The Impact of Corpus Quality and Type on Topic based Text Segmentation Evaluation

Alexandre Labadié, Violaine Prince

The impact of corpus quality and type on topic based text segmentation evaluation

Alexandre Labadié
LIRMM
161, rue Ada
34391 Montpellier Cedex 5
Email: labadie@lirmm.fr

Violaine Prince
LIRMM
161, rue Ada
34391 Montpellier Cedex 5
Email: prince@lirmm.fr

***Abstract*—In this paper, we try to fathom the real impact of corpus quality on methods performances and their evaluations. The considered task is topic-based text segmentation, and two highly different unsupervised algorithms are compared: $C99$, a word-based system, augmented with LSA, and $Transeg$, a sentence-based system. Two main characteristics of corpora have been investigated: Data quality (clean vs raw corpora), corpora manipulation (natural vs artificial data sets). The corpus size has also been subject to variation, and experiments related in this paper have shown that corpora characteristics highly impact recall and precision values for both algorithms.**

I. Introduction

There are several distinct tasks labeled as 'text segmentation', among which **topic-based text segmentation** is the particular process that tries to find the topical structure [7] of a text and thus provide a possible thematic decomposition of a given document [16]. It relies on the evidence that most texts do not talk about only one topic and the longer the documents, the more topics they include. There are several definitions of a topic in the relevant literature. 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)*. In this paper we do not consider the formal definition of topic, but the more commonly admitted definition that a topic of a text is *what talking is about*. So, the goal of an automated text segmentation could be simplified into dividing a text in segments, each sentence of which 'talks about' the same subject. Segments are supposed to be distinct, and coherent [13].

Text segmentation, whether topic-based or not, has been thoroughly evaluated in several campaigns (in several TREC (Text Retrieval Evaluation Conference) editions, as well as in its French equivalent DEFT (Défi Fouille de Textes), especially in the DEFT 2006 session devoted to text segmentation [1]). However, if protocols (concatenating texts like in [6]), methods [2], and measure metrics [15] have been evoked, the corpora features have not been at the center of the attention. Apart from some general characteristics such as corpus lengths (in words, sentences or MBytes), or origin (e.g. news [19], newspapers or magazine articles [8]), corpora seem to have escaped their own evaluation procedure.

As participants in several text retrieval evaluation campaigns, we have been accustomed to dealing with corpora that were provided by the challenge organizing committees, and thus, not tailored for our demonstration needs. We have noticed that, the same algorithm, run on different corpora, had its results highly impacted by two items:

- The 'cleanliness' of the corpus: Does it contain typos, unreadable words, several punctuation marks, unanticipated abbreviations, ill-formed sentences? All these elements could prevent some algorithms, especially those relying on NLP techniques such as syntactic and semantic analysis, from obtaining the results they are supposed to produce.

- The 'naturality' of the corpus: Is it an artificially generated set of words, obtained through concatenation of segments or texts of different origins, or is it a real document or collection of documents, written in a given style, addressing a given subject, and so forth? Artificial corpora could favor techniques sensitive to clean cuts in topics, whereas natural corpora would introduce a higher difficulty, since transitions are smoother, and so topic shifting more difficult to detect.

In this paper, we try to fathom the real impact of corpus quality on the performances of topic-based segmentation methods. We have set up an experiment using two unsupervised segmenting algorithms, Choi's $c99$ [6], a renowned segmenting method, and $Transeg$ [9], [17]. We chose unsupervised methods because they are not attuned to corpus idiosyncrasies as supervised methods are: Corpus adjustment by learning would have prevented us from studying the impact of its quality. The second reason for this choice is that both algorithms highly differ methodologically: The first is lexical, based on words frequencies, and neglects syntactical or rhetorical information, the second is more sentence or segment oriented, based on syntactic parsing and sentence semantics calculus, and is not sensitive to frequencies. The difference is crucial. A pending question is always the robustness issue: Would a lexical based system be less sensitive to an 'raw' corpus than a sentence based system? Would it be more sensitive to an artificial corpus of concatenated texts, where differences in topics are neat and precise? This would mark a differential approach to quality

sensitivity, and we wanted to know whether this feeling was justified or not.
In section 2, we present both methods, their features and characteristics. In section 3, we describe our experiment and the working hypothesis that we have tried to evaluate. As it will be seen in results comments(section 4) and in conclusion (section 5), the impact of on precision/recall performances of both algorithms isn't to be ignored. But interesting differences are pointed out when matching both items (i.e. clean/dirty vs natural/artificial) distinct values. We hope this will help evaluation campaigns organizers shape up test corpora that will be as reliable and as discriminant as possible for the running methods they intend to evaluate.

II. A BRIEF OVERVIEW OF $c99$ AND $Transeg$

In this section we present the two methods we compared during our experiments: the well known $c99$ algorithm [5] and $Transeg$ the method we are currently developing. Both are unsupervised and thus corpus free approaches.

A. C99

Developed by Choi, $c99$ is a text segmentation algorithm strongly based on the lexical cohesion principle [14]. It is, at this time, one the best and most popular algorithms in the domain [2], which convinced us to choose it as a baseline for our experiments.
$C99$ uses a similarity matrix of the text sentences. First projected in a word vector space, 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 vector space [6].
The author then builds a second matrix known as the *rank matrix*. The latter is computed by giving to each cell of the similarity matrix 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. By definition, $c99$ always retrieves the first sentence of a text as a the beginning of a new topic (which is obviously true). Choi's original experiments in both cited papers use an 'artificial' corpus, created by concatenating multiple texts. So, retrieving text boundaries in a concatenation and segmenting topically a text have been considered as equivalent tasks by the authors.
All experiments in this paper have been conducted using the original latest version of the algorithm (it can be found at http://myweb.tiscali.co.uk/freddyyychoi/) to avoid any implementation bias. So testing $c99$ on natural, non concatenated texts provides information about its behavior in an environment different from the original protocol.

B. Transeg

$Transeg$ is also based on a vectorial representation of the text and on a precise definition of what a *transition* between two text segments should be. It has been developed with a sentence parser, and until now, experiments and results have been obtained for the French language. A shifting to any other language is naturally possible, provided that a syntactic parsing occurs and the language words are dipped into a Roget-based representation. Since this system has not yet been as widely described as $c99$, we focus on its principles for the sake of information.

1) Text Vector Representation: The first step is to convert each text sentence into a semantic vector obtained using the French language parser SYGFRAN [3]. Vectors are mathematical representations of the Roget Thesaurus indexing each English word with a set of 1043 concepts [18], but 'exported' to French, through the local Roget equivalent, the Larousse thesaurus [10] . Sentence vectors are recursively computed by linearly combining sentence constituents, in turn computed by linearly combining word vectors. The weights of each word vectors are computed according to a formula relying on a constituents and dependencies syntactic analysis (The formula is given in [4]). So, sentence vectors bear both the semantic and the syntactic information of the sentence, but are not sensitive to words frequencies.

2) Transition zones and boundaries: How to detect topic shifts: In well written structured texts, the transition between a topic and the next one is not abrupt. An author should conclude one topic before introducing another.This specific part of text between two segments is waht we call a **transition zone**. Ideally, the transition zone should be composed of two sentences:

- The last sentence of the previous segment.
- The first sentence of the beginning new segment.

$Transeg$ tries to identify these two sentences in order to track topic boundaries.

a) Transition score and the beginning of a new segment: The **transition score** of a sentence represents its likelihood of being *the first sentence of a segment*. Each sentence of the text is assumed to be the first of a 10 sentences long segment. This 'potential segment' is then compared with another one composed by the 10 preceding sentences. The 10 sentences size was chosen by observing results on the training corpus of French political discourses in the DEFT06 competition, segmented by human experts. Competitors such as [9] noticed that the average size of a segment was around 10 sentences (10.16) with a σ of (3.26). So they decided to use this empirical value as the standard segment size. However, this value has no impact on boundaries detection. Any other might fit as well.
To compute the score of each sentence, $Transeg$ slides a

20 sentences long window along the text, considering each half of the window as a potential segment. The latter is then represented by one vector, calculated as a weighted barycenter of its sentence vectors (which are designated as centroid in figure 1). Stylistic information was added by giving a better weight to first sentences, relying on the fact that introductions bear important information [12],[11]. Then a 'thematic' distance is calculated between the two barycenters, and is considered as the window *central sentence transition score* (figure 1). It is computed according to the augmented concordance distance formula defined in next paragraph.

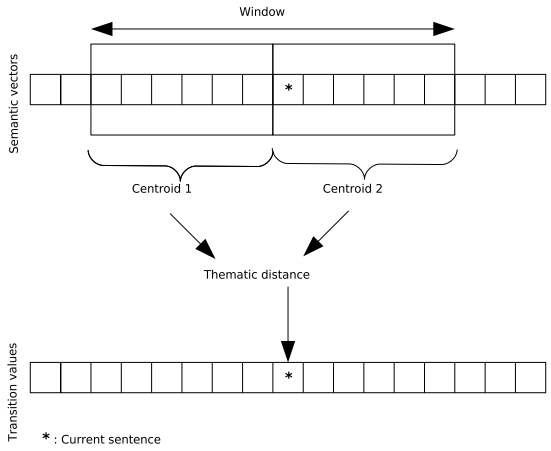

Fig. 1. The transition score of a sentence represent its likelihood of being the first sentence of a segment

b) Concordance distance: Semantic vectors resulting from parsing and semantic calculus have 873 components and most of which having null values. Therefore, either cosine or plain angular distance are not able alone to finely detect a shift in direction. The goal of the concordance distance is to be more discriminant by considering vectors components ranks as well as their values. For the purpose of being more discriminating, we developed the concordance distance, which is itself based based on the concordance measure presented in [**?**].

Considering $\vec{A}$ and $\vec{B}$ two semantic vector representing respectively two sentences A and B. Their values are sorted from the most activated to the least activated and only $\frac{1}{3}$ of the original vectors is kept. $\vec{A_{sr}}$ and $\vec{B_{sr}}$ are respectively the sorted and reduced versions of $\vec{A}$ and $\vec{B}$.

Obviously, if both vectors have no common components then their distance is set to 1. If $\vec{A_{sr}}$ and $\vec{B_{sr}}$ have common components, two differences are necessary to evaluate their 'distance' :

- THE RANK DIFFERENCE: if i is the rank of C_t a component of $\vec{A_{sr}}$ and $\rho(i)$ the rank of the same component in $\vec{B_{sr}}$, the rank difference is calculated as:

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

Where Nb is the number of values kept in the sorted vector.

- THE INTENSITY DIFFERENCE: One has to compare the intensity of common strong components. If a_i is the intensity of i rank component from $\vec{A_{sr}}$ and $b_{\rho(i)}$ the intensity of the same component in $\vec{B_{sr}}$ (its rank is $\rho(i)$), then intensity difference is given by the formula:

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

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

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

As P concentrates on components intensities and ranks, the overall components direction is introduced by mixing P with the classical vector angular distance. If $\delta(\vec{A}, \vec{B})$ is the angular distance between $\vec{A}$ and $\vec{B}$, then:

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

Where β is a coefficient used to give more weight (or less) to P. $\Delta(\vec{A_{sr}}, \vec{B_{sr}})$ is the *concordance value*, presented in [4]. It is easy to prove that neither P nor $\Delta(\vec{A_{sr}}, \vec{B_{sr}})$ are symmetric. But $Transeg$ needs a symmetric value (a distance). To have one, we just have to compute an average between $\Delta(\vec{A_{sr}}, \vec{B_{sr}})$ and $\Delta(\vec{B_{sr}}, \vec{A_{sr}})$:

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

c) Practical example of the concordance distance: : If we considerate these two sentences:

- "Car overuse has a disastrous impact on the environment and more precisely on the global warming."
- "Our new car model only emit $127g$ of CO_2 per miles, be kind to the environment, buy a X car !"

Both sentences speak about environment and cars, but the first sentence is about ecology, the second sentence is an advertisement for a car. If we represent both sentences with semantic vectors:

- $\vec{Ph_1} = [0,0,0,0,0,1,0,0,0,0,2,0,0,0,0]$ for the first sentence.
- $\vec{Ph_2} = [0,0,0,0,0,2,0,0,0,0,1,0,0,0,0]$ for the second sentence.

Where the sixth value of vectors is the CAR concept and the eleventh the ECOLOGY concept[1]. If we compute a normalized angular distance[2] between the two vectors, we obtain a value of $0,41$. Which means that the to sentences are close enough to be part of the same topic (see next paragraph for the threshold value).

If we compute a concordance distance between the two sentences, the result is 0.56 (rounded down). Such a value

[1]The representation as been greatly simplified for the purpose of the demonstration

[2]The result would be between 0 and 1 instead of 0 and $\frac{\pi}{2}$

undoubtly differentiate the two sentences.

d) Transition zones: Once each sentence has a transition score, parts of the text where boundaries are likely to appear are identified. These zones are successive sentences with a score greater than a determined threshold S (figure 2). Since the ideal transition zone is assumed to be a two sentences long text segment, isolated sentences are ignored. Distance helps

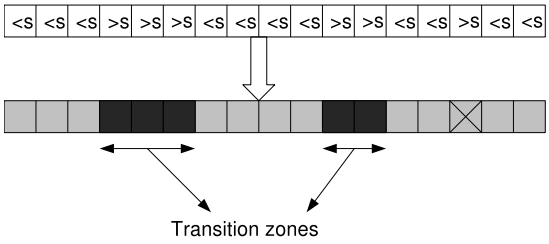

Fig. 2. Identifying transition zones

comparing sentences, but cutting a text into segments needs a maximal distance value acting as a threshold. The chosen one for D (formula 5) is 0.45, also empirically deduced, like the 10 sentences segment size. In order to know whether it is corpus dependent or not, $Transeg$ designers browsed two other corpora segmented by human experts, and belonging to the fields of computer science and law (available for the same edition of the DEFT06 competition mentioned before). The threshold seemed to remain constant on these data. This is not a proof that it is completely corpus independent, and needs to be further investigated. However, at a first attempt, it resisted variation, and authors assumed it to be representative of a 'natural trend' of topical discrimination, among other criteria, of course.

The augmented concordance distances were computed between all identified text segments and as a result, the average distance of 0.45 with a σ of 0.08 appeared to be clearly identified.

e) Ending sentences and breaking score : Identifying boundaries inside transition zones needs information about topic ending. The transition score of a sentence is defined as its likelihood of being the first sentence of a segment. The **breaking score** is a sentence likelihood of being *the last sentence of a segment*.

It is obvious that the last sentence of a topic should conclude the topic and more or less introduce the next topic. So the thematic distance of this sentence to its segment should be quite equal to the thematic distance of this sentence to the next segment. The breaking score B_i of the i sentence is:

$$B_i = 1 - |D_p - D_n| \tag{6}$$

Where D_p is the thematic distance of the sentence to the previous segment and D_n the thematic distance of the sentence to the next segment. The closer D_p and D_n are to each other, the closer to 1 B_i is.

The last step of $Transeg$ method consists in multiplying the transition score of each sentence of a transition zone with the breaking score of the previous sentence. The higher score has high probabilities of being the first sentence of a new segment.

III. Experiment

The two methods presented in the previous section are obviously quite different, even though tackling the same task and belonging to unsupervised methods. $C99$ concentrates on lexical cohesion to find topic segments by regrouping them. $Transeg$, on the opposite, concentrates on supposed characteristics of topic boundaries to identify them.

To test the incidence of corpus quality and origin on both $c99$ and $Transeg$ results, we used four different corpora during this experiment, impersonating the four variations of our pair of items (clean/raw, natural/artificial).

A. Corpus descriptions

- A 'clean and natural' corpus (Corpus CN). Consisting in 22 French political discourses extracted from the DEFT'06 training corpus [1]. These discourses have been segmented by human experts into topic segments, and have been cleaned from the noise (typo errors, full capital sentences, etc.). To impersonate this situation, we did not concatenate them. Corpus size: $54,551$ words, $1,895$ sentences.

- A 'raw and natural' corpus (Corpus RN). Consisting in many French political discourses extracted from the same data pool than the previous corpus. Also segmented by human experts, these discourses display lots of noisy items (many typo errors and full capital sentences for example, which in French often introduces diacritics errors). Corpus size: $69,643$ words, $2,214$ sentences.

- A 'clean and artificial' corpus (Corpus CA). Consisting on 134 concatenated short news from the French news paper "Le Monde". For this corpus each short news is considered as a topic segment. Corpus size: $50,691$ words, $1,574$ sentences.

- A 'raw and artificial' corpus (Corpus RA). Consisting on 131 concatenated laws extracted from another training corpus from DEFT'06. Corpus size: $53,919$ words, $2,310$ sentences.

As all the cleaning on corpora CN and CA has been hand made, there can still be some noise but far less than before the cleaning. This also explain the relatively short amount of text the experiment was running on. The effort of producing clean data is very heavy. Moreover, other experiments in the domain have been presented with corpora not considerably bigger than ours [6], [19]. In order not to introduce a size bias, we restricted the R corpora (RN and RA) to roughly the same size as the CA and CN corpora (results presented in subsections 4.1 and 4.2). However, since producing raw corpora is very easy, we wanted to know whether size could have an effect on performance results, therefore, we doubled RA and DN corpora sizes, and Tables V (RN: $160,524$ words, $5,445$ sentences) and VI (RA: $105,350$ words, $4,854$ sentences) have brought up interesting results about the impact of corpus lengths, commented in section 4.3.

B. Tolerant measures

We ran both methods on each corpus and evaluated the results using the DEFT'06 **tolerant recall and precision** described in [1]. They consider 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 DEFT'06 organizing committee noticed that the use of either strict or tolerant measures had no effect on the ranking of the submissions they had to evaluate, but gave better scores to all methods.

Note that both methods consider first sentences of texts as a boundary, so every first sentence of the CN and RN corpus texts is counted as a boundary (which means both algorithms have at least one good boundary per text). The results on other corpora are not too affected by this specificity as they are all one big text with several sentences and topic segments.

Even if $c99$ isn't corpus or language sensitive, it has been slightly optimized to English language by using a stop list. As our experiment hab to be the fairest possible, we used $TreeTagger$ to stem the corpora and eliminate tool words from it (based on their categories).

IV. Results

A. Clean Vs Raw: The Impact of Corpus Data Quality

The individual results of each of the 22 texts of CN Corpus have been combined in Table I into an average precision and an average recall summarizing the 22 individual values (too long to expose here). Table I presents the best conditions output

	C99	Transeg
Precision	**53.32**%	45.49%
Recall	23.12%	**38.76**%

TABLE I
AVERAGE PRECISION AND RECALL ON THE 22 TEXTS OF THE 'CLEAN AND NATURAL' (CN) CORPUS

of both methods. It highlights the differences between the two methods. With its default boundaries detection $c99$ has a better precision, but proposes less solutions and so has a worse recall. On the opposite $Transeg$, by actively searching for boundaries, is more sensitive to variation between sentences. It suggests more solutions than $c99$ and so worsens its precision for a better recall. These 'ideal' conditions clearly demonstrate the strengths and weaknesses of both approaches and give hints about means to improve them. As shown in next paragraphs, results considerably worsen whenever corpus quality drops. As soon as we deteriorate the quality of the corpus (Table II) results drop. At first sight, the natural tendency of $c99$ toward precision seems to be conserved, whereas its recall is dramatically affected (with a 1.54% value, one wonders whether it still has a meaning!). A deeper observation of results indicates that $c99$ brings only 3 sentences back as a boundaries, including the first one, which is a boundary

	C99	Transeg
Precision	35.14%	**43**%
Recall	1.54%	**9.85**%

TABLE II
PRECISION AND RECALL ON THE 'RAW AND NATURAL' (DN) CORPUS

per se. So its relatively good precision is mostly due to the experiment conditions. $Transeg$ seems to be sturdier: Its precision is almost untouched, but its recall has badly deteriorated. A counterintuitive output: $Transeg$ does better than $c99$ in conditions where we thought that the latter would be the most robust. $Transeg$ is considered to be more sensitive to ill-formed sentences, unknown or misspelled words. It was supposed to be distanced by $c99$ on corrupt data. It seems that it is not the case.

B. Artificial Corpora: Do Corpus Manipulations impact Methods Results?

	C99	Transeg
Precision	**30.77**%	22.3%
Recall	5.03%	**19.5**%

TABLE III
PRECISION AND RECALL ON THE 'CLEAN AND ARTIFICIAL' (CA) CORPUS

When coming to an artificial but clean corpus (Table III), we retrieve the original balance between the two methods: A precision oriented output for $c99$ and a recall oriented one for $Transeg$. When compared to Table I, the results range is far worse, and becomes difficult to interpret. $C99$ bad recall (around 5%) is quite surprising. One would also have expected the opposite: Concatenating distinct texts would make the segmenting task much easier to a word-based algorithm! The

	C99	Transeg
Precision	**42.86**%	8.02%
Recall	2.14%	**9.27**%

TABLE IV
PRECISION AND RECALL ON THE 'RAW AND ARTIFICIAL' (DA) CORPUS

'worst' conditions case, impersonated by the DA Corpus, have not been matched with the worst results by both methods: Only $Transeg$ seems to be very sensitive to this loss in quality and naturality! When compared with the best case, $c99$ precision is less by 11 points, whereas $Transeg$ precision looses 35. On the other hand, recalls are strongly impacted by both data corruption and manipulation (a 2.14% recall for $c99$ and a less than 10% one for $Transeg$ are very bad scores). But the orientation seems to be maintained: $C99$ is still leading in precision, and $Transeg$ in recall. However, with such low values, interpretation is risky. One cannot but acknowledge the impact of data reliability on performances degradation.

C. Complementary Experiment: The Impact of Size on the D Corpora

The impact of corpora length could be assumed to have two opposite effects.

- Either a biggest data would introduce more corruption cases, and thus worsens results (because of the D aspect)
- Or it would provide both algorithms with more opportunities to detect boundaries 'by chance' and thus augment their performances.

In order to see which of these two assumptions is more likely to be supported, we run the experiments on the doubled RN and RA corpora. Results are summarized in Tables V and VI. When comparing Tables II (simple RN corpus) and V

	C99	Transeg
Precision	33.33%	**35.7**%
Recall	0.12%	**20.28**%

TABLE V
PRECISION AND RECALL ON THE 'RAW AND NATURAL' (RN) BIGGER CORPUS

	C99	Transeg
Precision	**15.91**%	8.54%
Recall	5%	**19.29**%

TABLE VI
PRECISION AND RECALL ON THE 'RAW AND ARTIFICIAL' (RA) BIGGER CORPUS

(double RN corpus), we see that both methods 'orientation' is maintained: $Transeg$ does better, in both precision and recall, but if its precision has been reduced by 8 points, its recall has improved by 11!.It seems that 'chance' retrieved boundaries are rather significant. At the same time since precision drops, the number of corrupt data cases prevent good boundaries to be found. Let us notice that the Recall/Precision ratio is invariant with size. On the other hand, $c99$ recall continues to drop down to incredible values. One cannot risk an interpretation. Table VI has to be compared to Table IV. The better values in Table VI could be interpreted with a 'canceling' effect provided by size. More data corrupted cases, but also more boundaries to be retrieved by chance. This drives us to conclude that providing a bigger set of data adds more noise to algorithms performances interpretation.

D. Overall Results

In short, the rather unexpected results obtained in experiments could be summarized by the following statements.

- Data quality (i.e. clean vs dirty) has an impact on results of both methods: Best comparative results in recall and precision are achieved with the CN corpus. The dramatic fall in recall for $c99$ (Tables II, V, VI) is difficult to explain. But, surprisingly, a deterioration in quality seems to favor $Transeg$ over $c99$, provided that the corpora keep their 'natural' origin. This is highly counterintuitive.
- Corpus manipulations (i.e. natural vs artificial) has no impact on both methods orientation (a trend toward precision for $c99$ and one toward recall for $Transeg$), in Tables I, III, IV, VI. Values drop for both methods between the best case (Table I) and the worst case (Table IV), but $Transeg$ is more affected than $c99$.When comparing Table I and Table III, there is a loss of 20 points in general. So artificial corpora seem to lower results. This is also opposite to our previous intuition, in which a general improvement of $c99$ results was expected.
- The comparison between Tables II and IV shows that it is $Transeg$ which is the most affected by data manipulation (artificially built corpora) when corrupt data is present. Since RA corpora are what most evaluation campaigns provide, then $Transeg$ presents a liability which has to be improved.
- The comparison between Tables II and III, the 'diagonal values' is most interesting. The 'artificiality' of the CA corpus has a discriminant effect on $Transeg$ precision (from 43 in Table II down to 22 in Table III), but an improvement on its recall (it is doubled). $C99$ looses the 5 points in precision that it gains in recall. This confirms that $Transeg$ should not run on artificial corpora.
- Tables V and VI show the impact of corpora size. Recall and precision values improve with size for $Transeg$ but they drop down for $c99$. This means that $Transeg$ will resist better with bigger corpora, however, an artificial corpus is what handicaps it most.
- By choosing to show recall and precision values in call cases, and not an $FScore$ built on their ratio, we hope to have grasped the variations introduced by different corpus quality criteria.

V. CONCLUSION

The impact of copora constitution on methods evaluation cannot be neglected. The previous experiments and their results tend to show it. What is interesting is that given two quite different but corpus independent methods, they cannot resist a deterioration in corpora. The latter can be introduced by either data corruption or data handling. If one is more apt to stand data corruption ($Transeg$) and the other data handling ($c99$) both give their best when quality is granted. For us this means that: (1) If evaluation campaigns organizers do not test their corpora quality, they already handicap unsupervised methods, which will never achieve spectacular results. Supervised methods will tune to corpus bias and overcome them. However they won't be able to do so well with unlearnt data. (2) If they manipulate their data, they will favor lexical based methods over those which are not. (3) If they don't clean it, they will favor sentence or segment based methods. (4) A big corpus is not necessarily more informative on algorithms capabilities than a smaller one. If data is not of a high quality then chance and noise are more likely to temper with results. Of course, topic-based segmentation methods should be able to deal with any kind of text. But depending on what we want to evaluate the kind of the corpus could be very important. If the

goal is to evaluate a method in a practical applicative context, then any corpus should be used as in real conditions anything could happen. On the opposite, if the goal is to to evaluate the validity of a theory or to the feasibility of a task, the choice of the corpus become important. If we do not want to add the complexity of the corpus properties to the complexity of the task, then we should carefully choose our corpora depending on exactly what we want to evaluate. Otherwise the soundness of our evaluations will be jeopardized.

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*, vol. 1, pp. 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*, vol. 1, pp. 11–15, 1984.

[4] J. Chauché and V. Prince, "Classifying texts through natural language parsing and semantic filtering." *In Proceedings of LTC'07, third international Language and Technology Conference*, 2007.

[5] F. Y. Y. Choi, "Advances in domain independent linear text segmentation," *Proceedings of NAACL-00*, pp. 26–33, 2000.

[6] F. Y. Y. Choi, P. Wiemer-Hastings, and J. Moore, "Latent semantic analysis for text segmentation," *Proceedings of EMNLP*, pp. 109–117, 2001.

[7] 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*, pp. 59–68, 1993.

[8] S. Kaufmann, "Cohesion and collocation: using context vectors in text segmentation," in *Proceedings of the 37th annual meeting of the ACL*, 1999, pp. 591–595.

[9] A. Labadié and Chauché, "Segmentation thématique par calcul de distance sémantique," *Proceedings of DEFT'06*, vol. 1, pp. 45–59, 2006.

[10] Larousse, *Thésaurus Larousse - des idées aux mots, des mots aux idées*. Paris: Larousse, 1992.

[11] A. Lelu, C. M., and S. Aubain, "Coopération multiniveau d'approches non-supervises et supervises pour la détection des ruptures thématiques dans les discours présidentiels fran," *In Proceedings of DEFT'06*.

[12] C. Lin and E. Hovy, "Identifying topics by position," in *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP–97)*, 1997, pp. 283–290.

[13] K. McCoy and J. Cheng, "Focus of attention: Constraining what can be said next." *in C. Paris, W. Swartout, and W. Mann, ' Natural Language Generation in Artificial Intelligence and Computational Linguistics'*, 1991.

[14] J. Morris and G. Hirst, "Lexical cohesion computed by thesaural relations as an indicator of the structure of text," *Computational Linguistics*, vol. 17, pp. 20–48, 1991.

[15] L. Pevzner and M. Hearst, "A critique and improvement of an evaluation metric for text segmentation," *Computational Linguistics*, pp. 113–125, 2002.

[16] J. M. Ponte and W. B. Croft, "Text segmentation by topic," *European Conference on Digital Libraries*, pp. 113–125, 1997.

[17] V. Prince and A. Labadié, "Text segmentation based on document understanding for information retrieval." *In Proceedings of NLDB'07*, pp. 295–304, 2007.

[18] P. Roget, *Thesaurus of English Words and Phrases*. London: Longman, 1852.

[19] N. Stokes, "Spoken and written news story segmentation using lexical chains," in *Proceedings of NAACL '03*, 2003.