i.e., providing texts, and asking competitors to structure them into subtopical segments, occurs, then methods that improve their results

	Words	Sentences	Transeg			c99		
			Precision	Recall	FScore	Precision	Recall	FScore
Text 1	617	22	50	33.33	**20**	33.33	33.33	16.67
Text 2	3042	100	33.33	37.5	**17.65**	50	12.5	10
Text 3	2767	92	42.86	85.71	**28.57**	20	14.29	8.33
Text 4	1028	40	33.33	33.33	**16.67**	20	33.33	12.5
Text 5	4532	157	12.5	18.18	**7.41**	16.67	9.09	5.88
Text 6	5348	212	8.7	18.18	5.88	20	18.18	**9.52**
Text 7	1841	47	100	42.86	**30**	100	14.29	12.5
Text 8	1927	74	60	33.33	**21.43**	100	11.11	10
Text 9	1789	53	75	100	**42.86**	25	16.67	10
Text 10	1389	31	33.33	20	12.5	100	20	**16.67**
Text 11	2309	81	30	50	**18.75**	33.33	16.67	11.11
Text 12	7193	211	15.38	6.25	**4.44**	33.33	3.13	2.86
Text 13	6097	305	20.59	33.33	**12.73**	17.65	14.29	7.89
Text 14	1417	57	40	33.33	**18.18**	100	16.67	14.29
Text 15	3195	79	40	8	6.67	66.67	8	**7.14**
Text 16	1995	60	66.67	28.57	20	57.14	57.14	**28.57**
Text 17	558	16	33.33	33.33	16.67	50	66.67	**28.57**
Text 18	696	25	100	37.5	**27.27**	40	25	15.38
Text 19	678	26	33.33	33.33	16.67	50	66.67	**28.57**
Text 20	1388	57	50	66.67	**28.57**	100	16.67	14.29
Text 21	3127	110	62.5	25	**17.86**	40	10	8
Text 22	1618	40	60	75	**33.33**	100	25	20

Table 1. c99 and Transeg topic segmentation results

by using syntax and sentence semantics are likely to do better.

Thus, one cannot but notice the complementarity of both approaches. If it is hard consider a complete fusion of c99 and Transeg in order to maximize results, the development of an automatic process, choosing between methods based on text properties, should be a viable evolution of current topic based text segmentation methods.

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. Y. Bestgen and S. Piérard. Comment évaluer les algorithmes de segmentation automatiques ? essai de construction d'un matriel de référence. *Proceedings of TALN'06*, 2006.
3. J. Chauché. Un outil multidimensionnel de l'analyse du discours. *Proceedings of Coling'84*, 1:11–15, 1984.
4. J. Chauché and V. Prince. Classifying texts through natural language parsing and semantic filtering. *In Proceedings of LTC'03*, 2007.
5. 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.
6. F. Y. Y. Choi. Advances in domain independent linear text segmentation. *Proceedings of NAACL-00*, pages 26–33, 2000.
7. F. Y. Y. Choi, P. Wiemer-Hastings, and J. Moore. Latent semantic analysis for text segmentation. *Proceedings of EMNLP*, pages 109–117, 2001.
8. M. A. Hearst and C. Plaunt. Subtopic structuring for full-length document access. *Proceedings of the ACM SIGIR-93 International Conference On Research and Development in Information Retrieval*, pages 59–68, 1993.
9. D. Karatzas. *Text Segmentation in Web Images Using Color Perception and Topological Features*. ECS Publications, UK, 2003.
10. A. Labadié and Chauché. Segmentation thématique par calcul de distance sémantique. *Proceedings of DEFT'06*, 1:45–59, 2006.
11. Larousse. *Thésaurus Larousse - des idées aux mots, des mots aux idées*. Larousse, Paris, 1992.
12. A. Lelu, C. M., and S. Aubain. Coopération multiniveau d'approches non-supervises et supervises pour la detection des ruptures thématiques dans les discours présidentiels franais. *In Proceedings of DEFT'06*, 2006.
13. J. Morris and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17:20–48, 1991.
14. J. M. Ponte and W. B. Croft. Text segmentation by topic. *European Conference on Digital Libraries*, pages 113–125, 1997.
15. V. Prince and A. Labadié. Text segmentation based on document understanding for information retrieval. *In Proceedings of NLDB'07*, pages 295–304, 2007.
16. P. Roget. *Thesaurus of English Words and Phrases*. Longman, London, 1852.
17. L. Sitbon and P. Bellot. Evaluation de méthodes de segmentation thématique linéaire non supervisés après adaptation au franais. *Proceedings of TALN'04*, 2004.
18. Z. Wu and G. Tseng. Chinese text segmentation for text retrieval: Achievements and problems. *Journal of the American Society for Information Science*, 44:532–542, 1993.

19. C. C. Yang and K. W. Li. A heuristic method based on a statistical approach for chinese text segmentation. *Journal of the American Society for Information Science and Technology*, 56:1438–1447, 2005.