# Supervised and unsupervised Web-based language model domain adaptation

*Gwénolé Lecorvé*[1], *John Dines*[1,2], *Thomas Hain*[3], *Petr Motlicek*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Koemei, Martigny, Switzerland
[3]University of Sheffield, Sheffield, United Kingdom
glecorve@idiap.ch, dines@idiap.ch, t.hain@dcs.shef.ac.uk, motlicek@idiap.ch

## Abstract

Domain language model adaptation consists in re-estimating probabilities of a baseline LM in order to better match the specifics of a given broad topic of interest. To do so, a common strategy is to retrieve adaptation texts from the Web based on a given domain-representative seed text. In this paper, we study how the selection of this seed text influences the adaptation process and the performances of resulting adapted language models in automatic speech recognition. More precisely, the goal of this original study is to analyze the differences of our Web-based adaptation approach between the supervised case, in which the seed text is manually generated, and the unsupervised case, where the seed text is given by an automatic transcript. Experiments were carried out on data sourced from a real-world use case, more specifically, videos produced for a university YouTube channel. Results show that our approach is quite robust since the unsupervised adaptation provides similar performance to the supervised case in terms of the overall perplexity and word error rate.

**Index Terms**: Language model, domain adaptation, supervision, Web data

## 1. Introduction

The $n$-gram language model (LM) of most automatic speech recognition (ASR) systems is usually trained on a large multi-topic text collection. As a consequence, this LM is not optimal to transcribe spoken documents dealing with a given specific domain. To solve this problem, domain LM adaptation seeks to re-estimate the $n$-gram probabilities of the baseline LM in order to fit the specifics of the considered domain. The ultimate goal of this adaptation is to improve the quality of ASR transcripts.

Nowadays, a standard approach for LM domain adaptation consists of using the Web as an open corpus in order to retrieve domain-specific data providing accurate statistics for $n$-gram re-estimation [1, 2, 3, 4, 5]. The process of the Web-based adaptation can be split into the following steps: first, one has to extract queries from a given text that is representative of the domain of interest—this text is called the *seed text* ; then Web pages are retrieved by submitting the queries to a Web search engine ; finally, an adapted LM is built by integrating the retrieved adaptation data with background training material.

The seed text is a key aspect of this process since it is supposed to provide a good characterization of the domain in order to extract meaningful information for the adaptation. In the literature, two main approaches are commonly known: either the adaptation is supervised, i.e., the domain is known *a priori* and the considered seed text is a manually generated reliable text, typically a manual transcript [3, 6], or the adaptation is unsupervised where the seed text is obtained from ASR on spoken documents [5, 7, 8].

Obtaining large amount of seed text is desirable since large texts are assumed to more widely characterize the encountered domain. However, the feasibility of supervised adaptation depends on the size of the seed text, since the level of human effort required to produce this text manually is significant. Thus, automation of this process could provide important savings in cost and effort for the development of domain specific LMs in real-life applications.

One would naturally think that supervised approaches based on a very large seed text produce better performance than equivalent unsupervised approaches, but to the knowledge of the authors very few study has yet been conducted to verify this. Only [9] carefully examined the effect of supervision and non supervision on the performance of LM adaptation. However, the studied adaptation approach was not based on the Internet. Hence, this paper aims at comparing the Web-based domain LM adaptation process using different levels of supervision. More precisely, we seek to understand the impact of recognition errors in the seed text on speech recognition accuracy gains resulting from LM adaptation and the dependence on the size of the seed text. Since the paper focuses on LM adaptation, the problem of vocabulary adaptation is not considered here.

The paper is organized as follows: Section 2 presents the LM adaptation used in the experiments. Section 3 describes the experimental setup and introduces different adaptation scenarios for the seed text. Finally, Section 4 studies the effect of these scenarios on various aspects of our LM adaptation technique.

## 2. LM adaptation technique

The strategy of our LM adaptation technique is three-fold. Given a seed text which is assumed to be representative of the domain of interest, queries are first extracted. Then, Web pages are retrieved by submitting the queries to a Web search engine from which we construct an adaptation corpus. Finally, an adapted LM is trained by linearly interpolation statistics from the adaptation corpus with the set of background texts previously used to train the baseline LM. Such an adapted LM is supposed to provide higher speech recognition accuracy than the baseline LM when applied to recordings from the domain of interest. This section describes the query extraction method before explaining how Web pages are retrieved and how the adapted LM is effectively trained in our experiments.

### 2.1. Extracting queries from the seed text

The principle of our query extraction method, as introduced in [3], is to determine which $n$-grams of the baseline LM are not well enough modeled according to the given seed text and then to directly use these $n$-grams as queries. Given the seed text $T$, this principle is driven by the search for an adapted LM

This work is funded by the project CTI 12189.2 PFES-ES of the *Comité pour la Technologie et l'Innovation* (Switzerland) and was carried out in collaboration with Koemei.

whose likelihood on the seed text is greater than the one using the baseline LM, i.e.,:

$$P_A(T) > P_B(T) \;, \qquad (1)$$

where $P_A$ and $P_B$ respectively refer to the probability distribution of target adapted LM and of the baseline LM. This inequality can be guaranteed by decomposing the likelihood onto every $n$-gram $(h, w)$ from $T$, where $w$ is a word and $h$ is a word history, leading to the following set of constraints:

$$P_A(w|h) > P_B(w|h), \quad \forall(h, w) \in T \;. \qquad (2)$$

Then, extracting queries consists in finding out which $n$-grams in $T$ are the most likely to satisfy (2). To do so, $P_A$ can be first assumed to be a linear interpolation of $P_B$ and probability distribution $P_C$ trained on the corpus $C$ of retrieved Web pages. Second, we postulate that $P_C$ can be modeled as another linear interpolation of $P_B$ with the probability distribution $P_T$ trained on seed text $T$. Hence, (2) can be greatly simplified, as follows:

$$\lambda P_T(w|h) + (1-\lambda)P_B(w|h) > P_B(w|h), \forall(h, w) \in T \quad (3)$$

$$P_T(w|h) > P_B(w|h), \forall(h, w) \in T \;. \quad (4)$$

In practice, we approximate (4) by arbitrarily considering as queries the sole trigrams from the seed text which have not been observed during the baseline LM training, i.e., trigrams whose probability is computed by backing off. However, these $n$-grams may be numerous, depending on the size of the seed text $T$, thereby leading to a very long retrieval process and most of these $n$-grams are just sequences that are not specific to the domain of interest. Hence, the set of these $n$-grams is finally filtered by discarding any $n$-gram containing a stopword[1]. In our experiments, this query extraction strategy leads to a few hundred queries for a given seed text.

### 2.2. Web pages retrieval and adapted LM training

To retrieve domain-specific adaptation data, the queries are submitted to a Web search engine. The returned hits are downloaded following a round-robin algorithm, i.e., the $i$-th hits of each query are downloaded successively before downloading the $(i + 1)$-th hits, and so on. Web pages are cleaned and normalized before gathering them into an adaptation corpus. This process stops as soon as a selected number of words is reached. In our experiments, this number is set to 5 million words. On average, this threshold is reached after downloading about 20-40 pages per query.

To train the domain adapted LM, the process initially developed for the baseline LM is then re-used. More precisely, the adaptation corpus is added to the set of background corpora used to train the baseline model, and compound LMs are trained using each corpus. Then, these LMs, including the adaptation LM, are linearly interpolated such that their combination minimizes the perplexity on the seed text. Finally, the resulting LM is pruned in order to reach the same size as the baseline LM. This strategy enables to determine the relative importances of the various background corpora according to the seed. Thus, it is supposed to be better than directly linearly interpolating the baseline LM with the adaptation LM.

## 3. Experimental setup and adaptation scenarios

Before presenting the impact of the seed text on the adaptation process, this section presents the experimental setup, i.e., the ASR system and experimental data. Then, adaptation scenarios are introduced.

### 3.1. Experimental setup

The recognition system used in the experiments is a two-pass system for English. In brief, it uses individual head-mounted microphones (IHM) based acoustic models, a lexicon of $50,000$ words and a 4-gram LM trained on various corpora (AMI corpus, ICSI meeting corpus, *etc.*) for a total amount of about one billion words. The decoder is based on weighted finite state transducers. The first decoding pass relies on generic acoustic models whereas the second is performed after speaker adaptation. All details about the system architecture and the training setups can be found in [10].

The domain is represented by $57$ videos produced for a university YouTube channel. While the broad domain is centered on the course content offered, these videos are of various types (faculty teaching, self-promotion, conferences, interviews, *etc.*). They have been recorded in different acoustic conditions, are of varying duration and some stakeholders are non-native English speakers. The reference transcript represents a total of $40,000$ words. The data was split into two sets: a development set of 29 videos that can be considered as the seed information source to characterize the target domain ; and a test set of 28 held-out videos. The length of the reference transcription is the same for both sets, i.e., about $20,000$ words. Out-of-vocabulary rates are $0.65\,\%$ and $0.59\,\%$ on the development set and on the test set respectively.

### 3.2. Adaptation scenarios

The aim of this paper is to study the importance of the seed text in achieving an effective domain LM adaptation. In fact, this adaptation may be applied within two main scenarios. Either adaptation is meant to be used in a multi-pass recognition process where spoken documents are first transcribed using the baseline LM, before adapting the LM using the first pass output as seed text with which we perform a subsequent decoding pass—we denote this as *self adaptation*. Or it is dedicated to a longer term application where the domain of documents to be transcribed in the future will remain the same—we denote this as *long term adaptation*.

Considering the development and test sets as independent, but covering the same domain, the nature of seed texts within these scenarios can vary according to two aspects: their origin and their size. Regarding the origin, the supervised case consists in considering the reference of the development set. This case is costly in terms of money and time since it requires manual transcription. Conversely, the unsupervised situation relies on the noisy transcript generated by the baseline ASR system. The word error rate (WER) of the baseline ASR is $29.6\,\%$ on the development set. Further, the levels of supervision and non supervision can be modulated by varying the seed text size. In our experiments, this is done by subsampling the seed text.

### 3.3. Evaluation

Effect of the domain adaptation is mainly evaluated by comparing the perplexities of the baseline LM with those of adapted LMs, on the reference transcriptions of the development set and of the test set. For most interesting settings, WERs are also reported. Results on the development set may be considered representative of a self adaptation scenario while those on the evaluation set stand for long term adaptation. Furthermore, let us notice that results for self adaptation using the reference as a seed are “cheating experiments” whose goal is to exhibit optimal (oracle) results. Finally, let us recall that no vocabulary adaptation is performed during the experiments since the paper is focusing on the sole LM adaptation task.

The next section investigates the adaptation scenarios within the two steps of the process involving the seed text.

[1] The list of stopwords is about 600 words.

# 4. Experiments and results

The seed text plays an important role during two steps of the domain adaptation process: it is used to extract domain-specific queries, and it helps determine the importance of the adaptation data when combining domain-specific $n$-gram probabilities with those obtained from the background training texts. This section thus first studies the effect of the seed text on query extraction before analyzing its role in the final linear interpolation step. Finally, the dependence on the seed text size on both steps is presented.

## 4.1. Effect of the seed text on query extraction

As described in Section 2, query extraction is the first step of the adaptation process. Hence, the quality of the seed text is probably crucial. To assess this hypothesis, this section compares the use of the reference and the ASR transcript of the development set ($20,000$ words each) in order to investigate the effect of recognition errors on query extraction.

Table 1 compares perplexities obtained using the baseline LM and LMs adapted from supervised and unsupervised seed texts. For every adapted LM, linear interpolation is carried out using the reference transcript in order to train optimal LMs and, thus, to highlight lower bounds of perplexity for each seed used for query extraction. It appears that, on the development set, the largest improvement is obtained when using the reference as the seed text. This is quite logical since this setting (in italic) represents an artificial case where the seed text is similar to the text modeled by the LM. It is thus common sense to observe that the improvement is less significant on the evaluation set. Interestingly, when using the ASR transcript as seed text we do not observe such differences in perplexity between the development and test data.

To better understand these first results, a second series of evaluations have been carried out whereby we isolate the correctly and incorrectly recognized parts (words) of text in the reference and in the ASR transcripts and use these sole parts as new seed texts for query extraction. Recognition errors are spotted by aligning ASR transcripts with the reference. The results of these experiments are presented in the three last rows of Table 1, where "misrecognized reference" denotes the parts of the reference which have been misrecognized using the baseline LM, "incorrect ASR" denotes what the ASR system has returned for these parts, and "correct ASR" stands for the correctly transcribed parts in the ASR. One can notice that the perplexity improvements on the development set mainly come from the misrecognized portions of the reference. This seems to be logical since it represents the word sequences which are the most inaccurately modeled by the baseline LM. However, such a conclusion is not observed on the evaluation set since the perplexity improvement obtained using "misrecognized reference" is almost the same as when only relying on the correctly recognized portions (correct ASR). Moreover, it appears that the use of "incorrect ASR" still results in perplexity improvements, though these improvements are lower. This surprising result can probably be explained by the fact that Web search engines attempt to automatically transform unlikely queries into more common word sequences while untransformed queries simply result in no hit. Further, some recognition errors may still be domain-specific words. Therefore, the use of ASR transcript is not as bad as expected since it seems that most recognition errors are harmless for query extraction, be it for long term adaptation or for self adaptation.

Table 1: *Perplexities of the development and evaluation sets using different seed texts for query extraction.*

| Query extraction | Linear interp. | Dev. | Test |
|---|---|---|---|
| Baseline LM | | 165 | 170 |
| Reference | Reference | *119* | 139 |
| ASR | Reference | 133 | 143 |
| Correct ASR | Reference | 134 | 143 |
| Incorrect ASR | Reference | 142 | 150 |
| Misrecognized reference | Reference | 120 | 140 |

Table 2: *Perplexities on the development and evaluation sets using different texts to estimating the linear interpolation weights.*

| | Query extraction | Linear interp. | Dev. | Test |
|---|---|---|---|---|
| | Baseline LM | | 165 | 170 |
| (a) | Reference | Background text | 159 | 168 |
| | ASR | Background text | 163 | 169 |
| (b) | No data | Reference | 154 | 159 |
| | No data | ASR | 155 | 161 |
| (c) | Reference | | 119 | 139 |
| | ASR | | 136 | 145 |
| (d) | Correct ASR | | 135 | 143 |

## 4.2. Choice of the seed text for linear interpolation

The second aspect involving the seed text is the estimation of linear interpolation weights. Table 2 presents the results of experiments conducted. In addition to the seed texts previously presented, the text initially used to build the baseline LM, referred to as "background", is introduced. As shown in rows (a), where the linear interpolation is based on the background text, it is clear that the use of adaptation data is completely inefficient if the interpolation text is disconnected from the domain. Moreover, the rows (b) show that re-interpolation of the background training texts, i.e., when no adaptation corpus is retrieved, leads to modest improvements when considering a domain-specific text to estimate the linear interpolation weights. Moreover, in this case there is nearly no difference between the use of the reference against the ASR transcript, meaning that recognition errors do not bias the interpolation weight estimation.

The set of rows (c) denotes the settings where the same text is used for both query extraction and linear interpolation, as this would probably be the case in a real application. On the whole it appears that the use of noisy seed text for interpolation as well as query generation is not significantly worse than the query generation scenario alone. Finally, the row (d) shows that by focusing on the sole correctly transcribed ASR parts linear interpolation does not perform better[2], further reinforcing previous observations. In summary it would appear that recognition errors do not bias the interpolation weight estimation (at least at the error rates that we have observed).

Achieved error rates for the settings (c) and (d) are reported in Table 3. In general, the relative trends are the same as observed for perplexity measures. More precisely, it appears that all the settings lead to significantly outperform the baseline results, even when using the ASR as a seed. Furthermore, it is clear that the recognition errors do not have any significant impact on the system performance, as was already evident from the perplexity results.

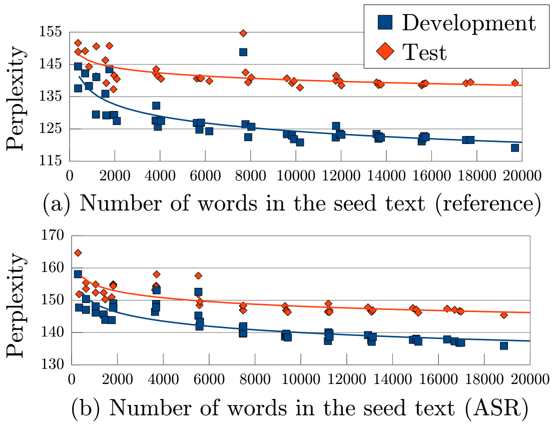
## 4.3. Dependence on the size of the seed text

The size of the seed text may change the conclusions drawn above concerning the low impact of recognition errors on final LM perplexities. Indeed, one would naturally assume that shorter the seed text, more variable we would expect the results of the adaptation. This is due to the fact that the domain of

[2] This is done by replacing recognition errors by out-of-vocabulary words while minimizing the perplexity of the interpolated LM.

Table 3: *WERs (%) obtained with or without domain adaptation. In brackets, relative variations w.r.t. baseline are given.*

| Query extraction and linear interpolation | Development | Test |
|---|---|---|
| Baseline LM | 29.6 | 25.8 |
| Reference | 26.3 (-11.1 %) | 24.1 (-6.6 %) |
| ASR | 27.3 (-7.8 %) | 24.6 (-4.7 %) |
| Correct ASR | 27.5 (-7.1 %) | 24.4 (-5.4 %) |



(a) Number of words in the seed text (reference)

(b) Number of words in the seed text (ASR)

Figure 1: *Perplexity of adapted LMs versus the size of the seed text by subsampling the reference (a) or the ASR transcripts (b).*

interest cannot be characterized so well. In our last series of experiments, we studied the influence of the seed text size on LM domain adaptation. Both reference and ASR transcripts from the development data were randomly subsampled on a sentence basis with different rates and these subsamples were used as new seed texts, both for query extraction and linear interpolation.

Figure 1 reports perplexities of the adapted LMs w.r.t. the size of the seed text when relying on the reference or the ASR transcripts. Firstly, it appears that the perplexity improvements decrease and their variability increases with the size of the seed text in all cases. However, this decrease is very gradual until reaching $2,000$-$4,000$ words, i.e., only 10-20 % of the original seed text size. This tends to show that the efforts spent in generating a seed text can be quite limited. Finally, it is interesting to note that the trends of the curves are the same whether the seed text is derived from the reference or from the ASR transcripts. This means that recognition errors do not appear to have strong influence on LM adaptation when reducing the seed text size.

Decoding experiments were carried out by only considering about 10-20 % of the full seed texts for LM adaptation. Resulting WERs are presented in Table 4. Regarding the reference transcriptions, WERs are quite similar to those reported in Table 3. This is very interesting from a practical point of view since it shows that in the supervised case we can annotate less data without degrading the performance. Some slight improvements even show that better adaptations can be performed with less queries, meaning that some parts of the reference are more important than others for domain adaptation. On the contrary, considering 10-20 % of the ASR transcripts leads to average increase in the WER of 0.5 % absolute compared to the use of the full development set transcript. We assume that this comes from the fact that decreasing the seed text size not only limits the ability of the text to characterize the domain but increases the impact of queries containing transcription errors. Nevertheless, WER gains w.r.t. the baseline are still significant.

Table 4: *WERs (%) obtained when reducing the size of the seed text derived from the reference or from the ASR transcripts. In brackets, relative variations w.r.t. the baseline are given.*

| Query extraction and linear interpolation | Development | Test |
|---|---|---|
| Baseline LM | 29.6 | 25.8 |
| Reference (∼20 % words) | 26.2 (-11.4 %) | 24.4 (-5.4 %) |
| Reference (∼10 % words) | 26.5 (-10.5 %) | 24.1 (-6.6 %) |
| ASR (∼20 % words) | 28.2 (-4.7 %) | 25.0 (-3.1 %) |
| ASR (∼10 % words) | 28.2 (-4.7 %) | 24.8 (-3.9 %) |

## 5. Conclusion

In this paper, we have conducted an investigation of supervised and unsupervised Web-based LM domain adaptation. Various scenarios have been explored to highlight the influence of the seed text used to extract queries and to perform the final linear interpolation step leading to the adapted LM. Obviously, it appears that using manual transcripts brings the greatest improvements of perplexity and ASR accuracy, but other interesting conclusions can be drawn. Firstly, the recognition errors do not bias LM adaptation, as can be seen for query extraction or for linear interpolation. This is very interesting due to the fact that error spotting in ASR outputs is a complex task. Instead, the main effect of recognition errors is a loss of information which prevents us from achieving an optimal characterization of the domain. Nevertheless, relative improvements of 7.8 % and 4.7 % over the baseline WER are achieved using the ASR transcript, depending on the adaptation scenario. Secondly, reducing the size of the seed text does not change this conclusion. Rather, the experiments have shown that decreasing the seed text size reduces both the gains in perplexity and in word error rates consistently for both supervised and unsupervised cases, though in the unsupervised case this is more pronounced.

Further aspects of supervision could be studied in the future work. For example, it would be interesting to know what is the influence of the baseline word error rate on the adaptation process. Furthermore, while having voluntarily left the problem of vocabulary adaptation aside, it would be interesting to know the influence of supervision on the recovery of domain-specific out-of-vocabulary words.

## 6. References

[1] X. Zhu and R. Rosenfeld, “Improving trigram language modeling with the World Wide Web,” in *Proc. of ICASSP*, 2001, pp. 533–536.

[2] A. Kilgarriff and G. Grefenstette, “Introduction to the special issue on the Web as corpus,” *Computational Linguistics*, vol. 29, no. 3, pp. 333–347, 2003.

[3] V. Wan and T. Hain, “Strategies for language model Web-data collection,” in *Proc. of ICASSP*, 2006, pp. 1520–6149.

[4] I. Bulyko, M. Ostendorf, M. Siu, T. Ng, A. Stolcke, and O. Çetin, “Web resources for language modeling in conversational speech recognition,” *ACM Trans. on Speech and Language Processing*, vol. 5, no. 1, pp. 1–25, 2007.

[5] G. Lecorvé, G. Gravier, and P. Sébillot, “An unsupervised Web-based topic language model adaptation method,” in *Proc. of ICASSP*, 2008, pp. 5081–5084.

[6] A. Sethy, P. G. Georgiou, and S. Narayanan, “Building topic specific language models from Webdata using competitive models,” in *Proc. of Eurospeech*, 2005, pp. 1293–1296.

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

[8] A. Ito, Y. Kajiura, S. Makino, and M. Suzuki, “An unsupervised language model adaptation based on keyword clustering and query availability estimation,” in *Proc. of ICALIP*, 2008, pp. 1412–1418.

[9] G. Tür and A. Stolcke, “Unsupervised language model adaptation for meeting recognition,” in *Proc. of ICASSP*, 2007, pp. 173–176.

[10] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, D. van Leeuwen, M. Lincoln, and V. Wan, “Transcribing meetings with the AMIDA systems,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, pp. 486–498, 2012.