Tools and methods for objective or contextual evaluation of topic segmentation

Laurianne Sitbon, Patrice Bellot

Laboratoire d'Informatique d'Avignon - Université d'Avignon
339, chemin des Meinajaries - Agroparc BP 1228
84911 AVIGNON Cedex 9 - FRANCE
Tel : +33 (0) 4 90 84 35 09
{laurianne.sitbon, patrice.bellot}@univ-avignon.fr

Abstract

In this paper we discuss the way of evaluating topic segmentation, from mathematical measures on variously constructed reference corpus to contextual evaluation depending on different topic segmentation usages. We present an overview of the different ways of building reference corpora and of mathematically evaluating segmentation methods, and then we focus on three tasks which may involve a topic segmentation : text extraction, information retrieval and document presentation. We have developped two graphical interfaces, one for an intrinsec comparison, and the other one dedicated to an evaluation in an information retrieval context. These tools will be very soon distributed under GPL licences on the Technolangue project web page.

1. Introduction

Topic segmentation is a task with multiple applications and multiple methods. It is well known that segmentation can improve information retrieval significantly by giving the specific part of a document corresponding to the query as a result and by indexing documents more precisely. By subdividing texts into topically coherent segments, a segmentation stage allows a better estimation of the relevance compared to the query. A second role is described by the task of TDT (Topic Detection and Tracking). Finally, segmentation can be part of a summary process, for example in TXTRACTOR (McDonald and Chen, 2002).

This diversity of uses leads us to consider a specific class of tools and objectives, in order to compare them. In this paper we propose to discuss the adequacy between the evaluation measure and the application employing the segmentation method. The evaluations are based on a comparison between the unsupervised segmenter we developped and state of the art results. This work is done in the context of the project « Technolangue AGILE - OURAL » financed by the French minister of research. We focused on unsupervised algorithms based on linear segmentation, by placing boundaries between segments, as in C99 (Choi, 2000). We have developped a new method for linear topic segmentation, based on weighted lexical chains. An implementation called LIA_topic_seg is distributed under GPL licence.

In literature there are several mathematical measures for evaluating segmentation algorithms. These measures all operate with a reference corpus, which may be built in different ways. However in this process, the task the topic segmentation is involved in is not taken into account. In a first part we present an overview of the different ways of building reference corpora and of mathematically evaluating segmentation methods. In a second part we focus on three applications of segmentation less common than TDT or a summary. Text extraction, information retrieval and document presentation may all involve a topic segmentation process in order to improve their results, but the need of boundaries strictly corresponding to a reference is not obvious.

2. Objective evaluation measures

2.1. Reference corpora building

2.1.1. Use of text structure

Many experiments in topic segmentation evaluate the capacity of systems to retrieve original paragraphs, in texts where titles and paragraphs are originally printed. This is the aim of DEFT'06 campaign (`http://www.lri.fr/ia/fdt/DEFT06/`). All titles and empty lines are removed in order to build test documents. This method can be applied to scientific documents as well as to the chapters of a book, as in the experiments done by (Ji and Zha, 2003) on a novel called *Mars*. The main disadvantage is that sometimes authors may divide topically coherent segments only for eye comfort, or on the contrary they may not separate many sentences from various topics, in introduction or conclusion paragraphs for example. There remains an ambiguity concerning transition sentences, which is usually linked to both the previous and the next paragraph.

2.1.2. Manual decision

An other way of a building reference corpus is to ask human judges to put subtopic frontiers inside large texts without any typographic cues. Such a corpus, called *Stargazers*, has been done by (Hearst, 1997) in order to evaluate TextTiling. The main issue of this corpora constitution is the disagreement betwen human judges, particularly for transition sentences.

2.1.3. Automatic construction

Many evaluations operate with reference corpora automatically built according to the method proposed by (Choi, 2000), adapted in french by (Sitbon and Bellot, 2004), in portugese by (Dias and Alves, 2005). Each document is composed of 10 extracts from several newspaper articles chosen among different thematic categories randomly. Extract sizes are not constrained and may be different. This corpus of documents is a reference, as the segments of the documents are chosen in different texts to ensure thematic variability. The main advantage of this method, compared to a manually annotated corpus, is the amount of data we can test. Moreover, the manual annotation is subjective and takes more human ressources.

2.2. Mathematical measures

With the help of reference corpora, evaluation of new algorithms is generally based on mathematical measures recently refined.

2.2.1. Classical recall/precision

Both standards recall and precision, classically used in information retrieval, detailed in (Baeza-Yates and Ribeiro-Neto, 1999), were often employed to evaluate segmentation algorithms. For several reasons, they are not very relevant. First, they are too related to each other, and we are not looking for supporting more one or the other, but for evaluating the algorithms globally. To solve this problem, (Jr. et al., 1997) proposed an other measure, the Harmonic Mean, which gives only one result, combining both recall and precision. Another problem is that they do not weight errors because they are locally binary measures. It means that the error is the same if there is one sentence between the found boundary and the real boundary, or if there are 5 sentences. Lastly, for algorithms like dot plotting where the number of boundaries to find is predefined, recall and precision are equal. This is because the number of boundaries to find (used for recall) and the number found by the algorithm (used in precision) is the same in those cases.

2.2.2. Beeferman measure : P_k

In order to go over those problems, (Beeferman et al., 1997) proposed an other measure that takes into account the distance (computed in number of words for example) between a boundary found and the right boundary to find. The first proposed measure was the probability for any pair of sentences to be in the same segment in the hypothetical segmentation (**hyp**) if they are in the same segment in the reference segmentation (**ref**), and to be in different segments in **hyp** if they are in **ref**. Formally that gives :

$$P_D(\textbf{ref}, \textbf{hyp}) = \sum D(i,j).(\delta_{\textbf{ref}}(i,j) \overline{\bigoplus} \delta_{\textbf{ref}}(i,j)) \quad (1)$$

where $\overline{\bigoplus}$ means XNOR (both or neither), and $\delta_{\mathbf{x}}(i,j)$ is a boolean set to 1 if sentences i and j are in the same segment in segmentation **x** and 0 if they are in different segments. D is a distance probability distribution for each set of distances between random sentences. Then, they proposed a simplification of this measure in [Beef99], fixing the distance between both sentences to a fixed number, which is half the average number of words in a segment. A probability of "error" is then calculated on the segmentation :

$$\begin{aligned}
&p(error|\textbf{ref}, \textbf{hyp}, k) = \\
&\quad p(miss|\textbf{ref}, \textbf{hyp}, different\ \textbf{ref}\ segments, k) \times \\
&\quad p(different\ \textbf{ref}\ segments|\textbf{ref}, k) + \\
&\quad p(false\ alarm|\textbf{ref}, \textbf{hyp}, same\ \textbf{ref}\ segment, k) \times \\
&\quad p(same\ \textbf{ref}\ segment|\textbf{ref}, k) \qquad (2)
\end{aligned}$$

An hypothesized boundary which is far from the reference one at a greater distance than k is considered as false alarm where it is and as missing where it should be (the same place as the reference one). It is also considered as 2 errors instead of 1.

2.2.3. Hearst measure : WindowDiff

(Pevzner and Hearst, 2002) shows that Beeferman measure, even if it is better than recall and precision, presents some failures. Five bad points have been highlighted in (Pevzner and Hearst, 2002) :

- Missing boundaries are more penalized than false alarms.
- When a boundary is added implying a new segments of size smaller than k, it is not detected and so not added in the score.
- The algorithm is more lenient when there are strong variations in segment sizes.
- Near-miss errors are penalized too much compared to false alarms and missing boundaries.
- The meaning of the score is not clear : it looks like a percentage but it is not.

They proposed a new way of scoring segmentation algorithms, called Window Diff. It is almost identical to P_k, except that instead of $\overline{\bigoplus}$ which is evaluated to 0 or 1, they use the difference between the number of boundaries between positions i and i+k in both **ref** and **hyp**. If this difference is null so the sentences i and i+k are locally in the same segments of **ref** and **hyp**. This let small added segment in **hyp** be penalized.

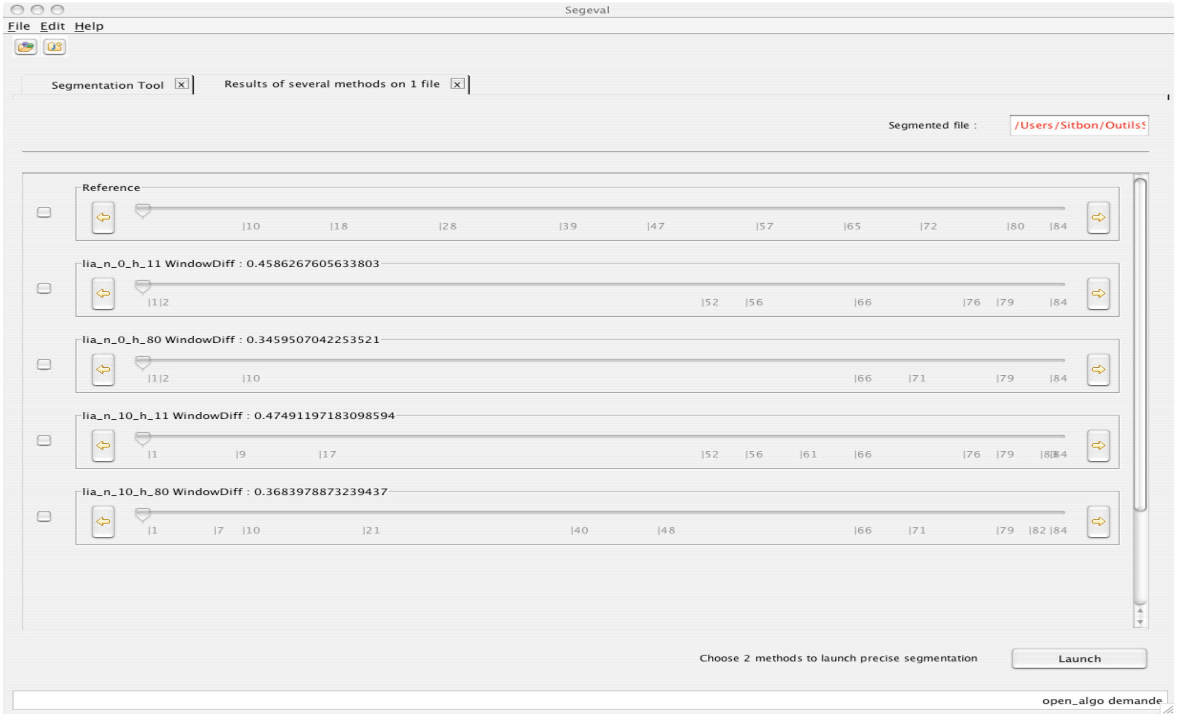
$$\begin{aligned}
\textit{WindowDiff(ref, hyp)} \quad = \quad & \frac{1}{N-k} \sum (|b(ref_i, ref_{i+k}) \\
& -b(hyp_i, hyp_{i+k})|) \qquad (3)
\end{aligned}$$

where $b(x_i, x_j)$ represents the number of boundaries between positions *i* and *j* in the text *x*, and N represents the number of sentences of the text.

Experimentally, they show that this measure is stable with variation of segment sizes and equivalent for false alarms and missing boundaries. This new measure presents the inconvenient that the score can now be greater than 1, so it can no longer be assimilated to a percentage. Thus, it is now clear that the measure is only for comparison, and does not evaluate the quality of a segmentation algorithm directly.

2.3. Segeval, a graphical interface for objective evaluation

In order to simplify massive objective comparisons between different methods (or different parameters applied to a method), for variously featured corpora, we have developped a graphical tool based on WD measure. This way of evaluating has been used in (Choi, 2000) and (Sitbon and Bellot, 2004) for comparing methods. Segeval lets a user make a very precise comparison, inside the text. It means many methods can be runned on the same document, and a graphical view of each set of hypothesized boundaries in



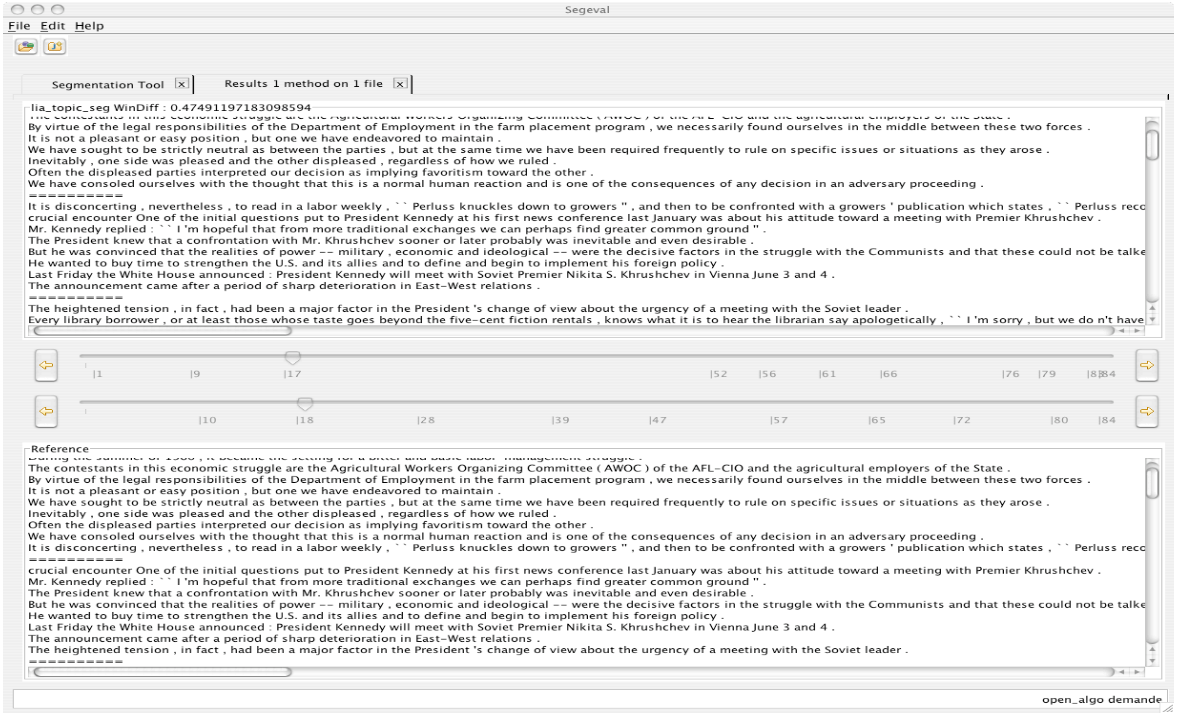
lia_n_0_h_11 WindowDiff : 0.4586267605633803

lia_n_0_h_80 WindowDiff : 0.3459507042253521

lia_n_10_h_11 WindowDiff : 0.47491197183098594

lia_n_10_h_80 WindowDiff : 0.3683978873239437

Figure 1: Segeval : graphical comparison between referenced and hypothesised boundaries computed with varying parameters of LIA_topic_seg on one text. One may select two segmentations for a text comparison as shown on figure 2



lia_topic_seg WinDiff : 0.47491197183098594

```
By virtue of the legal responsibilities of the Department of Employment in the farm placement program , we necessarily found ourselves in the middle between these two forces .
It is not a pleasant or easy position , but one we have endeavored to maintain .
We have sought to be strictly neutral as between the parties , but at the same time we have been required frequently to rule on specific issues or situations as they arose .
Inevitably , one side was pleased and the other displeased , regardless of how we ruled .
Often the displeased parties interpreted our decision as implying favoritism toward the other .
We have consoled ourselves with the thought that this is a normal human reaction and is one of the consequences of any decision in an adversary proceeding .
==========
It is disconcerting , nevertheless , to read in a labor weekly , `` Perluss knuckles down to growers '' , and then to be confronted with a growers ' publication which states , `` Perluss reco
crucial encounter One of the initial questions put to President Kennedy at his first news conference last January was about his attitude toward a meeting with Premier Khrushchev .
Mr. Kennedy replied : `` I 'm hopeful that from more traditional exchanges we can perhaps find greater common ground '' .
The President knew that a confrontation with Mr. Khrushchev sooner or later probably was inevitable and even desirable .
But he was convinced that the realities of power -- military , economic and ideological -- were the decisive factors in the struggle with the Communists and that these could not be talke
He wanted to buy time to strengthen the U.S. and its allies and to define and begin to implement his foreign policy .
Last Friday the White House announced : President Kennedy will meet with Soviet Premier Nikita S. Khrushchev in Vienna June 3 and 4 .
The announcement came after a period of sharp deterioration in East-West relations .
==========
The heightened tension , in fact , had been a major factor in the President 's change of view about the urgency of a meeting with the Soviet leader .
Every library borrower , or at least those whose taste goes beyond the five-cent fiction rentals , knows what it is to hear the librarian say apologetically , `` I 'm sorry , but we do n't have
```

Reference

```
The contestants in this economic struggle are the Agricultural Workers Organizing Committee ( AWOC ) of the AFL-CIO and the agricultural employers of the State .
By virtue of the legal responsibilities of the Department of Employment in the farm placement program , we necessarily found ourselves in the middle between these two forces .
It is not a pleasant or easy position , but one we have endeavored to maintain .
We have sought to be strictly neutral as between the parties , but at the same time we have been required frequently to rule on specific issues or situations as they arose .
Inevitably , one side was pleased and the other displeased , regardless of how we ruled .
Often the displeased parties interpreted our decision as implying favoritism toward the other .
We have consoled ourselves with the thought that this is a normal human reaction and is one of the consequences of any decision in an adversary proceeding .
It is disconcerting , nevertheless , to read in a labor weekly , `` Perluss knuckles down to growers '' , and then to be confronted with a growers ' publication which states , `` Perluss reco
==========
crucial encounter One of the initial questions put to President Kennedy at his first news conference last January was about his attitude toward a meeting with Premier Khrushchev .
Mr. Kennedy replied : `` I 'm hopeful that from more traditional exchanges we can perhaps find greater common ground '' .
The President knew that a confrontation with Mr. Khrushchev sooner or later probably was inevitable and even desirable .
But he was convinced that the realities of power -- military , economic and ideological -- were the decisive factors in the struggle with the Communists and that these could not be talke
He wanted to buy time to strengthen the U.S. and its allies and to define and begin to implement his foreign policy .
Last Friday the White House announced : President Kennedy will meet with Soviet Premier Nikita S. Khrushchev in Vienna June 3 and 4 .
The announcement came after a period of sharp deterioration in East-West relations .
The heightened tension , in fact , had been a major factor in the President 's change of view about the urgency of a meeting with the Soviet leader .
==========
```

Figure 2: Segeval : comparison between an automatic segmentation with LIA_topic_seg and the reference. The small vertical lines on the central sliders give an overview of the boundaries in the text and offer a quick navigation system from one boundary to another.

a linear representation as shown in figure 1 is available. Then a comparison can be made between two segmentations and an in depth analysis can be made as in figure 2. It allows a qualitative evaluation of the differences. Also, Segeval allows a comparison (based on the average WindowDiff measure) between different nature of data. For example one can compare the efficiency of a method applied to a corpus of small segments (3 sentences) and the same method applied to a corpus of large segments (10 sentences). This tool will be available very soon (`http://`

www.technolangue.net/article79.html) with our home made segmenter called LIA_topic_seg plugged in. The design enables other tools to be plugged in only with a command line and a set of parameters.

3. Contextual evaluation

Mathematical measures tend to look for a perfect alignment between hypothesized boundaries and referenced ones, which can vary depending on the subjectivity with wich they are built or on the granularity of their definition. In this section we discuss the fact that WindowDiff presents the advantage of being objective, but it is not task-oriented enough. Each existing evaluation measure computes a distance between existing boundaries and computed boundaries. Now we will present some applications which need to focus on precise features of segmentation, and not necessarily on each additional, missing and displaced boundaries.

3.1. Segmentation for text extraction

A topic segmentation may be used in combination with a classifier in text extraction tasks. For example, in the context of DEFT'05 clustering evaluation campaign, participants had to find extracts from F. Mitterrand about national politics in French political discourses from J. Chirac about international politics. In this context, additional boundaries are not mistakes as long as both segments before and after are long enough to be well classified. The highlight must be on missing boundaries which lead segments with both discourses to be classified in one. However, an error of only one sentence must not be considered as a pure oversight. In order to be able to evaluate the quality of a segmenter in this task, we propose an evaluation of the segmentation method with an "ideal-case F-measure" computing "ideal-case F-scores". This is a classical F-measure, which evaluates well classified sentences, but the results of clustering are computed with a simulation of a perfect clustering tool runned after the segmenter. According to the reference classification, the class of each sentence is known and one can affect to each segment the same class as the majority of sentences it contains.

As for a comparison of the efficiency of segmenters, table 1 shows the results of "ideal-case F-scores" for C99 and LIA_topic_seg runned with default parameters, for various fixed average sizes of segments (in number of sentences). The differences between two sizes of segments are more important than between both methods, which lets us to conclude that the main issue of text extraction is not the comparison between two segmenting algorithms, but the determination of the better size to choose regarding the classifier.

method	7 sent.	8 sent.	10 sent.
Lucene + C99	0.9183	0.9003	0.8636
Lucene + LIA_topic_seg	0.9128	0.8915	0.8532

Table 1: F-scores for an ideal classification of segments of various sizes computed by means of the C99 implementation and of LIA_topic_seg

3.2. Segmentation for information retrieval

Topic segmentation can also be dedicated to information retrieval, as (J.Callan, 1994) shows it is better to retrieve a part of a large document, and that sentence is too small a text unit for indexing. This also higlight that best results are obtained with passage sizes around 250 words. Even if topic segmentation can produce denser information units than whole documents, it presents the possibility of splitting coherent sequentially written information, discarding complex requests. The main problems may be with the use of anaphoras or abbreviations revealed in the first paragraph of a long text.

In order to evaluate segmentation in the context of information retrieval, we have developped a specific graphical interface of LUCENE search engine. It allows to retrieve topically coherent text segments instead of documents as a whole. We use it to make some tests on TREC 8 data (Voorhees and Harman, 1999). We runned Lucene with default parameters on title queries (containing approximtively two words) and description queries (usually one sentence long). Three indexes were established, one on whole documents, the second on segments of documents obtained with C99, and the third with segments from LIA_topic_seg.

Table 2 shows that results are lower than without segmentation. It can be explained by the fact that additional boundaries may split information. We now need to go further with this and run some tests with bigger segments, and some tests by means of other evaluation campaigns data.

Segmentation	title queries	desc. queries
None	0.1679	0.1271
C99	0.1305	0.1051
LIA_topic_seg	0.1143	0.0711

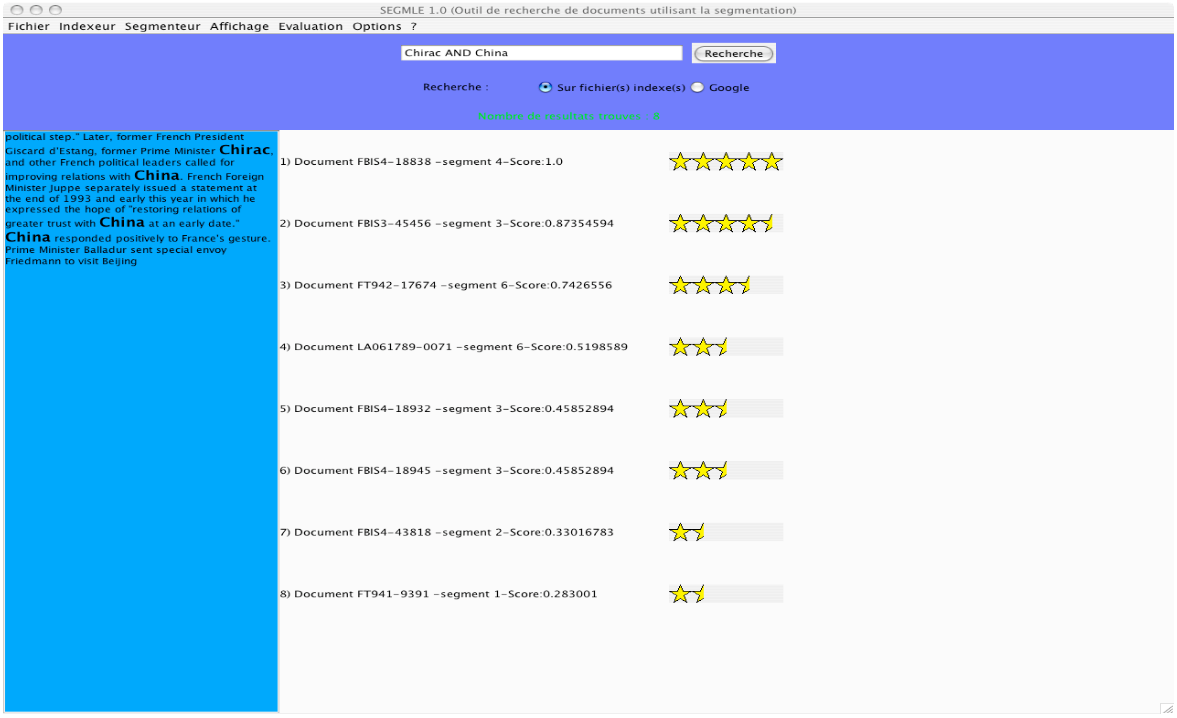
Table 2: Mean average precision measures on TREC with LUCENE and different segmentations of documents.

If one wants to accurately evaluate an hypothesized segmentation in this context with a mathematical measure, the focus must be on added boundaries and absolutly not on missing boundaries.

3.3. Segmentation for improving readability

Presenting a segmented text improves readability by highlighting the thematic structure of the document. This also alleviates visual deficiency aspects as dyslexic people, the succession of lines is a problem. Gaps between paragraphs are visual anchors. (Nordbrock et al., 2004) show that such spaces improve comprehension.

A fixed segmentation based on a number of sentences per paragraph could be an answer, but topic coherence is necessary for an accurate understanding of the document. Moreover, a focus on a specific part of the text may help in an information retrieval process. This feature has also been implemented in our software previously used to test segmentation on TREC ad-hoc data by using our Lucene-based search engine. The application can use segmentation from the indexation or not, and in both cases provides most similar segments for a query. This allows a direct access to the


SEGMLE 1.0 (Outil de recherche de documents utilisant la segmentation)
Fichier Indexeur Segmenteur Affichage Evaluation Options ?
Recherche

Chirac AND China

Recherche : ● Sur fichier(s) indexe(s) ○ Google

Nombre de resultats trouves : 8

political step." Later, former French President Giscard d'Estang, former Prime Minister **Chirac**, and other French political leaders called for improving relations with **China**. French Foreign Minister Juppe separately issued a statement at the end of 1993 and early this year in which he expressed the hope of "restoring relations of greater trust with **China** at an early date." **China** responded positively to France's gesture. Prime Minister Balladur sent special envoy Friedmann to visit Beijing

1) Document FBIS4–18838 –segment 4–Score:1.0

2) Document FBIS3–45456 –segment 3–Score:0.87354594

3) Document FT942–17674 –segment 6–Score:0.7426556

4) Document LA061789–0071 –segment 6–Score:0.5198589

5) Document FBIS4–18932 –segment 3–Score:0.45852894

6) Document FBIS4–18945 –segment 3–Score:0.45852894

7) Document FBIS4–43818 –segment 2–Score:0.33016783

8) Document FT941–9391 –segment 1–Score:0.283001

Figure 3: Human Interface of our Lucene based search engine that uses segments instead of whole documents for the indexation as long as for results printing. The results are ordered by relevance in the middle of the window, and by one click on a title the corresponding segment appear on the left, with words from the query highlighted.

important information and gives the user the possibility to not read the whole document.

The main graphical interface is presented in figure 3. The answer is focused on the target segment (the most similar according to the query) and there is an access to the original document wih the page-setting using the segmentation of the whole document.

In this context, additional boundaries do not decrease readability even if both created segments deal with the same topic. On the contrary, if two segments with totally different topics are merged, it is confusing for the reader.

4. Conclusion

The currently used framework of evaluation of the segmentation tools is reliable but not adapted to many uses of topic segmentation. We proposed new ways of evaluating tasks like clustering, improving readability, and retrieving information. We have developped two graphical interfaces for two contexts of segmentation tools evaluation. One was for an intrinsec comparison, the other one was dedicated to an evaluation in an information retrieval context. These tools will be very soon distributed under GPL licences on the Technolangue project web page.

5. References

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison-Wesley.

Douglas Beeferman, Adam Berger, and John Lafferty. 1997. Text segmentation using exponential models. In *Proceedings of the 2nd conf. on Empirical Methods in Natural Language Processing*, USA.

Freddy Y. Y. Choi. 2000. Advances in domain independent linear text segmentation. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, USA.

Gael Dias and Elsa Alves. 2005. Unsupervised topic segmentation based on word co-occurrence and multi-word units for text summarization. In In association with ACM editions, editor, *Proceedings of the ELECTRA Workshop associated to 28th Annual International ACM SIGIR Conference*, pages 41–48, Salvador, Brazil, August 19.

Marti A. Hearst. 1997. Text-tiling : segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, pages 59–66.

J.Callan. 1994. Passage-level evidence in document retrieval. In *Proccedings of the ACM/SIGIR Conference of Research and Development in Information Retrieval*, pages 302–310.

X. Ji and H. Zha. 2003. Domain-independent text segmentation using anisotropic diffusion and dynamic programming. In *in Proceedings of the 26 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages pp.322–329, Toronto, Canada.

W. M. Shaw Jr., R. Burgin, and P. Howell. 1997. Performance standards and evaluations in ir collections: Cluster-based retrieval models. *Information Processing and management*, pages 1–14.

Daniel McDonald and Hsinchun Chen. 2002. Using sentence selection heuristics to rank text segments in txtractor. In *Proceedings of the 2nd ACM/IEEE Joint Confer-*

ence on Digital Libraries, pages 25–38.

Gabriele Nordbrock, Henrike Gappa, Yehya Mohamad, and Carlos A. Velsaco. 2004. Automatic modification of text for people with learning disabilities using internet services. In *Proceedings of ICCHP*, pages 995–998, July.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, pages 19–36.

L. Sitbon and P. Bellot. 2004. Adapting and comparing linear segmentation methods for french. In *Proceedings RIAO'04*, Avignon, France.

Ellen M. Voorhees and Dona Harman. 1999. Overview of the eighth text retrieval conference (trec-8). In *proceedings of the eighth Text REtrieval Conference*, pages 1–24, Gaithersburg, Maryland, USA, November.