A bootstrapping approach for thematic analysis

Olivier Ferret, Brigitte Grau

A bootstrapping approach for thematic analysis

Olivier Ferret and Brigitte Grau
LIMSI-CNRS
BP 133
91403 Orsay cedex
[ferret,grau]@limsi.fr

Abstract

Thematic analysis is important for a lot of applications dealing with texts, such as text summarisation or information extraction. But it can be done with a great precision only if it relies on structured knowledge, which is difficult to produce. In this paper, we propose using bootstrapping in order to solve this problem: a first thematic analysis based on a weakly structured source of knowledge, a collocation network, is used for learning explicit topic representations that then support a more precise and a more reliable thematic analysis.

1 Introduction

Applications on large text bases, as information retrieval systems, have to deal with topic identification and tracking if they aim at improving their results by presenting retrieved texts in a better way, by highlighting the most relevant passages or the thematic structure of the texts. Information extraction systems would also benefit of a precise topic identification in order to delimit a context for the searched information that gives a mean to reduce ambiguities. A problem such systems have to face consists of detecting thematically coherent pieces of text and identifying their topic, without limitation about the topics that are likely to be found. Detecting coherent pieces of text refers to text segmentation, and when it is applied to large text collections, segmentation method are based on lexical knowledge (Hearst, 1997; Salton, 1996) or on non specialized knowledge as collocation networks, thesaurus or dictionary (Kozima, 1993) when dealing with narratives. However, these methods cannot identify topics as this problem requires specialized knowledge.

In order to develop methods keeping the same coverage, such knowledge bases have to result from a learning module. The main works in this field have been realized for the TDT (Topic Detection and Tracking) task and are generally based on a probabilistic approach trained on a tagged corpus as in (Bigi et al., 1998; Beeferman et al., 1999). However, we wanted to develop a method without constraints about resources to possess or to build manually, other than having texts and a general lexicon that is quite easy to obtain. Thus, our system implements an incremental clustering process to learn topic representation. These representations are called semantic domains, and are made of weighted words (Ferret and Grau, 1998). Domains result from the aggregation of similar thematic units, made of a set of words. This units are provided by a first segmentation method that relies on lexical cohesion computed from a collocation network. In a further step, domains are used to develop a topic analyzer (Ferret and Grau, 2000), that is more precise in its segmentation capacity than the first process, and, overall, that is able to identify topics and track them along the texts if they are not linearly developed.

Thus, the approach we propose consists of learning a first kind of knowledge from texts and then using this knowledge to develop a better thematic analysis. We claim that it is possible to

improve knowledge and processes in an incremental way: a weakly structured knowledge and a shallow analysis both lead to bootstrap a better analysis by providing more structured knowledge. By applying this method, we are able to go towards in-depth analysis of text, while keeping the great coverage of word-based methods. In this paper, we detail the implementation of this process by the ROSA system and we will show its performances on a classical task of evaluation for topic segmentation algorithms in order to give evidence of the contribution of such an approach.

2 Overview of the ROSA system

The ROSA system (see Figure 1) has two main components: SEGCOHLEX (Ferret, 1998), which segments texts by relying on lexical cohesion and SEGAPSITH, that incrementally learns topic representations from the most cohesive segments (Ferret and Grau, 1998) from SEGCOHLEX and use them for supporting another thematic analysis module

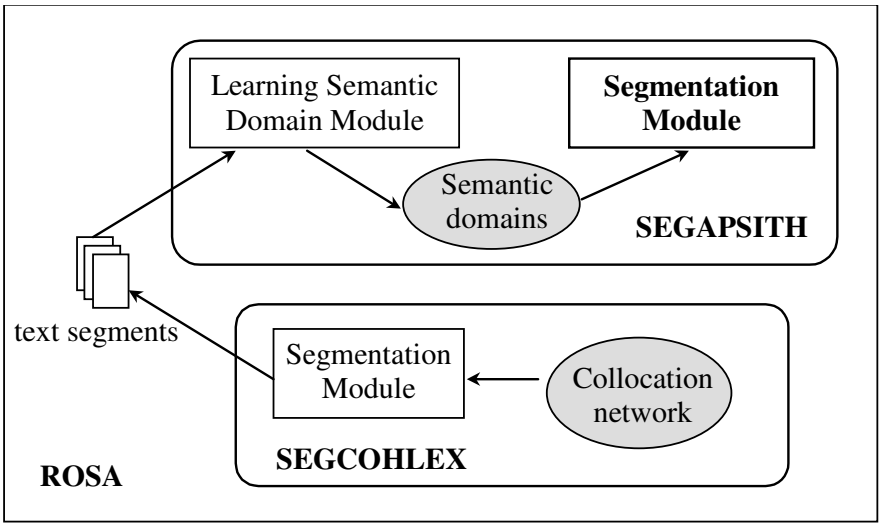

Figure 1: The architecture of the ROSA system

SEGCOHLEX only uses a general knowledge base that is automatically built, the collocation network. This kind of knowledge allows the development of a process that only delimits text segments with rather average performances and does not identify topics. However, this rough analysis generates text segments made of weighted words that are aggregated altogether when they refer to the same topic. This aggregation process produces representations of semantic domains, also made of weighted words. This new knowledge, more precise relative to the kind of analysis, allows the elaboration of a more competent segmentation module, able to identify topics and to follow topic developments. The performances of this second module are also better than SEGCOHLEX performances, as it is grounded on more structured knowledge.
We will also show that grounding SEGAPSITH on SEGCOHLEX, i.e. on a first segmentation process, is better than learning domains from the whole texts.

3 SEGCOHLEX

The topic segmentation of SEGCOHLEX, as the one of SEGAPSITH, handles texts that have been pre-processed to only keep their lemmatized content words (adjectives, single or compound nouns and verbs). This segmentation is based on a large collocation network, built

from 24 months of *Le Monde* newspaper, where a link between two words aims at capturing semantic and pragmatic relations between them. The strength of such a link is evaluated by the mutual information between its two words. The segmentation process relies on these links for computing a cohesion value for each position of a text. As in Kozima's work, this computation operates on words belonging to a focus window that is moved all over the text. It assumes that a discourse segment is a part of text whose words, referring to the same topic, are strongly linked to each other in the collocation network and yield a high cohesion value. In contrast, low cohesion values indicate topic shifts. Segments are finally delimited by an automatic analysis of the resulting cohesion graph (see (Ferret 1998) for details about the whole process and its evaluation). Such a method leads to delimit small segments, whose size is equivalent to a paragraph, i. e. capable of retrieving topic variations in short texts, as newswires for example. Nevertheless, that does not mean that our topic segmentation aim at retrieving paragraph marks. These kinds of marks are often non relevant to segment texts in their different topics. Table 1 shows an extract of the words belonging to a cohesive segment about a dedication of a book.

Segment			
strider	0.683	entourer (to surround)	0.368
toward	0.683	signature (signature)	0.366
dédicacer (to dedicate)	0.522	exemplaire (exemplar)	0.357
apposer (to append)	0.467	page (page)	0.332
pointu (sharp-pointed)	0.454	train (train)	0.331
relater (to relate)	0.445	centaine (hundred)	0.330
boycottage (boycotting)	0.436	sentir (to feel)	0.328
autobus (bus)	0.435	livre (book)	0.289
enfoncer (to drive in)	0.410	personne (person)	0.267

Table 1: Extract of a segment about a dedication

4 Semantic Domain learning in SEGAPSITH

Learning a semantic domain consists of aggregating all the most cohesive thematic units that are related to a same subject, i. e. a same kind of situation. We only retain segments whose cohesion value is upper than a threshold, in order to ground our learning on the more reliable units. Similarity between thematic units is evaluated from their common words. Each aggregation of a new TU increases the system's knowledge about one topic by reinforcing recurrent words and adding new ones. Weights on words represent their importance relative to the topic and are computed from the number of occurrences of these words in the TUs.

Inferred words of a TU			
paraphe (paraph)	0.522	imprimerie (press)	0.418
presse_parisien (parisian-press)	0.480	éditer (to publish)	0.407
best_seller (best_seller)	0.477	biographie (biography)	0.406
maison_d'édition (publishing_house)	0.450	librairie (bookshop)	0.405
libraire (bookseller)	0.447	poche (pocket)	0.389
tome (tome)	0.445	éditeur (publisher)	0.363
Grasset	0.440	lecteur (reader)	0.355
rééditer (to republish)	0.428	israélien (Israeli)	0.337
parution (appearance)	0.427	édition (publishing)	0.333

Table 2: Extract of words selected in the collocation network for the segment Table 1

Units related to a same topic are found in different texts and often develop different points of view of a same type of subject. To ensure a better similarity between them, SEGAPSITH enriches a particular description given by a text segment by adding to these units those words of the collocation network that are particularly linked to the words found in the segment. Table 2 gives an extract of the words added to the segment of Table 1.
This method leads SEGAPSITH to learn specific topic representations (see Table 3) as opposed to (Lin, 1997) for example, whose method builds general topic descriptions as for economy, sport, etc. Moreover, it does not depend on any *a priori* classification of the texts.

words	occurrences	weight
juge d'instruction (examining judge)	58	0.501
garde_à_vue (police custody)	50	0.442
bien_social (public property)	46	0.428
inculpation (charging)	49	0.421
écrouer (to imprison)	45	0.417
chambre_d'accusation (court of criminal appeal)	47	0.412
recel (receiving stolen goods)	42	0.397
présumer (to presume)	45	0.382
police_judiciaire (criminal investigation department)	42	0.381
escroquerie (fraud)	42	0.381

Table 3: The most representative words of a domain about justice

We have applied the learning module of SEGAPSITH on one month (May 1994) of AFP newswires, corresponding to 7823 TUs. The learning stage has produced 1024 semantic domains. Table 1 shows an example of a domain about justice that gathers 69 TUs. The segmentation module of SEGAPSITH only works with the most reliable of these domains. Thus, we have selected those whose number of aggregations is upper than 4. Moreover, we have selected in these 193 domains words whose weight is greater than 0.1, as below this limit a lot of words only represent noise.

5 The topic segmentation in SEGAPSITH

In accordance with works on discourse segmentation as (Grosz and Sidner, 1996), this module processes texts linearly and detects topic shifts without delaying its decision, i.e. by only taking into account the data extracted from the part of text already analyzed. A window that delimits the current focus of the analysis is moved over the text to be segmented. The segmentation algorithm locates topic shifts by detecting a significant difference between the set of domains selected for each position of the text – this set defines the topic context of the window – and the set of domains associated with the segment that is currently active at this time, which defines the topic context of the segment.

5.1 Topic contexts

A topic context aims at characterizing the entity it is associated with from the thematic point of view and is represented by a vector of weighted semantic domains. The weight of a domain expresses the importance of this domain with regard to the other domains of the vector. A context contains several semantic domains because domains are rather specific. Thus, having several of them that are close to each other allows us to cover a larger thematic field. Secondly, SEGAPSITH handles representations made of words whose meaning may be

ambiguous and refer to different topics. By putting several domains into a context, we cope with ambiguity without having to explicitly choose one interpretation.

5.1.1 Topic context of the focus window

The topic context of the focus window is made of the N^{th} semantic domains that are activated the most strongly by the words of the window, N being a fixed size for all the contexts. The activation process is equivalent to a one-step activity propagation based on the number and the weight of the words common to the window and a domain. The activation value of a semantic domain is given by:

$$activ(dom_i) = \sum_j wght(dom_i, w_j) \cdot nbOcc(w_j) \quad (1)$$

where the first factor is the weight of the word w_j in the domain dom_i (see (Ferret and Grau, 1998) for more details) and the second one is the number of occurrences of w_j in the focus window. The weight of a domain in this context is then equal to its activation value.

5.1.2 Topic context of a segment

The topic context of a segment contains the semantic domains that were activated the most strongly when the focus window was moving in the segment space. This is achieved by combining the contexts associated with each position of the focus window inside the segment (see Figure 2).

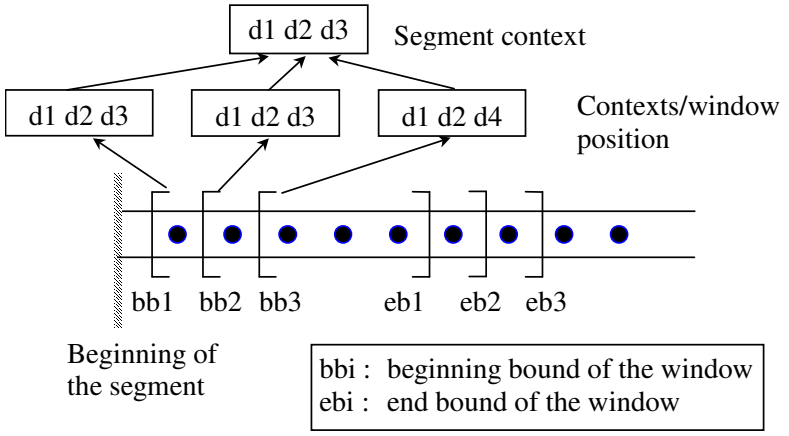

Figure 2: Building of the context segment

This fusion is done incrementally: the domains associated with each new position of a segment are merged with those of the current context of the segment; then their weights are re-evaluated according to the following general formula:

$$wght(dom_i, Cs, t+1) = \alpha(t) \cdot wght(dom_i, Cs, t) + \beta(t) \cdot wght(dom_i, Cw, t) \quad (2)$$

with Cw, the context of the window, Cs, the context of the segment and $wght(dom_i, Cx, t)$, the weight of the domain dom_i in the context Cx for the position t. The results we present in section 6 were obtained with $\alpha(t)=1$ and $\beta(t)=1$. These functions are a solution halfway between a fast and a slow evolution of the context of segments. The context of a segment has to be stable because if it follows too narrowly the thematic evolution given by the context of the window, topic shifts could not be detected. However, it must also adapt itself to small

variations in the way a topic is expressed when progressing in the text in order not to detect false topic shifts.
After weight reevaluation, the joined domains are sorted in decreasing order of weight and finally, the N^{th} first of them are selected for building the new version of the segment context.

5.2 Similarity of a window context and a segment context

In order to determine whether the content of the focus window is thematically coherent with the currently active segment, the topic context of the window is compared with the topic context of the segment. This comparison is achieved using a similarity measure between the two contexts taking into account the following four factors (see (4)):

- The significance of the domains shared by the two contexts (dom_c) with regard to those of the window (C_w) in terms of weight;
- The significance of the domains shared by the two contexts (dom_c) with regard to those of the segment (C_s) in terms of weight;
- The significance of the number of domains shared by the two contexts with regard to the size of contexts. This ensures that a high similarity is not found with only a few common domains having a very high weight (term p/N in (4));
- The difference in order among the domains shared by the two contexts. This difference is given by:

$$rankDiff\,(Cw,Cs) = \frac{\sum_{c=1}^{p} \left| rank(dom_c,Cw) - rank(dom_c,Cs) \right|}{(N-1)\cdot p} \quad (3)$$

with p, the number of common domains, dom_c, one of these common domains and $rank(dom_c,Cx)$, the rank of this domain in the context Cx. The sum of the rank differences in the two contexts is normalized by an upper bound assuming that the difference in rank is maximal ($N-1$) for each common domain.
These four factors are combined by using a geometric mean (the four terms that represent them in (4) appear in the same order):

$$sim(Cw,Cs) = \left(\frac{\sum_{c=1}^{p} wght(dom_c,Cw)}{\sum_{i=1}^{N} wght(dom_i,Cw)} \cdot \frac{\sum_{c=1}^{p} wght(dom_c,Cs)}{\sum_{i=1}^{N} wght(dom_i,Cs)} \right)^{1/4} \cdot \left(\frac{p}{N} \cdot (1 - rankDiff\,(Cw,Cs)) \right)^{1/4} \quad (4)$$

The last term is the complement of the fourth factor, as two contexts are more similar if they share domains in the same order. Two contexts are similar if the value of the similarity measure is above a fixed threshold, equal here to 0.5.

5.3 Topic shift detection

The basic algorithm that detects topic shifts computes the similarity between the context of the window and the context of the current segment at each position of a text. If this value is lower than a fixed threshold, a topic shift is assumed and a new segment is opened. Otherwise, the active segment is extended up to the current position.

This basic algorithm assumes that the transition phase between two segments is punctual. Because such a precision is not achieved, it must in fact rely on a short delay before deciding that the active segment really ends and similarly, before deciding that a new segment with a stable topic begins. Thus, the segmentation algorithm takes the form of an automaton (see Figure 3) whose transitions between its four states are controlled by three parameters:

- The current state of the algorithm;
- The similarity between the context of the focus window and the context of the current segment: *Sim* or *non Sim*;
- The number of successive positions of the focus window for which the current state stays the same: *confirmNb,* which must be above the $T_{confirm}$ threshold for going away from the states *NewTopicDetection* and *EndTopicDetection*.

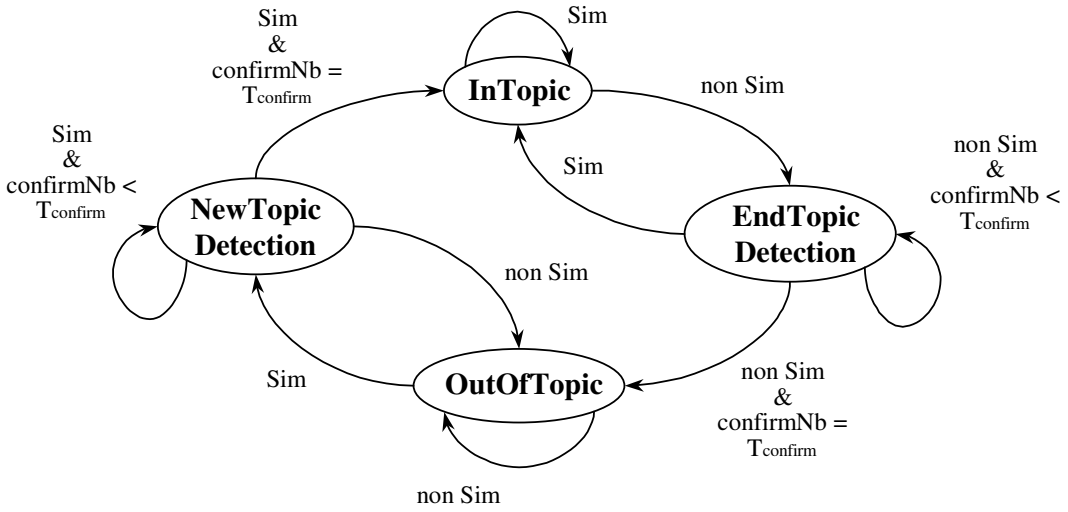

Figure 3: The automaton for topic shift detection

The usual processing of a segment starts with the *OutOfTopic* state, after the end of the previous segment or at the beginning of the text. As soon as the set of semantic domains of the focus window does not change too much between two successive positions, the topic segmenter enters into the *NewTopicDetection* state. The *InTopic* state can then be reached only if the same stability among the domains of the window context is found for the next *confirmNb-1* positions. Otherwise, the segmenter assumes that it is a false alarm and returns to the *OutOfTopic* state. The detection of the end of a segment is symmetrical to the detection of its beginning. The segmenter goes into the *EndTopicDetection* state as soon as the content of the window context begins to significantly change between two successive positions; but the transition towards the *OutOfTopic* state is done only if this change is confirmed for the *confirmNb-1* next positions.

This general algorithm is completed by two specific mechanisms. The first of them takes into account the fact that several segments of a text may refer to the same topic, which is interesting to detect for making the structure of a text explicit. Hence, when the topic segmenter goes from the *NewTopicDetection* state to the *InTopic* state, it first checks whether the current context of the new segment is similar, according to (4), to one of the contexts of the previous segments. If such a similarity is found, the new segment is linked to the

corresponding segment and it takes the context of this one as its own context. It assumes that the new segment continues to develop a previous topic.
The second mechanism is related to the *OutOfTopic* state. When the topic segmenter stays in this state for too long (10 positions of the focus window in our experiments), it assumes that the topic of the current part of text is not represented among the available semantic domains and creates a new segment with an unknown topic that covers all the positions concerned. Of course, this mechanism can not separate several connected segments of this kind but it allows us to segment texts without having all the topic representations that should be necessary.

6 Experiments

6.1 Qualitative results and discussion

A first qualitative test of the second segmentation method was done with a small set of texts and without a formal protocol as in (Passonneau and Litman, 1997). We have tested several range of values for the different parameters of the method and have found that for the kind of texts as the one given in Figure 4, the best results are obtained with a size of 19 words for the focus window and a value of 3 positions for the *confirmNb* parameter. Furthermore, results are rather stable around these values. Figure 4 shows the value of the similarity measure between the context of the focus window and the context of the current segment for each position of the given text. The two topic shifts, from the Miss Universe topic to the terrorism topic and then the return to the Miss Universe topic, are clearly detected through significant falls of the similarity values (positions 62-63 for the first et 89 to 91 for the second; these shifts are marked in bold in the text). On the other hand, the method misses the last topic shift (from the Miss Universe topic to the demonstration topic) because it is expressed very shortly and not in a very specific way. Note that in the text, the foreseen topic shifts, marked by (<ST>, </ST>) tags do not correspond to paragraph boundaries.

<ST> An 18 year old Indian model, Sushmita Sen, caused a surprise on Sunday in Manilla when winning the Miss Universe 1994 title ahead of two South-American beauties, Miss Colombia, Carolina Gomez Correa, and above all Miss Venezuela, Minorka Mercado, who appeared as favorite in the competition.
The young Indian, a brown beauty, hazel eyed and 1.75 meters tall, is the first candidate of her country to win this title. She succeeds to Miss Porto Rico, Dayanara Torres, 22, who gave her her crown in front of a television hearing estimated to six hundred million people all over the world. Among the six finalists also appeared Miss United States, Frances Louis Parker, Miss Philippines, Charlene Gonzales, and Miss Slovak Republic, Silvia Lakatosova. The new miss was chosen among a group of ten finalists that also included the **representatives** of Italy, Greece, Sweden and Switzerland. *</ST>*
<ST> A few hours before the ceremony, a man was killed by the explosion of an appliance he carried, at about one kilometer from the Congress Center where the beauty competition was being held, in front of the Manilla bay. The police was not immediately able to establish if this incident was in relation with this competition.
On Thursday, a weak-power craft bomb had exploded in a garbage can of the **congress** center without any damages. *</ST>*
<ST> The new Miss Universe, who won more than 150,000 Dollars in different prizes, declared that she intended to do theater, publicity or writing. However, her most cherished wish, she assured, was to meet Mother Teresa because she was “a perfect example of a person totally devoted, unselfish and completely involved”. *</ST>*
<ST> During the election, about a hundred feminists demonstrated pacifically in front of the Congress Center to denounce the competition, stating that it promoted sexual tourism in Philippines. *</ST>*

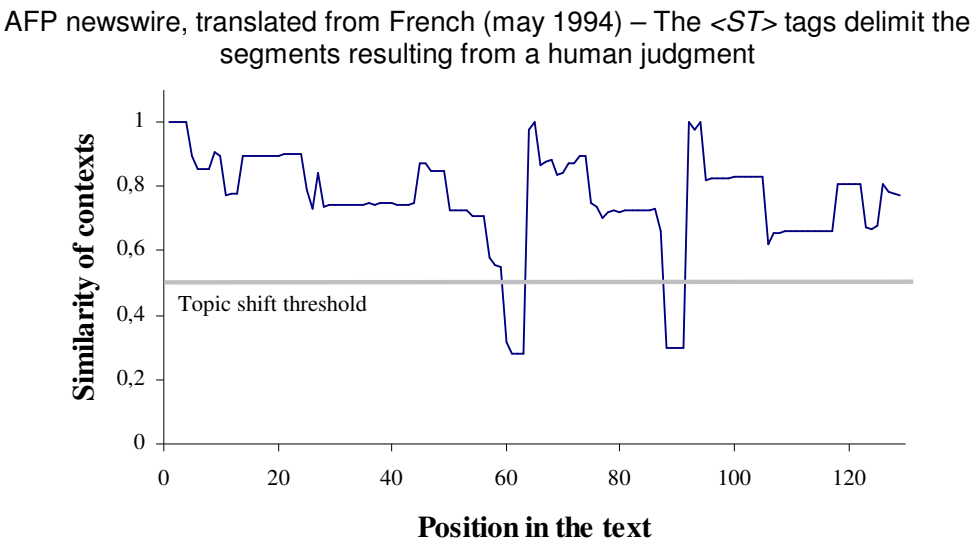

Figure 4: A text and its context similarity graph (for the French text)

The analysis of this example also illustrates two important characteristics of our method. As it makes use of an explicit representation of topics, it allows us to recognize that two disconnected segments are related to the same topic, as it is done here for the segments 1 and 3 about the Miss Universe topic.

Our method also segments texts without having an exact representation of the topics of the texts. Thus, the newswire above was segmented without having a semantic domain related to beauty competitions. This topic was represented here by only one of its dimensions, competition, through a set of domains about sport competitions. More generally, as a context is a set of domains, a topic representation can be dynamically built by associating several domains related to different dimensions of this topic.

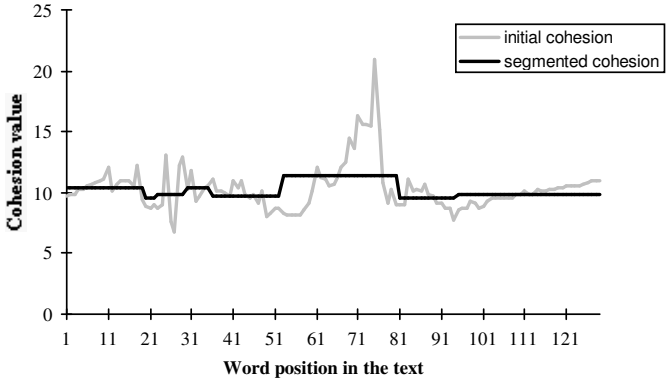

Figure 5: SEGCOHLEX results on the same text than in Figure 4

If we now compare the results given by SEGAPSITH and SEGCOHLEX when analyzing the same text, we see in Figure 5 that SEGCOHLEX detects 6 small units, made of few words, within the first part of the text (until the 52nd word), where SEGAPSITH only finds one segment. The boundaries of its main segment (words 53 to 80) recognized by SEGCOHLEX are also less precise. When evaluating SEGCOHLEX (Ferret 1998), we have shown that this

method was comparable to the other works in the field and can be considered as a baseline. Thus, the better results of SEGAPSITH can be attributed to the exploitation of a dedicated source of knowledge.

6.2 Quantitative evaluation

In order to evaluate the interest of bootstrapping objectively, we have applied the segmentation methods of SEGCOHLEX and SEGAPSITH to the classical task of discovering boundaries between concatenated texts. As we are interested in segmenting texts at the paragraph level, our evaluation has been performed with short texts, more precisely 49 texts from *Le Monde* newspaper that are 133 words long on average. Like (Hearst, 1997), we have defined precision as N_t / N_b and recall as N_t / D, where N_t is the number of boundaries that match with document breaks (within an interval of 9 words around each document break, after pre-processing), N_b, the number of found boundaries, and D, the number of document breaks. Results shown in table 2 are means computed from 10 trials, the order of texts being changed from one trial to another. These results clearly show that SEGAPSITH has better performances than SEGCOHLEX. Furthermore, they show the interest of bootstrapping: SEGAPSITH(1), which relies on semantic domains built from TUs, has better results – its little inferiority in precision is greatly balanced by its superiority in recall – than SEGAPSITH(2), whose domains were built from texts without any segmentation or enrichment.

segmentation procedure	recall	precision	f-measure
SEGOHLEX	0.675	0.374	0.481
SEGAPSITH(1)	0.920	0.523	0.666
SEGAPSITH(2)	0.810	0.535	0.644

Table 2: Results of the evaluation for the different methods

Although a meaningful comparison is not easy to do, our best results seem at least comparable to the other works in the field. In a similar evaluation with 44 texts (average length of 16 paragraphs), Hearst (Hearst, 1997) gets 0.95 as precision and 0.59 as recall. However, a direct comparison is not completely relevant as Hearst's method cannot be applied to texts as smaller as ours. The work in (Bigi et al., 1998) is more similar to ours although its evaluation method is slightly different. The differences between its results – 0.75 as precision and 0.80 as recall – and ours can be explained by the nature of topics: this work focuses on a small set of very general topics (business, politics, ...) while we focus on a large set of specific topics. As our topic representations are closer to the topics of texts, we have a better recall; but we also are likely to have more noise, which explains our lower precision.

7 Related works

Thematic analysis able to identify topics must rely on explicit topic representations, as (Bigi et al., 1998) or works done in the Topic Detection and Tracking (TDT) framework (Fiscus et al., 1999). Like these methods, the one of SEGAPSITH makes use of explicit topic representations but it exploits them with tools similar to text segmentation works based on lexical cohesion as (Kozima, 1993) or (Hearst, 1997) and not with the probabilistic approach generally found in TDT or in (Beeferman et al., 1999). Works done in the TDT framework

also differ from ours in the delay for deciding if a topic shift occurs: from 100 up to 10000 words in TDT but only 3 content words in SEGAPSITH. Moreover, topics in TDT represent events. On the contrary, they are very general in (Bigi et al., 1998). Semantic domains in SEGAPSITH are halfway between these two extremes: they aim at describing specific topics but not events.
Topic representations clearly allow working at a finer grain than methods based on lexical cohesion. But on a large scale, they also require to be automatically built, preferably in an unsupervised way. This problem is tackled to some extent in the Detection task of the TDT evaluation but not in the segmentation one. On the contrary, SEGAPSITH includes an unsupervised learning of the topic representations that support its segmentation module. Moreover, its association with SEGCOHLEX allows to progressively bootstrap a fine-grained segmentation module based on topic representations from a rougher segmentation module based on lexical cohesion.

8 Conclusion

Developing reliable NLP processes requires large bases of structured knowledge that are very difficult to build. In order to overcome this problem for thematic analysis, we adopted a bootstrapping approach: we implemented a first process, based on automatically acquired knowledge, used its results to build finer topic representations and then developed another process that exploits the new base in order to get better performances. The evaluation of this method shows that results obtained with structured and specialized knowledge are better than with general one and moreover, that learning knowledge from the results of text segmentation is more reliable than learning it from non processed texts.

References

D. Beeferman, A. Berger and J. Lafferty. 1999. Statistical Models for Text Segmentation. Machine Learning, 34 (1/3), 177-210.

B. Bigi, R. de Mori, M. El-Bèze and T. Spriet. 1998. Detecting topic shifts using a cache memory, In *Proceedings of 5th International Conference on Spoken Language Processing*, Sydney, Australia.

O. Ferret. 1998. How to thematically segment texts by using lexical cohesion? In *Proceedings of ACL-COLING'98 (Student Session)*, Montréal, Canada.

O. Ferret and B. Grau. 1998. A Thematic Segmentation Procedure for Extracting Semantic Domains from Texts. In *Proceedings of ECAI'98*, Brighton, UK.

O. Ferret and B. Grau. 2000. A Topic Segmentation of Texts based on Semantic Domains. To appear in *Proceedings of ECAI'2000*, Berlin.

J. Fiscus, G. Doddington, J. Garofolo and A. Martin. 1999. NIST's 1998 Topic Detection and Tracking Evaluation (TDT2), In *Proceedings of DARPA Broadcast News Workshop*, Herndon, Virginia.

B.J. Grosz and C.L. Sidner. 1986. Attention, Intentions and the Structure of Discourse. *Computational Linguistics*, 12: 175-204.

M.A. Hearst. 1997. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics*, 23(1): 33-64.

H. Kozima. 1993. Text Segmentation Based on Similarity between Words. In *Proceedings of the 31th Annual Meeting of the ACL (Student Session)*, Colombus, Ohio, USA.

C.-Y. Lin. 1997. *Robust Automated Topic Identification*, Doctoral Dissertation, University of Southern California.

R.J. Passonneau and D.J. Litman. 1997. Discourse Segmentation by Human and Automated Means, *Computational Linguistics*, 23 (1): 103-139.

G. Salton, A. Singhal, C. Buckley and M. Mitra. 1996. Automatic Text Decomposition Using Text Segments and Text Theme*s,* In *Proceedings of Hypertext'96*, Washington, D.C.