# Dynamic Topic-Based Adaptation of Language Models: A Comparison Between Different Approaches

Julian Echeverry-Correa, Beatriz Martínez-González, Rubén San-Segundo, Ricardo Córdoba, and Javier Ferreiros-López

Speech Technology Group, Universidad Politécnica de Madrid, Spain
{jdec,beatrizmartinez,lapiz,cordoba,jfl}@die.upm.es

**Abstract.** This paper presents a dynamic LM adaptation based on the topic that has been identified on a speech segment. We use LSA and the given topic labels in the training dataset to obtain and use the topic models. We propose a dynamic language model adaptation to improve the recognition performance in 'a two stages' ASR system. The final stage makes use of the topic identification with two variants: the first one uses just the most probable topic and the other one depends on the relative distances of the topics that have been identified. We perform the adaptation of the LM as a linear interpolation between a background model and topic-based LM. The interpolation weight is dynamically adapted according to different parameters. The proposed method is evaluated on the Spanish partition of the EPPS speech database. We achieved a relative reduction in WER of 11.13% over the baseline system which uses a single background LM.

**Keywords:** language model adaptation, topic identification, automatic speech recognition, information retrieval

## 1 Introduction

The performance of a speech recognition system depends significantly on the similarity between the language model (LM) used by the system and the context of the speech that is being recognized. This similarity is even more important in scenarios where the statistical properties of the language fluctuates throughout the time, for instance, in application domains involving spontaneous speech from multiple speakers on different topics. One representative example of this kind of domain is the automatic transcription of *political speeches*. Within this domain, the usage of content words (i.e. those that convey information and have a specific meaning rather than indicating a syntactic function) depends on several factors, such as the topic the speaker is addressing, the style of the speech, the vocabulary used by the speaker and the scenario in which the speech is taking place. Regarding these factors, in this paper we are focusing on studying the identification of the topic and its application in the adaptation of language

models. The performance of the speech recognition system will depend, among other elements, on its capacity to update or dynamically adapt the LMs. In this paper we propose a dynamic LM adaptation based on an information retrieval (IR) approach. We used IR techniques for identifying the topics that are related to the content of the speech segment under evaluation. This information enables the system to perform an adaptation of the language model. We explore different approaches for the dynamic language model adaptation. These approaches are based on the interpolation between a background model and topic-based LMs.

## 2 General Overview

In this paper two major tasks can be distinguished: **topic identification** and **dynamic LM adaptation**. Both tasks pursue one common goal, that is improving the performance of an automatic speech recognition system for multitopic speech. We integrate these tasks in 'a two stages' ASR framework. In the first stage, an initial speech recognition of an audio segment is performed using a background LM built from the entire training set. Then, an IR module automatically identifies the topic based on the results of the initial recognition pass. This module uses topic models that have been previously trained for each of the topics available in the database. Using the information provided by the topic identification system and topic-specific LMs, a dynamic adaptation of the background LM is performed. In this paper we present different approaches for the dynamic adaptation of LMs. In the final stage of the framework, the adapted LM is used to re-decode the utterance.

## 3 Related Work

The task of topic identification (TI) falls at the intersection of information retrieval and machine learning systems. In the last years a growing number of statistical learning methods have been applied in TI from these research fields [1]. Common approaches includes Latent Semantic Analysis [2], Rocchio's method [3], Decision Trees [4] and Support Vector Machines [5]. TI has been successfully applied in many contexts and disciplines, ranging from topic detection [6], automated metadata generation [7], document and messages filtering [8] and the recently developed area, sentiment analysis [9], among many other fields of application. Nevertheless it is interesting to review the influence of TI in the field of language model adaptation. Within this field, TI has been used to study the changes that the language experiences when moving towards different domains [10]. In that sense, TI is able to contribute to LM adaptation by adding new sources of information to previously existent models with the objective of enriching them. This leads to a diversity of approaches in the field of LM adaptation that can be distinguished regarding the origin of the new sources of information. Some LM adaptation approaches are based on the specific context of the task that they are addressing. In these approaches, the new data is used to generate a context-dependent LM which is then merged with a

static LM. These new sources of information can proceed, for instance, from text categorization systems as in [11], from speaker identification systems [12], from linguistic analysis systems [13] or from the application context itself [14]. Other approaches are based on analysis and extraction of semantic information. Latent Semantic Analysis (LSA) is an example of this type of approach. In [15], the use of LSA is proposed to extract the semantic relationships between the terms that appear in a document and the document itself. More robust techniques in the field of information retrieval, as Latent Dirichlet Allocation (LDA) [16], have also been used for adapting LMs [17]. When using data available online it is possible to find information related to a large variety of topics. In this regard, clustering algorithms have been proposed to group together those elements that share some properties. Topic-based language modeling is an example of this clustering criterion [18, 19].

## 4 Topic Identification

In a broad sense, topic identification is the task of automatically identifying which of a set of predefined topics are present in a document. To perform topic identification some steps must be followed. These steps are: preprocessing, document representation, term weighting and topic modeling and identification.

### 4.1 Preprocessing

The preprocessing stage allows us to to convert both, documents and queries, to a more precise and concise format. This stage has a substantial impact on the success of the topic identification process [20]. Typical preprocessing steps include: structural processing, lexical analysis, tokenization, stopwords removal, stemming and term categorization. We provide a small description of the steps in which we made special considerations:

• *Stopwords removal.* There are several stopwords lists available online for different languages and for general applications in IR systems. However, generic stopwords lists do not contemplate terms, that in fact, are very frequent in domain specific documents. For that reason, we performed the evaluation using two lists: a generic list with 421 stopwords (*List-1*) and a domain specific stopword list (*List-2*), that we created by adding, to the generic list, those terms with an Inverse Document Frequency (IDF) value below a threshold. The IDF measures how common a term is in the whole document collection; we computed it by using a Term-Document Matrix composed of the 1802 documents in the training dataset and the 16528 terms in the word inventory. We performed different experiments on the Development set in order to find the optimal threshold. The lowest topic identification error on this dataset was obtained by setting the threshold to 0.4435, which means removing the terms that appear, at least, in 649 documents. The List-2 has 446 stopwords.

• *Stemming.* This step refers to the transformation of a word to its stem or root form. For this step, we have used the Freeling Toolkit [21]. Due to few errors in

the original stemming process, we have modified some of the stemming rules for the Spanish language of the toolkit.

### 4.2 Document Representation

The document representation is based on the widely known bag-of-words model. In this model the relationships between the index-terms and each of the documents in the collection are represented by a Term-Document Matrix, that describes the frequency of occurrence of the index-terms in the documents.

### 4.3 Term Weighting

To improve the capacity of discrimination of the index-terms, weights can be applied to the elements of the Term-Document Matrix by associating the occurrence of an index-term with a weight that represents its relevance with respect to the topic of the document. We have selected the combination of TF (*Term Frequency*) and IDF as the baseline weighting scheme for comparing the results obtained for the topic identification task in this paper. Among the most common weighting schemes, *term entropy* ($te$) is based on an information theory approach and it exploits the distribution of terms over documents [22]. For the index-term $t_i$ in the document $d_j$, it is defined as follows:

$$te_{i,j} = 1 - \sum_{j=1}^{N} \frac{p_{i,j} \cdot \log(p_{i,j})}{\log(N)}, \text{ where } p_{i,j} = \frac{c_{i,j}}{gf_i} \tag{1}$$

Where $c_{i,j}$ represents the term frequency of the index-term $t_i$ in the document $d_j$. $gf_i$ is the global frequency of the index-term $t_i$ measured over the $N$ documents in the collection. This scheme may lead to a log zero calculation if an index-term is not present in a document. It has been suggested to include a smoothing parameter $a$, resulting in $p_{i,j} = (a + c_{i,j})/gf_i$. Indeed, it solves the log zero calculation, but the evaluation that we have performed on the combination of TF and this scheme has shown that it does not significantly improve the TF-IDF baseline weighting scheme. We propose a *pseudo term entropy* calculation based on the *term entropy* formula. Our idea is to assign less weight to the terms that are equally distributed over the documents in the collection and assign more weight to terms that are concentrated in a few documents. In this *pseudo term entropy* the parameter $p_{i,j}$ is calculated as the weighted sum of $c_{i,j}$ and the inverse of $gf_i$.

$$p_{i,j} = \beta \cdot c_{i,j} + \frac{\gamma}{gf_i} \tag{2}$$

The proposed scheme not only solves the log zero problem, but also improves the topic identification accuracy as shown in section 6. We performed different experiments on the Development set in order to adjust the parameters $\beta$ and $\gamma$. For the evaluation proposed in this paper, the best results were obtained by adjusting $\beta = 1.5$ and $\gamma = 2.1$.

### 4.4 Topic Models

In this paper we compare two topic models: the Generalized Vector Model (GVM) and Latent Semantic Analysis (LSA) [2]. Both models represent documents and queries as vectors in a multi-dimensional space, in which the number of dimensions is determined by the number of index-terms in the GVM or the number of latent dimensions in the LSA approach.

In both models, the similarity $sim(\boldsymbol{d}, \boldsymbol{q})$ between a document $\boldsymbol{d}$ and a query $\boldsymbol{q}$ can be computed using the cosine distance. According to this distance, each document is ranked on how close it is to the query. In our approach, we have gathered all documents in the collection belonging to the same topic into one document. We have done the same for all the topics. By doing this, each document represents a distinct topic. So, when computing the similarity between the query and a document, we are actually computing the similarity between the query and a topic.

## 5 Topic-based Language Model Adaptation

Topic-based LM adaptation becomes a strategy to lower the word error rate of the transcription given by the ASR by providing language models with a higher expectation of words and word-sequences that are typically found in the topic or topics of the story that is being analyzed. LM interpolation is a simple and widely used method for combining and adapting language models [23, 24].

### 5.1 Language Model Interpolation

Let us consider probabilistic language modeling and let $P(w|h)$ be the probability of word $w$ given the previous sequence of words $h$. Then, given a background model $P_B(w|h)$ and a topic-based model $P_T(w|h)$ it is possible to obtain a final model $P_I(w|h)$, to be used in the second decoding pass, as

$$P_I(w|h) = (1 - \lambda)P_B(w|h) + \lambda P_T(w|h) \tag{3}$$

where $\lambda$ is the interpolation weight between both models, which has to fulfill the condition $0 \leq \lambda \leq 1$. The topic-based LM is generated by combining several topic-specific LMs $P_t(w|h)$ in general. In our case, the background model, as well as the topic-specific models are static models. They are trained once and remain unchanged during the evaluation. The topic-based LM could be either static or dynamic. It depends on the adaptation scheme followed, as we will see later in this paper. This model, as well as the final model $P_I(w|h)$, are generated during the evaluation of each audio segment.

### 5.2 Interpolation Schemes

Two questions arise at this point. How to generate the topic-based model $P_T(w|h)$? and, how to determine the interpolation weight $\lambda$ with the background

model? For solving these questions, we propose different approaches:

• **Hard approach.** In this approach, the topic-based LM $P_T(w|h)$ is built by considering only one of the topic-specific language models ($P_t(w|h)$). This model is selected as the one related to the topic ranked in the first position by the TI system. For estimating the interpolation weight $\lambda$ we define a distance measure $\delta$ between this LM and the background LM. In this approach, our hypothesis is that the greater the distance between both models, the greater the contribution of the topic specific model to the final one. This distance is computed by considering the average difference in the unigram probabilities of both models.

$$\delta_T = \frac{1}{N} \sum_{\forall w_i \in P_T} |P_T(w_i) - P_B(w_i)| \tag{4}$$

Where $N$ is the number of unigrams in the topic-based LM $P_T(w|h)$. To ensure the interpolation weight fulfills the condition $0 \leq \lambda \leq 1$, we include the summation of the distances of all the topic-specific LMs to the background model as a normalization constant. Then, the interpolation weight is computed as the relative distance between $\delta_T$ and this normalization constant.

$$\lambda = \frac{\delta_T}{\sum_{j=1}^{n} \delta_j} \tag{5}$$

Where $n$ is the number of topics and $\delta_j$ the distance of the *j-th* topic-specific LM to the background LM.

• **Soft-1 approach.** In this case, instead of using only one specific-topic LM for generating the topic-based LM, this model is built on a dynamic basis by the interpolation of a different number of topic-specific LMs. The **Soft-1 approach** tries to gather the dynamic of the specific-topic models $P_t(w|h)$ depending on the similarity of the audio segment to each of the topics. By doing this, more relevance is given to the topics ranked in the first positions by the TI system. The topic-based LM is then computed as follows:

$$P_T(w|h) = \alpha_1 P_{t_1}(w|h) + \alpha_2 P_{t_2}(w|h) + \cdots + \alpha_k P_{t_k}(w|h) \tag{6}$$

where $k$ is the number of models considered for obtaining the topic-based LM. The interpolation weight $\alpha_i$ is calculated as the normalized value of the similarity measure of the TI system.

$$\alpha_i = \frac{sim(\boldsymbol{d}_i, \boldsymbol{q})}{\sum_{j=1}^{k} sim(\boldsymbol{d}_j, \boldsymbol{q})} \tag{7}$$

The interpolation weight $\lambda$ between the background LM and the topic-based LM was set experimentally in this case.

• **Soft-2 approach.** This approach is similar to the previous one, but instead of setting $\lambda$ experimentally, we have computed it by weighting the relevance of the topic-specific LMs according to the cosine distance. That is:

$$\lambda = \sum_{i=1}^{k} \frac{sim(\boldsymbol{d}_i, \boldsymbol{q})}{\sum_{j=1}^{k} sim(\boldsymbol{d}_j, \boldsymbol{q})} \cdot \frac{\delta_i}{\sum_{j=1}^{k} \delta_j} \tag{8}$$

In Soft-1 and Soft-2 approaches, we have considered two additional possibilities: a) to create the topic-based LM using all the topic-specific LMs, that is by setting $k$ as the total number of topics, and b) to create the topic-based LM by selecting the 10 topics with higher positions in the TI ranking.

## 6 Experimental Evaluation

Our evaluation focuses in two aspects: the evaluation of the topic identification approach and the evaluation of the dynamic language model adaptation by means of evaluating the performance of the speech recognition system. Before discussing the results obtained, we describe the dataset used for the evaluation.

### 6.1 Dataset

We have used the Spanish partition of the EPPS Database (*European Parliament Plenary Sessions*) of the TC-STAR Project to study the performance of the proposed system. Due to the fact that the training dataset of the database is the only one that includes distinct labels for the topics, we used it for training, development and evaluation purposes. The topics have been manually labeled according to the main discussion subject of each session [25]. We believe that identifying the topic on short sentences can be ambiguous because few words do not provide semantic information about the topic that is being addressed. For that reason we decided to perform the evaluation over segments of audio with a length no less than a minute. We extracted these segments from turns of intervention of just one speaker. By this criterion, we obtained 252 audio segments for the evaluation. Some details of the corpus: The language of the corpus is Spanish. There are both male and female speakers (approx. 75% - 25% distributed). The domain of the corpus is political speeches. Training set is composed of 21127 sentences grouped in 1802 speaker turns of intervention. Development set is composed of 2402 sentences grouped in 106 speaker turns. The lexicon size is 16.5k words and the Test set is composed of 3738 sentences grouped in 252 speaker interventions. Each of the speaker interventions belongs to one of 67 different topics. We also use the EUROPARL [26] text database for training both background and topic-specific LMs.

### 6.2 Topic Identification Evaluation

For the topic identification task, the initial performance of the system was obtained by using the Generalized Vector Model, a classic TF-IDF weighting scheme and a general domain stopwords list (SW *List-1*). We will use this configuration as the baseline to discuss the improvements in the different approaches that we have applied. We compared the two different lists of stopwords. We also compared different weighting schemes and the influence of preprocessing stages like stemming in the topic identification error. Table 1 shows the results obtained in topic identification using both GVM and LSA approaches.

**Table 1.** Topic Identification error (T.ID. error) using GVM and LSA topic models approaches

| Topic identification approach | T.ID. error (%) |
|---|---|
| GVM + TF-IDF + SW (*List-1*) | 35.32 ± 5.90 |
| GVM + TF-IDF + SW (*List-2*) | 34.13 ± 5.85 |
| GVM + TF-IDF + SW (*List-2*) + Stemming | 36.11 ± 5.93 |
| GVM + TF-Entropy + SW (*List-2*) | 34.52 ± 5.87 |
| GVM + TF-PseudoEntropy + SW (*List-2*) | 33.33 ± 5.82 |
| LSA + TF-IDF + SW (*List-1*) | 32.94 ± 5.80 |
| LSA + TF-IDF + SW (*List-2*) | 30.95 ± 5.70 |
| LSA + TF-IDF + SW (*List-2*) + Stemming | 32.14 ± 5.76 |
| LSA + TF-Entropy + SW (*List-2*) | 29.76 ± 5.64 |
| **LSA + TF-PseudoEntropy + SW (*List-2*)** | **27.38 ± 5.50** |

In general, LSA outperforms the Generalized Vector Model. In both topic models, the combination of TF and *pseudo term entropy* (TF-PseudoEntropy) reduces the topic identification error when compared to TF-*entropy* and to TF-IDF weighting schemes, nevertheless this reduction is not statistically significant. For both models, Stemming does not significantly contribute in error reduction. The criterion that we followed for creating the *List-2* of stopwords contributes in most of the cases in reducing topic identification error. The best combination of parameters is obtained for the LSA model, using the *List-2* of stopwords and weighting the terms with TF-PseudoEntropy scheme. This configuration presents a relative improvement of 22.48% when compared to the baseline approach.

### 6.3 Dynamic LM Evaluation

For the evaluation of the dynamic LM adaptation we have used the best configuration of parameters obtained in the previous section. The initial performance of our baseline system (i.e. without the dynamic LM adaptation) achieved a WER of 21.75. In Table 2 the results of the speech recognition performance when using the proposed approaches for the dynamic LM adaptation are compared. Although there is no significant difference between the **Soft-1** and the **Soft-2** approaches when comparing both variants (all topics and top-10), there is, in fact, a significant difference between the results obtained by the **Soft-1 - top 10** and the **Hard** approach, and even better results can be found when compared to the baseline approach. In general, with this soft integration we manage to reduce 11.13% of the initial WER.

## 7 Conclusions

In this paper we have presented a framework for dynamic language model adaptation based on topic identification. The results in the ASR task have

**Table 2.** Comparison between the word error rate obtained for different LM adaptation approaches

| LM Adaptation approach | WER | Relative Improvement |
|---|---|---|
| Baseline (no adaptation) | 21.75 ± 0.26 | |
| Hard | 19.90 ± 0.25 | 8.51 |
| Soft 1 - all | 19.61 ± 0.25 | 9.84 |
| Soft 1 - top 10 | **19.33 ± 0.25** | **11.13** |
| Soft 2 - all | 19.65 ± 0.25 | 9.66 |
| Soft 2 - top 10 | 19.50 ± 0.25 | 10.34 |

shown that a small but statistically significant improvement in word error rate can be obtained by the adaptation strategy that has been proposed. Adapting the LM by taking only into consideration the closest topic, improves the baseline performance, but does not take advantage of all the sources of information available. The proposed criterion for selecting stopwords and the proposed weighting scheme have contributed in reducing the topic identification error.

**Acknowledgements.** The work leading to these results has received funding from TIMPANO (TIN2011-28169-C05-03), and MA2VICMR (Comunidad Autonóma de Madrid, S2009/TIC-1542) projects. It has also been supported by the European Union under grant agreement number 287678. Julian Echeverry-Correa also acknowledges the support from Colciencias and from the Universidad Tecnológica de Pereira, Colombia.

## References

1. Sebastiani, F. "Machine learning in automated text categorization". ACM Computing Surveys (CSUR), 34(1):1–47, 2002.
2. Deerwester, S. et al. "Indexing by latent semantic analysis". Journal of the American Society for Information Science, 41 (6): 391–407, 1990.
3. Rocchio, J. "Relevance Feedback in Information Retrieval", in G. Salton [Ed], The SMART retrieval system: experiments in automatic document processing. Prentice-Hall, Inc., 1971.
4. Lewis, D., and Ringuette, M. "A comparison of two learning algorithms for text categorization". In Proc. of 1994 Symposium on Document Analysis and Information Retrieval, pages 81–93. 1994.
5. Joachims, T. "Text categorization with Support Vector Machines: Learning with many relevant features". Machine Learning: ECML-98. Springer Berlin Heidelberg, pages 137–142. 1998.
6. Qiu, Y. "A keyword based strategy for spam topic discovery from the internet". In Proc. of 2010 Fourth International Conference on Genetic and Evolutionary Computing (ICGEC)., pages 260–263. 2010.
7. Cheng, N. and Chandramouli, R. and Subbalakshmi, K.P. "Author gender identification from text". Digital Investigation, 8 (1):78–88, 2011.
8. Günal, S. et al. "On feature extraction for spam e-mail detection". Multimedia Content Representation, Classification and Security. Springer Berlin Heidelberg, 2006

9. Maks, Isa and Vossen, Piek. "A lexicon model for deep sentiment analysis and opinion mining applications". Decision Support Systems, 53: 680–688, 2012.
10. Bellegarda, J. "Statistical language model adaptation: review and perspectives". Speech communication, 42 (1): 93–108, 2004.
11. Seymore, K. and Rosenfeld, R. "Using story topics for language model adaptation". In Proc. of EUROSPEECH, 1997.
12. Nanjo, H. and Kawahara, T. "Unsupervised language model adaptation for lecture speech recognition". In ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, 2003.
13. Liu, Y. and Liu, F. "Unsupervised language model adaptation via topic modeling based on named entity hypotheses". In Proc. of IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, ICASSP 2008., pages 4921–4924, 2008.
14. Lucas-Cuesta, J. et al. "On the dynamic adaptation of language models based on dialogue information". Expert Syst. Appl., 40 (4): 1069–1085, 2013.
15. Bellegarda, J."Exploiting latent semantic information in statistical language modeling". Proceedings of the IEEE, 88 (8): 1279–1296, 2000.
16. Blei, D. and Ng, A. and Jordan, M. "Latent dirichlet allocation". Journal of Machine Learning Research, 3: 993–1022, 2003.
17. Chien, J.T. and Chueh, C.H. "Dirichlet class language models for speech recognition". Audio, Speech, and Language Processing, IEEE Transactions on, 19 (3): 482–495, 2011.
18. Florian, R. and Yarowsky, D. " Dynamic nonlocal language modeling via hierarchical topic-based adaptation". In Proc. of the ACL, pages 167–174, 1999.
19. Iyer, R. and Ostendorf, M. "Modeling long distance dependence in language: Topic mixtures versus dynamic cache models". Speech and Audio Processing, IEEE Transactions on, 7 (1): 30–39, 1999.
20. Uysal, A. and Günal, S. "The impact of preprocessing on text classification". Information Processing and Management, 50: 104–112, 2014.
21. Padró, L. and Stanilovsky, E. "Freeling 3.0: Towards Wider Multilinguality". Proceedings of the Language Resources and Evaluation Conference (LREC 2012), Istanbul, Turkey, May 2012. ELRA.
22. Dumais, S. " Improving the retrieval of information from external sources". Behavior Research Methods, Instruments, & Computers, 23 (2): 229–236, 1991.
23. Federico, M. and Bertoldi, N. "Broadcast news LM adaptation over time". Computer Speech & Language, 18 (4): 417–435, 2004.
24. Chiu, H. and Chen, B. "Word topical mixture models for dynamic language model adaptation". In Proc. of IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, ICASSP 2007, volume 4, pages 169–172, 2007.
25. Mostefa, D. and Hamon, O. and Moreau, N. and Choukri, K. "Evaluation Report for the Technology and Corpora for Speech to Speech Translation (TC-STAR Project). Deliverable N. 30"
26. Koehn, P. "Europarl: A Parallel Corpus for Statistical Machine Translation". In Proc. of the 10th Conference on Machine Translation (MT Summit'05), 2005.