Natural Language Engineering **8** (2/3): 209–233. © 2002 Cambridge University Press
DOI: 10.1017/S1351324902002929 Printed in the United Kingdom

A bootstrapping approach for robust topic analysis

OLIVIER FERRET, BRIGITTE GRAU
LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France

(*Received 10 July 2001; revised 7 February 2002*)

Abstract

Topic analysis is important for many applications dealing with texts, such as text summarization or information extraction. However, it can be done with great precision only if it relies on structured knowledge, which is difficult to produce on a large scale. In this paper, we propose using bootstrapping to solve this problem: a first topic 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 reliable topic analysis.

1 Introduction

The problem we address in this paper is the topic analysis of texts on a large scale. Topic analysis is a kind of paradox: it is a very intuitive process that seems familiar to every reader, but it is also very difficult to define precisely, especially in the field of linguistics. In our opinion, topic analysis covers three kinds of problems: topic segmentation, topic identification and topic structuring of texts. This division is closely akin to the division into the following three dimensions: syntagmatic, paradigmatic and functional. Topic segmentation consists of delimiting parts of texts that are thematically coherent. Topic identification associates topic representations to the parts of texts delimited by topic segmentation, and topic structuring makes the underlying topical structure of a text explicit by finding the relations between its segments.

Much work has already tackled these three problems (see section 2), but the Topic Detection and Tracking (TDT) evaluations from DARPA (Fiscus *et al.* 1999) have recently brought topic analysis to the fore. The TDT framework mainly addresses the segmentation and the identification problems. The segmentation task of TDT consists of segmenting a stream of text into a set of documents. Hence, it is not a fine-grained segmentation. The identification problem is represented by two specialized tasks: the detection task, which aims at detecting the occurrence of new topics and the tracking task, which recognizes the occurrences of already known topics. A more recent task, called link detection, is dedicated to detecting whether two stories are topically linked, which represents a first attempt towards the identification of the topic structuring of texts.

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

The recent interest given to topic analysis is mainly motivated by its applications. As it is shown by TDT studies, topic analysis has an interest from that viewpoint on its own. More generally, it is useful for most applications in information retrieval and information extraction. For example, it supports passage retrieval, which improves indexing by producing more homogeneous text units and makes it possible to present retrieved texts in a better way, by highlighting their most relevant passages or the words that characterize their topics. In information extraction, a precise topic analysis enables systems to delimit a context for searched information, which narrows the search area and contributes to reduce ambiguities. This capability is also used in Question Answering systems (Hovy *et al.* 2000). In the field of automatic speech recognition, topic analysis is exploited for adapting language models according to the current topic (Lau 1994). More generally, topics are interesting content units for multimedia processing as they are not linked to a specific medium.

One important feature of recent applications in information retrieval and extraction, such as the question answering systems evaluated in the TREC evaluation (Voorhees 2000), is their broad coverage. Possibly because of the influence of the WEB, they are supposed to be open-domain, which is a kind of robustness. As a consequence, the linguistic analysis upon which they rely, such as topic analysis, must also be open-domain. Methods exhibiting this kind of robustness exists for topic analysis (e.g. see Hearst (1997) or Kozima (1993)), but as they only make use of weakly structured knowledge or basic text features, they get poor results and are limited to topic segmentation. Methods that exploit more elaborated knowledge also exist (e.g. see Grosz and Sidner (1986) or Grau (1984)), but they can only be applied in restricted domains.

The approach we propose aims at overcoming the deadlock between knowledge and coverage. It consists of learning a first kind of knowledge from texts and then using this knowledge to develop a better topic 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 texts while keeping the wide coverage of shallow methods. In this paper, we detail the implementation of this process by the ROSA system and we present its results on a classical task for evaluating topic segmentation algorithms in order to show evidence of the contribution of such an approach.

Our paper is organized as follows: in section 2, we review some previous works about topic segmentation and topic analysis. In section 3, we give an overview of the ROSA system. Then, we describe its components: the segmentation module that depends on weak knowledge in section 4, the learning of topic representations in section 5, and finally, the topic analysis module that exploits them in section 6. In section 7, we report on a series of qualitative and quantitative experiments and we compare our results to those of others works in the domain. Finally, in section 8, we mention some extensions to our work.

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

2 Previous work about topic analysis

Our use of a bootstrapping method aims at combining the two main approaches in the field of topic analysis in one framework.

2.1 Approach based on structured knowledge

The first approach is represented by work such as Grosz and Sidner (1986), Grau (1984) or Hahn (1990), that uses high level knowledge, schemata for instance (Schank 1982), in order to achieve precise and complete topic analysis. This analysis delimits topical segments, identifies the topic of each one and finds the relations between them that structure a text in a thematic way. In Grosz and Sidner (1986), the topic analysis is only a particular case of a more general model about discourse analysis. The main drawback of this approach is that it requires a large amount of work to represent each considered domain. Hence, it can only be applied in restricted domains.

2.2 Quantitative approach

The second approach is quantitative, and relies on the topical information that can be found at the lexical level. Such a topic analysis is generally restricted to segmenting texts into adjacent units. Three kinds of work can be distinguished according to the type of resources they use.

2.2.1 Exploiting intrinsic characteristics of texts

In a first category, systems only make use of the intrinsic characteristics of texts. The topic shifts are detected either by identifying cues that mark a new topic or the end of the current topic (Passonneau and Litman 1997), or by exploiting the shifts of the lexical cohesion, as it is defined by Halliday and Hasan (1976). The lexical cohesion is characterized in this case by the distribution of words in texts. This method was first experimented by Youmans (1991), and improved by further work: Hearst (1997), Nomoto and Nitta (1994) and Masson (1995) depended upon a narrow context delimited by a focus window; Salton, Singhal, Buckley and Mitra (1996) and Yaari (1997) looked at finding a global topical structure of texts; Kan, Klavans and McKeown (1998) evaluated the interest of using complex terms; Reynar (1994) and Choi (2000) built a global map of similarity between sentences and found topically coherent regions by divisive clustering; and Utiyama and Isahara (2001) designed a statistical model for text segmentation.

2.2.2 Exploiting knowledge about lexical cohesion

A second set of methods exploit knowledge that is not related to the topics texts are about. They rely more precisely on sources of knowledge about lexical cohesion: a network of words built from an electronic dictionary for Kozima (1993), a thesaurus for Morris and Hirst (1991), or a collocation network for Ferret (1998), Ferret,

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

Grau and Masson (1998), Kaufmann (1999) and Choi, Wiemer-Hastings and Moore (2001), whose work was done in the context of Latent Semantic Analysis (Deerwester *et al.* 1990). These methods particularly suit texts with a general vocabulary, where an idea is expressed in many different forms, as in narrative texts. Their results obviously depend upon the presence of the vocabulary of the processed texts in the lexical network.

2.2.3 Exploiting topical knowledge

A last set of methods exploit knowledge about topics, mainly in the form of probabilistic language models. Each topic that could be present in a text is represented by a specific model, built from a set of texts that are manually selected as representative of this topic. Most of the work done in the *Topic Detection and Tracking* (TDT) framework (Fiscus *et al.* 1999) follows that line of conduct.[1] The work of Beeferman, Berger and Lafferty (1999) is also a typical example of it. In both cases, segmentation aims at finding text bounds, which is a rather high level of granularity. On the contrary, topic identification focuses on very specific topics that are similar to events. Nevertheless, the system of Bigi, de Mori, El-Bèze and Spriet (1998) shows that similar means can be used for segmenting texts at a fine-grained level and identifying more general topics.

2.3 Our approach

In comparison with previous work, the ROSA system we present in this paper is a hybrid system. It resorts to a quantitative method of segmentation that uses a knowledge source about lexical cohesion to build topic representations. These representations are then used by a topic analysis that performs topic segmentation, topic identification and link detection. Thus, we aim at progressively implementing on a large scale the first approach presented above, i.e. achieving a precise topic analysis on a large scale that can find the topic structure of texts.

3 Overview of the ROSA system

The ROSA system (Figure 1) has two main components: SEGCOHLEX (Ferret 1998), that segments texts by relying on lexical cohesion and SEGAPSITH, that incrementally learns topic representations from the most cohesive segments (Ferret and Grau 1998) produced by SEGCOHLEX, and exploits them for supporting a more elaborate topic analysis module.

SEGCOHLEX relies on a general knowledge source about lexical cohesion – a collocation network[2] – that is automatically built. This kind of knowledge only

[1] Segmentation systems in TDT are often based on several kinds of models: some of them represent topics but others try to characterize more directly the topic shifts by using cues.

[2] A collocation network is a large set of collocations between words, i.e. word cooccurrences, collected from a corpus. This set can be viewed as a network whose nodes are words and edges are collocations.

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

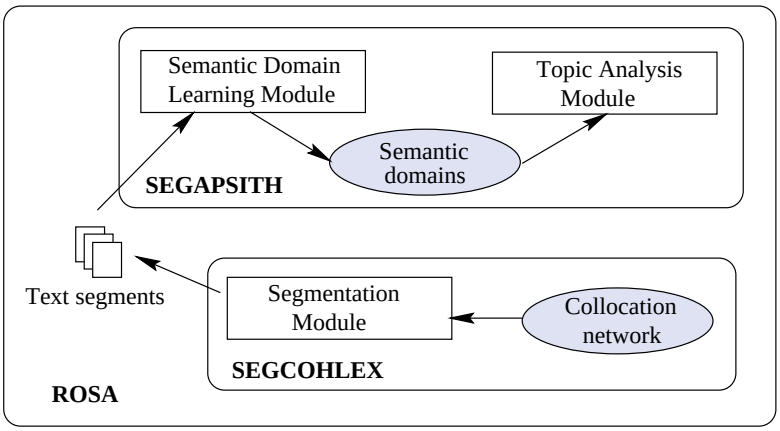


Fig. 1. Architecture of the ROSA system.

enables a system to delimit text segments with rather average results, and does not enable it to achieve topic identification. However, this rough analysis produces text segments, made of weighted words, that are aggregated altogether by the learning module of SEGAPSITH when they refer to the same topic. This aggregation process, which is incremental and unsupervised, produces topic representations called semantic domains, also made of weighted words. This new knowledge, which is more specific and more structured from a topical viewpoint than a collocation network, supports the topic analysis module of SEGAPSITH. This module is able to segment texts with better results than those of SEGCOHLEX, and is also more complete than SEGCOHLEX, since it is able to identify topics and to characterize the topical structure of texts through the detection of links between segments that refer to the same topics. Moreover, we will show that grounding SEGAPSITH on SEGCOHLEX, i.e. on a first segmentation process, gets better results than learning domains directly from texts.

4 SEGCOHLEX

First we present SEGCOHLEX, as it is the basis of ROSA, and more specifically its collocation network, which is its reference about lexical cohesion.

4.1 The collocation network of SEGCOHLEX

The collocation network of SEGCOHLEX was built from a large corpus, whose size was around 39 million words and which was made up of 24 months of the *Le Monde* newspaper taken from 1990 to 1994. The corpus was pre-processed so as to characterize texts by their significant words from the topical point of view. Thus, we retained only the canonical form of plain words, i.e. nouns, verbs and adjectives. We also kept compound nouns recognized by using a list of the 2300 most frequent

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

Table 1. *Sample of collocations*[3]

Word1	Word2	Occurrences	Cohesion
printer	computer	13	0.227
ship	yacht	125	0.224
priest	parish priest	44	0.209
surgeon	hospital	87	0.195
policeman	burglary	41	0.190
unemployment	employment	1985	0.167
to take	root	120	0.110
collision	frank	7	0.076

compound nouns found in 11 years of the *Le Monde* newspaper. This selection induces a 63% cut.

Collocations were extracted according to the method described in Church and Hanks (1990) by moving a window on texts. Parameters were chosen to catch topical relations: the window was quite large, 20 words wide, and took into account the boundaries of texts; moreover, collocations here are indifferent to word order.

After filtering the less significant collocations (collocations with less than six occurrences, which represent two-thirds of the whole), we got a network with approximately 31,000 words and seven million collocations. As in Church and Hanks (1990), we adopted an evaluation of mutual information as a measure of the cohesion between two words. The finite size of the corpus used to constitute such a network allows us to normalize this measure according to the maximal mutual information relative to the corpus, given by:

$$I_{max} = \log_2(N^2 \cdot (S_w - 1)) \quad (1)$$

with N as the corpus size and S_w as the window size.

Table 1 gives examples of different kinds of collocations. These examples demonstrate that the network accounts for lexico-syntactic (take-root), semantic (boat-sailboat or employment-unemployment) and topical relations (police-burglary). They also show the presence of noise (collision-frank), which corresponds to a large number of collocations.

4.2 Topic segmentation of SEGCOHLEX

In accordance with work described in section 2.2.2, we hypothesize that the lexical cohesion of a text is representative of its topical consistency, and that zones with a weak lexical cohesion can be viewed as topic shifts. Hence, the segmentation algorithm of SEGCOHLEX includes two stages. First, it evaluates the cohesion of the different parts of the text to be segmented. Then it exploits the significant breaks

[3] Words of the examples are translations of the results obtained for French words.

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

in this cohesion to detect the topic changes and to create segments. We only give an overview of this method here. More details can be found in Ferret (1998).

4.2.1 Evaluating the cohesion of a text

The evaluation of the cohesion of a text relies on the following hypothesis: the greater the number of words of the window related to the same topic is, the greater the number of links they have in the collocation network is (either directly or through intermediary words) and the greater the computed cohesion value is.

More concretely, the evaluation is achieved by sliding a fixed size window over the text to be segmented and by computing, at each position of the window, the cohesion of words that are present within it using the collocation network presented above. A cohesion value is calculated for each position of the text after it has been pre-processed, as explained in section 4.1.

The first step of this computation consists of selecting words of the network that are thematically close to those in the window. We assume that this closeness is related to the number of links that exist between a word of the network and the words of the window. Then, a word of the network is selected if it is linked to at least WST words of the window.[4] Each selected word, whether coming from the network or from the window, is assigned a weight. This weight is equal to the initial weight of the word plus the contribution of the selected words it is linked to. The contribution of a word to the weight of another word is equal to its initial weight, 1 for words of the window and 0 for words of the network, modulated by the cohesion measure between these two words in the network (see Figure 2).

The last stage is computation of the cohesion value associated to the current position p in the text. This value is equal to the sum of the weights of the selected words, each of them being modulated by its significance:

$$cohesion(p) = \sum_i signif(w_i) \cdot weight(w_i) \tag{2}$$

where $weight(w_i)$ is the weight of the word w_i (belonging to the window or added from the network), calculated according to the principles described above, and $signif(w_i)$ is the significance of w_i.

The significance of a word is defined as in Kozima (1993) as its normalized information in a reference corpus:[5]

$$signif(w) = \frac{-\log_2(freq(w)/Sc)}{-\log_2(1/Sc)},\ with\ signif(w) \in [0,1] \tag{3}$$

where $freq(w)$ is the number of instances of the word w in the reference corpus and Sc is the size of the reference corpus.

[4] The Word Selection Threshold (WST) parameter is equal to 3 in our experiments.
[5] The reference corpus here is the corpus that was used for building the collocation network.

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

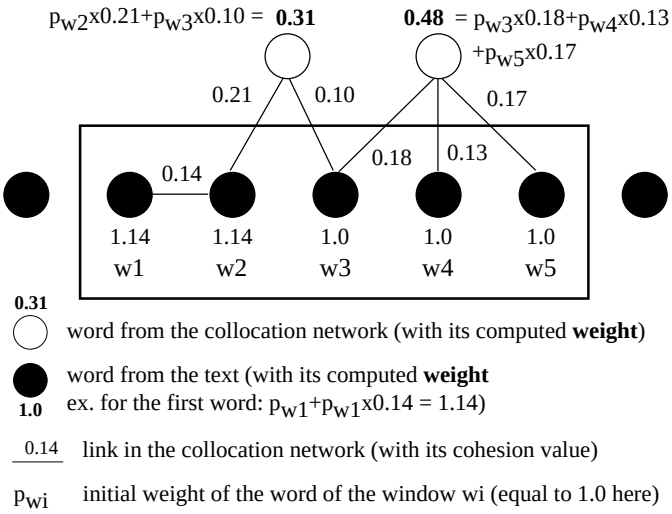


Fig. 2. Computation of the weight of selected words.

4.2.2 Segmenting the cohesion graph

To delimit segments from the cohesion curve, SEGCOHLEX first performs a smoothing of the curve to facilitate detection of maxima and minima. This is done by sliding a window on the text and re-evaluating the cohesion associated to the center of the window as the mean of all the cohesion values in the window. Considering the interpretation associated with cohesion, minima are assumed to correspond to topic shifts. They are detected by deriving the smoothed curve. A segment is then defined from each sequence of type minimum – maximum – minimum. A global cohesion value of each segment is computed by averaging the cohesion values for all the positions it includes.

4.3 Building thematic units

SEGCOHLEX delimits segments that are thematically homogeneous. These segments are the basic elements to build Thematic Units (TUs). Only the most cohesive segments (i.e. segments whose global cohesion is high enough) are turned into TUs. A TU is the representation of a topic built from a text. It represents a partial point of view on the considered topic. In order to limit noise, a TU is not made of the words of a segment but contains words that are selected from the collocation network during the delimitation of the segment. As mentioned above, we assume that these words refer in majority to the topic of the segment. We strengthen this trend by only taking the words that are selected for a significant proportion of the positions of a segment. More precisely, we keep the words from the network involved in the calculation of at least 75% of the cohesion values inside the segment

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

Table 2. *Some of the words of a segment and its TU*

Segment	Weight	TU	Weight
to dedicate	0.522	paraph	0.522
append	0.467	parisian-press	0.480
sharp-pointed	0.454	best seller	0.477
to relate	0.445	publishing house	0.450
boycotting	0.436	bookseller	0.447
bus	0.435	tome	0.445
to plunge	0.410	grasset	0.440
to surround	0.368	to republish	0.428
signature	0.366	appearance	0.427
exemplar	0.357	press	0.418
page	0.332	to publish	0.407
train	0.331	biography	0.406
hundred	0.330	bookshop	0.405
feel	0.328	pocket	0.389
book	0.289	publisher	0.363
person	0.267	reader	0.355
shrill	0.683	israeli	0.337
toward	0.683	publishing	0.333

under consideration. Within a TU, words are weighted according to their significance value.

Table 2 shows the most representative words in terms of weight of both a segment and the TU built from it. The segment is about a booksigning event and the topic represented by its TU is clearly related to the publishing world, even if some words, as "israeli", are not linked to this topic.[6]

5 Learning of semantic domains in SEGAPSITH

In SEGAPSITH, learning a complete description of a topic consists in merging all successive points of view (i.e. similar TUs) into a single memorized thematic unit, called a semantic domain. This process is progressive and unsupervised. We here only give an overview of it. More details can be found in Ferret and Grau (1998).

[6] Words such as "train", "shrill" or "toward" result from errors done by the morpho-syntactic tagger we used.

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

Table 3. *The most representative words of a domain about justice*

Words	Occurrences	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

5.1 Representation of semantic domains

A semantic domain is the result of the aggregation of several TUs. As a consequence, its structure is identical to the structure of a TU. Thus a domain is a set of weighted words. Only the weighting of words is different:

$$(4) \qquad weight(w_i, dom_j) = \frac{nbOcc(w_i, dom_j)}{agrNb(dom_j)} \cdot signif(w_i) \cdot \frac{agrNb(dom_j)^4}{(agrNb(dom_j)+1)^4}$$

where $nbOcc(w_i, dom_j)$ is the number of occurrences of the word w_i in the domain dom_j and $agrNb(dom_j)$ is the number of aggregations that produced the domain dom_j.

The first factor takes into account the importance of the word in relation to the domain, while the last factor is a modulator preventing recently-created domains from being overly favored *vis-à-vis* the oldest domains.

Table 3 shows the most representative words (i.e. words with the highest weights) of a domain about justice. Each word is characterized by its weight (see (4)) and its number of occurrences, which is also the number of TUs in which the word is present.

Considering the method for constructing TUs, the words of a domain necessarily belong to the collocation network, and thus domains constitute a structuring of this network according to the topical point of view.

5.2 Building semantic domains

After a new TU has been built, it is memorized either by aggregating it to an existing similar domain or by creating a new domain. Hence, it is first necessary to determine which domains within memory are similar to the new TU. The first stage of this operation is performed by an one-step propagation of activation, followed by the selection of the most activated domains. The activation of a domain dom_i is

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

given by the following function:

$$activation(dom_i) = \sum_j weight(w_j, dom_i) \cdot weight(w_j, TU) \tag{5}$$

where the first factor is the weight of word w_j in the domain dom_i and the second factor represents the weight of the same word, but in the TU to be memorized.

The selection of the most activated domains is then achieved by comparing their activation value with a threshold based on the distribution of all the activation values: domains are kept only if their activation value is greater than the average of the activation values plus their standard deviation.

This selection process can be seen as a first similarity measure, not very elaborate but widely applicable due to its reasonably low cost. After this first restriction, it becomes possible to apply a more complex similarity measure in order to determine if the new TU can be joined to one of the selected domains, or when the similarity is below a fixed threshold for all of them, if it is the starting point of a new domain.

This similarity measure only relies on the common words between a domain and a TU, because our learning method is a source of significant noise, even though we precisely aim at reducing it. In fact, the relations in the collocation network are not only topical ones. Since the type of collocations remains implicit, some words are kept on the basis of criteria other than topical proximity, and represent a source of noise from the viewpoint of our task.

The similarity measure between a TU and a domain combines the significance, for each of these two entities, of their common words in relation to the set of words that constitute them. This significance is itself a combination of the weight of these words and their number of occurrences. We avoid in this manner having a strong similarity between a TU and a domain that only share a small number of common words with a strong weight. Each of these two combinations is performed using a geometric average. More formally, the similarity measure is given by:

$$\begin{aligned} ratio_{\{d,tu\}} &= \sqrt{\frac{\sum_c weight(w_c, \{d, tu\})}{\sum_t weight(w_t, \{d, tu\})} \cdot \frac{\sum_c nbOcc(w_c, \{d, tu\})}{\sum_t nbOcc(w_t, \{d, tu\})}} \\ similarity(d, tu) &= \sqrt{ratio_d \cdot ratio_{tu}} \end{aligned} \tag{6}$$

where the index c refers to the words common to the TU tu and to the domain d, while the index t designates the set of words constituting the TU and the domain, respectively. Words with too weak a weight in domains (weight < 0.1) are not kept for computing the similarity, as they are assimilated to noise. The threshold above which a new domain is created was experimentally set to 0.25.

When a similar domain is found, the new TU is aggregated to it. This operation is very simple, since these two entities have the same structure. It mainly consists in merging two weighted lists of words. As the weight of a word in a domain is dynamically computed from its number of occurrences, the aggregation can be reduced to an additive operation: if a word of the TU is not present in the domain, it is added to it with an occurrence of 1; if it already exists in the domain, its number of occurrences is increased by 1.

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

This method leads SEGAPSITH to learn specific topic representations, as opposed to Lin (1997), for example, whose method builds general topic descriptions as for economy, sport, etc. We applied the learning module of SEGAPSITH on one month (May 1994) of *AFP* newswires, corresponding to 7823 TUs. The learning stage produced 1024 semantic domains. The domain shown in Table 3, which gathers 69 TUs about justice, is one of them. The topic analysis of SEGAPSITH only works with the most reliable of these domains. Thus, we selected those domains whose number of aggregations is superior to 4. Moreover, we selected in these 193 domains words whose weight is superior to 0.1, since below this limit a lot of words only represent noise.

6 The topic analysis of SEGAPSITH

In accordance with work on discourse segmentation, such as Grosz and Sidner (1986), the topic analysis module of SEGAPSITH (Ferret and Grau 2000) 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 processed. A window that delimits the current focus of the analysis is moved over the text to be analyzed. The segmentation algorithm 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. As for SEGCOHLEX, texts are first pre-processed in order to select their significant words (see section 4.1).

6.1 Topic contexts

A topic context aims at characterizing the entity to which it is associated 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 and having several domains that are close to each other enables 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 ambiguousness without having to choose explicitly one interpretation.

6.1.1 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 domains among them. The activation value of a semantic domain is given by:

$$activ(dom_i) = \sum_j weight(dom_i, w_j) \cdot nbOcc(w_j) \quad (7)$$

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

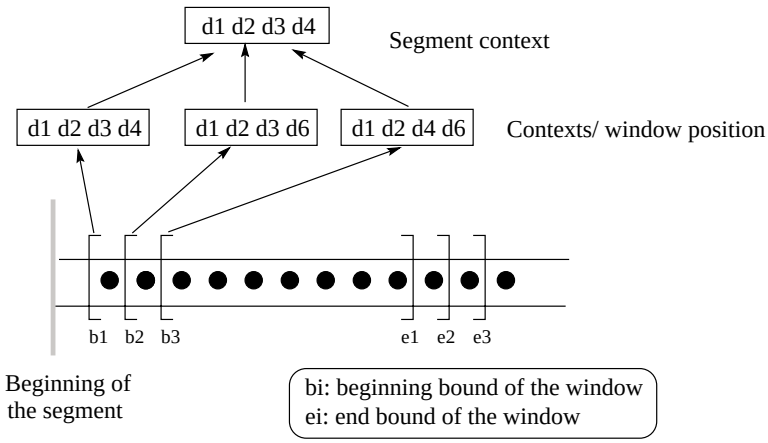


Fig. 3. Building of the context segment.

where the first factor is the weight of the word w_j in the domain dom_i (see (4)), and the second factor 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 first semantic domains according to their activation value. Their weight in the context is equal to their activation value. N is the fixed size of all the contexts.

6.1.2 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. It is built by combining the contexts associated with each position of the focus window inside the segment (see Figure 3). 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:

$$(8) \qquad wght(dom_i, Cs, t+1) = \alpha(t) \cdot wght(dom_i, Cs, t) + \beta(t) \cdot wght(dom_i, Cw, t)$$

with Cw being the context of the window, Cs the context of the segment, and $wght(dom_i, Cx, t)$ is the weight of the domain dom_i in the context Cx for the position t.

The results presented in the following 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 the thematic evolution given by the context of the window too narrowly, topic shifts cannot be detected. However, it must also adapt itself to small variations in the way in which a topic is expressed when progressing in the text so as not to introduce false topic shifts.

After weight revaluation, the joined domains are sorted in decreasing order of weight, and finally the N first of them are selected for building the new version of the segment context.

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

6.2 Evaluating the similarity of two contexts

To determine whether the content of the focus window is thematically coherent 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:

$$rankDiff(Cw, Cs) = \frac{\sum_{c=1}^{p} |rank(dom_c, Cw) - rank(dom_c, Cs)|}{(N-1) \cdot p} \quad (9)$$

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 (see (10)). 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.

$$sim(Cw, Cs) = \left(\frac{\sum_{c=1}^{p} wght(dom_c, Cw)}{\sum_{i=1}^{N} wght(dom_i, Cw)} \cdot \frac{\sum_{c=1}^{p} wght(dom_c, Cs)}{\sum_{i=1}^{N} wght(dom_i, Cs)} \right)^{1/4} \cdot \left(\frac{p}{N} \cdot (1 - rankDiff(Cw, Cs)) \right)^{1/4} \quad (10)$$

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.

6.3 Topic shift detection

The algorithm that detects topic shifts is based on the following principle: for each position of a text, if the value of the similarity measure between the topic context

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

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 principle assumes that the transition phase between two segments is punctual. The algorithm 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 does not change too much 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 (1), 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 one or two positions, but when the semantic domains that should be related to the current topic of the text do not exist in the database of domains, this number of positions may be equal to the size of a segment.

The segmentation algorithm follows the transitions of the automaton of Figure 4 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 *nonSim*;
3. the number of successive positions of the focus window for which the current state does not change: *confirmNb*, which must be above the $T_{confirm}$ threshold for going away from the states *NewTopicDetection* and *EndTopicDetection.*

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

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

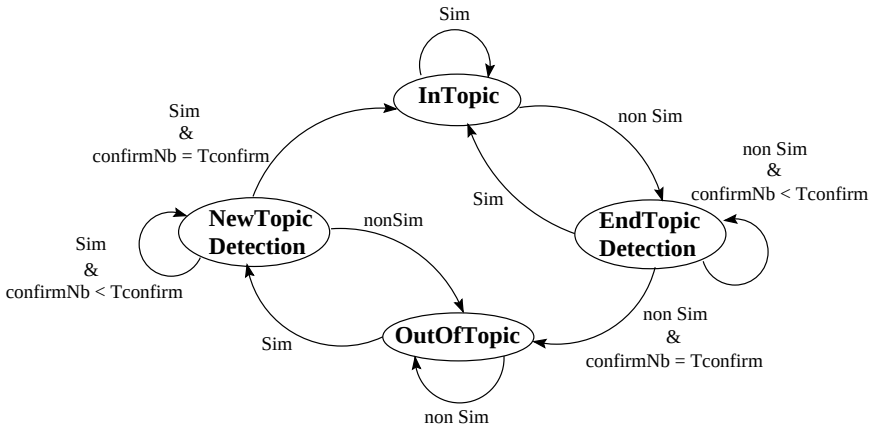


Fig. 4. Automaton for topic shift detection.

between two successive positions; 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 (10), 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 enables the system to segment texts without having all relevant topic representations.

7 Experiments

Before presenting the results of our experiments, it is important to note that the ROSA system, as most systems that implement quantitative methods, can be tuned by adjusting its parameters. As we aim at implementing a robust system rather than a high-performance system, we chose not to optimize the values of these parameters. Such an optimization would be costly and could be done only with a reference corpus. Hence, it could not systematically be done and would be in contradiction with our initial purpose.

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

7.1 Qualitative results and discussion

A first qualitative test of the topic analysis of SEGAPSITH was done with a small set of texts and without a formal protocol, as in Passonneau and Litman (1997). We relied on the set of 193 semantic domains described in section 5.2. We tested several ranges of values for the different parameters of the method and found that, for the kind of texts as that given in Figure 5a, the best results are obtained with a size of 19 words for the focus window, 10 domains for contexts (*N* parameter) and a value of three positions for the *confirmNb* parameter. Furthermore, results are rather stable around these values. Figure 5b 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 89–91 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.

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

<ST> An 18 year old Indian model, Sushmita Sen, caused a surprise on Sunday in Manila 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 audience 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 Manila 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>

Fig. 5a. AFP newswire, translated from the French (May 1994) – The <ST> tags delimit the segments resulting from a human judgment.

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

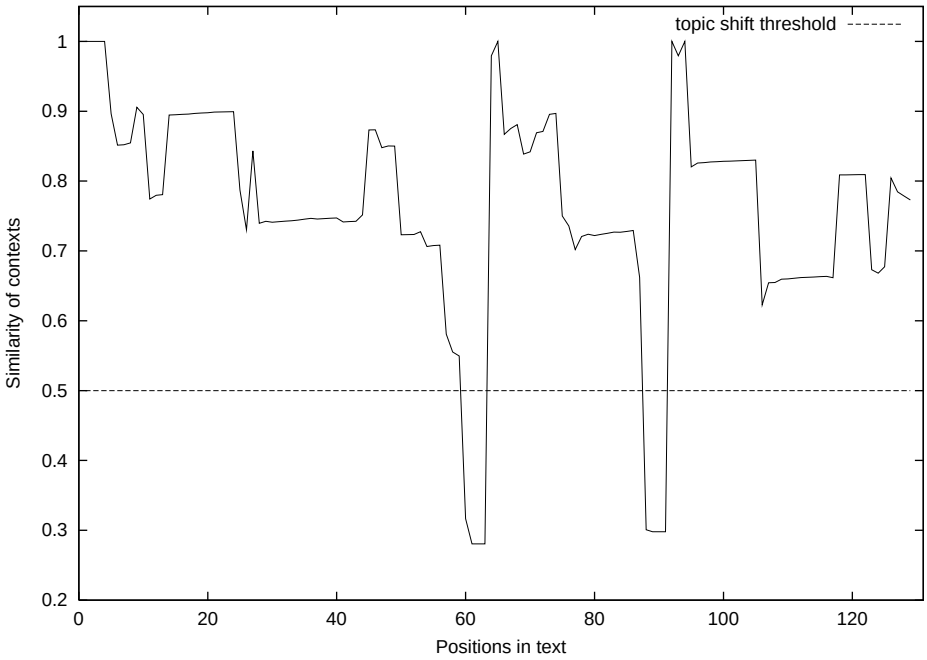


Fig. 5b. The context similarity graph (for the text of Figure 5a French text).

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 only represented here by one of its dimensions, competition, through a set of domains about sport competitions. Moreover, 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.

7.2 Quantitative evaluation

7.2.1 Evaluation of ROSA

To have a more objective evaluation, we applied the segmentation algorithm 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 was performed with short texts in French, precisely 49 texts from *Le Monde* newspaper of 133 words long on average. As in Hearst (1997), we resorted to the precision and the recall measures to characterize the ability of our systems to find text boundaries.[7] These measures are defined as follows:

(11) $$recall = \frac{Nc}{D}$$

[7] In TDT and other recent work, segmentation accuracy is measured by a probabilistic error metric proposed in Beeferman *et al.* (1999) for overcoming some limitations of the recall-precision metric. Since our experiments with ROSA, we have used this error metric on the same corpus for evaluating other segmentation methods and we have noted that results got with the error metric are strongly correlated with those got with the recall-precision metric.

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

Table 4. *Evaluation of the topic segmentation methods of ROSA*

Segmentation methods	Recall	Precision	F1-measure
Random	0.513	0.282	0.364
SEGCOHLEX	0.675	0.374	0.481
SEGAPSITH(1)	0.920	0.523	0.666
SEGAPSITH(2)	0.810	0.535	0.644

(12) $$precision = \frac{Nc}{Nb}$$

with Nb as the number of boundaries found by a segmentation system, D as the number of document breaks, and Nc as the number of boundaries that match document breaks. The match between a boundary and a document break was accepted if the boundary was not further than nine words (after pre-processing).

We classically use the f1-measure, which is the harmonic mean of precision and recall, to gather these two measures into one. Results shown in Table 4 are average values from 10 tests, with a change in the order of the texts from one test to another.

As a baseline, we implemented a segmentation procedure, called *Random* in Table 4, that randomly chooses Nb positions as document boundaries. Nb was fixed to the number of boundaries found by SEGCOHLEX, that is supposed to be the baseline for SEGAPSITH in ROSA. Results of *Random* are average values from 1000 tests.[8]

The first observation from Table 4 is that all the segmentation algorithms of ROSA obtained far better results than the random procedure. Among ROSA's methods, SEGAPSITH obtained better results than SEGCOHLEX. Using explicit topic representations actually seems to be more effective than using a knowledge source, as a collocation network, that is not structured on the topical point of view. Moreover, the comparison between SEGAPSITH(1) and SEGAPSITH(2) shows that the bootstrapping mechanism of ROSA is actually interesting. SEGAPSITH(1) is the version of the SEGAPSITH topic analysis that relies on semantic domains built from the TUs of SEGCOHLEX. SEGAPSITH(2) is the version of the SEGAPSITH topic analysis that relies on semantic domains built from the texts of the *AFP* corpus, without any processing by SEGCOHLEX. For building these domains, each text was directly turned into a TU by taking its content words. The precision of SEGAPSITH(1) is slightly lower than SEGAPSITH(2)'s precision but the difference about recall greatly corrects the global result in favor of SEGAPSITH(1).

7.2.2 Quantitative comparison with related work

To complete the evaluation of ROSA, we compared its results to those of a well-known segmentation method that is generally used as a reference by other work

[8] In practice, results quickly became stable and there was no significant difference between the results from 100 and those from 1000 tests.

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

Table 5. *Results of* TextTiling *on our test set*

Segmentation methods	Recall	Precision	F1-measure
TextTiling(1)	0.718	0.805	0.758
TextTiling(2)	0.806	0.847	0.826

in the field: *TextTiling* (Hearst 1997). We reimplemented *TextTiling* with the pre-processing of texts applied in ROSA and evaluated it with the test set presented in section 7.2.1. Two versions of *TextTiling*, that differed about their parameters, were tested. The first, TextTiling(1), corresponds to the parameters given in Hearst (1997): 20 words by pseudo-sentence and six pseudo-sentences by block for the main parameters. The second one, TextTiling(2), results from the optimization of these parameters in relation to our test set: 10 words by pseudo-sentence and 10 pseudo-sentences by block. Results are shown in Table 5. For information, Hearst (1997) 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 0.73 as the f-measure.

First, Table 5 shows that the results of *TextTiling* on our test set are better than those given in Hearst (1997), while *a priori*, *TextTiling* seems to be suited to long texts. More precisely, the recall measure is lower but recall and precision are closer to each other, which is better. The lower precision in Hearst (1997) may be explained by the fact that long texts are more likely to be split than short texts, whereas the evaluation only takes into account document boundaries.

Table 5 also shows that the results of *TextTiling* are globally better than those of ROSA. In fact, it is not very surprising, since our evaluation corpus is very favorable to *TextTiling*: it is made of a sequence of texts such as the topics of two adjacent texts are very different. Methods that rely on word reiteration, such as *TextTiling*, have an advantage over methods that use knowledge, when topic shifts are so clear, since the last ones are more suited to find links between parts of text.

The results of SEGAPSITH(1) have the same characteristics as those of *TextTiling* in Hearst (1997), which according to the interpretation presented above, means that SEGAPSITH tends to delimit smaller segments than *TextTiling*. Hence, it also tends to produce more segments and have a lower precision. This interpretation is confirmed by the analysis of the short text of Figure 5a. SEGAPSITH is able to detect the two most important topic shifts (see section 7.1), while TextTiling(1) and TextTiling(2) find only one of them. TextTiling(1) detects this shift at the beginning of the second segment (at the word "carried"), and TextTiling(2) is more precise and locates this shift at the boundary between the two first segments (at the word "Switzerland"). In the case of such a text, *TextTiling* can detect that a significant topic shift occurs and, depending upon its parameters, it can locate its position more or less precisely. However, it is not able to follow all the topic changes

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

(especially the return to the first topic that ends the second segment) because the size of some segments is too low in relation to the resolving power of this kind of methods.[9]

The comparison between SEGAPSITH(1) and TextTiling(2) is not itself very interesting, since the cost of the optimization work done for getting TextTiling(2) is too high for doing it for each text to analyze. Nevertheless, this work has shown that the results of a method such as *TextTiling* are quite sensitive to the value of its parameters: a small difference in one of these values can strongly debase its results. From that viewpoint, the results of SEGAPSITH are more stable, which is not very surprising: the knowledge of a system gives to it a force of inertia. As a consequence, SEGAPSITH is less sensitive than *TextTiling* to the particularities of each text, which is also a form of robustness.

The comparison we have done between *TextTiling* and SEGAPSITH globally shows that the two methods do not have the same characteristics and are complementary. In Ferret *et al.* (1998), we have already shown that designing an all-purpose method is certainly not the best solution. The problem rather consists of choosing the most suitable method according to the type of the text to analyze. From that point of view, SEGAPSITH is a tool that can operate at a fine-grained level and when the variability of the vocabulary is rather high. Moreover, SEGAPSITH can identify the topic of the segments it delimits and the links between them, which is not done by a method such as *TextTiling*.

Among the systems that achieve both segmentation and identification, that of Bigi *et al.* (1998) is most similar to ours, although its learning of topic representation is supervised, and its evaluation method is slightly different from ours. In this case, the comparison is only based on the results presented in Bigi *et al.* (1998). Its test corpus is different from ours, even if it is also extracted from the *Le Monde* newspaper. The differences between its results – in the best case, 0.75 as precision, 0.80 as recall and 0.77 as f1-measure – can be explained by the nature of topics: Bigi *et al.* (1998) focus on a small set of very general topics (such as business, politics), while we focus on a large set of specific topics. Since our topic representations are closer to the topics of texts, our recall is higher. However, this closeness also makes SEGAPSITH more sensitive to the local topical variations, and leads it to delimit more segments. As all these boundaries are not document boundaries – they represent noise according to the kind of evaluation we consider (see section 7.2.1). As a consequence, the precision of SEGAPSITH is lower than the precision of Bigi *et al.* (1998).

7.3 Comparison with related work

A topic analysis that achieves topic identification generally relies on topic representations, such as Bigi *et al.* (1998) or as the work done in the TDT framework (Fiscus, Doddington, Garofolo and Martin (1999)). SEGAPSITH also exploits topic representations, but with an approach that is similar to work using lexical cohesion (as

[9] SEGAPSITH(1) may also fail when a segment is very short (see last segment of the text of Figure 5a), but it has a larger resolving power than *TextTiling*.

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

Hearst (1997) or Kozima (1993)), and not with the probabilistic approach generally found in TDT or in Beeferman *et al.* (1999), for instance. 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 10,000 words while this parameter is equal to only three content words in our method. Moreover, the purpose of the Segmentation task of TDT is to segment a stream of text into documents,[10] and not to segment a document according to its sub-topics.

For achieving such a segmentation, several types of language models are often combined: some of them are topic representations, but others are models for specifically detecting topic shifts. These latter ones are often over-specialized in relation to the corpus that was used to build them. Beeferman *et al.* (1999) mention that the most effective cues for segmenting a set of articles from the *Wall Street Journal* are very specific to this newspaper: the word "incorporated", for instance, often appears at the beginning of the articles because these articles often refer to companies and the complete name of a company, which includes the word "incorporated", is only given once, at the beginning of the articles.

This kind of over-specialization is not favorable to robustness. Therefore, although SEGAPSITH depends upon knowledge built from corpora, it aims at being as general as possible. The fact that the words of semantic domains come from a collocation network and not directly from texts is an example of this concern. It is a means to reduce the dependency upon the particularities of texts. The same viewpoint was adopted concerning the specificity of the topics handled by SEGAPSITH. In TDT, topics are very specific and often comparable to events. On the contrary, they are very general in Bigi *et al.* (1998). Domains in SEGAPSITH are halfway between these two extremes: they aim at describing specific topics, but not events.

Having topic representations clearly enables SEGAPSITH to work at a finer grain than methods based on lexical cohesion. However, on a large scale, it also requires us 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 task. On the contrary, SEGAPSITH includes a module that learns the topic representations that support its segmentation module in an unsupervised way. Moreover, its association with SEGCOHLEX enables ROSA to progressively go from a segmentation module based on lexical cohesion to a segmentation module based on topic representations.

8 Conclusion

Developing robust and deep NLP processes requires large bases of structured knowledge that are very difficult to build. To overcome this problem for topic analysis, we adopted a bootstrapping approach: we implemented a first topic analysis, based on automatically acquired knowledge, used its results to build topic representations, and then developed another analysis that exploits these representations in order to

[10] Documents are called stories according to the TDT terminology. The stream of text is made up of newspaper articles but also of broadcast news transcripts.

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

get better performances. The evaluation of this method shows that results obtained with structured and specialized knowledge are better than results obtained with general knowledge and moreover, that learning this knowledge from the results of text segmentation is more reliable than learning it from non processed texts.

The work we have presented points out the soundness of our approach, but it still may be improved. We consider three kinds of extensions. First, the evaluation of ROSA must be extended along two axes. The first one consists of evaluating the ability of a topic segmenter to find not only document boundaries, but also topic shifts in documents. It requires building a reference corpus with topic shifts located by a set of human annotators, as it was done in Hearst (1997), Passonneau and Litman (1997) or Kan *et al.* (1998). The second axis is concerned with the design of an evaluation framework that integrates topic segmentation, topic detection and link detection in an unified way, which have not existed to-date, as far as we are aware.[11]

The second extension to our present work aims at testing the topic analysis of SEGAPSITH on a large scale, especially to use it for applications in information retrieval or information extraction, as question answering systems, for instance. It requires building a large base of semantic domains. This base need not contain all the possible topics, but it must have a wide enough topic extent to make it possible to represent each new topic as a set of the topics already represented in the base. This approach relies on the ability we have seen in section 7.1 to characterize a topic as a composition of more elementary topics.

The last extension to our work is linked to the previous one. For exploiting a large base of semantic domains, this base must have a more elaborated structure than only a set. More precisely, we consider structuring domains hierarchically, which implies to modify the learning procedure of semantic domains. Furthermore, such a hierarchy will make it possible to distinguish different levels of topic in texts, and thus to go further in the topic structuring of texts.

References

Beeferman, D., Berger, A. and Lafferty, J. (1999) Statistical models for text segmentation. *Machine Learning* **34**(1/3): 177–210.

Bigi, B., de Mori, R., El-Bèze, M. and Spriet, T. (1998) Detecting topic shifts using a cache memory. *Proceedings 5th International Conference on Spoken Language Processing*, pp. 2331–2334. Sydney, Australia.

Choi, F., Wiemer-Hastings, P. and Moore, J. (2001) Latent semantic analysis for text segmentation. *Proceedings Conference on Empirical Methods in Natural Language Processing*, pp. 109–117. Pittsburgh, PA.

Choi, F. (2000) Advances in domain independent linear text segmentation. *Proceedings NAACL'00*. Seattle, WA.

Church, K. W. and Hanks, P. (1990) Word association norms, mutual information, and lexicography. *Computational Linguistics* **16**(1): 22–29.

[11] TDT proposes the same kind of metrics for the three tasks but does not propose a metric for gathering them.

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer T. K. and Harshman R. (1990) Indexing by latent semantic analysis. *J. Am. Soc. Infor. Sci.* **41**: 391–407.

Ferret, O. (1998) How to thematically segment texts by using lexical cohesion? *Proceedings ACL-COLING'98 (Student Session)*, pp. 1481–1483. Montréal, Canada.

Ferret, O. and Grau, B. (2000) A topic segmentation of texts based on semantic domains. *Proceedings ECAI 2000*, pp. 426–430. Berlin, Germany.

Ferret, O. and Grau, B. (1998) A thematic segmentation procedure for extracting semantic domains from texts. *Proceedings ECAI'98*, pp. 155–159. Brighton, UK.

Ferret, O., Grau, B. and Masson, N. (1998) Thematic segmentation of texts: two methods for two kinds of texts. *Proceedings ACL-COLING'98*, pp. 392–396. Montréal, Canada.

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

Grau, B. (1984) Stalking coherence in the topical jungle. *Proceedings 5th Generation Computer*, pp. 652–659. Tokyo, Japan.

Grosz, B. J. and Sidner, C. L. (1986) Attention, intentions and the structure of discourse. *Computational Linguistics* **12**: 175–204.

Halliday, M. A. K. and Hasan, R. (1976) *Cohesion in English*. Longman, London.

Hahn, U. (1990) Topic parsing: accounting for text macro structures in full-text analysis. *Infor. Process. & Manage.* **26**(1): 135–170.

Hearst, M. A. (1997) TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* **23**(1): 33–64.

Hovy, E. H., Gerber, L., Hermjakob, U., Junk, M. and Lin, C.-Y. (2000) Question answering in Webclopedia. *Proceedings 9th Text Retrieval Conference (TREC9)*, pp. 655–664. Gaithersburg, MD.

Kan, M.-Y., Klavans, J. L. and McKeown, K. R. (1998) Linear segmentation and segment significance. *Proceedings 6th International Workshop of Very Large Corpora (WVLC-6)*, pp. 197–205. Montréal, Canada.

Kaufmann, S. (1999) Cohesion and collocation: Using context vector in text segmentation. *Proceedings 37th Annual Meeting of the ACL (Student Session)*, pp. 591–595. College Park, OH.

Kozima, H. (1993) Text segmentation based on similarity between words. *Proceedings 31st Annual Meeting of the ACL (Student Session)*, pp. 286–288. Colombus, OH.

Lau R. (1994) *Adaptive Statistical Language Modelling*. Doctoral Dissertation, Carnegie Mellon University.

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

Masson, N. (1995) An automatic method for document structuring. *Proceedings 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 372–373. Seattle, WA.

Morris, J. and Hirst, G. (1991) Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics* **17**(1): 21–48.

Nomoto, T. and Nitta, Y. (1994) A grammatico-statistical approach to discourse partitioning. *Proceedings 15th International Conference on Computational Linguistics (COLING)*, pp. 1145–1150. Kyoto, Japan.

Passonneau, R. J. and Litman, D. J. (1997) Discourse segmentation by human and automated means. *Computational Linguistics* **23**(1): 103–139.

Reynar, J. C. (1994) An automatic method of finding topic boundaries. *Proceedings 32nd Annual Meeting of the ACL (Student Session)*, pp. 331–333. Las Cruces, NM.

Salton, G., Singhal, A., Buckley, C. and Mitra, M. (1996) Automatic text decomposition using text segments and text themes. *Proceedings Hypertext'96*, pp. 53–65. Washington, DC.

Schank, R. C. (1982) *Dynamic Memory: A theory of reminding and learning in computers and people*. Cambridge University Press, New York.

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press

Utiyama, M. and Isahara, H. (2001) A statistical model for domain-independent text segmentation. *Proceedings 39th Annual Meeting of the ACL*, pp. 491–498. Toulouse, France.
Voorhees, E. M. (2000) The TREC-8 question answering task report. *Proceedings 8th Text Retrieval Conference (TREC8)*, pp. 77–82. Gaithersburg, MD.
Yaari, Y. (1997) Segmentation of expository texts by hierarchical agglomerative clustering. *Proceedings 2nd Conference on Recent Advances in Natural Language Processing (RANLP'97)*, pp. 59–65. Tzigov Chark, Bulgaria.
Youmans, G. (1991) A new tool for discourse analysis: the vocabulary-management profile. *Language* **67**(4): 763–789.

https://doi.org/10.1017/S1351324902002929 Published online by Cambridge University Press