AN UNSUPERVISED WEB-BASED TOPIC LANGUAGE MODEL ADAPTATION METHOD

Gwénolé Lecorvé, Guillaume Gravier and Pascale Sébillot

IRISA, 263 av. Gén Leclerc Campus universitaire de Beaulieu, 35042 RENNES, France
{gwenole.lecorve,guillaume.gravier,pascale.sebillot}@irisa.fr

ABSTRACT

This paper focuses on a solution to better adapt ASR systems, whose language models (LM) are usually trained on topic-independent corpora, to new topics, in particular in the case of broadcast news. We propose a new complete and fully unsupervised technique that selects keywords from each segment using information retrieval methods, to build a thematically coherent adaptation corpus from the Internet. The LM used for the initial transcription is then adapted before rescoring word lattices. Experimental results demonstrate the validity of the proposed adaptation technique with a significant reduction of the perplexity after LM adaptation. Word error rates are also improved in some cases though to a lesser extent.

***Index Terms*—** Speech recognition, natural languages, Internet

1. INTRODUCTION

Most current ASR systems rely on n-gram language models (LM) where the word sequence probabilities are estimated once and for all on a large topic-independent text corpus. However, these probabilities change between topics, which is frequent in a long speech stream, especially in broadcast news (BN). A solution to circumvent this problem consists in modifying the vocabulary and adapting the LM to each topic found in a document. In this paper, we focus on LM adaptation and disregard the complex problem of vocabulary adaptation.

Most works on topic LM adaptation aim at getting a topic-specific LM whose n-gram probabilities are then mixed with a general purpose LM. Two different methods are proposed: some works adopt a supervised approach which consists in selecting a specific LM in a set of pre-calculated thematic LMs [1], whereas, in other studies, a topic-specific LM is trained on a dynamically built thematic corpus [2].

This paper focuses on LM adaptation for thematically coherent segments of transcript. These segments may come from thematically segmented multimedia streams, as in [3], or from shorter multimedia documents dealing with one single topic, like audio or video podcasts. The proposed approach seeks to build thematic corpora by collecting texts from the Internet, which is an interesting source to model spoken language [4]. Moreover, the use of the Internet as an open linguistic resource does not imply any limitation on the number and on the nature of topics that can be processed—contrary to the use of static collection of texts, as in [2, 3].

Benefits of Web data for LM adaptation have already been studied in several works. In [5], interesting Word Error Rate (WER) gains are obtained by collecting texts from the Internet using both stochastic and Information Retrieval (IR) methods. However, this technique is supervised since an initial preexistent topic-specific set of documents is required, contrary to our work. An unsupervised approach is presented in [6]. Nevertheless, this work studies the use of Web data jointly for the LM adaptation task and the ASR system vocabulary enrichment. Hence, no clear conlusion can be drawn concerning the sole LM adaptation task, on which we focus. More generally, related topic adaptation techniques usually seek to retrieve several thousands of Web pages, whereas our method handles much less adaptation data. Furthermore, their experiments are often led on a few very specific topics or segments[1], which prevents any deep analysis of the results. In this paper, experiments are carried out on a large variety of transcripts dealing with various topics.

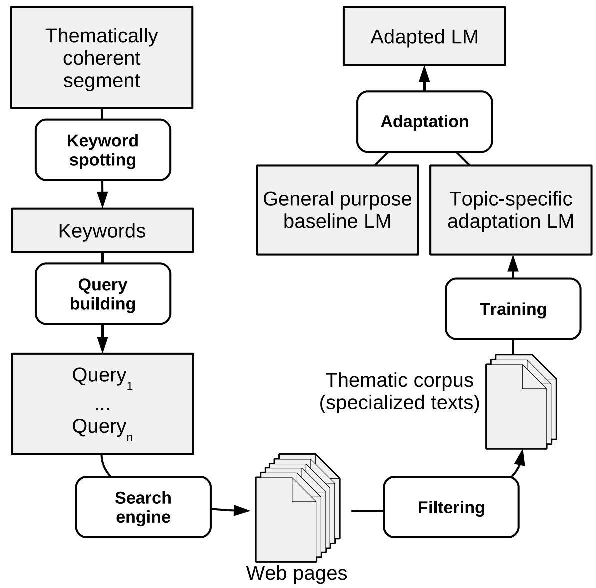

Fig. 1. Overview of the Web-based LM adaptation process.

The paper is organized as follows: Section 2 details the key points of the adaptation technique while Section 3 presents experimental results.

2. ADAPTATION METHOD

Our topic-based adaptation technique is presented in Fig. 1. First, for a given segment transcribed with the general purpose LM, keywords are selected in order to characterize the topic. These terms are then used to form queries that are submitted to a Web search engine (*Yahoo!*). The retrieved pages are browsed to form a corpus from which topic-specific LM probabilities are estimated before combining the topic-specific LM with a general purpose one.

[1]In [5], LM adaptation is processed for the sole health care domain and, in [6], only 5 segments are studied.

Several questions arise to implement this adaptation scheme. First, keywords must be significant enough to characterize well all the thematic aspects of a segment. However, they must not be too precise if we want to form queries able to retrieve enough Web pages. Furthermore, possible transcription errors have to be considered. Second is the question of how to derive one or more efficient queries from the selected discriminating terms. Then, one has to define a strategy to select relevant pages among those returned in order to get a large enough, yet thematically homogeneous, adaptation corpus. Finally, the combination of the general purpose LM with the topic-specific LM is also an open problem. In this study, focused on the feasability of the proposed approach, we have chosen to use linear interpolation which is a well known technique though far from optimal.

In the remainder of this section, we discuss our solutions for each step of the process. For illustration purposes, we report for the different parameters values that were either empirically set on a collection of 22 segments or found to be optimal on the development set described in Section 3.

2.1. Keyword spotting

Our basic idea to extract keywords from a transcript is based on the *tf* ∗ *idf* criterion widely used in IR. For each word w, we measure the normalized frequency

$$\mathit{tf}(w) = \frac{\mathit{freq}(w)}{\max_{x \in t} \mathit{freq}(x)} \text{ with } \mathit{freq}(w) = \frac{|w|_t}{|t|} \tag{1}$$

of a word w in a transcript t, and the inverse document frequency

$$\mathit{idf}(w) = log\left(\frac{|\mathcal{C}|}{Card\,\{d \in \mathcal{C} | w \in d\}}\right) \tag{2}$$

which is related to the number of documents containing w in a reference corpus $\mathcal{C}$[2]. In the above equations, $|w|_t$ 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 practice, lemmas are considered rather than words. Words sharing the same lemma are therefore gathered into a class ℓ for which a score $S(\ell)$ is computed. As we need words rather than lemmas to build queries, each class is represented by its most frequent constituent word. Moreover, the *tf* ∗ *idf* scores are normalized to provide scores ranging from 0 to 1, high scores corresponding to characteristic terms.

Finally, the scores $S(\ell)$ are modified to take into account specificities of the transcriptions. First, proper names which do not appear in the reference corpus $\mathcal{C}$ are very salient according to the *tf* ∗ *idf* criterion. However, they tend to yield very specific adaptation corpora, too small and too specific to be valuable. Hence, penalties are applied to each proper name by multiplying its score by a coefficient empirically set to 0.75. In our experiments, proper names are detected based on morpho-syntactic tags and a dictionary, *i.e.* nouns with no definition in the dictionary are considered as proper names. Even if no precise tuning of this factor has been done, current results show that this strategy leads to improve by $6.5\,\%$ relative the decrease of perplexity with respect to the use of the classical *tf* ∗ *idf* criterion. Second, possible transcription errors may bias the computation of some word class scores, possibly resulting in irrelevant documents in the topic-specific corpus. Contrary to [6] where no approach is proposed to overcome this problem, we modify the initial scores according to the confidence measures provided by the ASR system. The new score for a lemma ℓ is defined as

$$\sigma(\ell) = \alpha\ S(\ell) + (1-\alpha)\ c_\ell\ S(\ell) \text{ with } c_\ell = \frac{1}{|\ell|}\sum_{w\in\ell} c_w \tag{3}$$

where c_w is the average confidence measure over all the occurrences of w and $|\ell|$ is the number of words in the class ℓ. The parameter α, empirically set to 0.25, limits the influence of c_ℓ since confidence measures are not always reliable [7]. By including confidence measures into the computation of *tf* ∗ *idf*, we observed a $8.5\,\%$ relative gain of the decrease of perplexity as opposed to the standard technique.

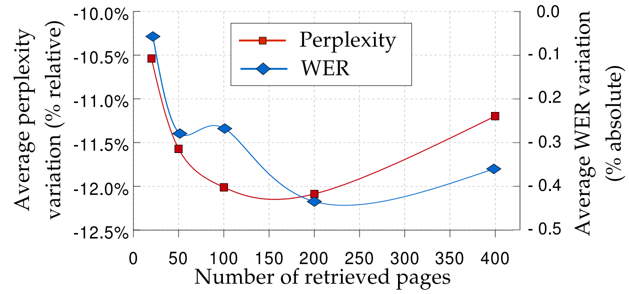

Fig. 2. Influence of the number of pages on perplexity and WER.

2.2. Querying

Based on (3), a sorted list of keywords is obtained from which we wish to derive one or several queries to gather Web pages related to the current topic. Combining the best keywords in a single query which both describes entirely a topic and always returns enough pages turns out impossible. As mentioned in [5], we also observed that queries containing more than six keywords resulted in too few hits to build an adaptation corpus. Hence, the use of several simple queries, combining 2 or 3 keywords, has been studied in preliminary experiments. The set of keywords selected is therefore limited to the few first words and various simple queries are made by combining subsets of the selected keywords. For example, the first query is composed of the two best keywords while another one is composed of the first and the third keywords. It was observed that successively using 1, 3, 5 and 15 simple queries resulted in a decrease of perplexity as the number of queries increased: a $30\,\%$ relative improvement of the decrease of perplexity was obtained with 15 queries. This can be explained by the fact that this strategy offers the advantage of maximizing the probability of having at least one relevant query, even when transcription errors are present.

2.3. Document selection

Using multiple simple queries usually returns a large number of hits, frequently over a million, from which only a fraction is relevant. Therefore, some filtering strategy is necessary to select a sufficient amount of relevant pages.

In a preliminary experiment, we measured perplexity and WER variations with respect to the number of retrieved Web pages (Fig. 2). The results show that between 50 and 400 pages are needed to get a good topic-specific corpus, 200 pages being a good compromise. This corresponds in average to 240k words[3]. Notice that this value

[2] 800k articles from the French newspaper *Le Monde*, 1987–2003.

[3] Since the text of a Web page is noisy, we do not simply remove all HTML

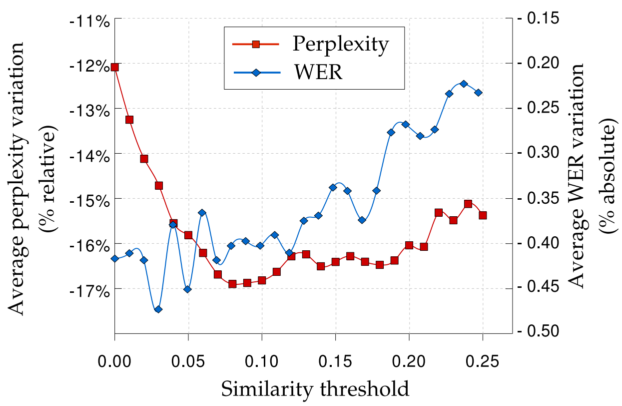

Fig. 3. Influence of the similarity threshold on perplexity and WER.

is much lower than the one used in related works [5, 6]. As a consequence, runtime keeps reasonable eventhough it has not been optimized as this is not a primary issue.

Pages are selected based on their similarity with the segment. A thematic similarity $\mathrm{sim}(t,p)$ is computed between a segment t and the content p of each Web page. Considering t and p as vectors of scores $S(\ell)$, this measure is computed by a cosine distance:

$$\mathrm{sim}(t,p) = \frac{\sum\limits_{\ell \in t \cap p} S_t(\ell) \times S_p(\ell)}{\sqrt{\sum\limits_{\ell \in t} S_t(\ell)^2 \times \sum\limits_{\ell \in p} S_p(\ell)^2}} \qquad (4)$$

where $S_t(\ell)$ and $S_p(\ell)$ are the scores of the word class ℓ in t and p respectively. Pages whose similarity is below a threshold are discarded from the topic-specific corpus. Fig. 3 presents perplexity and WER variations computed with different thresholds. These preliminary results highlight that this parameter must not be too high to be able to estimate n-gram probabilities in a stable and reliable way on the corpus. In the back-end experiments, threshold has been set to 0.08.

2.4. Adaptation of the general purpose LM

The linear interpolation technique used to obtain the adapted LM is driven by a factor giving more or less importance to the topic-specific LMs. The impact of the interpolation weight on the perplexity has been measured on a small set of transcripts. As an example, Fig. 4 shows the perplexity gain as a function of the interpolation weight for two similarity thresholds. Results reported with constant interpolation weights across documents (solid lines) demonstrate that carefully choosing the weight is crucial as a variation of 0.1 of this parameter can lead to a decrease of perplexity of over 20 %. However, the optimal interpolation weight clearly depends on the quality of the adaptation corpus gathered and should be determined separately for each segment. If optimal weights could be derived for each segment, this would result in a 18 % decrease of the average perplexity. Unfortunately, automatically determining the optimal interpolation weight for a given segment is not an easy task. In the future, different adaptation techniques, such as MDI [8], could be used to overcome this limitation.

tags but also automatically seek for irrelevant text segments, like advertisements, copyright notifications, menus. . . , and remove them.

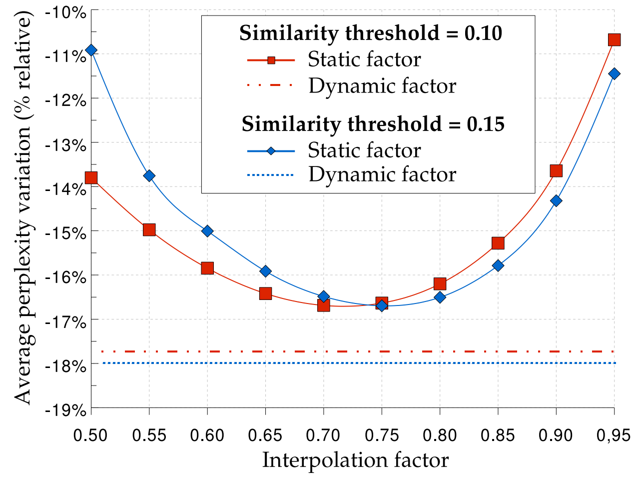

Fig. 4. Influence of the interpolation factor on the perplexity with an interpolation factor constant over all documents (solid lines) or optimally set for each document (dotted lines).

3. EXPERIMENTS AND RESULTS

Experiments are carried out on 172 segments from 6h of BN shows from the French radio BN corpus ESTER [9]. These segments, coming from 3 different broadcasters and all dated from the same period of time, are spread over diversified topics (war in Iraq, national politics, sports, wheather. . .) and lengths (from 30 to 2,000 words). This collection is divided into a development set and a test set of respectively 91 and 81 segments. For each segment, an adapted LM is trained by executing the whole adaptation process, as described in Section 2. Perplexities and WERs are then measured and compared with those obtained before adaptation.

Our ASR system, briefly described in [7], is a multipass system based on a 4-gram general purpose LM over a 64k word vocabulary. Word lattices generated by the last pass are reevaluated using the adapted LM.

Results presented in Table 1 show constant and significant perplexity reductions. Word error rates are also improved in almost all cases, particularly on *RFI* which deals with topics absent from the initial LM training corpus[4]. However, the WER improvement is clearly less on the test set than on the development set. Topic-based adaptation of the baseline LM even results in a WER increasing by 0.3 on *France Info* in the test set, whereas a gain of 0.23 is reported for the same broadcaster in the development set. This is all the more surprising since the time shift between these two broadcasts is merely 5 hours. Several possible explanations are contemplated. First, since previous experiments have highlighted the crucial impact of the interpolation weight factor, the LM interpolation weight optimized on the development set might be unadapted to the test set. Diagnostic experiments tend to prove that this is not the only explanation. However, we believe that more sophisticated adaptation techniques could improve this point. Second, the word lattices that are rescored are rather small and some words of the reference transcript are not included. As a consequence, rescoring these lattices with the adapted LM cannot improve the quality of the transcription, even if the adapted LM gives more importance to these

[4] *RFI* includes many news related to Africa whereas the training corpus mainly deals with national matters and major international news.

		Development set			Test set		
		Baseline	Adapted	Variation	Baseline	Adapted	Variation
PPL	*France Inter*	188.6	150.1	-20.0%	185.3	162.3	-12.3%
	France Info	183.7	137.6	-21.1%	186.7	140.3	-22.3%
	RFI	157.8	126.0	-21.1%	173.7	148.0	-14.4%
	Average	178.2	138.7	-20.7%	182.3	148.1	-17.2%
WER	*France Inter*	20.3	20.0	-0.33	19.8	19.8	-0.06
	France Info	21.0	20.7	-0.23	21.7	22.1	$+0.30$
	RFI	24.8	24.0	-0.75	23.2	22.9	-0.29
	Average	21.9	21.5	-0.43	21.6	21.5	-0.03

Table 1. Perplexities (PPL) and word error rates (WER) measured on the development set and on the test set, for each broadcaster and on average.

Reference	`de la gorge et des bronches` (*of the throat and the bronchi*)
Baseline	`- l' accord - de branche` (*the agreement of branch*)
Adapted	`- la gorge - des bronches` (*the throat the bronchi*)

Table 2. Comparison of an utterance transcribed with our baseline LM and our adapted LM according to the reference transcription.

absent words. To avoid these cases, it would be interesting to use the adapted LMs on the larger first pass word lattices, which include more transcription possibilities. This idea is reinforced by measuring oracle WERs: with the word lattices currenlty used, the *minimum a posteriori* WER is $12.6\,\%$ as opposed to $8.7\,\%$ when considering the first pass lattices.

A more detailed analysis of the results shows that our technique improves the transcription of utterances with thematic words, *i.e.*, words which are frequently used in a given domain. Table 2 illustrates this idea through an example taken from a transcript dealing with viral diseases and for which the extracted keywords were *flu*, *virus*, *pneumopathy*, *who* (World Health Organization) and *cold*. In the example, the transcript hypothesis returned by the baseline LM was completely independent of the topic of the report, whereas the use of the adapted LM leads to decode perfectly the thematic words. However, a deeper analysis of other results shows that the gain implied by the corrections of badly transcribed thematic words is frequently offset by grammatical errors that do not appear when using only the generic LM. This can be explained by the small size of thematic corpora which leads to badly estimate the probability of n-grams containing generic terms, like auxiliary verbs or ordinary adjectives. Even if a morphosyntactic processing could probably correct these grammatical errors, a more clever use of the topic-based corpora should first and foremost be done.

4. CONCLUSION

In this paper, we have presented a complete and unsupervised technique to adapt automatically a generic LM to the several topics that can be encountered during a transcription, especially in broadcast news. This technique mainly relies on the use of the Internet to build thematic corpora. Experiments lead to contrasting results which oppose constant and significant perplexity reductions to variable WER gains. The detailed analysis of these results reveals mainly that the thematic corpora are too small to infer correct probability estimations of the n-grams containing generic words. As a consequence, the corresponding probabilities should not be modified from the generic LM in the adapted LM.

In future work, it would then be interesting to improve the combination of these 2 LMs, for example, by building an intermediary topic-specific LMs with only n-grams containing thematic words or by using a more sophisticated adaptation technique such as MDI. Another step would also be to aim at factorizing data. First, it may be more effective to cluster thematically similar segments before launching any adaptation process. Second, thematic corpora could be considered as linguistic resources for new adaptations instead of always retrieving pages only from the Internet. Finally, the vocabulary update issue, voluntarily disregarded in this study, should also be investigated.

5. REFERENCES

[1] K. Seymore and R. Rosenfeld, "Using story topics for language model adaptation," in *Proc. Eurospeech*, 1997, pp. 1987–1990.

[2] D. Klakow, "Selecting articles from the language model training corpus," in *Proc. ICASSP*, 2000, vol. 3, pp. 1695–1698.

[3] L. Chen, J.-L. Gauvain, L. Lamel, and G. Adda, "Unsupervised language model adaptation for broadcast news," in *Proc. ICASSP*, 2003, vol. 1, pp. 220–223.

[4] D. Vaufreydaz, M. Akbar, and J. Rouillard, "Internet documents: A rich source for spoken language modeling," in *Proc. Workshop ASRU*, 1999, pp. 277–280.

[5] A. Sethy, P. G. Georgiou, and S. Narayanan, "Building topic specific language models from webdata using competitive models," in *Proc. Interspeech*, 2005, pp. 1293–1296.

[6] M. Suzuki, Y. Kajiura, A. Ito, and S. Makino, "Unsupervised language model adaptation based on automatic text collection from WWW," in *Proc. Interspeech*, 2006, pp. 2202–2205.

[7] S. Huet, G. Gravier, and P. Sébillot, "Morphosyntactic processing of N-best lists for improved recognition and confidence measure computation," in *Proc. Interspeech*, 2007, pp. 1741–1744.

[8] M. Federico, "Efficient language model adaptation through MDI estimation," in *Proc. Eurospeech*, 1999, vol. 4, pp. 1583–1586.

[9] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ESTER phase II evaluation campaign for the rich transcription of French broadcast news," in *Proc. Interspeech*, 2005, pp. 1149–1152.