Enhancing lexical cohesion measure with confidence measures, semantic relations and language model interpolation for multimedia spoken content topic segmentation

Camille Guinaudeau, Guillaume Gravier, Pascale Sébillot

Enhancing lexical cohesion measure with confidence measures, semantic relations and language model interpolation for multimedia spoken content topic segmentation

Camille Guinaudeau[1], Guillaume Gravier[2], Pascale Sébillot[3]

INRIA Rennes[1] & IRISA (CNRS[2], INSA[3]), France

camille.guinaudeau@inria.fr, guillaume.gravier@irisa.fr, pascale.sebillot@irisa.fr

Abstract

Transcript-based topic segmentation of TV programs faces several difficulties arising from transcription errors, from the presence of potentially short segments and from the limited number of word repetitions to enforce lexical cohesion, i.e., lexical relations that exist within a text to provide a certain unity. To overcome these problems, we extend a probabilistic measure of lexical cohesion based on generalized probabilities with a unigram language model. On the one hand, confidence measures and semantic relations are considered as additional sources of information. On the other hand, language model interpolation techniques are investigated for better language model estimation. Experimental topic segmentation results are presented on two corpora with distinct characteristics, composed respectively of broadcast news and reports on current affairs. Significant improvements are obtained on both corpora, demonstrating the effectiveness of the extended lexical cohesion measure for spoken TV contents as well as its genericity over different programs.

Keywords: Topic segmentation, lexical cohesion, confidence measures, semantic relations, language model interpolation, TV broadcasts

1. Introduction

Structuring video feeds has become a requirement and is a highly challenging issue. Indeed, with the proliferation of video-sharing websites and the increasing number of television channels, the quantity of video that users can access has become so important that it is necessary to develop methods to structure this material and enable users to navigate inside the stream. Since videos available are of different kinds (movies, news, talk shows, etc.) and in order to avoid the use of several domain specific methods, such structuring approaches must necessarily be generic enough to treat various types of video. A crucial structuring stage is the segmentation into successive TV shows on the one hand, and of the shows into topically homogeneous segments on the other hand. Topic segmentation of informative content TV shows can rely on the speech pronounced in the programs. In this case, the segmentation is based on the analysis of the distribution of words within the speech, a topic change being detected when the vocabulary changes significantly. With the improvement of automatic speech recognition (ASR) systems in recent years [1], topic segmentation of spoken documents can now be performed on automatic transcripts of the speech material. However, most of the work in this direction directly apply methods developed for textual topic segmentation to automatic transcripts of spoken language. These methods are often based on the notion of lexical cohesion which corresponds to the lexical relations that exist within a text, mainly enforced by reiterations, i.e., repetitions of the same words [2, 3]. Alternately, discourse markers, obtained from a preliminary expert or learning process, can also be used to identify topic boundaries [4, 5].

Christensen *et al.* [5] have established that transcription errors have little effect on the performance of a supervised segmentation algorithm using discourse markers. However, we have observed a large gap of performance between manual and automatic transcripts in previous work on topic segmentation of radio broadcasts using an unsupervised approach based on lexical cohesion [6]. This difference is mostly due to the specifics of the material on which we focus, i.e., TV shows. Indeed, automatic transcripts of TV shows have certain peculiarities that are detrimental to topic segmentation and, in general, to natural language processing. Firstly, the error rate of the ASR system used, even if it remains reasonable for news, can be as high as 70% for challenging programs such as talk shows or debates. Moreover, TV programs are composed of topic segments that can be very short and contain few

repetitions of vocabulary, particularly in news where journalists make use of synonyms to avoid reiterations. In our corpus, we have measured that a word occurs on average 1.8 times in a topically coherent segment in broadcast news and 2.0 times in reports on current affairs (for more details, cf. Section 4.3). In order to overcome difficulties related to transcription errors, some studies have suggested to add features specific to spoken documents to the sole concept of lexical cohesion. For example, [7] exploits speaker detection to locate the anchor speaker in news program, relating anchor speaker occurences with new reports and hence with topic changes. In [8], prosody is used in addition to automatic transcription. However, such clues are seldom used in practice because their automatic extraction is difficult. Moreover, they are highly dependent on the type of documents and therefore imply domain and genre specific knowledge.

The aim of this paper is to propose a segmentation method able to deal with peculiarities of professional videos (transcription errors, possibly short segments and limited number of repetitions) while remaining generic enough. This method is based on the criterion of lexical cohesion which is not dependent on a type of document since it is mainly enforced by word repetitions. However lexical cohesion is not efficient when the number of reiterations is low—i.e., when synonyms are used or topic segments are very short—and is sensitive to transcription errors, two characteristics of our video material. We propose several extensions to a measure of lexical cohesion, based on generalized probabilities using a unigram language model, in order to make this criterion more robust to spoken content. On the one hand, the measure of the lexical cohesion is modified by an original technique that incorporates two sources of additional information: semantic relations, highlighting the semantic proximity between words, and confidence measures. On the other hand, we propose to use language model interpolation techniques so as to provide better estimates of the lexical cohesion on short segments.

The paper is organized as follows: we first present the topic segmentation method, based on lexical cohesion, developed for the segmentation of textual documents and used as a baseline in this work. In Section 3, extensions of the probabilistic lexical cohesion measure to improve robustness to TV program specifics are described. The experimental setup is presented in Section 4. Finally, experimental results are extensively discussed in Section 5, before the presentation of future work.

2. Topic segmentation

Within the framework of TV stream structuring, the objective of topic segmentation is to split relevant shows (broadcast news and reports on current affairs in this work) into segments that deal with a single topic. Topic segmentation algorithms can be based on the criterion of lexical cohesion. In this case, the segmentation relies on the analysis of the distribution of words within the text, a topic change being detected when the vocabulary changes significantly [3, 9].

In this section, the lexical cohesion criterion is first described, as it is usually used in the context of topic segmentation, before the presentation of the topic segmentation method of Utiyama and Isahara [3] which serves as a baseline in this work.

2.1. Lexical cohesion

The notion of lexical cohesion refers to lexical relations that exist within a text to provide a certain unity. Lexical cohesion is created by repetitions of the same words, co-references, and the use of sets of semantically related words [10]. As lexical cohesion is a guide to the organization of the flow of ideas in the text, this criterion is widely used in many fields of the natural language processing domain. In discourse analysis, [11] studies the relationship between cohesion and coherence in texts while [12] makes a stylistic analysis of political speeches. Others studies use lexical cohesion for word sense disambiguation [13], and automatic summarization [14, 15]. Error identification and correction applications can also rely on lexical cohesion to detect errors by identifying tokens that are semantically unrelated to their context. These methods can be applied on regular text, as in [16], or on automatic transcripts [17].

In [3], the lexical cohesion computation of a segment S_i consists in the evaluation of the ability of a language model Δ_i whose parameters are estimated from words in S_i to predict the words in the segment. In this framework, two important steps are needed: the estimation of the language model Δ_i and the computation of the generalized probability of words in S_i, reflecting the ability of the language model Δ_i to predict the words of S_i.

Language model. The language model Δ_i estimated on S_i is a unigram language model [3] which specifies a distribution over all the words of the text

(or transcript) to be segmented. The calculation of the language model of the segment S_i is formalized, for a Laplace smoothing, by

$$\Delta_i = \left\{ P_i(u) = \frac{C_i(u) + 1}{z_i},\ \forall\ u \in V_K \right\} \ , \tag{1}$$

with V_K the vocabulary of the text of size K and $C_i(u)$ the count of word u in S_i corresponding to the number of occurrences of u in S_i. The probability distribution is smoothed by incrementing the count of each word by 1. The normalization term z_i ensures that Δ_i is a probability mass function and, in the particular case of Eq. 1, $z_i = K + n_i$ with n_i the number of words in S_i.

Generalized probability. The language model Δ_i is used to compute the generalized probability of the words in S_i as a measure of lexical cohesion according to

$$\ln P[S_i; \Delta_i] = \sum_{j=1}^{n_i} \ln P[w_j^i; \Delta_i] \ , \tag{2}$$

where w_j^i denotes the j^{th} word in S_i. Intuitively the probability favors lexically consistent segments since its value is greater when words appear several times within the segment and decreases if many words are different.

2.2. Topic segmentation

The topic segmentation method used, introduced by Utiyama and Isahara [3] for textual documents, was chosen in this context of transcript-based TV program segmentation for two main reasons. It is currently one of the best performing method that makes no assumption on a particular domain (no discourse markers, no topic models, etc.). Moreover, contrary to many methods based on local measures of the lexical cohesion, the global criterion used in [3] makes it possible to account for the high variability in segment lengths.

The method consists in searching the segmentation that produces the most consistent segments from a lexical point of view, while respecting a prior distribution of the segment lengths. Cast in a probabilistic framework, the principle is to find the most probable segmentation of a sequence of l basic units (words or sentences) $W = W_1^l$ among all possible segmentations,

$$\hat{S} = \underset{S_1^m}{\operatorname{argmax}}\, P[W|S]\ P[S] \ . \tag{3}$$

Assuming that $P[S_1^m] = n^{-m}$, with n the number of words in the text and m the number of segments, and that segments are independent, the probability of a text W for a segmentation $S = S_1^m$ is given by

$$\hat{S} = \underset{S_1^m}{\operatorname{argmax}} \sum_{i=1}^{m} \left(\ln(P[W_{a_i}^{b_i}|S_i]) - \alpha \ln(n) \right) \quad , \tag{4}$$

where $P[W_{a_i}^{b_i}|S_i]$ denotes the generalized probability of the sequence of basic units corresponding to S_i as given by Eq. 2. The parameter α allows for different trade-offs between lexical cohesion and segment lengths.

In this paper, the basic units used are utterances as given by the partitioning step of the ASR system, thus limiting possible topic boundaries to utterance boundaries, and the text W is only composed of lemmatized[1] nouns, adjectives and non modal verbs.

3. Robust topic segmentation for spoken multimedia contents

The language model based lexical cohesion measure, as defined in Section 2.1, relies only on word repetitions. However, this can turn out to be insufficient in the case of automatically transcribed video material. Indeed, two occurrences of a word can be erroneously recognized as two different words and therefore not considered for lexical cohesion. Moreover, due to potentially small segment lengths and to the use of synonyms, the number of repetitions can be very low.

To adjust the computation of lexical cohesion to spoken documents, we propose several extensions. In the first ones, additional information is incorporated to the lexical cohesion measure while, in the second one, language model interpolation techniques are used so as to povide better language model estimates.

3.1. Integration of additional sources of information

Different kinds of additional information, such as prosody or confidence measures, can be used to improve the generalized probability measure of lexical cohesion. In this work, confidence measures and semantic relations

[1] A lemma is an arbitrary canonical form grouping all inflections of a word in a grammatical category, e.g., the infinitive form for verbs, the masculine singular form for adjectives, etc.

are employed. Confidence measures, associated with each word by the ASR system, correspond to (an estimation of) the probability that a word has been correctly transcribed. They are considered to reduce the contribution of non properly transcribed words in the lexical cohesion calculation. Semantic relations, which represent the semantic proximity between words, are used to consider the fact that two different words can be semantically related and seen as a repetition. Such information can be accounted for either during the language model estimation step or during the calculation of the generalized probability.

3.1.1. Using confidence measures

Confidence measures can be accounted for at the language model level by replacing the count $C_i(u)$ by the sum of the confidences over all occurrences of u, i.e.,

$$C'_i(u) = \sum_{w_j^i = u} c(w_j^i)^{\delta_1} \quad , \tag{5}$$

where $c(w_j^i) \in [0, 1]$ corresponds to the value of the confidence measure for the j^{th} word in S_i and δ_1 is a parameter used to reduce the weight of words whose confidence measure is low. Indeed, the higher δ_1, the lower the impact of words with a low confidence measure.

Confidence measures can also be accounted for during the generalized probability computation. In this case, the log-likelihood of the occurrence of a word in a segment is multiplied by the confidence measure of the word occurrence,

$$\ln P[S_i; \Delta_i] = \sum_{j=1}^{n_i} c(w_j^i)^{\delta_2} \ln P[w_j^i; \Delta_i] \quad , \tag{6}$$

with δ_2 equivalent to δ_1. Eq. 6 allows to reduce the contribution to the lexical cohesion of a word whose confidence measure is low. In this case, the language model Δ_i can be either estimated from the counts $C_i(u)$, thus limiting the use of confidence measures to the probability computation, or from the modified counts $C'_i(u)$. Note that in Eq. 6, $\ln P[S_i; \Delta_i]$ is not strictly speaking a log probability. However, if $\delta_2 = 1$, $\ln P[S_i; \Delta_i]$ can be seen as the joint probability for the word w_j^i to be correctly transcribed and to be represented by the language model Δ_i.

3.1.2. Using semantic relations

Many topic segmentation techniques, based on the lexical cohesion criterion, use semantic relations as additional information to take into account the semantic links that exist between words. Indeed, the primary goal of semantic relations is obviously to ensure that two semantically related words, e.g., "car" and "drive", contribute to the lexical cohesion, thus avoiding erroneous topic boundaries between two such words. These different methods can use semantic relations that are manually defined by experts, as in [18], or extracted automatically from corpora [19, 20]. For example, in [19], Ferret uses a network of lexical co-occurrences built from a large corpus to improve the results of a topic segmenter based on lexical reiteration. The algorithm in [20] compares adjacent windows of sentences and determines their lexical similarity, based on repetitions of words and collocations, to detect topic boundaries.

In our work, as in [18, 19, 20], semantic relations are employed to overcome the limited number of vocabulary repetitions. But semantic relations are also expected to limit the impact of recognition errors. Indeed, contrary to correctly transcribed words, misrecognized words are unlikely to be semantically linked to other words in a topic segment [17]. As a consequence, the contribution of non properly transcribed words will be less important than the one of correct words.

As for confidence measures, accounting for semantic relations can be achieved by modifying the counts in the language model estimation step. Counts, which normally reflect how many times a word appears in a segment, are extended so as to emphasize the probability of a word based on its number of occurrences as well as on the occurrences of related words. More formally, the counts C_i in Eq. 2 are amended according to

$$C_i''(u) = C_i(u) + \sum_{j=1, w_j^i \neq u}^{n_i} r(w_j^i, u) \ , \tag{7}$$

where $r(w_j^i, u) \in [0, 1]$ denotes the semantic proximity between w_j^i and u, close to 1 for highly related words and null for non related words. More details about the computation of r are given in Section 4.2.

Unlike confidence measures, semantic relations cannot be accounted for during generalized probability computation. Indeed, in this case, it does not really make sense to multiply the probability for a word to appear in a

segment S_i by the sum of the relations the word maintains with other words in S_i.

To conclude this section, it is important to note that, in general, semantic relations are obtained from domain specific corpora. They can also be learned on a general purpose corpus but will in any case fail to address all domains. One strong point of our technique is that when semantic relations are not adequate for a particular document—e.g., when segmenting a document from a domain not in the semantic relations training data—$C_i''(u)$ will remained unchanged with respect to $C_i(u)$ since $r(u, v)$ is null between any two words of the document. Put differently, out of domain relations will have no impact on topic segmentation, contrarily to latent semantic approaches [21, 22].

3.2. Combining global and local measurements of lexical cohesion

In TV programs, and especially in news, topic segments can be very short. An easy, and unfortunately accurate, criticism of Eq. 1 is that for small segments, the language model Δ_i is poorly estimated. Therefore, in order to deal with short segments, the use of a more sophisticated estimation method is needed. To skirt this problem, we propose to interpolate the language model at the segment level with a more robust language model estimated on the entire transcript. This allow to take into account the whole text to be segmented in order to have a better language model estimation for short segments. Two interpolation strategies are studied—interpolation of the probabilities [23] and interpolation of the counts [24].

3.2.1. Linear interpolation of probabilities

The first interpolation technique is a basic probability interpolation. In this case, the lexical cohesion of a segment S_i, given the segment S_i and the text T, is measured according to

$$\begin{aligned} \ln P[S_i; S_i, T] &= \sum_{j=1}^{n_i} \ln(\lambda P[w_j^i; \Delta_i] + (1-\lambda) P[w_j^i; \Delta_t]) \\ &= \sum_{j=1}^{n_i} \ln \left(\lambda \frac{C_i(w_j^i) + \xi}{\sum_{u \in V_T} C_i(u) + \xi} + (1-\lambda) \frac{C_t(w_j^i)}{\sum_{u \in V_T} C_t(u)} \right) , \end{aligned} \tag{8}$$

where Δ_i is the language model estimated on S_i and Δ_t the one estimated on T. $C_t(u)$ is the count of word u in T and $C_i(u)$ the count of that word in

S_i. ξ is a count smoothing bias that corresponds to the Laplace smoothing when $\xi = 1$.

3.2.2. Count interpolation

Rather than interpolating probabilities, language model interpolation can be based on the interpolation of counts. In this case, the lexical cohesion of a segment S_i is defined as

$$\begin{aligned} \ln P[S_i; S_i, T] &= \sum_{j=1}^{n_i} \ln P[w_j^i; \Delta_{it}] \\ &= \sum_{j=1}^{n_i} \ln \left(\frac{\lambda(C_s(w_j^i) + \xi) + (1-\lambda)C_t(w_j^i)}{\sum_{u \in V_T} \lambda(C_i(u) + \xi) + (1-\lambda)C_t(u)} \right) , \end{aligned} \tag{9}$$

where Δ_{it} is the interpolated language model of the segment S_i and the text T. As for linear interpolation, frequent words in T will get a high probability regardless of their frequency in S_i while non frequent ones will always get a low probability—depending on λ. However, because of the renormalization by the sum of all the counts, this fact might be less detrimental than for probability interpolation and the behavior of this interpolation technique more likely to be consistent with what is expected.

4. Experimental setup

Experiments were carried out with our speech recognition system on a comprehensive corpus. Before presenting the corpus, a brief description of the ASR system is provided and the acquisition and selection of semantic relations is discussed.

4.1. ASR system and confidence measures

All TV programs were transcribed using our Irene ASR system, originally developed for broadcast news transcription. Irene implements a multiple pass strategy, progressively narrowing the set of candidate transcripts—the search space—in order to use more complex models. In the final steps, a 4-gram language model over a vocabulary of 65,000 words is used with context-dependent phone models to generate a list of 1,000 sentence transcription hypotheses. Morphosyntactic tagging, using a tagger specifically

designed for ASR transcripts, is used in a post-processing stage to generate a final transcription by consensus from a confusion network, combining the acoustic, language model and morphosyntactic scores [25]. Confusion network posterior probabilities are used directly as confidence measures.

Acoustic models were trained on about 250 hours of broadcast news material from the French ESTER 2 data [26]. The language model probabilities were estimated on 500 million words from French newspapers and interpolated with language model probabilities estimated over 2 million words corresponding to the reference transcription of radio broadcast news shows. The system exhibits a word error rate (WER) of 16 % on the non accented news programs of the ESTER 2 evaluation campaign. As far as TV contents are concerned, we estimated word error rates ranging from 15 % on news programs to more than 70 % on talk shows or movies.

4.2. Semantic relations

Semantic relations were automatically extracted from text corpora and relevant relations were selected so as to avoid abusive enforcement of the lexical cohesion.

4.2.1. Acquisition

Automatic extraction of semantic relations from text corpora has been widely studied [27, 28, 29]. Two main types of relations might be of interest for lexical cohesion, namely syntagmatic and paradigmatic relations.

Syntagmatic relations correspond to relations of contiguity that words maintain within a given syntactic context (sentence, chunk, fixed length window, etc.), two words being related if they often appear together. A popular criterion to measure the degree of syntagmatic proximity between two words u and v is the mutual information (MI) which compares the probability of observing the two words u and v together with the probability of observing these two words separately [29]. Several variants of the mutual information criterion have been proposed. For example, in [30], the mutual information is cubed to avoid emphasizing rare associations. In [31], order between words is taken into account.

Paradigmatic relations link two words having an important common component from a meaning point of view. These relations, corresponding to synonyms, hyperonyms, antonyms, etc., are typically calculated by means of context vectors for each word, grouping together words that appear in the same contexts. The context vector of a word u describes the distribution of

Table 1: *Words with the highest association scores, in decreasing order, for the word "cigarette", as extracted automatically.*

syntagmatic	paradigmatic
fumer (*to smoke*)	cigare (*cigar*)
paquet (*pack*)	gitane[2]
allumer (*to light*)	gauloise[2]
contrebande (*smuggling*)	clope (*ciggy*)
fabricant (*producer*)	tabac (*tobacco*)

words in its vicinity and contains, for each word v, its frequency of occurrence in the neighborhood of u, possibly normalized by its average frequency in the neighborhood of any word. The semantic proximity between two terms can be defined thanks to the Jaccard index as in [32] or as the angular distance between their respective context vectors.

In this work, the MI^3 criterion for syntagmatic relations and the cosine distance between normalized context vectors for paradigmatic ones were used. Relations were extracted from a corpus of articles from French newspapers (about 800 million words) and from reference transcripts of radio broadcast news shows (about 2 million words). All the data were lemmatized, keeping only nouns, adjectives, and non modal verbs. Semantic proximity scores given either by the MI^3 criterion or by the angular distance were normalized in [0,1]. Table 1 shows, for the word "cigarette", the five related words with the highest semantic proximity score, for syntagmatic and for paradigmatic relations.

4.2.2. Selection

To prevent the creation of too many links between words, a selection step is implemented to choose relevant syntagmatic and paradigmatic relations among all the existing ones so as to introduce only the best ones in the segmentation algorithm. Two selection strategies were evaluated, either retaining globally the N relations with the highest scores (global selection strategy) or retaining for each word the M best relations (word level selection strategy). Moreover, common usage words such as "go" or "year" were

[2]Brand name.

Table 2: Comparison of the *news* and *reports* corpora

	average #repetitions	average confidence measure	average #words per segment	average #words per file	#topic boundaries
news	1.82	0.62	106.4	2,243	1,202
reports	2.01	0.57	424.1	2,495	86

found to be related to many words, thus jeopardizing topic segmentation by abusively enforcing lexical cohesion. Semantic relations for words whose number of related words exceeds a given threshold are therefore discarded, the threshold being equal to the average number of relations associated with a word multiplied by a parameter $\gamma \in [0, 1]$.

4.3. Corpus

Results are reported on two distinct corpora. The first one, a *news* corpus, is made up of 57 news programs ($\approx 1/2$ hour each) broadcasted in February and March 2007 on the French television channel France 2, and the second one is a *reports* corpus composed of 16 reports on current affairs "Sept à Huit" (≈ 1 hour each) transmitted on the French channel TF1 between September 2008 and February 2009. In the *reports* corpus, reports are longer (around 10-15 minutes) and possibly on non news topics, while the *news* corpus follows the classical scheme of rather short reports (usually 2-3 minutes). Having two distinct corpora makes it possible to study the behavior of topic segmentation on data sets with different characteristics. Indeed, in addition to different durations, the average number of topics and the average number of segments per show vary greatly between news and reports. Moreover, the number of repetitions is less important in news programs than in reports ones, as reported in Table 2, while the transcription error rate is higher on the latter due to a larger amount of non professional speakers.

In each show, headlines and closing remarks were removed, these two particular parts disturbing the segmentation algorithm and being easily detectable from audiovisual clues. A reference segmentation was established by considering a topic change associated with each report, the start and end boundaries being respectively placed at the beginning of the report's introduction and at the end of the report's closing remarks. Note that in the *news*

corpus, considering a topic change between each report is a choice that can be questioned as, in most cases, the first reports all refer to the main news of the day and are therefore dealing with the same broad topic.

5. Results

The goal of the article is to study to which extent confidence measures, semantic relations and interpolation techniques can help to make the lexical cohesion criterion more robust for professional video processing. We therefore study the influence of each technique and each parameter on the entire data set, reporting results for optimal parameter values. Even though the resulting figures are optimistic and do not reflect real-life behavior, their comparison with a baseline system clearly demonstrates the impact of each technique on topic segmentation.

Recall and precision on topic boundaries are considered for evaluation purposes after alignment between reference and hypothesized boundaries, with a tolerance on boundary locations of respectively 10 and 30 s for *news* and *reports*, while different trade-offs between precision and recall are obtained by varying α in Eq. 4. To compare the different parameters (δ, γ, etc.), the tables in this section contain the best F1-measure obtained for an optimal value of α, i.e., the one leading to the segmentation with average segment length closest to that of the reference segmentation.

5.1. Confidence measures

Results for the integration of confidence measures (CMs) are presented in Tables 3 and 4 for the *news* and *reports* corpus respectively. In these tables, the grey line stands for the values obtained when CMs are introduced during the language model estimation alone, with δ_1 varying from 0 to 4, while the grey column represents their integration during the generalized probability computation alone. The dark grey cell therefore corresponds to the baseline method ($\delta_1 = \delta_2 = 0$). Recall and precision curves are reported on Figure 1.

We can observe that confidence measures integration leads to an improvement in the topic segmentation quality. Indeed, it can be seen that the best value of the F1-measure is improved by 2 points for the *news* corpus and by 3.7 points for the *reports* corpus.

Moreover, results show that the behavior of the integration of confidence measures is different for the two corpora depending on whether CMs are integrated during generalized probability or language model computation. For

Table 3: Integration of confidence measures - *news* corpus - best F1-measure

$\delta_2 \setminus \delta_1$	0	0.5	1	1.5	2	2.5	3	3.5	4
0	59.7	60.4	60.7	61.0	**61.1**	60.9	60.9	59.7	59.1
0.5	60.4	60.7	61.5	61.5	60.9	60.9	60.8	59.5	58.8
1	61.1	61.5	**61.7**	61.1	60.9	60.9	60.6	58.9	58.8
1.5	61.2	61.6	61.3	60.9	60.5	60.1	60.2	58.9	58.8
2	61.3	61.4	61.4	60.8	60.2	59.9	59.9	58.9	58.7
2.5	61.2	61.5	61.1	60.5	60.1	59.6	59.6	58.6	58.1
3	**61.6**	61.5	60.8	60.4	60.0	59.5	59.4	58.1	57.9
3.5	60.6	60.5	59.3	58.8	58.8	58.3	58.1	57.8	57.8
4	60.9	60.6	59.3	59.0	58.9	58.1	57.8	57.8	57.7

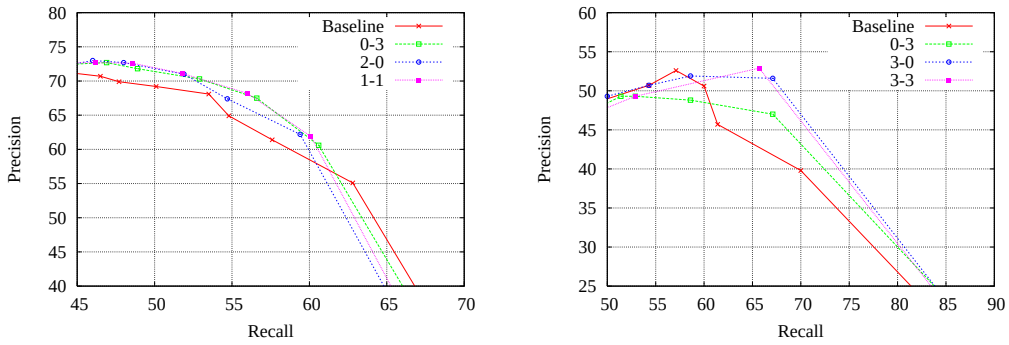
the *news* corpus, results are better when confidence measures are integrated during the generalized probability computation alone rather than during the language model computation alone; while for the *reports* corpus, experiments lead to the opposite conclusion. This phenomenon is also observed on the recall/precision curves, where the integration during the generalized probability calculation (green curve) gives better results for the *news* corpus, while integration during the language model computation is more efficient for *reports*.

Another difference between the two corpora is the optimal values of δ_1 and δ_2. For the *news* corpus, the highest F1-measure is obtained when δ_1 and δ_2 are quite small (both equal to 1) while for the *reports* corpus the optimal values for δ_1 and δ_2 are greater. This difference can be explained by the fact that for high values of δ, $c(w_j^i)^\delta$ becomes negligible except for words whose confidence measure is very close to 1. As the proportion of words with a CM less than 0.9 is more important in the *reports* data, the impact of the confidence measures is more perceptible on this data set and higher values of δ lead to a greater improvement.

Similarly, the impact of the transcription quality on the efficiency of the integration of confidence measures is also highlighted by the difference between the improvements observed for the two corpora. Figure 1 shows that confidence measures have more impact on the *reports* corpus than on the *news* corpus. This difference shows that confidence measures are of utmost importance when transcription quality decreases.

Table 4: Integration of confidence measures - *reports* corpus - best F1-measure

$\delta_2 \setminus \delta_1$	0	0.5	1	1.5	2	2.5	3	3.5	4
0	54.9	55.0	56.8	54.6	54.7	55.6	58.3	**59.9**	59.1
0.5	53.4	56.1	55.3	56.1	54.7	54.7	55.0	57.9	58.2
1	**55.3**	56.5	55.4	55.6	54.7	55.0	55.0	59.1	58.2
1.5	54.4	55.6	55.6	56.3	54.7	55.0	55.0	56.9	58.6
2	54.9	56.0	56.3	55.0	55.3	55.3	57.0	56.1	56.1
2.5	55.1	55.7	56.3	55.0	55.3	57.0	57.0	56.1	56.5
3	**55.3**	56.8	56.3	55.0	56.7	57.0	58.6	57.7	58.1
3.5	55.3	56.8	56.3	56.3	58.3	58.6	58.6	57.7	57.7
4	54.5	56.8	56.3	**59.9**	59.5	58.6	57.7	57.7	57.1



(a) *news* corpus (b) *reports* corpus

Figure 1: Integration of confidence measures - Recall/Precision curves (figures in the legend correspond to resp. δ_1 and δ_2)

From a more qualitative point of view, we also observed that accounting for confidence measures not only increases the number of correct boundaries detected but also improves boundary locations. Indeed, boundary locations are more precise when using confidence measures, even if this fact does not show in the recall/precision curves because of the tolerated gap on the boundary positions. We also observed that the quality of confidence measures have an impact on boundary precision. Indeed, using confidence measures improved thanks to high-level linguistic features with a classifier [33] resulted in more accurate boundary locations.

To conclude, experiments carried out on the two corpora confirm the hy-

pothesis that the integration of confidence measures in the lexical cohesion computation allows the criterion to be less sensitive to transcription errors. Indeed, amending the calculation method of lexical cohesion to include CMs provides a significant improvement in the quality of topic segmentation. Confidence measures seem also to lead to an increase in the number of correct boundaries detected but also to the displacement of previously recognized borders, moving them closer to reference boundaries.

5.2. Semantic relations

Various tests, on the choice of the type and number of semantic relations introduced and on the use of the selection strategy (global or word level) or the filtering technique have produced a large number of results. The most convincing ones are discussed here.

Tables 5 and 6 summarize the results on the *news* and *reports* corpus respectively, with values for syntagmatic relations in the leftmost part of the tables and results for paradigmatic ones in the rightmost part. For each kind of semantic relations, the two selection strategies are evaluated for different numbers of semantic relations, 2 to 10 per word for the word-level strategy and 5,000 to 90,000 for the global strategy. Finally, the results for the filtering technique, used to discard semantic relations for words whose number of related words exceeds a given threshold, are presented for values of parameter γ between 0.2 and 1. In Figure 2, recall/precision curves for the integration of paradigmatic relations are presented.

For the *news* corpus, semantic relations can help the lexical cohesion criterion to be more robust to the limited number of vocabulary repetitions due to short segments and/or the use of synonyms or related words. Indeed, the best F1-measure is improved by 1.4 when syntagmatic relations are introduced and by 0.6 for paradigmatic relations (Table 5). Moreover, even if the best values of the F1-measure are reached for syntagmatic relations, the global improvement is higher with paradigmatic ones. For the *reports* corpus, the improvement is much smaller than for the *news* one. Indeed, Figure 2 shows that results for the integration of paradigmatic relations—which are comparable to those for syntagmatic relations, cf. Table 6—are almost equivalent to the baseline. The difference in results for both corpora can be explained by the fact that in the *reports* corpus the number of reiterations and the segments lengths are more important than in the *news* corpus. Thus, the use of semantic relations is not as effective for this corpus. Moreover, in this work, semantic relations were extracted from a corpus composed

of news articles and are therefore less suited for the *reports* corpus. However, it is interesting to note that using non appropriate relations does not degrade results.

Table 6 shows that the integration of semantic relations damages the segmentation quality when a too large number of relations is introduced—i.e., when the parameter γ is high—degradation which is proportional to the number of relations included. This behavior can also be observed for the *news* corpus but for higher values of γ. Therefore, the filtering technique is crucial to avoid considering relations that seem less adapted in the context of topic segmentation because they introduce some noise (i.e., they connect words and segments that should not be). The stronger effect of the filtering technique on the *reports* corpus can be explained by the fact that many relations are out of domain for this corpus. Therefore, more general relations—associated with words such as "year" and "go"—that introduce noise appear in a higher proportion and have more impact on reports than on news.

Finally, concerning the technique used for the semantic relations selection, no difference was found between the global selection strategy and the word level selection strategy for the *news* corpus. In contrast, for the *reports* corpus, the global selection strategy is better, both for paradigmatic and syntagmatic relations. We believe that the word level strategy selects relations which are the most characteristic of the corpus they are learned from, while the global strategy is more likely to pick more general relations. Thus, as relations are extracted from a corpus composed of news articles, the ones selected by the word level strategy are more suitable for the *news* corpus but less adapted for the *reports* one.

To conclude, the introduction of semantic relations can improve topic segmentation when the number of vocabulary repetitions is low. However, it is essential to limit the number of relations considered.

5.3. Interpolation

Finally, results for language model interpolation are given in Table 7. Values for the count interpolation strategy are presented in the leftmost part of the table while results for the linear interpolation of probabilities are located in the rightmost one.

For the two corpora, we can observe that the improvements in the topic segmentation quality are higher with the interpolation of the counts than with the linear interpolation of probabilities. Indeed, the best value for the F1-measure is increased by 4.9 against 2.3 for the *news* corpus and by 0.7

Table 5: Integration of semantic relations - *news* corpus

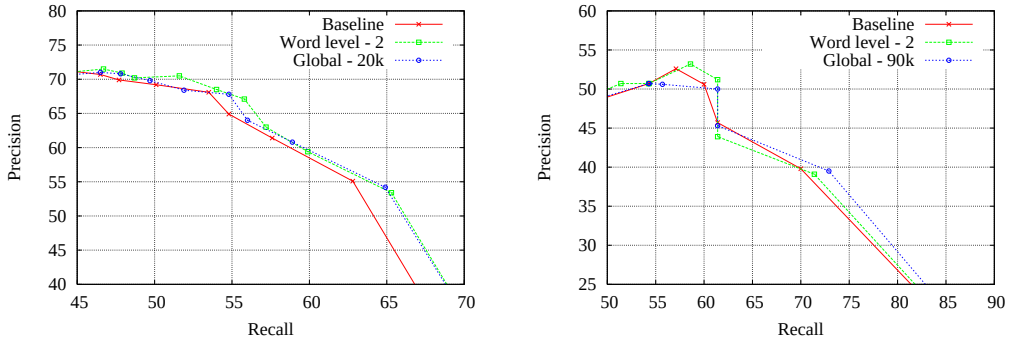
	Syntagmatic						Paradigmatic					
	Word level			Global			Word level			Global		
γ	2	3	10	5k	20k	90k	2	3	10	5k	20k	90k
0.2	59.7	59.7	59.7	59.7	59.7	59.7	59.4	59.7	59.7	59.7	59.7	59.7
0.4	59.7	59.7	60.8	59.7	59.8	59.7	59.7	59.7	59.7	59.7	59.7	60.0
0.5	59.7	59.7	60.1	59.7	60.3	60.4	59.7	59.7	60.0	59.7	59.7	60.3
0.6	59.7	59.6	60.5	59.7	60.0	60.1	59.7	59.7	**60.2**	59.7	60.1	60.1
0.8	**60.0**	59.8	**60.9**	**60.5**	60.6	60.6	59.8	60.3	60.0	59.9	**60.6**	**60.3**
1	58.9	**61.1**	59.8	**60.5**	**60.8**	**61.0**	**60.9**	**60.4**	**60.2**	**60.0**	60.3	**60.3**

Table 6: Integration of semantic relations - *reports* corpus

	Syntagmatic						Paradigmatic					
	Word level			Global			Word level			Global		
γ	2	3	10	5k	20k	90k	2	3	10	5k	20k	90k
0.2	54.9	54.9	54.9	54.9	54.9	54.9	54.9	54.9	54.9	54.9	54.9	54.9
0.4	54.9	54.9	**55.7**	54.9	**55.5**	**55.5**	54.9	54.9	**55.5**	54.9	54.9	54.9
0.5	54.9	54.9	54.0	54.9	**55.5**	**55.5**	54.9	54.9	55.1	54.9	54.9	**55.1**
0.6	54.9	54.9	53.3	54.9	54.4	54.4	54.9	54.9	54.0	54.9	54.9	54.4
0.8	**56.5**	**55.7**	53.4	**55.5**	53.1	53.1	**55.8**	53.2	53.6	54.9	**55.8**	54.8
1	54.7	54.6	51.6	55.1	54.0	54.0	55.2	53.0	53.0	54.9	52.3	54.0

against 0 for the *reports* corpus. This behavior is explained by the fact that, as mentioned in Section 3.2, for the linear interpolation of probabilities, frequent words in the text get a high probability regardless of their frequencies in the segment, while non frequent ones always get a low probability. However, because of the renormalization in Eq. 9, this fact is less detrimental for count interpolation than for probability interpolation.

Table 7, as well as the recall/precision curves in Figure 3, shows that the interpolation techniques are much less effective for the *reports* corpus than for the *news* corpus. This observation can be easily explained by the fact that segments in reports are four times longer than the ones in the news. Thus, the use of interpolation has no effect for the calculation of lexical cohesion in the *reports* segments since the initial language models



(a) *news* corpus (b) *reports* corpus

Figure 2: Integration of paradigmatic relations - Recall/Precision curves

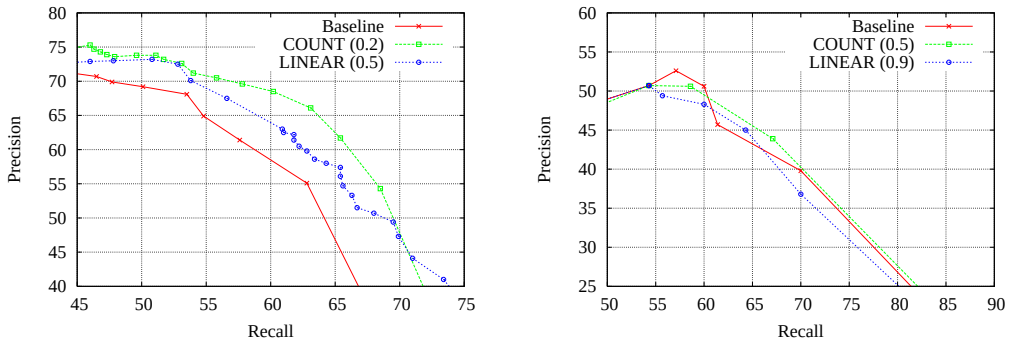
Table 7: Interpolation

	COUNT		LINEAR	
λ	news	reports	news	reports
Baseline	59.7	54.9	59.7	**54.9**
0	3.8	14.8	12.2	10.4
0.1	63.7	54.0	60.8	47.8
0.2	**64.6**	53.4	61.0	47.3
0.3	64.2	52.8	61.6	50.0
0.4	64.2	**55.6**	61.7	50.3
0.5	64.1	54.3	**62.0**	51.1
0.6	64.1	52.4	61.9	51.7
0.7	63.7	53.1	61.9	50.9
0.8	62.7	53.3	61.8	52.7
0.9	61.6	54.9	60.6	53.5
1	59.9	54.9	59.7	**54.9**

were already well estimated. This can also be seen from the optimal value of parameter λ, representing the weight of the language model of the segment in the interpolation, that is greater for reports than for news.

5.4. Combination

Confidence measures, semantic relations and language model interpolation can be used in conjunction. In this part, the joint integration of these



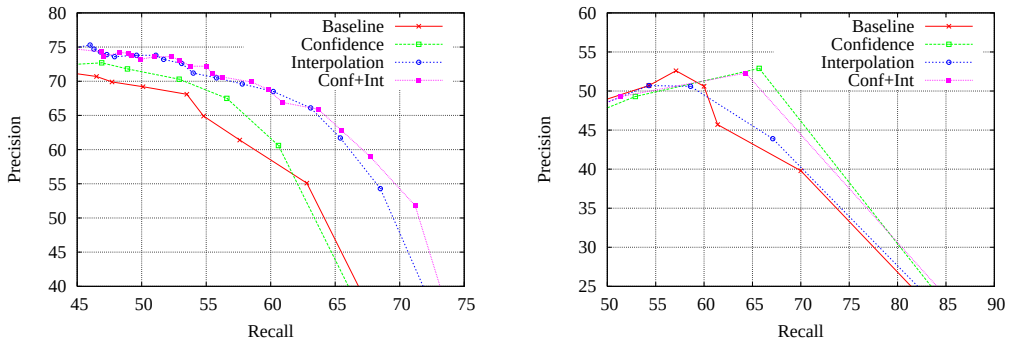
(a) *news* corpus (b) *reports* corpus

Figure 3: Interpolation - Recall/Precision curves

three elements is evaluated to check if their advantages can be combined.

Confidence + Interpolation. When confidence measures and interpolation techniques are integrated together, the improvements observed for both are combined for the *news* corpus. In this case, the CMs introduced are the ones that exhibit the most important amelioration (with $\delta_1 = \delta_2 = 1$) and the interpolation technique is the interpolation of counts with λ equals to 0.2. Figure 4 shows that the combination of the two cues leads to a greater improvement in the topic segmentation quality than the sole interpolation, especially for high recall values. Thus, we can conclude that the problems inherent in TV programs, i.e., transcription errors, potentially short segments and limited number of reiterations, are partially handled for the *news* corpus. For *reports*, unsurprisingly, comparable results are obtained with or without interpolation when confidence measures are used (cf. Fig. 4), interpolation alone yielding no improvement on this corpus. However it is important to note that the combination does not damage topic segmentation compared to the sole use of CMs.

Confidence + Semantics. Confidence measures can also be combined with semantic relations, whether syntagmatic or paradigmatic. For the corpus composed of broadcast news, the relations introduced are paradigmatic ones, selected thanks to the word level selection technique (2 relations per word). In this case, topic segmentation is slightly improved compared to the use of confidence measures or semantic relations alone. For reports, the combination does not lead to a better topic segmentation, compared with the use of



(a) *news* corpus (b) *reports* corpus

Figure 4: Integration of confidence measures and interpolation - Recall/Precision curves

Table 8: Integration of confidence measures and semantic relations

	news	*reports*
Baseline	59.7	54.9
Confidence	61.6	**58.3**
Semantics	60.9	55.7
Conf+Sem	**62.0**	**58.3**

confidence measures alone (cf. Table 8). This observation is not surprising as the semantic relations do not give real improvements for this corpus as explained in Section 5.2. However, as before, the use of two different cues does not damage the topic segmentation quality.

Semantics + Interpolation. Finally, semantic relations can also be combined with interpolation techniques. This combination, possibly associated with the use of confidence measures, does not improve the topic segmentation quality, either for the *news* or the *reports* corpus. As semantic relations and especially interpolation techniques both allow to overcome problems related to small segments, their combination is redundant.

6. Conclusion and future work

In this paper, we have improved and extended a probabilistic measure of lexical cohesion to include confidence measures and semantic relations, making use of language model interpolation. This results in a lexical cohesion

measure more robust to TV program specifics while being generic enough to be effective on different kinds of programs. First, it has been shown that the use of semantic relations and interpolation techniques improve the topic segmentation quality of TV programs divided in short segments and in which lexical repetition rates are low. It was also pointed out that the integration of confidence measures has more impact when the transcription quality is lower, as for the *reports* corpus. Finally, we demonstrated that these different elements can be used together to combine their advantages. Interestingly, it was also found that this combination never leads to a deterioration of topic segmentation.

These results clearly lead to the integration of other features, such as prosodic ones, to detect lexical stresses or speaker changes for example. Moreover, as the integration of confidence measures has a positive impact, we think it would be interesting to investigate the application of the topic segmentation method on intermediary outputs of the ASR system (such as word graphs or confusion networks) rather than on the final transcripts. Finally, as mentioned in Section 4.3, the structure of news programs is clearly hierarchical as, in most cases, the first reports all refer to the main news of the day and are therefore dealing with the same broad topic. However, in this work, only linear topic segmentation has been considered. Therefore, a long view prospect of this paper is to develop a method that can handle a hierarchical topic segmentation of our data.

Acknowledgments

This work was partially funded by OSEO, French state agency for innovation, in the framework of the Quaero project.

References

[1] M. Ostendorf, B. Favre, R. Grishman, D. Hakkani-Tür, M. Harper, D. Hillard, J. B. Hirschberg, H. Ji, J. G. Kahn, Y. Liu, E. Matusov, H. Ney, E. Shriberg, W. Wang, C. Wooters, Speech segmentation and spoken document processing, Signal Processing Magazine 25 (3) (2008) 59–69.

[2] P. V. Mulbregt, I. Carp, L. Gillick, S. Lowe, J. Yamron, Segmentation of automatically transcribed broadcast news text, in: Proceedings of DARPA Broadcast News Workshop, 1999.

[3] M. Utiyama, H. Isahara, A statistical model for domain-independent text segmentation, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2001.

[4] D. Beeferman, A. Berger, J. Lafferty, Statistical models for text segmentation, Machine Learning 34 (1-3) (1999) 177–210.

[5] H. Christensen, B. Kolluru, Y. Gotoh, S. Renals, Maximum entropy segmentation of broadcast news, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2005.

[6] S. Huet, G. Gravier, P. Sébillot, Un modèle multi-sources pour la segmentation en sujets de journaux radiophoniques, in: Proceedings of the 15e Conférence sur le Traitement Automatique des Langues Naturelles, 2008.

[7] R. Amaral, I. Trancoso, Topic indexing of TV broadcast news programs, in: Proceedings of the 6th International Workshop on Computational Processing of the Portuguese Language, 2003.

[8] A. Stolcke, E. Shriberg, D. Hakkani-Tür, G. Tür, Z. Rivlin, K. Sönmez, Combining words and speech prosody for automatic topic segmentation, in: Proceedings of DARPA Broadcast News Workshop, 1999.

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

[10] M. Halliday, R. Hasan, Cohesion in English, Longman, 1976.

[11] M. Xingwei, The relationship between cohesion and coherence, Journal of Foreign Languages 4.

[12] B. B. Klebanov, D. Diermeier, E. Beigman, Lexical cohesion analysis of political speech, Political Analysis 16 (4) (2008) 447–463.

[13] O. Manabu, H. Takeo, Word sense disambiguation and text segmentation based on lexical cohesion, in: Proceedings of the 15th International Conference on Computational Linguistics, 1994.

[14] R. Barzilay, M. Elhadad, Using lexical chains for text summarization, in: Proceedings of the Association for Computational Linguistics Intelligent Scalable Text Summarization Workshop, 1997.

[15] B. K. Boguraev, M. S. Neff, Lexical cohesion, discourse segmentation and document summarization, in: Proceedings of the 6th International Conference on Content-Based Multimedia Information Access, 2000.

[16] G. Hirst, A. Budanitsky, Correcting real-word spelling errors by restoring lexical cohesion, Natural Language Engineering 11 (2005) 87–111.

[17] D. Inkpen, A. Desilets, Semantic similarity for detecting recognition errors in automatic speech transcripts, in: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005.

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

[19] O. Ferret, Approches endogène et exogène pour améliorer la segmentation thématique de documents, Traitement Automatique des Langues 47 (2) (2006) 111–135.

[20] A. C. Jobbins, L. J. Evett, Text segmentation using reiteration and collocation, in: Proceedings of the 17th International Conference on Computational Linguistics, 1998.

[21] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, R. Harshman, Indexing by latent semantic analysis, Journal of the American Society for Information Science 41 (6) (1990) 391–407.

[22] T. Landauer, P. Foltz, D. Laham, Introduction to latent semantic analysis, Discours Processes 25 (1998) 259–284.

[23] F. Jelinek, R. L. Mercer, Interpolated estimation of Markov source parameters from sparse data, Pattern Recognition in Practice (1981) 381–397.

[24] M. Bacchiani, B. Roark, Unsupervised language model adaptation, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2003.

[25] S. Huet, G. Gravier, P. Sébillot, Morpho-syntactic post-processing of N-best lists for improved French automatic speech recognition, Computer Speech and Language 24 (4) (2010) 663–684.

[26] S. Galliano, G. Gravier, L. Chaubard, The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts, in: Proceedings of Conference of the International Speech Communication Association, 2009.

[27] C. D. Manning, H. Schütze, Foundations of statistical natural language processing, MIT Press, 1999.

[28] V. Claveau, P. Sébillot, From effiency to portability: Acquisition of semantic relations by semi-supervised machine learning, in: Proceedings of the 20th International Conference on Computational Linguistics, 2004.

[29] G. Grefenstette, Corpus-derived first, second and third-order word affinities, in: Proceedings of Euralex, 1994.

[30] B. Daille, Study and implementation of combined techniques for automatic extraction of terminology, MIT Press, 1996, pp. 49–66.

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

[32] G. Grefenstette, Sextant: Exploring unexplored contexts for semantic extraction from syntactic analysis, in: Proceedings of the 30st Annual Meeting of the Association for Computational Linguistics, 1992.

[33] J. Fayolle, F. Moreau, C. Raymond, G. Gravier, P. Gros, CRF-based combination of contextual features to improve a posteriori word-level confidence measures, in: Proceedings of the 11th International Conference on Speech Communication and Technologies, 2010.