A Topic Segmentation of Texts based on Semantic Domains

Olivier Ferret[1] and **Brigitte Grau**[1]

Abstract. Thematic analysis is essential for many Natural Language Processing (NLP) applications, such as text summarization or information extraction. It is a two-dimensional process that has both to delimit the thematic segments of a text and to identify the topic of each of them. The system we present possesses these two characteristics. Based on the use of semantic domains, it is able to structure narrative texts into adjacent thematic segments, this segmentation operating at the paragraph level, and to identify the topic they are about. Moreover, semantic domains, that are topic representations made of words, are automatically learned, which allows us to apply our system on a wide range of texts in varied domains.

1 INTRODUCTION

Thematic analysis covers different kinds of problems. Text segmentation consists of delimiting parts of texts related to different topics. Topic structuring makes the underlying thematic structure of a text explicit while topic identification associates topic representations to parts of texts. Depending on the kind of texts to be analyzed and the problem to be solved, methods require background knowledge or not. Scientific or technical texts may be segmented by examining how words are distributed in the texts [7] [11]. Within this paradigm, Salton [11] proposed a method to structure a text based on a graph of links between segments. However, when dealing with narratives, as newspaper articles, such methods are inappropriate and require some domain knowledge.

Approaches for analyzing narratives can roughly be categorized in two classes, knowledge-based approaches and word-based approaches. Knowledge-based systems [6] lead to a precise decomposition and identification of discourse topics by using in-depth understanding processes and a knowledge base about situations (their characters, their events chronologically ordered, their consequences). As such bases do not exist, apart possibly for small domains, word-based approaches have been developed [7] [8] [3] allowing to overcome this limitation and process texts regardless of the topic. The purpose of these systems is to segment texts, but not to recognize topics. Improving such methods requires going towards in-depth understanding and thus using knowledge about topics, but without the former limitation.

The thematic analysis we propose relies on specific semantic domains, automatically learned from texts. These topic representations, described by sets of words, allowed us to develop a segmentation process at the paragraph level. We detail in this paper the implementation of this process and we show its performances on a classical evaluation task for topic segmentation algorithms.

[1] LIMSI-CNRS. BP 133, 91403 Orsay Cedex, France.
email: [ferret,grau]@limsi.fr

2 OVERVIEW OF THE SEGAPSITH SYSTEM

SEGAPSITH has two main components: a module that learns semantic domains from texts and a topic segmentation module that relies on these domains. Semantic domains are topic representations made of weighted words. They are built by incrementally aggregating a large number of similar text segments. These segments are delimited by a rough segmentation module called SEGCOHLEX [3]. Figure 1 shows the overall architecture of the system.

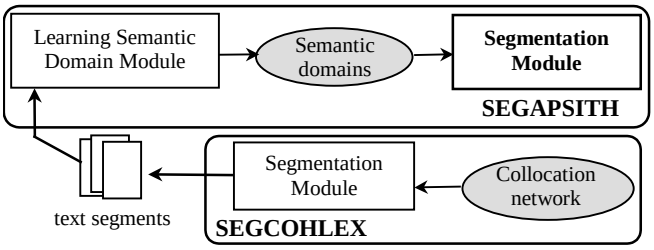

Figure 1. The architecture of the SEGAPSITH system

The segmentation modules operate on pre-processed texts in order to keep only their lemmatized content words. In accordance with works on discourse segmentation as [6], the SEGAPSITH one 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 each text to be segmented. A topic context is associated to this window. It is made up of the semantic domains that are the most activated by the content of this window. The current segment is also given a topic context. This context results from the fusion of the contexts associated to the focus window when this window was in the segment space. A topic shift is then detected when the context of the focus window and the context of the current segment are not similar any more for several successive positions of the focus window. This process not only locates topic shifts but by identifying topics of segments, also links non adjacent segments related to the same subject.

3 SEMANTIC DOMAIN LEARNING

As the focus of this paper is the segmentation component of the SEGAPSITH system, we give here only a brief overview of the

learning semantic domain module. This module is described more precisely in [4]. It builds topic representations incrementally; these representations are made of weighted words, from discourse segments delimited by SEGCOHLEX [3]. The module works without any *a priori* classification or hand-coded pieces of knowledge. Processed texts are typically newspaper articles. They are pre-processed to only keep their lemmatized content words (adjectives, single or compound nouns and verbs).

The topic segmentation implemented by SEGCOHLEX is based on a large collocation network, built from 24 months of the French newspaper *Le Monde*, 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. It assumes that a discourse segment is a part of text whose words refer to the same topic, that is, words are strongly linked to each other in the collocation network and yield a high cohesion value. On the contrary, low cohesion values indicate topics shifts. After delimiting segments by an automatic analysis of the graph resulting from the cohesion computation, only highly cohesive segments, named Thematic Units (TUs), are kept to learn topic representations. This segmentation method entails a text to be decomposed in small TUs, whose size is equivalent to a paragraph. To enrich the particular description given by a text, we add to TUs those words of the collocation network that are particularly linked to the words found in the corresponding segment.

Table 1. The most representative words of a domain about justice

words	occ.	weight
examining judge	58	0.501
police custody	50	0.442
public property	46	0.428
indictment	49	0.421
to imprison	45	0.417
court of criminal appeal	47	0.412
receiving stolen goods	42	0.397
to presume	45	0.382
criminal investigation department	42	0.381
fraud	42	0.381

Learning a complete description of a topic consists of merging all successive points of view, i.e. similar TUs, into a single memorized thematic unit, called a semantic domain. 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 the importance of each word relative to the topic and is computed from the number of occurrences of these words in the TUs. This method leads SEGAPSITH to learn specific topic representations as opposed to [9] for example whose method builds general topic descriptions as for economy, sport, etc.

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 superior to 4. Moreover, we have selected in these 193 domains words whose weight is superior to 0.1, since below this limit a lot of words only represent noise.

4 THE TOPIC SEGMENTATION

The segmentation algorithm of SEGAPSITH locates topic shifts by detecting when a significant difference is found between the set of semantic domains selected for each position of a text and the set of domains associated to the segment that is currently active at this time. The first set of domains defines the window context and the second set the segment context.

4.1 Topic contexts

A topic context aims at characterizing the entity it is associated to 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 domains because domains are rather specific. Thus, having several domains that are close to each other allows the system 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 this ambiguity without having to choose explicitly one interpretation.

4.1.1 Building the topic context of the focus window

The topic context of the focus window is built first, by activating the semantic domains that are available from the words of the focus window and then, by selecting the most activated of these domains. The activation value of a semantic domain is given by:

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

where the first factor is the weight of the word w_j in the domain dom_i (cf. [4] for more details) and the second one is the number of occurrences of w_j in the focus window.

After this activation step, the context of the focus window is set by selecting the N^{th} first semantic domains according to their activation value. Their weight in this context is equal to their activation value. N is the fixed size of all the contexts.

4.1.2 Building the topic context of a segment

The topic context of a segment contains the semantic domains that were the most activated when the focus window was moving in the segment space. This is achieved by combining the contexts associated to each position of the focus window inside the segment (see Figure 2). This fusion is done incrementally: the context of each new position of a segment is combined with the current context of the segment. First, the semantic domains of both contexts are joined. Then, their weight is revalued according to this formula:

$$wght(dom_i, Cs, t+1) = \alpha(t) \cdot wght(dom_i, Cs, t) + \beta(t) \cdot wght(dom_i, Cw, t) \qquad (4)$$

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 the next sections are 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 cannot 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.

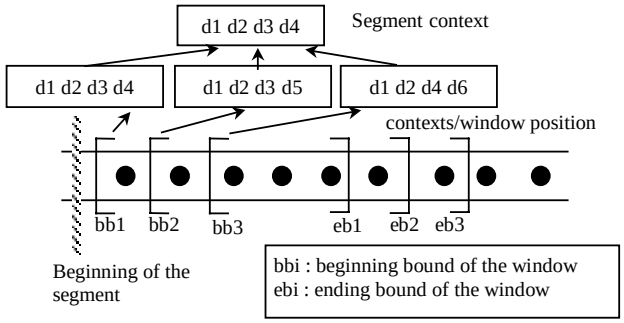

Figure 2. Building of the context segment

After weight revaluation, 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.

4.2 Evaluating the similarity of two contexts

In order to determine if the content of the focus window is thematically coherent or not with the segment that is currently active, the topic context of the window is compared to the topic context of the segment. This comparison is achieved by a similarity measure taking into account the following four factors:

1. The significance of the domains shared by the two contexts with regard to the window domains in terms of weight;
2. The significance of the domains shared by the two contexts with regard to the segment domains in terms of weight;
3. 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 domains in common having a very high weight;
4. The difference of order among the domains shared by the two contexts. This difference is given by:

$$rankDiff(Cw,Cs) = \frac{\sum_{c=1}^{p} \left| rank(dom_c,Cw) - rank(dom_c,Cs) \right|}{(N-1)\cdot p} \quad (5)$$

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. In this factor, the sum of the rank differences of domains in the two contexts is normalized by an upper bound assuming that the difference of rank is maximal (N-1) for each common domain.

These factors are combined in a geometric mean (6). The first two factors are gathered in the first term of the global product. The term p/N corresponds to the third factor and the last term is the complement of the fourth factor, as two contexts are more similar if they share domains in the same order. The values of this similarity measure are in the interval [0,1] since the values of each of its four components are also in the same interval.

$$sim(Cw,Cs) = \left(\frac{\sum_{c=1}^{p} wght(dom_c,Cw)}{\sum_{i=1}^{N} wght(dom_i,Cw)} \cdot \frac{\sum_{c=1}^{p} wght(dom_c,Cs)}{\sum_{i=1}^{N} wght(dom_i,Cs)} \right)^{1/4} \cdot \left(\frac{p}{N} \cdot (1 - rankDiff(Cw,Cs)) \right)^{1/4} \quad (6)$$

Two contexts are considered as similar if the value of the similarity measure is above a fixed threshold. In all the experiments we present here, this threshold was set to 0.5.

4.3 Topic shift detection

The basic algorithm that detects topic shifts is the following: for each position of a text, if the value of the similarity measure between the topic context of the focus window and the topic context of the current segment 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. It actually must be more complex because of the lack of precision of SEGAPSITH. This imprecision makes it necessary to set a short delay before deciding that the active segment really ends and similarly, before deciding that a new segment with a stable topic begins. Hence, the algorithm for detecting topic shifts distinguishes four states:

1. The *NewTopicDetection* state. This state takes place when a new segment is going to be opened. This opening will be confirmed provided that the content of the focus window context stays mainly the same for several positions. Moreover, the core of the segment context is defined when the topic segmenter is in the *NewTopicDetection* state;
2. The *InTopic* state, which is active when the focus window is inside a segment with a stable topic;
3. The *EndTopicDetection* state. This state is active when the focus window is inside a segment but a difference between the context of the focus window and the context of the current segment suggests that this segment could end soon. As for 2, this difference has to be confirmed for several positions before a change of state is decided;
4. The *OutOfTopic* state. This state occurs between two segments. Most of the time, the segmentation algorithm stays in this state no longer than 1 or 2 positions but when the semantic domains that should be related to the current topic of the text are not available, this number of positions may be equal to the size of a segment.

The segmentation algorithm follows the transitions of the automaton of the Figure 3 according to three parameters:

1. the current state of the algorithm;
2. the similarity between the context of the focus window and the context of the current segment: *Sim* or *non Sim*;
3. 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*.

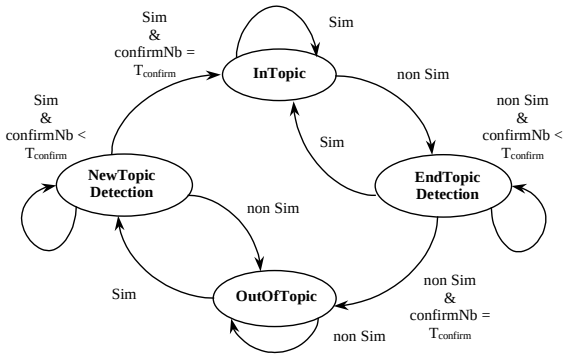

Figure 3. The automaton for topic shift detection

The 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 is stable enough between two successive positions, the topic segmenter enters into the *NewTopicDetection* state. The *InTopic* state can then be reached only if the same stability 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 change significantly between two successive positions but the transition towards the *OutOfTopic* state is done only if this change is confirmed for the next *confirmNb*-1 next positions.

This algorithm is completed by two specific mechanisms. First several segments of a text may refer to the same topic, which is necessary 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 (6), 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 too long in this state (this time is defined as 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 domains and it creates a new segment with an unknown topic that covers all the concerned positions. Of course, this mechanism cannot separate several connected segments of this kind but it allows the system to segment texts without having all the topic representations that should be necessary.

5 EXPERIMENTS

5.1 Qualitative results and discussion

A first qualitative test of the segmentation method was done with a small set of texts and without a formal protocol as in [10]. 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.

<*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*>

AFP newswire, translated from French (may 1994) – The <*ST*> tags delimit the segments resulting from a human judgment

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.

5.2 Quantitative evaluation

In order to have a more objective evaluation, we applied our method to the "classical" task of discovering boundaries between concatenated texts. As we are interested in segmenting texts at the paragraph level, our evaluation was performed with short texts, precisely 49 texts from *Le Monde* newspaper of 133 words long on average. Results are shown in Table 2. As in [7], boundaries found by the method are sorted in decreasing order of their probability to be document breaks (we rely for this on the similarity between contexts). For the first N_b boundaries, N_t is the number of boundaries that match with document breaks. Precision is given by N_t / N_b and recall, by N_t / D, with D the number of document breaks. The f-measure is the harmonic mean of precision and recall. The match between a boundary and a document break was accepted if the boundary was not further than 9 words (after pre-processing).

Table 2. Results of the evaluation

N_b	N_t	recall	precision	f-measure
10	7	0.146	0.636	0.237
20	11	0.229	0.524	0.319
30	18	0.375	0.581	0.456
40	24	0.5	0.585	0.539
50	28	0.583	0.549	0.566
60	33	0.688	0.541	0.606
70	38	0.792	0.535	0.639
77	43	0.896	0.551	0.683

Globally, our results are comparable to the other studies in the field although a significant comparison is not easy to do. In [7], Hearst reports a similar evaluation but with much larger texts (average length of 16 paragraphs). For 44 texts, she gets 0.95 as precision, 0.59 as recall and 0.728 as f-measure (comparison must be done with the last line of Table 2). However, a direct comparison is not completely relevant as Hearst's method cannot be applied to texts as smaller as ours. The work in [2] is more similar to ours. Although its evaluation method is slightly different, its results can be compared to ours: in the best case, 0.75 as precision, 0.80 as recall and 0.774 as f-measure. The differences can be explained by the nature of topics: [2] focuses on a small set of very general topics (such as business, politics) while we focus on a large set of specific topics. As a consequence, our recall is better – our topic representations are closer to the topics of the texts – while [2] shows a better precision – we are likely to have more noise. However, it is important to note that our precision does not decrease as more document bounds are found.

6 RELATED WORKS

Our method is a synthesis between methods that rely on lexical cohesion for segmenting texts and use word recurrence [7] or a lexical network [8], and methods that rely on explicit topic representations, as [2] or studies done in the Topic Detection and Tracking (TDT) framework [5]. As these last methods, ours makes use of explicit topic representations but it exploits them with the same tools as [7] or [8] and not with the probabilistic approach generally found in TDT or in [1]. In [2], topics are very general. On the contrary, they are very specific in [5] and often comparable to events. Domains in SEGAPSITH are halfway between these two extremes: they aim at describing specific topics but not events. Studies done in the TDT framework also differ from ours in the delay for deciding if a topic shift occurs. They take a decision after a deferral period going from 100 up to 10000 words while this parameter is equal to only 3 content words in our method.

Having topic representations clearly allows us to work at a finer grain than methods based on lexical cohesion. But on a large scale, it also requires to automatically build these representations, 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 a module that learns in an unsupervised way the topic representations that support its segmentation module. Moreover, its association with SEGCOHLEX allows our global system to progressively go from a segmentation module based on lexical cohesion to a segmentation module based on topic representations.

7 CONCLUSION

The segmentation module of SEGAPSITH implements a topic segmentation of texts at the paragraph level. Using semantic knowledge about domains provides a mean to identify these topics. SEGAPSITH was also designed to recognize links between non adjacent segments. Furthermore, it is able to manage lack of knowledge about some domains. The evaluation of this system gives good results on narrative texts. However, we envisage to structure domains hierarchically and to take into account this hierarchy for going further in the thematic structuring of texts.

REFERENCES

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

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

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

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

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

[6] B.J. Grosz and C.L. Sidner. Attention, Intentions and the Structure of Discourse. *Computational Linguistics*, **12**, 175-204, (1986).

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

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

[9] C.-Y. Lin. *Robust Automated Topic Identification*, Doctoral Dissertation, University of Southern California, (1997).

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

[11] G. Salton, A. Singhal, C. Buckley and M. Mitra. Automatic Text Decomposition Using Text Segments and Text Themes, In *Proceedings of Hypertext'96*, Washington, D.C., (1996).