A Thematic Segmentation Procedure for Extracting Semantic Domains from Texts

Olivier Ferret, Brigitte Grau

A Thematic Segmentation Procedure for Extracting Semantic Domains from Texts

Olivier Ferret[1] and Brigitte Grau[1]

Abstract. Thematic analysis is essential for a lot of Natural Language Processing (NLP) applications, such as text summarization or information extraction. It is a two-dimensional process which has both to identify the thematic segments of a text and to recognize the semantic domain concerned by each of them. This second task requires having a representation of these domains. Such representations are built in Information Retrieval or Text Categorization fields by grouping together the words of a set of texts which have been manually linked to the same domain. We claim that this kind of method can only be apply to characterize very general topics. We propose here a method for building the representation of narrower semantic domains without any manual intervention. First, we present a procedure for the thematic segmentation of texts which relies on lexical cohesion evaluated from a collocation network. This procedure allows us to have basic units that are more thematically coherent than a whole text. Then, we show how these units can be aggregated together, according to a similarity measure, to build the representation of semantic domains in an incremental and unsupervised way.[1]

1. INTRODUCTION

Thematic analyzis is a necessary process when analyzing discourse. Different tasks as story understanding, text summarization or text classification for example requires to recognize discourse topics, as a step towards their achievement, or as their final purpose. Thematic analyzis involves decomposing a text in parts relative to a same topic and to identify these topics.

Approaches to solve this problem can roughly be categorized in two classes, knowledge-based approaches or word-based approaches. Knowledge-based systems [Grau, 1984 #42; Grosz, 1986 #52] lead to a precise decomposition and identification of the discourse topics, by using in-depth understanding processes. They required an extensive manual knowledge engineering effort to create the knowledge base, represented by semantic network and/or frames, and this is only possible in very limited and well-known domains.

To overcome this limitation, and be able to treat a large amount of texts, word-based approaches have been developed [Kozima, 1993 #72; Morris, 1991 #73; Ferret, 1997 #121]. The purpose of these systems is to segment texts, but not to recognize topics in terms of associating discourse segments to classes. Such a problem is close to text categorization, where a system must find the appropriate domain of a text. Domain descriptions are either hand-coded or computed by the systems [Lehnert, 1994 #164], but it always requires an a priori classification of the texts constituting the training corpus. Furthermore, these systems consider texts as a whole, and do not proceed to a finer analyzis for identifying different topics inside the texts.

At the opposite, automated summarization requires to segment texts and to recognize their related topics. For this latter task, Hovy and Lin [Hovy, 1997 #136] are developing automatic construction of text signatures, based on a classification of newspaper texts in very general domains. But a precise analyzis of text topic requires more specialized classes difficult to establish by advance. The definition of such classes result from a study of a lot of text contents and this task is too tedious and time consuming to be done by human beings.

So, even if systems perform robust text analyzis and can be applied to a wide range of texts, their performance always depends on human interventions to define domain representations or at least to classify texts. In order to go towards a finer analyzis of texts without restriction about domains, the system we present here aims at learning specialized semantic domains. It automatically segments texts, and topic representations, described by weighted words, are incrementally built from these discourse segments. It works without any a priori classification or hand-coded pieces of knowledge. Our approach merges statistical techniques and the use of knowledge about word proximity that are also learned from a corpus.

2. OVERVIEW OF THE SYSTEM

Studied texts are newspaper articles coming from two corpora: "Le Monde" and "AFP". They are pre-processed to only keep lemmatized content words (adjective, single or compound nouns and verbs). A part of these texts has been used to build a lexical network where links between two words represent an evaluation of their mutual information to capture semantic and pragmatic relations between them, computed from their co-occurrence number.

Text segmentation is based on the use of this network. A discourse segment is a part of text whose words refer to the same topic. A topic is detected by computing a cohesion value for each word resulting from the relations found in the network between these words and their neighbours in a text. These values lead to build a graph and by successive transformations applied to it, texts are automatically divided in discourse segments. Only highly cohesive segments are kept to learn topic representations.

Discourse segments in texts, even related to the same topic, often develop different points of view. To incrementally learn a complete description of a topic, all successive points of view have to be merged in a single memorized thematic unit. As each segment contains from twenty to forty words, which is low according to the number of words belonging to a same domain, we have to face here two problems : a) recognizing the similarity of two units, even described by few identical words in the original texts and b) building thematic unit complete enough without having to process too numerous texts. A solution is to go beyond the words really used in the processed texts by inferring missing information. So, correlated words coming from the network are selected and added to the representation of a discourse segment; this leads to the formation of a discourse unit described by weighted words. Weights represent the importance of each word relative to the topic, and result from the number of occurrences of a word in the segment and the weights found in the network.

The memorization process then selects memorized units. If one is sufficiently close to the current discourse unit, topic descriptions are aggregated, otherwise a new unit is created. Each aggregation leads to augment the system's knowledge about one topic by reinforcing

[1] LIMSI-CNRS. BP 133 - 91403 Orsay Cedex. France

recurrent words and adding new ones. An aggregated thematic unit is then also represented by weighted words. Similarity is based on the number of common words and their weights. The retrieval process is akin to a propagation in one step, departing from the words belonging to a discourse unit towards the memorized topics.

The whole process has been applied to 5949 texts and we detail in this paper each of its components and an evaluation of the results.

3. THE THEMATIC SEGMENTATION

3.1. Preprocessing of the texts

As we are interested in the thematic dimension of the texts, texts have to be represented by their significant features from that point of view. So, we only hold for each text the lemmatized form of its nouns, verbs and adjectives. This has been done by combining existing tools. MtSeg from the Multext project is used for segmenting the raw texts. As compound nouns are less polysemous than single ones, we have added to MtSeg the ability of identifying 2300 compound nouns. We have retained the most frequent compound nouns in 11 years of the *Le Monde* newspaper. They have been collected with the INTEX tool [Silberztein, 1994 #165]. The part of speech tagger TreeTagger [Schmid, 1994 #163] ,as for it, is applied to disambiguate the category of the words and to provide their lemmatized form. The selection of the meaningful words, which do not include proper nouns and abbreviations, ends the pre-processing. This one is applied to the texts both for building the collocation network and for their thematic segmentation.

3.2. Building the collocation network

Our segmentation mechanism relies on lexical cohesion. In order to evaluate it, we have built a network of lexical collocations from a large corpus. Our corpus, whose size is around 39 million words, is made up of 24 months of the *Le Monde* newspaper taken from 1990 to 1994. The collocations have been calculated according to the method described in [Church, 1990 #67] by moving a window on the texts. The corpus was pre-processed as described above, what induces a 63% cut. The window in which the collocations have been collected is 20 words wide and takes into account the boundaries of the texts. Moreover, the collocations here are indifferent to order.

These three choices are motivated by our task point of view. We are interested in finding if two words belong to the same thematic domain. As a topic can be develop in a large textual unit, it requires a quite large window to detect these thematic relations. But the process must avoid jumping across the texts boundaries as two adjacent texts from the corpus are rarely related to a same domain. Lastly, the collocation w1-w2 is equivalent to the collocation w2-w1 if we only try to characterize a thematic relation between w1 and w2.

After filtering of the non-significant collocations (collocations with less than 6 occurrences, which represent 2/3 of the whole), we obtain a network with approximately 31000 words and 14 million relations. The cohesion between two words is measured as in [Church, 1990 #67] by an estimation of the mutual information based on their collocations frequency. This value is normalized by the maximal mutual information with regard to the corpus, which is given by :

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

N : corpus size
S_W : window size

3.3. The segmentation algorithm

The segmentation algorithm we propose includes two steps. First, the evaluation of the cohesion of the different parts of a text, and second, the location of the major breaks in this cohesion to detect the thematic shifts and to select the most coherent segments useful for domain learning.

3.3.1. Evaluating the cohesion of a text

The method for evaluating textual cohesion is close to Kozima's work about this matter [Kozima, 1993 #72]. A cohesion value is computed at each position of a window in a text from words in this window. The collocation network is used for determining how close together the words in the window are. We suppose that if these words are strongly connected in the network, it means that they belong to the same domain and so, that the cohesion in this part of text is high. On the other hand, if they are not very much linked together, we can infer that the words of the window belong at least to two different domains. It means that the window is located across the transition from one theme to another.

In practice, the cohesion inside the window is evaluated by the sum of the weights of the words in this window and the words selected from the collocation network as well. For each word in the window, the system collects the words in the network linked to it and it only selects those that are common to at least one other word of the window. Thus, it makes words related to the same topic as the one referred by the words in the window explicit and produces a more stable description of this topic when the window moves.

Words are weighted by combining the cohesion values of the words they are linked to and their initial weight, equal to their number of occurrences in the window, as shown in figure 1. The more the words belong to a same topic, more they are linked together and the higher their weights are.

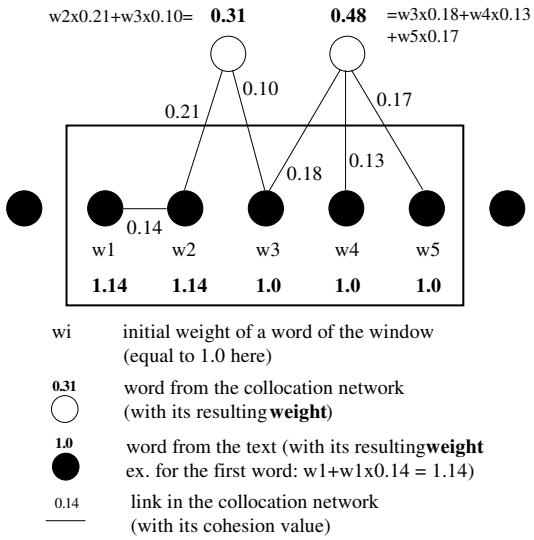

Figure 1 Computation of word weight

Finally, the value of the cohesion for one position of the window is the result of the following weighted sum:

$$coh(p) = \sum_i sign(w_i) \cdot wgth(w_i)$$

where

wgth(w_i) is the weight of the word w_i computed as described above,

sign(w_i) is the significance of the word w_i.

The significance of a word is defined as in [Kozima, 1993 #72] as its normalized information in a corpus (here, the corpus used for building the collocation network):

$$sign(w) = \frac{-\log_2(f_w / S_c)}{-\log_2(1/S_c)}, \; with \; sign(w) \in [0,1]$$

where

f_W is the number of instances of the word w in the corpus,

S_C is the number of words in the corpus.

The figure 2 shows the result of the cohesion computation for the text below with a window of 19 words wide.

A few years ago, I was in a department store in Harlem with a few hundred people around me. I was signing copies of my book "Stride toward Freedom" which relates the boycott of buses in Montgomery in 1955-56. Suddenly, while I was appending ma signature to a page, I felt a **pointed** thing sinking brutally into my chest. I had just been stabbed with a paper knife by a woman who was acknowledged as mad afterwards. I was taken immediately to the Harlem Hospital where I stayed on a bed during long hours while many preparations were made in order to remove the weapon from my body. Far later, when I was in condition to **converse** with Dr Aubrey Maynar, the chief surgeon who had carried out this delicate and dangerous operation, I learnt of the reasons of this long wait before the operation. The blade of the instrument had touched the aorta and, for extracting it, it was necessary to open up all the rib cage.

from Révolution Non-Violente by Martin Luther King (based on a French version of the original text)

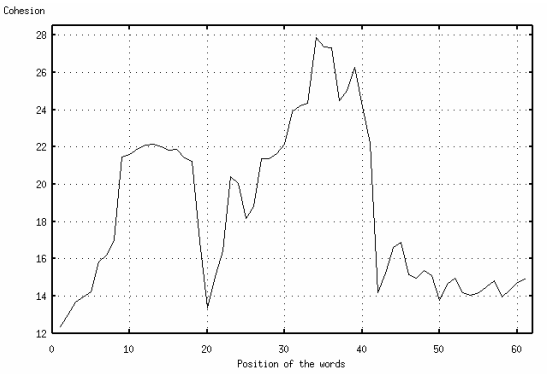

Figure 2 A text and its cohesion graph (computed for the French text)

A manual analyzis of the graph shows three zones. A first thematic segment goes from the beginning of the text to approximately the word 'pointed'. It is related to the dedicate situation mentioned by the text. The second segment, which stops around the word 'converse' is about the murder attempt and its direct consequences. The last segment, which is only about a detail of the second situation, can not be considered as interesting for us because of its too low cohesion. In fact, this means that this detail is too specific in relation to the thematic knowledge implicitly held by the collocation network.

3.3.2. Detecting the thematic shifts and selecting the most coherent segments

Our method for building thematic segments from the cohesion graph is simple. First, the graph is smoothed to more easily detect the main minima and maxima. This operation is done again by moving a window on the text. At each position, the cohesion associated to the center of the window is re-evaluated as the mean of all the cohesion values in the window.

After this smoothing, the derivative of the graph is calculated to locate the maxima and the minima. We consider that a minimum marks a thematic shift. So, a segment is characterized by the following sequence: minimum - maximum - minimum. The final step is to transform the graph so that each segment is represented by a plateau with its cohesion value equal to the value of the maximum between the two minima that surround it. The figure 3 shows the results for the above text. In order to make the delimitation of the segments more acute, a segment can be stopped before the next (or the previous) minimum if there is a brutal break of the graph and after this, a very slow descent. This is done by detecting that the cohesion values fall under a given percentage of the maximum value.

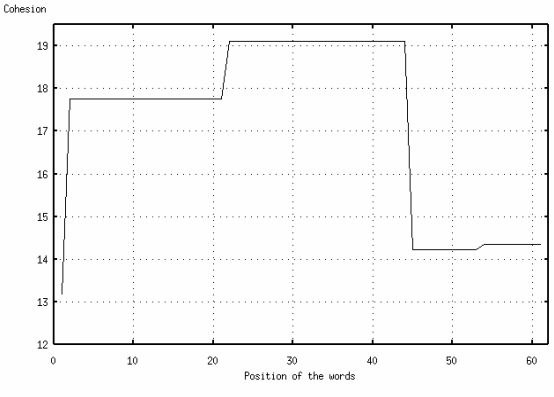

Figure 3 The segmented graph

On the figure 3, we see the same three parts we had previously found by a manual analyzis of the initial cohesion graph. For eliminating segments such as the third one here, which are not coherent enough to be considered in the learning stage, we have defined an adaptive threshold: the cohesion of a segment must not be under the mean of the cohesion values of all the segments minus their standard deviation.

4. EMERGENCE OF SEMANTIC DOMAINS

4.1. Thematic Unit and domain representation

The segmentation process produces a set of segments, that are lists of words. In the purpose of learning semantic domain representations, each segment leads to a description of the topic it refers to, named Thematic Unit (TU). Each TU is then memorized, according to the existing domains previously created. If it is similar to an existing domain, then an aggregation occurs between them, otherwise a new domain appears.

A TU is represented by a set of words, weighted by the following product:

$$wght(wi) = \sqrt{nbOcc(wi)} \cdot sign(wi)$$

where nbOcc(w_i) is the number of occurrences of the word w_i in the segment.

The square root of the number of occurrences is taken because, without this kind of modulation, the importance of this factor is too high in the similarity measure in comparison with its true meaning. Moreover, a TU does not only contain the words coming from the text but also some of the words from the collocation network that

have been used to compute the cohesion during the segmentation process. For every position in the text, all the words selected from the collocation network are retained but, for building a TU, only the words common to at least 75% of all the positions within the segment are hold. These words, also called inferred words, maintain their special status in a TU, but a weight is assigned to them similarly. In computing the similarity between a TU and a domain or in selecting domains from a TU, they are useful to bypass the fact that a theme can be expressed by different sets of words.

A domain is the result of the aggregation of several TUs. So, it is also called an aggregated TU. Its structure is exactly the same as the TU structure. The only difference between them lies in the weight of their words. This one for an aggregated TU is given by:

$$wght(w_i) = \frac{nbOcc(w_i)}{nbAgr} \cdot sign(w_i) \cdot \frac{nbAgr^4}{(nbAgr+1)^4}$$

where nbOcc(w_i) is the number of occurrences of the word w_i in the TU and nbAgr, the number of the aggregations that have produced the TU.

The first factor takes into account the importance of the word in the TU. The last one is a modulator that prevents from favouring too much the new aggregated TUs in similarity or in selection.

4.2. Selecting domains

After a new TU has been built, the system has to search which already existing domain the TU can complete. As the memory may contain a very large number of domains, it has first to efficiently select those for which a more comprehensive similarity will be evaluated. Our selection method is equivalent to a one step activation propagation. The activation of an aggregated TU which has at least one word in common with the new TU is given by:

$$activ(agrTU_i) = \sum_j wght(agrTU_i, w_j) \cdot wght(TU, w_j)$$

where the first factor is the weight of the word w_j in the aggregated TU and the second one is the weight of the same word in the new TU.

Because of their status, the influence of the inferred words is voluntarily reduced to the half of the influence of the other words. Moreover, the words which have too low a weight (under 0.1) are not used for activation because they are supposed to represent only noise. After this activation step, the selected aggregated TUs are those whose activation is greater than the average of all the activation values plus their standard deviation.

4.3. Similarity and aggregation

Once the domains have been selected, they can be compared to the new TU. In order to do that, a similarity measure is applied between the new TU and each of the selected aggregated TUs. If one of these similarity values is greater than a given threshold, the new TU is aggregated to the domain which is the most similar to it. Otherwise, a new domain is created.

The similarity measure takes into account only the words that are shared by the TU and the domain. It does not look at the differences between them because our learning method intrinsically generates a lot of noise. A TU contains many words that are not peculiar to its theme and a domain results from the aggregation of a lot of TUs. Only the recurrence of a word through these aggregations shows its importance for characterizing the domain. More precisely, the similarity measure is a combination of the importance of those common words in the new TU and in the aggregated TU. Unlike the activation process, the evaluation of these ratios does not only rely on the weights of the words. It also makes use of the number of common words between the new TU and the domain. This ensures that a high similarity is not found with only two or three words in common having a very high weight in comparison to the others. Actually, the similarity is given by:

$$ratio_{agrTU} = \sqrt{\frac{\sum_c wght(w_{agrTU}(c))}{\sum_t wgth(w_{agrTU}(t))} \cdot \frac{\sum_c nbOcc(w_{agrTU}(c))}{\sum_t nbOcc(w_{agrTU}(t))}}$$

$$ratio_{TU} = \sqrt{\frac{\sum_c wght(w_{TU}(c))}{\sum_t wgth(w_{TU}(t))} \cdot \frac{\sum_c nbOcc(w_{TU}(c))}{\sum_t nbOcc(w_{TU}(t))}}$$

$$sim(TU, agrTU) = \sqrt{ratio_{TU} \cdot ratio_{agrTU}}$$

where the *c* index is used for indicating common words between the TU and the aggregated TU and the *t* index, for indicating all the words of the TU or the aggregated TU.

The square roots aim at offsetting the effect of the products not to have too small values. We apply the same principles as for the activation process concerning the inferred words and the words with a low weight. The value of the threshold under which a new domain is created is 0.25.

The aggregation of a TU to a domain is a very simple operation since both have the same structure. It mainly consists in merging two lists of weighted words. As the weight of a word is dynamically evaluated from a number of occurrences, the aggregation can be viewed as an addition. If a word of the TU does not already exist in the domain, it is added to it with its modulated number of occurrences in the TU. If it already exists, its number of occurrences is added to the domain's one. This is done separately for the words from the texts and for the inferred words.

5. EXPERIMENTS

5.1. Experiments with a large set of texts

The validation of our method has been done by processing one month (may 1994) of AFP news wires, that is to say a set of 5949 texts that are 190 words long on average. Following the segmentation stage, 8601 TUs have been built. As for the example in the figure 2, we have used a window of 19 words wide. Experiments with a small set of texts has showed that the segmentation is quite the same for windows from 11 to 19 words wide. The learning stage has produced 3240 aggregated TUs. 691 are the result of at least two aggregations. The more aggregated TU gathers 413 TUs but most of the significant aggregated TUs results from 10 to 100 aggregations. Table 1 gives an example of one of these aggregated TUs, which gathers 61 text segments about terrorism.

Table 1 The most representative words of an aggregated TU about terrorism

text words	inferred words
attack (0.435)	trapped car (0.551)
bomb (0.244)	bomb attack (0.441)
police (0.226)	security forces (0.416)
explosion (0.222)	grenade (0.407)
to claim responsibility (0.209)	curfew (0.364)

We have observed that such domains seem to be reliable as their most representative words are coherent from a thematic point of view and stable after some aggregations. More precisely, the domains become stabilized after about twenty aggregations. This is illustrated by the figure 4 which shows the change rate of the first 30 words of the aggregated TU in table 1. This rate takes into account at the same time the words, their weight and their rank.

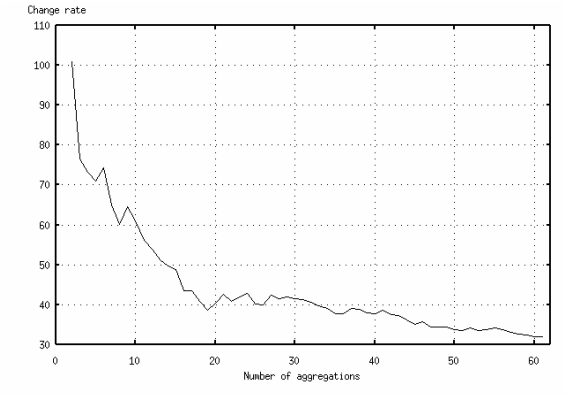

Figure 4 The evolution of the head of the aggregated TU in table 1

If we look more closely at the evolution of individual words in a domain, we recognize the same trends as for the whole domain. We can see on the figure 5 that during a first stage, the evolution is rather erratic and fast. After about twenty aggregations, it is more steady. Moreover, the significance of meaningful words such as 'kill' or 'attack' increases while more general words such as 'night' or 'Sunday' that have no special role in the domain become less and less important.

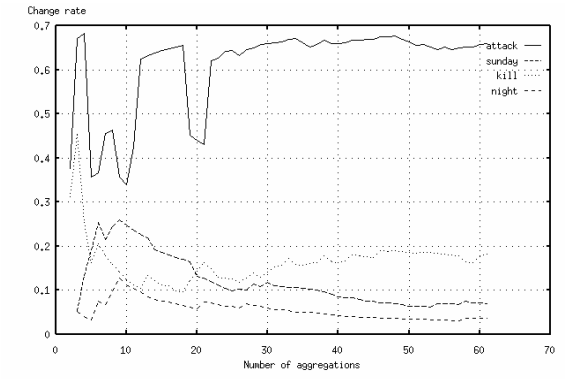

Figure 5 The evolution of some words of the aggregated TU in table 1

We have observed, as it was expected, that the aggregated TUs contain a lot of noise. The aggregated TU in table 1 for example holds 705 text words and 401 inferred words[2] but only 12 text words and 66 inferred words have a weight high enough to be used for domain activation or similarity computing.

5.2. Evaluation and discussion

We have evaluated more particularly the three main principles of the method. First, as learning is incremental, we have tried to study the influence of the order of the texts. As a first attempt in this way, we have extracted from the initial set of texts a subset corresponding to some of the domains that had been built and we have processed them as it had previously been done with different changes in the order of the texts. Of course, we have not obtained exactly the same results but these differences mainly concern the domains that do not result from many aggregations. For those which are better confirmed, no significant difference can be observed. So, these elements lead to think that this method is not too sensitive to the order of the texts.

The second point we have tested is the benefit of the inferred words. Globally, when no inferred word is considered, we get many more aggregated TUs, that, of course, gather less TUs. For example, the aggregated TU in table 1 is still built but gathers only 34 TUs even though the 61 TUs grouped by taking the inferred words into account were relevant. Moreover, most of the 27 TUs that are not aggregated any more together are memorized each one as a new aggregated TU. So, as expected, we can say that the inferred words help in having a better similarity between TUs.

The last point we have tested, but perhaps the most important, is to confirm that segmenting texts leads to build finer domain representations. In order to prove this, we have processed all the texts of our corpus without segmenting any one of them. So, each TU was produced from a whole text. The same kind of phenomenon got with the second test about the inferred words has been also obtained, but with a stronger effect. The aggregated TU in table 1 was thus reduced to only 21 TUs. So, by segmenting texts, the system builds a stabilized representation of a domain that can be detected as a good one (cf. figure 4 and 5) when without this segmentation, the doamin representation is less distinct and less stabilized, so more difficult to detect as a good one.

6. RELATED WORKS

Although similar to Kozima's work, our method to segment texts, that consists in moving a window through a text and computing cohesion values from a knowledge source, differs in its application. As Kozima uses word definitions and a propagation of activation process to select related words, we use a collocation network with a selection of the neighbours. In fact these two kinds of networks do not really encode the same knowledge. Our system goes further and automatically segments texts. This enables us to run it on a large amount of texts.

For the learning part, Hovy and Lin propose a method to learn signatures of domains. It requires a pre-classification of texts in the foreseen domains. As this kind of classification is only possible for very large domains, as banking, environment, etc., they form 32 classes made of 300 terms with their relative frequency in the corpus. As the authors say, hundreds or even thousands of different topics are needed for a robust summarization. With our system, signatures for more specialized topics, described by 80 significant words in average, are learned. The number of topics only depends on the subjects developed in the corpus. Descriptions are stable when around 20 discourse segments about a topic are processed. By this way, no a priori classification is needed. As the system is incremental, when new texts are processed, topics are learned or completed without having to process again previous texts.

7. CONCLUSION

[2] On average, the inferred words are as numerous as the text ones but there may be quite large variations.

We have developed a complete system that segments texts in thematic units and learn semantic description of specialized domain from the higher cohesive units. We have shown that steady units emerge quite rapidly with an incremental learning. Thematic units can also be viewed as an upper level to the collocation network that enables to structure it. We envisage to pursue on this way by studying the abstraction of the aggregated units and their formation into a hierarchy to build a more powerful knowledge source. Another project is to improve the thematic analyzis process by the feedback of the learned topics. By this way, the system would be able to better analyze currently low cohesive segments and also to recognize and learn domains that were less present in the texts used to build the collocation network.

REFERENCES

[1] Kenneth Ward Church and Patrick Hanks. Word Association Norms, Mutual Information, And Lexicography. Computational Linguistics, 16(1):22-29, 1990.

[2] Olivier Ferret, Brigitte Grau and Nicolas Masson. Utilisation d'un réseau de cooccurrences lexicales pour améliorer une analyse thématique fondée sur la distribution des mots. In *Proceedings of the 1ères Journées du Chapitre Français de l'ISKO*, Lille, France, 1997.

[3] Brigitte Grau. Stalking Coherence in the Topical Jungle. In *Proceedings of the FGCS, Fifth Generation Computer System*, Tokyo, 1984.

[4] Barbara J. Grosz and Candace L. Sidner. Attention, Intentions and the Structure of Discourse. *Computational Linguistics*, 12:175-204, 1986.

[5] Eduard Hovy and Chin Yew Lin. Automated Text Summarization in SUMMARIST. In *Proceedings of the ACL 97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Espagne, 1997.

[6] Hideki Kozima. Text Segmentation Based on Similarity between Words. In *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics (Student Session)*, Colombus, Ohio, USA, 1993.

[7] Ellen Riloff and Wendy Lehnert. Information Extraction as a basis for High-Precision Text Classification. *ACM Transactions on Information Systems*, 12(3):296-333, 1994.

[8] Jane Morris and Graeme Hirst. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics*, 17(1):21-48, 1991.

[9] Helmut Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

[10] Max D. Silberztein. INTEX: A Corpus Processing System. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING)*, Kyoto, Japan, 1994.