On the use of Web resources and natural language processing techniques to improve automatic speech recognition systems

Gwénolé Lecorvé, Guillaume Gravier, Pascale Sébillot

IRISA
Campus de Beaulieu, 35042 RENNES, France
{gwenole.lecorve,guillaume.gravier,pascale.sebillot}@irisa.fr

Abstract

Language models used in current automatic speech recognition systems are trained on general-purpose corpora and are therefore not relevant to transcribe spoken documents dealing with successive precise topics, such as long multimedia streams, frequently tackling reports and debates. To overcome this problem, this paper shows that Web resources and natural language processing techniques can be effective to automatically collect a topic specific corpora from the Internet in order to adapt the baseline language model of an automatic speech recognition system. We detail how to characterize the topic of a segment and how to collect Web pages from which a topic-specific language model can be trained. We finally present experiments where an adapted language model is obtained by combining the topic-specific language model with the general purpose one to obtain new transcriptions. The results show that our topic adaptation technique leads to significant transcription quality gains.

1. Introduction

Using speech transcriptions is an effective way for the indexing of long multimedia streams, like 24h of TV or radio broadcast. To generate these transcriptions, current automatic speech recognition (ASR) systems are based on language models (LM) which gather word sequence probabilities, typically n-gram probabilities, and assist the system in discriminating utterances with the highest likelihood. In practice, these n-gram probabilities are estimated globally once and for all on large multi-topic corpora. However, since n-gram probabilities change with topics, these multi-topic LMs are not accurate to transcribe spoken documents successively tackling various topics, like broadcast news or debates.

To circumvent this problem, this paper proposes to use natural language processing techniques to automatically adapt a general purpose LM to any topic, using the Internet as an open resource to dynamically gather an adaptation corpus. The final goal in this paper is to improve the transcription quality of ASR systems on topic specific segments. To this end, experiments are carried out on a large set of various radio broadcast news shows.

The paper is organized as follows: Section 2 presents an overview of the complete LM adaptation technique. Works related with all the steps of the adaptation process are presented in Section 3. Sections 4 to 6 detail each key point of the proposed approach and experimental results are given in Section 7.

2. Overall approach

As presented in Fig. 1, the basic idea of our topic adaptation technique is to use the Internet as an open linguistic resource from which topic-specific texts can be retrieved to estimate new n-gram probabilities for almost any topic. The process described below is applied to single topic segments, previously transcribed with a baseline general-purpose LM, where such segments may come either from the thematic segmentation of a long speech stream or from smaller documents like broadcast news shows, as considered in this paper, or podcasts. For a segment, keywords are extracted based on information retrieval techniques in order to characterize the topic. These terms are used to form queries that are submitted to a Web search engine (*Yahoo!*). Retrieved pages are then browsed and filtered to make a corpus from which topic-specific LM probabilities are estimated before being combined with the probabilities from the general-purpose LM. Finally, new, and hopefully better, transcriptions are obtained from the ASR system using the adapted LM.

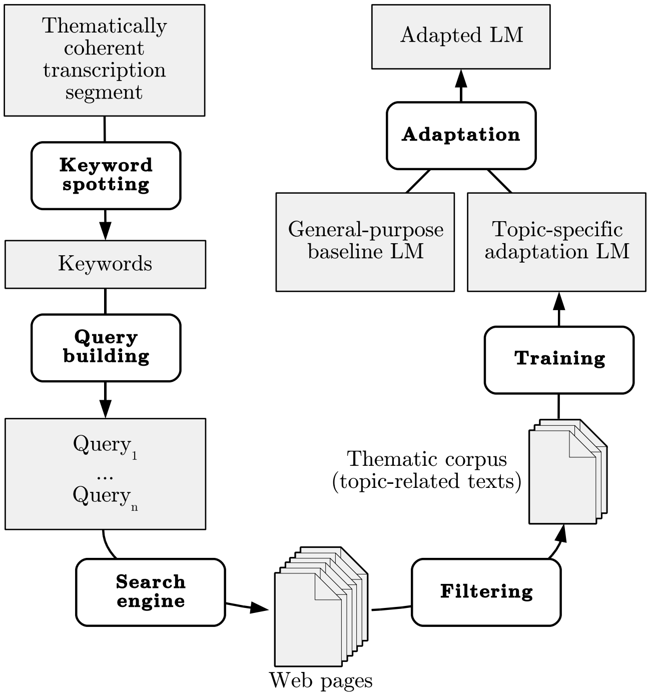
Thematically coherent transcription segment
Keyword spotting
Keywords
Query building
Query$_1$
...
Query$_n$
Search engine
Web pages
Filtering
Thematic corpus (topic-related texts)
Training
General-purpose baseline LM
Topic-specific adaptation LM
Adaptation
Adapted LM

Figure 1: Overview of our Web-based language model adaptation process.

This approach raises several questions at each step of the process. First, how to extract keywords which, on the one hand, characterize well all the thematic aspects of a segment and, on the other hand, are not too precise to return enough Web pages? Furthermore, possible transcription errors have to be considered in the keyword selection step. Then, how to combine these keywords into queries? Third, how many pages have to be retrieved and how to guarantee that these pages are related to the topic tackled in a segment? Finally, the combination of the general-purpose LM with the topic-specific LM is also an open problem that, however, will not be discussed in the scope of this paper which focuses primarily on the impact of Web-based resources to build topic-specific corpora.

Before discussing these many questions, we propose a rapid overview of the related works in the literature.

3. Related works

Since LMs are widely used tools, solutions to the problem of training topic specific LM are proposed in several domains. This section draws an outline of the most interesting related propositions.

In speech recognition, most current works aim at re-estimating n-gram probabilities from topic specific texts selected automatically. In several works, texts are selected among static collections of texts (Klakow, 2000; Chen et al., 2004). However, since these collections are limited, there are always topics for which no, or few, adaptation data can be found. These methods are therefore rather dedicated to cases where all the topics are known beforehand, which is a restrictive assumption. To overcome this limitation, the idea of using the Internet—an open linguistic resource—was introduced by Berger and Miller (1998). This resource is all the more interesting to process speech as it has been shown that texts coming from the Internet are more suited for spoken language modeling than classical texts such as newspapers (Vaufreydaz et al., 1999). Among the recent Web-based approaches, Sethy et al. (2005) propose to collect topic-specific texts from the Internet using both stochastic and information retrieval methods in order to update iteratively a background LM and a topic-specific LM. Even if interesting transcription quality gains are reported, this technique requires to train an initial topic-specific LM from a small collection of texts, which is not always possible. Moreover, experiments are carried out in a very specific and technical domain, namely the health care domain. In Suzuki et al. (2006), topic-specific corpora are jointly used to adapt n-gram probabilities and to enrich the vocabulary of an ASR system. This prevents from any conclusion on the influence of Web-based corpora for the sole LM adaptation task. Furthermore, results are presented on a very limited set of 5 segments of convenient sizes and from very specific domains quite different from the general-purpose baseline LM.

Apart from the ASR domain, the problem of gathering documents from the Web to build language models is also discussed in the field of information retrieval with text collections. The problem here is to sample a subset of a text collection using queries in order to build a LM as close as possible to the one that would be obtained from the entire collection (Callan et al., 1999; Monroe et al., 2002). Though not directly related to the problem of LM adaptation, these works interestingly raise questions about the best way to build queries and sample documents which best represent a topic or a domain from a text collection.

This literature survey points out that most related works focus on precise aspects of the topic LM adaptation task or restrict themselves to a few themes, whereas this paper presents a whole LM adaptation process, coming along with large scale experiments. Moreover, current works are rather based on information theory criteria, like entropy and mutual information, while interesting tools coming from the natural language processing domain are efficiently able to process themes. This paper precisely aims at using some of these tools to build topic-specific corpora.

4. Keyword spotting

In our approach, the first step consists in characterizing precisely the topic of a segment in order to retrieve texts from the same topic on the Internet. To do this, our idea is to extract keywords based on the *tf* $*$ *idf* criterion widely used in information retrieval (Salton, 1989). However, this criterion has been designed for regular texts rather than for automatic transcriptions, the latter containing misrecognized words, being case insensitive and lacking punctuation. The standard *tf* $*$ *idf* criterion is therefore adapted to take into account these specificities as well as those of the targeted application.

This section first introduces a general scheme for keyword spotting based on the *tf* $*$ *idf* criterion before presenting improvements for our framework.

4.1. *tf* $*$ *idf* criterion

Basically, the *tf* $*$ *idf* criterion aims at seeking words which discriminate well a text with respect to a reference set of various texts. To do this, two values are computed for each word w[1] of a transcript t: its frequency *tf*(w) in t and its inverse document frequency *idf*(w) which is related to the number of documents containing w in a reference corpus $\mathcal{C}$. The product *tf*(w) × *idf*(w) results in a score which is high for discriminating words. In practice, we compute these two values by

$$\textit{tf}(w) = \frac{\textit{freq}(w)}{\max\limits_{x \in t} \textit{freq}(x)} \text{ with } \textit{freq}(w) = \frac{|w|_t}{|t|} \qquad (1)$$

and

$$\textit{idf}(w) = \log \frac{|\mathcal{C}|}{\text{Card}\,\{d \in \mathcal{C} | w \in d\}} \ , \qquad (2)$$

where $|w|_t$ is the number of occurrences of w in t, $|t|$ is the number of words in t and $|\mathcal{C}|$ is the number of documents in $\mathcal{C}$. In our case, $\mathcal{C}$ is made of 800,000 articles from the French newspaper *Le Monde*, between 1987 and 2003. Furthermore, a normalization factor is applied for each word, resulting in a score $S(w)$ normalized between 0 and 1.

Since segments may be small and contain few word repetitions, *tf* $*$ *idf* word scores may not be satisfactory. A lemmatization technique is therefore applied to gather words sharing a same canonical form into a same word class. For example, plural names are reduced to their singular form,

[1] With the exception of stop words (prepositions, articles...).

$S(\ell)$	Label word	Word class ℓ
1.000	voile	voile
0.756	**adda**	adda
0.521	**bernadette**	bernadette
0.501	laïcité	laïcité
0.483	musulmans	musulmans, musulmane
0.449	photo	photo, photos
0.429	**sarkozy**	sarkozy
0.387	**chirac**	chirac
0.372	préfecture	préfecture
0.364	**serge**	serge

Table 1: List of the 10 keywords with the highest scores when using the standard *tf* ∗ *idf* criterion.

$S'(\ell)$		Label word	Word class ℓ
–	1.000	voile	voile
–	0.567	**adda**	adda
△	0.501	laïcité	laïcité
△	0.483	musulmans	musulmans, musulmane
△	0.449	photo	photo, photos
▼	0.391	**bernadette**	bernadette
△	0.372	préfecture	préfecture
△	0.330	**mimosa**	mimosa
△	0.329	tchador	tchador
△	0.326	carmélites	carmélites
▼	0.322	**sarkozy**	sarkozy
▼	0.290	**chirac**	chirac
▼	0.273	**serge**	serge

Table 2: List of the 10 keywords with the highest scores after having applied penalties on proper names.

conjugated verbs are reduced to their infinitive form... For each class ℓ, the score $S(\ell)$ is then computed on the lemmatized segment and with a lemmatized version of the reference corpus $\mathcal{C}$. Nevertheless, as queries will be formed with words from the segment rather than with lemmas, each class is represented by its most frequent constituent word.

In Table 1, we report an example of the 10 keywords with the highest scores extracted from a segment dealing with the problem of wearing a veil on identity photos in France. It can be observed that important words like *voile* (veil), *laïcité* (secularism), *musulmans* (Muslims) and photos are reported in the top 10 keywords. However, other well-ranked words (in bold) are related to the close context, or story, of the segment rather than to its topic. Though considering the story rather than the topic might be of interest, it often results in a very limited amount of adaptation data since keywords are too specific, thus making it impossible to reliably estimate an adapted language model. We therefore wish to discard such words from the keyword list which is achieved by modifying the standard *tf* ∗ *idf* criterion.

4.2. Modified *tf* ∗ *idf* criterion

To take into account specificities of the transcriptions, scores $S(\ell)$ are modified in several ways. First of all, the score for proper names is penalized. Then, word level confidence measures provided by the ASR system are considered to bias the *tf* ∗ *idf* scores.

The case of proper names

With the standard *tf* ∗ *idf* scores, proper names which do not describe the topic, *e.g.*, journalists names or trade marks, exhibit high scores due to the fact that these words are not frequent in the reference corpus $\mathcal{C}$. However, as mentioned previously, these terms are very specific and often yield too small corpora. One possibility would be to use a named entity detection program to remove proper names from the keyword lists. However, in some cases, keeping proper names as keywords might prove useful. For example, in a segment dealing with 9/11 terrorist attacks, the word *New-York* would probably help in collecting a valuable topic-specific corpus. Rather than the mere removal of proper names from the keyword list, we propose a smoother approach which consists in applying a penalty $p \in [0, 1]$ to the term frequency of proper names. Formally, the term frequency of a word class ℓ is given by

$$tf'(\ell) = \frac{\sum\limits_{w \in \ell} p_w}{|\ell|} \times tf(\ell) \quad , \quad (3)$$

$$\text{with } p_w = \begin{cases} 1 - p & \text{if } w \text{ is a proper name} \\ 1 & \text{else} \end{cases}$$

where $|\ell|$ is the number of lemmas ℓ. A new score $S'(\ell)$ is then obtained using the new term frequency $tf'(\ell)$. Since transcribed texts are lowercase and proper names cannot be directly detected, we rely on morphosyntactic tagging, which bind to each term its grammatical class, and on a dictionary to identify them: Common nouns with no definition in the dictionary are considered as proper names.

Table 2 presents the effect of the penalty on the keyword list previously given in Table 1, with p empirically set to 0.25. It appears clearly that proper names (in bold) are now less present among the best-scored keywords. However, this result is still imperfect. First, some proper names are not detected since they can also be considered as common nouns. For example, *mimosa* can refer to a plant, a record label, a micro-satellite acronym, and so on. Even if morphosyntactic tagging aims at resolving these ambiguities, it is not absolutely reliable. Furthermore, some proper names have such a high baseline score that the penalty has no real impact. As an example, *adda*, which is a misrecognized proper name, is still ranked second. These results might highlight that the best solution would rather be to keep only common nouns, adjectives and non-modal verbs. However, as discussed previously, this would be too restrictive since proper names can also be keywords. To overcome this problem, the idea of using a penalty could be generalized by applying different penalties to each grammatical class. However, tuning these penalties would be burdensome since no clear criterion to compare the quality of two keywords can be defined, apart from the final word error rate which requires the estimation of the adapted LM and the generation of a new transcription.

$\sigma(\ell)$		Label word	Word class ℓ
–	0.992	voile	voile
△	0.500	laïcité	laïcité
△	0.458	musulmans	musulmans, musulmane
▼	0.454	**adda**	adda
–	0.428	photo	photo, photos
–	0.390	**bernadette**	bernadette
–	0.371	préfecture	préfecture
△	0.328	tchador	tchador
△	0.325	carmélites	carmélites
△	0.321	**sarkozy**	sarkozy
▼	0.294	**mimosa**	mimosa

Table 3: List of the 10 keywords with the highest scores after inclusion of confidence measures within the score computation.

The case of transcription errors

Automatic transcriptions inevitably contain errors whose impact on the LM adaptation process has to be carefully considered. Transcription errors may result in biased word class scores and in a keyword list thematically unrelated to the topic of the segment. We rely on word level confidence measures provided by the ASR system to dampen the impact of transcription errors. Confidence measures, in $[0, 1]$, are associated with each word at the output of the ASR system and indicate how confident is the system in its output, 1 indicating a total confidence in the decision. For each lemma, the score $S'(\ell)$ is modified based on the confidence measures of the word occurrences for ℓ according to

$$\sigma(\ell) = \big[\alpha + (1-\alpha)\, c_\ell\big] \times S'(\ell) \qquad \text{with } \; c_\ell = \frac{\sum\limits_{w\in\ell} c_w}{|\ell|} \tag{4}$$

where $c_w \in [0, 1]$ is the average confidence measure over all the occurrences of w. The parameter α, empirically set to 0.25, limits the influence of confidence measures since they are not absolutely reliable (Huet et al., 2007).

In our example, it appears that the two misrecognized words *adda* and *mimosa* are effectively pushed down, whereas the other words are less affected by the penalties, as illustrated in Table 3. Nevertheless, we can notice that the list of keywords with the highest scores still contains the misrecognized word *adda*. Due to the complexity of the whole adaptation process and to the number of its intrinsic parameters, no optimization has been done to improve the use of confidence measures. Thus, it could be interesting to study the relationship between the limiting factor α and reliability measures frequently used for confidence measures, *e.g.*, normalized cross entropy (Huet et al., 2007) or confidence error rates (Wessel et al., 2001).

We have observed that the proposed modified *tf* ∗ *idf* scores lead, on average, to consider keywords which characterize well most aspects of the topic of a segment, even though this opinion is subjective as no objective criterion can yet be defined. We present in the next section how queries can be formed from the selected keywords.

#1	voile				
#2		laïcité			
#3	voile	laïcité			
#4	voile	laïcité	musulmans		
#5	**voile**	**laïcité**		**adda**	
#6	voile	laïcité			photo
#7	**voile**		**musulmans**	**adda**	
#8	voile		musulmans		photo
#9	**voile**			**adda**	**photo**
#10		**laïcité**	**musulmans**	**adda**	
#11		laïcité	musulmans		photo
#12		**laïcité**		**adda**	**photo**
#13			**musulmans**	**adda**	**photo**
#14	**voile**	**laïcité**	**musulmans**	**adda**	
#15	**voile**	**laïcité**	**musulmans**	**adda**	**photo**

Table 4: Example of queries formed based on subsets of the 5 best-scored keywords. Queries in bold include misrecognized words.

5. Querying the Internet

Considering a sorted list of keywords, the problem is now to form one or several queries in order to gather Web pages related to the topic of the segment. Two opposite considerations must be taken into account: queries must be precise enough to return topic-specific pages while they must also return enough pages to estimate reliably n-gram probabilities.

As mentioned in Sethy et al. (2005), the number of keywords included in a query is crucial with regard to the number of matching pages found. In preliminary experiments, we have observed that, on average, a query returns 40,000 hits when using 3 keywords, 6,000 hits with 4 keywords, and 600 hits with 5 keywords. Thus, in practice, the set of keywords selected is limited to the five words used to represent the five lemmas with the highest scores $\sigma(\ell)$. As in query-based sampling techniques (Callan et al., 1999; Monroe et al., 2002), various simple queries are formed based on subsets on the selected keywords. For example, one query is composed of the two best keywords while another one is composed of the first and third keywords. In our experiments, fifteen simple queries containing one, two or three keywords are formed for each segment. This strategy offers the advantage of maximizing the probability of having at least one relevant query, even when transcription errors are present. For example, queries resulting from the top-5 keywords in Table 3 are listed in Table 4. In this example, the retrieval of a topic-specific corpus does not fail even if the misrecognized word *adda* is not discarded from the best keywords: out of the 15 proposed queries, 7 queries do not include the word *adda* and are well related to the topic.

This method could probably be improved. First, no deep experiment has been carried out to measure how effective a method can be when submitting even more queries. In our method, many more simple queries could be formed if more than 5 different keywords were selected. However, apart from the runtime issue, disregarded in this study, increasing the number of queries is probably balanced by the quality of keywords which may be uncertain after a given rank, leading possibly to Web pages thematically far from the segment. Hence, the problem is to know how many keywords have to be kept for a given topic rather than to determine the right number of queries. Moreover, it would be interesting to find out how to form simple queries more cleverly. For example, it could be interesting to study the effect of gathering words sharing a same meaning in the context of the segment. For example, in Table 3, *veil* and *chador* share a semantic link since a chador is kind of veil. The combination of these two words in a same query may strengthen the thematic quality of returned Web pages, whereas it may also be redundant. The question is still open.

6. Selection of Web pages

The querying strategy that we have just described results in a large number of hits, frequently over a million. Obviously, keeping all the matching Web pages is impossible. Furthermore, not all of them are relevant. We therefore develop a strategy to filter out the matching pages in order to select a sufficient amount of relevant pages.

6.1. Number of documents retrieved

In preliminary experiments, 20, 50, 100, 200 and 400 of the pages returned by the queries have successively been used to build topic-specific corpora and to train adaptation LMs, where 200 pages, about 800,000 words, seem like a good compromise between speed and accuracy. If the number of documents considered is close to what is reported in the information retrieval literature (Callan et al., 1999; Monroe et al., 2002), it is more unusual within the ASR domain. For example, in Suzuki et al. (2006), several thousands Web pages are retrieved. However, even if broader experiments should be carried out with larger numbers of pages, it is assumed that the relevance of a Web page decreases with its rank in the list of answers. Therefore, continuously increasing the number of documents retrieved from the Web is not a cure-all and the quality of a document with respect to the topic at hand must be considered.

6.2. Filtering documents

In many works, relevant texts are selected based on their accordance with topic-specific LMs (Nisimura et al., 2001; Sethy et al., 2005; Bulyko et al., 2007) already available. However, in our application framework, no initial topic specific LM is available. We therefore propose to make use of the *tf* ∗ *idf* scores to measure the similarity between documents found on the Web and the initial transcription available. However, this method requires some adaptation to deal with noisy Web pages.

First of all, the content of each page is filtered to remove not only HTML tags but also irrelevant text parts, like advertisements, copyright notifications, menus... Selection of the relevant pieces of texts in the Web pages mainly relies on statistics on punctuation frequencies, on average sentence lengths and on the frequency of non-alphanumeric characters.

Web pages are then compared with the initial transcription of the segment under consideration. Given the initial automatic transcription of the segment t as a vector of scores $\sigma_t(\ell)$ and the content of each Web page p as a vector of scores $S'_p(\ell)^2$, the similarity between t and p is the cosine distance defined as

$$\mathrm{sim}(t,p) = \frac{\sum\limits_{\ell \in t \cap p} \sigma_t(\ell) \times S'_p(\ell)}{\sqrt{\sum\limits_{\ell \in t} \sigma_t(\ell)^2 \times \sum\limits_{\ell \in p} S'_p(\ell)^2}} \ . \qquad (5)$$

Pages whose similarity is below a given threshold are discarded from the topic-specific corpus and the remaining pages are used to train a topic-specific LM which is linearly interpolated with the baseline LM in order to get an adapted LM. Preliminary experiments highlighted that the similarity threshold must not be too high to be able to reliably estimate n-gram probabilities on the corpus. Indeed, the pages filtering reduces significantly the average size of the adaptation corpora: by setting empirically the threshold to 0.08, after optimization on a development set, the adaptation corpora consist now on average in about 200,000 words as opposed to the initial 800,000 words.

6.3. Analysis of the resulting corpora

Given that the theoretical highest similarity is 1.00, this threshold may look low. However, it results in thematically relevant adaptation corpora. Figure 2 shows the average similarity between a segment and the documents selected for different numbers of pages retrieved and different similarity thresholds. These results show that, for each threshold value, the average similarity decreases when more Web pages are considered, as previously assumed in Section 6.1. From a different perspective, Figure 3 reports the average similarity measured between a segment and its corresponding adaptation corpora obtained with various settings. On the one hand, it appears that the adaptation corpora overall similarities are much better than the average similarities computed separately on each of their constituent pages. This conclusion tends to prove that, even if several simple queries are submitted, each one describing a part of the topic, the combination of their respective hits results in a corpus thematically consistent with the segment. On the other hand, one can notice that the average similarity of the adaptation corpora, seen as a function of the number of pages, behaves differently depending on the threshold used to select the relevant pages. Indeed, over a given threshold of about 0.04, the similarity of the adaptation corpora increases as the number of pages increases, whereas, below this threshold, the similarity decreases as in Figure 2.

In any case, it appears that the similarity tends to converge as the number of pages retrieved increases. The same trend was observed for the number of documents after filtering. This is explained by the fact that, pages being considered in the order assigned by the search engine, low rank pages correspond to poor matches and are therefore rarely selected as relevant with respect to the original transcription. Finally, let us note that the same conclusions have been made when

[2] Since no confidence measure exists for a Web page, the scores $\sigma_p(\ell)$ cannot be computed.

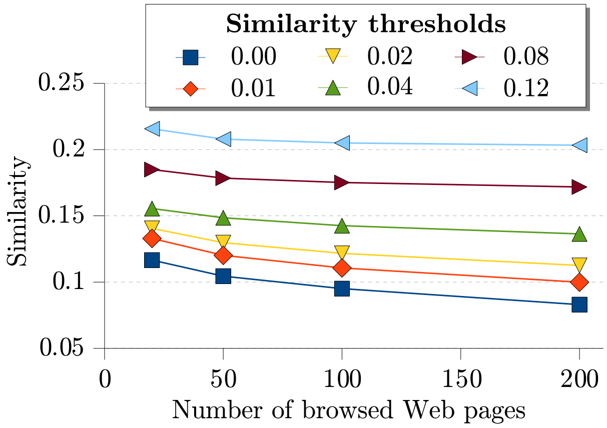

Figure 2: Average similarity between segments and each selected Web page as a function of the number of pages retrieved and of the similarity threshold.

replacing the original ASR transcription by the reference one, even though the similarity between the former and the latter is only 0.72.

7. Experiments

Experiments were carried out on about 6 hours of French radio broadcast news data coming from the ESTER evaluation campaign (Galliano et al., 2005). A total of 172 segments containing a single story were manually extracted from three different broadcasters, namely *France Inter*, *France Info* and *RFI*. These segments, whose respective size ranges from 30 to 2,000 words, cover various topics, such as the war in Iraq, national politics or sports. Furthermore, segments are divided into a development set (91 segments), used to tune the different parameters, and a test set, used for our back-end experiments.

Each segment was transcribed with our ASR system, described in Huet et al. (2007), which is based on a 4-gram general-purpose LM and a $64,000$ word vocabulary. For each segment, an adapted LM is computed following the adaptation technique previously exposed, while the vocabulary remains untouched.

Two performance measures are used to evaluate the impact of LM adaptation. On the one hand, we compare the perplexity of the reference transcription with the initial and adapted LMs respectively, in order to appraise the topic modeling quality of each LM. Since perplexity is linked to the n-gram probability distribution entropy, it is expected to be as low as possible. However, decreasing the perplexity of the LM does not necessarily translate into a better automatic transcription. Therefore, on the other hand, we compare word error rates (WER) of automatic transcriptions obtained respectively without and with LM adaptation. This last measure does not evaluate the quality of the sole LM as the first one but rather that of the entire ASR system.

Experimental results are presented in Table 5 (perplexity) and in Table 6 (WER). Language model adaptation based on text collections retrieved from the Web results in a significant perplexity improvement of about 14.5% relative and a WER gain of 0.2 absolute. This last result is statistically significant as it has been measured with p-values of respectively 0.0007 for the paired t-test and of 0.0001 for the paired Wilcoxon test. More precisely, the perplexity decreases after adaptation for 95% of the segments while the WER is improved for only 37% of the segments. Furthermore, one can notice that the best WER variation is obtained on shows coming from *RFI* while consistently lower variations are measured on *France Inter* and *France Info*. This can probably be explained by the fact that *RFI*, an international French speaking radio station, addresses many topics not included in the training data of the ASR system while the other (national) radio stations does not. In particular, *RFI* tackles many topics related to Africa, where this radio station is widely broadcasted, while national stations such as *France Inter* and *France Info* mostly deal with national and major international matters, these last matters being well represented in the training data of the baseline LM.

Table 7 presents word error rates measured on lemmatized lexical words (LER), *i.e.*, nouns, adjectives and non-modal verbs reduced to their respective lemmatized forms. Compared to the word-level results of Table 6, gains are more than doubled for each broadcaster and on average. A detailed analysis of the results highlights that adapted LMs mainly improves the transcription of "thematic" words, *i.e.*, terms which are frequently used in a given domain. This conclusion is illustrated by the example given in Table 8 which compares the transcription of an utterance, taken from the previous example dealing with secularism and wearing a veil, obtained without and with adaptation respectively. In this example, the small word sequence "*tête nue*" (bareheaded), misrecognized with the baseline LM, is correct after adaptation even though neither *tête* nor *nue* appeared in the keyword list. Nevertheless, we noticed that, in some cases, WER improvements due to adaptation are offset by grammatical errors that do not appear when using the baseline LM. This idea is backed up by the example reported in Table 9. In this second example, extracted from a segment dealing with strikes in transport services, the reference word *trams* was originally transcribed by the homophone *trames* (frames), grammatically correct but far from the topic. With the adapted LM, this mistake is corrected but the transcription remains erroneous: *trams* is now transcribed by its singular form *tram*, another homophone. This error is due to the fact that only the singular form *tram* can be found in the adaptation corpus. We also observed the frequent occurrence of new errors on ordinary words, like modal verbs, prepositions, conjunctions...This problem can be explained by the small size of thematic corpora which leads to badly estimate the probability of n-grams containing ordinary words. Even if using larger corpora may solve the problem, we rather think that a different adaptation method should be used instead of linear interpolation, *e.g.*, MDI adaptation (Federico, 1999).

8. Discussion

In this paper, we have demonstrated that Web resources and natural language processing techniques can be successfully used to improve an ASR system by adapting the language model. We have especially shown that, depending on few

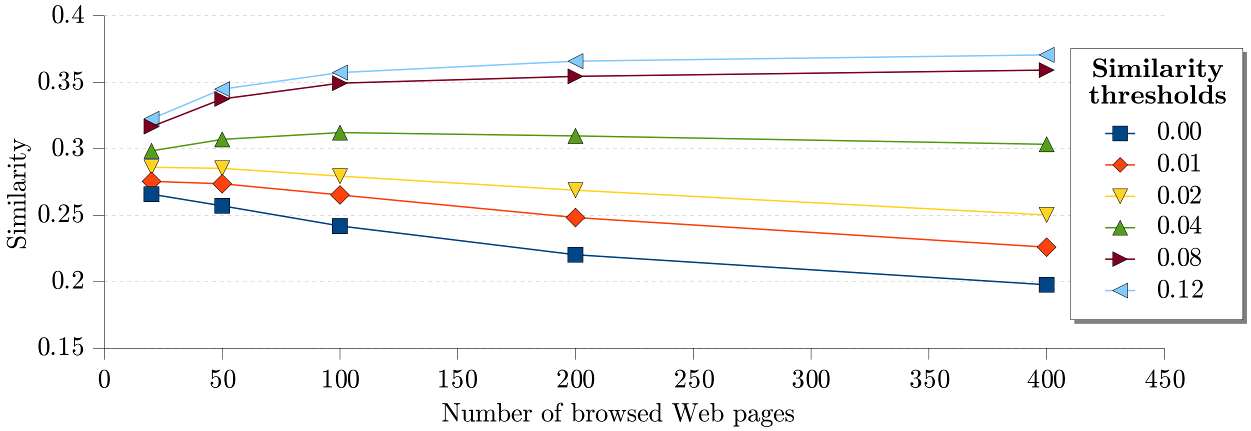

Figure 3: Thematic similarity between segments and their respective topic-specific corpora as a function of the number of pages retrieved and of the similarity threshold.

Broadcaster	Baseline LM	Adapted LM	Variation
France Inter	141.4	123.8	−9.9 %
France Info	138.7	106.9	−19.5 %
RFI	116.9	102.1	−11.8 %
Average	132.4	109.5	−14.5 %

Table 5: Perplexities measured on the test set with the baseline LM and with the adapted LM for each broadcaster and on average.

Broadcaster	Baseline LM	Adapted LM	Variation
France Inter	19.87	19.74	−0.13
France Info	21.74	21.59	−0.15
RFI	23.47	23.24	−0.23
Average	21.66	21.49	−0.17

Table 6: Word error rates observed on the test set when using the baseline LM and the adapted LM successively for each broadcaster and on average.

Broadcaster	Baseline LM	Adapted LM	Variation
France Inter	17.22	16.65	−0.57
France Info	19.91	19.50	−0.42
RFI	21.83	21.33	−0.50
Average	19.61	19.11	−0.50

Table 7: Word error rates measured on lemmatized lexical words when using the baseline LM and the adapted LM successively for each broadcaster and on average.

additional considerations like confidence measures, natural language processing techniques can be adapted to work on transcribed, unreliable, texts and that the innumerable linguistic resources present on the Internet can be effectively exploited to build topic-specific LMs by sampling queries based on automatically extracted keywords. Experiments result in a better thematic quality of the adapted LMs with respect to the baseline LM. This improvement is translated by encouraging WER gains, resulting mainly from a better transcription of thematic words. However, besides these results, discussions introduced along each key point highlight the remaining difficulties and outline future works.

(R)	`tête nue` (*bareheaded*)
(B)	**`sept mille`** (*seven thousands*)
(A)	`tête nue` (*bareheaded*)

Table 8: Comparison of an utterance transcribed with our baseline LM (B) and our adapted LM (A) according to the reference transcription (R).

(R)	`le service des trams est affecté` (*the tram service is affected*)
(B)	`le service des` **`trames`** `est` **`affectée`** (*the frame service is affected*)
(A)	`le service des` **`tram□`** `est` **`affectée`** (*the tram service is affected*)

Table 9: Comparison of a second utterance transcribed with our baseline LM (B) and our adapted LM (A) according to the reference transcription (R).

First, our approach rely on the assumption that segments are thematically coherent. However, for some segments, the adaptation process results in a deteriorated transcription. It therefore seems interesting to be able to diagnose, for any segment, whether topic adaptation is necessary or not. To do this, the link between the impact of LM on the WER and the segment thematic qualities should first and foremost be studied. However, this task is all the more difficult since the notion of theme is still not clear in our case: for example, a report on the war in Iraq, dealing with several aspects of the war (combats, the role of the non-profit organizations, etc.) can be considered as a single theme, or not.

Second, natural language processing techniques could still be more deeply integrated with speech processing algorithms, especially during the keyword spotting stage which appears to be crucial since the whole adaptation technique is based on keywords. On the one hand, extracted keywords are simple terms and are therefore sometimes not discriminative enough. For example, in a segment dealing with the 70's, the two words *flower* and *power* describe much more the topic when considered together than when considered separately. To do this, complex term extraction techniques (Daille, 2003) could be adapted in order to handle automatic transcriptions. On the other hand, despite the fact that words are grouped into word classes, the length and the style of some segments still imply only a few repetitions. This problem could be overcome with the help of information about semantic links between words (Grefenstette, 1994).

Finally, a deeper and better use of the thematic corpora is necessary. Given the sparseness of general-purpose n-grams in the adaptation corpora, it would be interesting to attach different importance levels to n-gram probabilities depending on their thematic specificity. For example, n-gram probabilities for general-purpose word sequences should be left unchanged from the baseline LM while only n-gram probabilities where topic specific words are involved should be modified. Topic-specific corpora could also be helpful to determine keywords. For example, in a segment dealing with a given sport, technical terms are frequently used, whereas more general terms, such as the name of the sport, are rarely pronounced, leading therefore to consider too specific keywords. An interesting solution to this problem consists in iterating the adaptation process: a first step consists in gathering a topic specific corpus as described in this paper from which new keywords are extracted to iterate the corpus gathering step.

9. References

A. Berger and R. Miller. 1998. Just-in-time language modelling. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 705–708.

I. Bulyko, M. Ostendorf, M. Siu, T. Ng, A. Stolcke, and Özgür Çetin. 2007. Web resources for language modeling in conversational speech recognition. *ACM Transactions on Speech and Language Processing*, 5(1):1–25.

J. Callan, M. Connell, and A. Du. 1999. Automatic discovery of language models for text databases. *SIGMOD Record*, 28(2):479–490.

L. Chen, J.-L. Gauvain, L. Lamel, and G. Adda. 2004. Dynamic language modeling for broadcast news. In *Proceedings of International Conference on Speech and Language Processing (ICSLP)*, pages 1281–1284.

B. Daille. 2003. Conceptual structuring through term variations. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment*, pages 9–16.

M. Federico. 1999. Efficient language model adaptation through MDI estimation. In *Proceedings of Eurospeech*, volume 4, pages 1583–1586.

S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier. 2005. The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In *Proceedings of Eurospeech*, pages 1149–1152.

G. Grefenstette. 1994. Corpus-derived first, second and third-order word affinities. In *Proceedings of EURALEX*, pages 279–290.

S. Huet, G. Gravier, and P. Sébillot. 2007. Morphosyntactic processing of N-best lists for improved recognition and confidence measure computation. In *Proceedings of Interspeech*, pages 1741–1744.

D. Klakow. 2000. Selecting articles from the language model training corpus. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 1695–1698.

G. Monroe, J. French, and A. Powell. 2002. Obtaining language models of Web collections using query-based sampling techniques. In *Proceedings of Hawaii International Conference on System Sciences (HICSS)*, volume 3.

R. Nisimura, K. Komatsu, Y. Kuroda, K. Nagatomo, A. Lee, H. Saruwatari, and K. Shikano. 2001. Automatic n-gram language model creation from Web resources. In *Proceedings of Eurospeech*, pages 2127–2130.

G. Salton. 1989. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc.

A. Sethy, P. G. Georgiou, and S. Narayanan. 2005. Building topic specific language models from Webdata using competitive models. In *Proceedings of Interspeech*, pages 1293–1296.

M. Suzuki, Y. Kajiura, A. Ito, and S. Makino. 2006. Unsupervised language model adaptation based on automatic text collection from WWW. In *Proceedings of Interspeech*, pages 2202–2205.

D. Vaufreydaz, M. Akbar, and J. Rouillard. 1999. Internet documents: A rich source for spoken language modeling. In *Proceedings of Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 277–280.

F. Wessel, R. Schlter, K. Macherey, and H. Ney. 2001. Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 9(3):288–298.