# Improving Spoken Document Retrieval by Unsupervised Language Model Adaptation Using Utterance-based Web Search

*Robert Herms, Marc Ritter, Thomas Wilhelm-Stein, Maximilian Eibl*

Technische Universität Chemnitz, Germany

firstname.lastname@informatik.tu-chemnitz.de

## Abstract

Information retrieval systems facilitate the search for annotated audiovisual documents from different corpora. One of the main problems is to determine domain-specific vocabulary like names, brands, technical terms etc. by using general language models (LM) especially in broadcast news. Our approach consists of two steps to overcome the out-of-vocabulary (OOV) problem to improve the spoken document retrieval performance. Therefore, we first separate the resulting transcript of a speech recognizer into blocks. Keywords are extracted from each transcribed utterance of a block for the search of web resources in an unsupervised manner in order to obtain adaptation data. These data are used to perform a block-specific adaptation of a general pronunciation dictionary and a general LM. The second step comprises the utilization of a certain adapted dictionary and LM in the speech recognizer to improve the vocabulary coverage and to regard the perplexity for a corresponding block at once. We evaluate this strategy on a dataset of summarized German broadcast news. Our experimental results show improvements of up to 11.7% for MAP of 18 different topics and 7.5% of WER in comparison to the base LM.

**Index Terms**: language modeling, unsupervised adaptation, out-of-vocabulary, spoken document retrieval

## 1. Introduction

Spoken document retrieval (SDR) has gained much interest during the last decades, especially in order to make data collections of audiovisual media searchable. A special case are archived materials with spoken content like broadcast news or talk shows that possess a varying temporal and versatile contextual scope. One of the main goals of SDR is to improve the access to content, which has not been adequately annotated or transcribed [1]. Consequently, the generation of high-quality transcriptions as well as the retrieval of relevant data from large datasets of heterogeneous audiovisual footage appears still as a challenging task.

In this context and in the area of automatic speech recognition (ASR) the out-of-vocabulary (OOV) problem has a serious impact. One way to cope with the massive heterogeneity of different topics is the usage of bulks of more domain-specific language models (LM). Despite fashionable annotation methods like crowd-sourcing, the transcription of diverse topics in supervised LM adaptation remains a time- and cost-intensive task particularly for huge amounts of data collections. In contrast, automatic and unsupervised learning methods provide a profound alternative. In this paper, we focus on the unsupervised adaptation of a general LM by using web-based data in combination with an ASR system.

As described by Chen et al. [2] within ASR hypotheses yields a major opportunity to draw conclusion from given information of the underlying content. Nonetheless, they are not rather being used as adaptation data, since transcripts contain recognition errors. Adaptation data can be acquired by using the hypotheses of the ASR as queries within an information retrieval (IR) system in order to specifically adapt the LM [2, 3]. A common way is to use web-ressources like HTML pages [3, 4, 5] or content from Web 2.0 like RSS feeds or Twitter [6]. Thus, a provision of additional textual data for the LM estimation may further reduce recognition errors and emerges as a major component to increase vocabulary for the ASR with specific data like names, brands, and technical terms to cope with the OOV problem. Anyhow, the enrichment of data, in particular with out-of-domain data, does not necessarily lead to improved LMs [7].

In general, topic-domain adaptation is a valuable mean to increase the knowledge about a specific topic. Hence, Lecorvé et al. [4] use the world wide web to build topic-adjusted adaptation corpora within an unsupervised approach. Notwithstanding, Meng et al. [3] mention that the application of topic specific models are not always best suited together with out-of-the-box systems, notably when the range of topics appears too heterogeneous or when the content is changing dramatically within short time spans. Saykham et al. [8] verified the change of subject in broadcast news over certain time spans and recommend modifications of the corresponding LM. This is based on the assumption that stories that occur closer in time tend to be more similar whereat more recent broadcast news should be used to cope with current content.

Our approach utilizes the previously mentioned methods within a two-pass based approach. Akin to Saykham et al. [8] the main assumption of our approach is that the contexts of utterances of a report that occur closer in time are more likely to be correlated. This is closely connected to a recurring vocabulary whereas the period and the appearance of similar content may be affected by the type of the contribution. In general, one can easily imagine that a specific topic in talk shows occurs in a much larger time span than in summarized broadcast news. By taking this fact into consideration, we separate the automatic generated transcript of a report into individual blocks and perform an unsupervised adaptation of the general pronunciation dictionary and the general LM for each block. Our experiments show a positive impact for the automatic speech recognition and thus for the SDR by applying the adapted models and dictionaries for the corresponding blocks.

Furthermore, the effectiveness of this domain-independent approach is investigated for different block sizes. Moreover, this method is not restricted to specific topics and can be applied to any other areas besides broadcast news or meeting recog-

nition. Again, we do not focus on the creation of error-free transcripts, but challenge the OOV problem with respect to improved spoken document retrieval.

This paper is organized as follows. In the next section we describe the system overview of our approach in detail for LM adaptation as well as the decoding strategy of the speech recognizer followed by the experimental setup and results including the description of the applied datasets in section 3. Finally, we conclude this paper in section 4 and give some future directions.

## 2. Adaptation Method

Our approach for LM adaptation works out-of-the-box in an unsupervised manner using a two-pass decoding strategy. In the following, we describe this strategy and the retrieval of adaptation data used for pronunciation dictionary and LM adaptation.

### 2.1. Two-pass Decoding

The proposed approach works out-of-the-box using a speech recognition system with a two-pass decoding strategy. A first-pass transcript is generated automatically by the speech recognition system. The segmentation of longer audio streams into individual transcripts is performed by the recognizer itself by means of voice and silence detection. The set of transcripts is then summarized into individual blocks of fixed sizes. The transcripts of each block are used for web search to retrieve web documents as adaptation data to build a block-specific adapted dictionary and LM.

Within a second-pass the adapted dictionaries and LMs are applied to the corresponding blocks enabling the reduction of the OOV rate while keeping the perplexity at a low rate.

### 2.2. Retrieval of Adaptation Data

At first, a hypothesis of an utterance and thus an associated transcript is produced by the automatic speech recognizer whereat utterances may range from short statements to more complex structures like whole sentences.

As shown in Figure 1, the next step applies some keyword extraction routines to allow the generation of queries to be used in a subsequent web search. In some cases there are too many words for realizing a web request or too many keywords may obstruct appropriate retrieval results. This can be avoided by the usage of additional prioritized constraints for building queries. Therefore, the NLP Stanford Parser enables us to extract parts of the generated transcript:

1. Nouns and named entities are directly used for a query.
2. If the first step does not result in any document, only named entities are directly used for a query.
3. If step 2 does not result in any document, the sequence of named entities is recursively decomposed into two parts until there is some retrieved result. Web search is then performed using each part.

The retrieved web documents are accumulated up to a specific defined amount. In order to be used within the adaptation corpus, the obtained web documents are normalized to match the conventions and restrictions of the pending adaptation procedure and the speech recognition system. This is achieved by removing HTML tags and the conversion of numbers, acronyms, and special characters. The process chain from speech recognition to the accumulation of the adaptation corpus is repeated until the number of transcripts of a block has been exhausted. Subsequent steps comprise the adaptation of the pronunciation dictionary and the LM for a certain block.

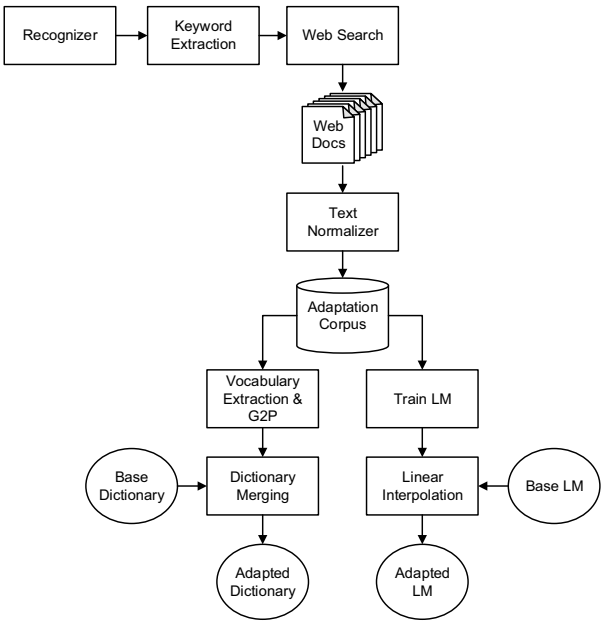


Figure 1: *Procedure of unsupervised pronunciation dictionary and LM adaptation using data from the web based on hypotheses of a speech recognizer.*

### 2.3. Dictionary and LM Adaptation

The adaptation of the pronunciation dictionary aims to enrich the base dictionary by additional words coming from the adaptation corpus. Hence, the vocabulary of the adaptation corpus is extracted and compared to the base dictionary to get the new vocabulary. We use the WFST-driven grapheme-to-phoneme (G2P) framework Phonetisaurus to phonetically transcribe additional vocabulary. For this purpose, we trained a model based on 140k words for German language. The model works quite well for most vocabulary. Some cases need to be considered even before like acronyms, numbers or anglicism. Finally, the vocabulary is merged with the base dictionary in order to obtain the adapted dictionary.

The training and adaptation of a LM is applied by using the MIT Language Modeling toolkit (MITLM). Within our work, we combine a temporary LM and the base LM using a linear interpolation whereat the first model is trained using the adaptation corpus and the latter one is a more general model which is provided for less context related data collections. The vocabulary of the resulting model emerges as a superset of both, the temporary as well as the base LM.

## 3. Experiments and Results

The experiments are divided into two parts by using an ASR and a representative IR system. In the next subsections, we describe the datasets, the experimental setup and summarize the experimental results for different block sizes.

### 3.1. Datasets

In our experiments we demonstrate how to manage the OOV problem while using distinct training and test datasets. Therefore, we are using training data from the years 2011 and 2012 and a more recent test data collection from late 2013 and the first quarter of 2014.

Both datasets have been acquired from the nationwide well-known German *Tagesschau*. Albeit the public newscast Tagesschau from the television station ARD airs multiple episodes in different lengths a day, we concentrate on a contextual summarized version of 100 seconds with approximately 15 to 20 utterances. These special episodes cover a variety of topics from politics over economy to sports and wheather and are available everyday via webcast ("http://www.tagesschau.de/") and contain heterogeneous content from the most important cutting-edge news clips. The training dataset comprises 208 clips with audio tracks of a total length of more than 6 hours, accompanied by about 46.2k words covering a vocabulary of 8.3k different words. In turn the test set consists of 30 chronologically combined clips that lasts 1 hour with 7.3k words and a vocabulary size of 2.5k in total.

### 3.2. Experimental Setup

We used CMU Sphinx to perform ASR as well as acoustic modeling (AM) and trained gender-dependent triphone Hidden-Markov-Models (HMMs) together with 8 Gaussian mixture models (GMMs). The application of gender detection right before the speech recognition event allows us to apply the appropriate AM. To train a temporary and the base LM, the MITLM toolkit with Kneser-Ney smoothing was used. The resulting LMs are trigram models. The LM adaptation was performed by a linear interpolation of the temporary and the base LM. For both LMs, equal interpolation weights were assigned.

The accumulation of the adaptation corpus was achieved by parsing web articles from the ARD news portal itself. However, these articles neither include any videoclips nor transcripts of the broadcasted news and can therefore guaranteed to be regarded as different. To acquire the articles, the search function of the website was used and the results were prioritized with respect to relevance and date according to the requirements of the test set in order to get more recent contributions. Preliminary test runs showed an improvement of the WER by increasing the adaptation corpus, where the number of retrieved articles was limited to a maximum of 990 entries.

Although our method is presented as unsupervised, the automatic detection of the block size is not part of this work and is to be investigated in the future. We assume that the appropriate block size depends highly on the type of a report. Since the test set belongs to only one type of a report, we used fixed block sizes in each test scenario. The optimal suited block size for the test set of the summarized German broadcast news Tagesschau is determined by a step-wise approach while incrementally increasing the block size by 10. The experiments are repeatedly applied until no improvement in the WER can be estimated or a decrease below the baseline is registered.

To evaluate our approach for improving spoken document retrieval we indexed all generated statements using the Xtrieval Framework [10]. As search and retrieval component Apache Lucene was utilized. During the pre-processing step 231 common German stop words were removed and the German Snowball stemmer was applied to reduce the number of forms of words. Each transcript of an utterance was indexed as a single document, hence the resulting index contained 583 documents. For the retrieval test we created 18 topics using the knowledge of the documents and some were especially aimed at out-of-vocabulary issues. The topics consist of short phrases with two to four words and cover different scopes like political news, sports, and weather (e.g., "edathy affäre", "olympische winterspiele in sotschi", and "regen oder schnee").

Table 1: *Mean perplexity (MPPL), out-of-vocabulary (OOV), word error rate (WER), and mean average precision (MAP) for the test set (Reference), Baseline, and different block sizes (BS).*

| Configuration | MPPL | OOV (%) | WER (%) | MAP |
|---|---|---|---|---|
| Reference | - | - | - | 0.9014 |
| Baseline | 103.7 | 13.4 | 45.6 | 0.5567 |
| BS10 | 116.6 | 8.9 | 39.2 | 0.6073 |
| BS20 | 117.7 | 7.1 | 38.4 | 0.6484 |
| BS30 | 119.1 | 6.3 | 38.1 | 0.6572 |
| BS40 | 119.5 | 6.1 | 38.9 | 0.6686 |
| BS50 | 120.1 | 5.7 | 39.3 | 0.6733 |
| BS60 | 118.9 | 5.4 | 39.3 | 0.6171 |
| BS70 | 120.8 | 5.3 | 41.0 | 0.6056 |
| BS80 | 119.1 | 5.0 | 41.3 | 0.6329 |
| BS90 | 122.1 | 4.9 | 39.7 | 0.5809 |
| BS100 | 120.6 | 4.4 | 41.0 | 0.5834 |
| BS110 | 122.1 | 4.7 | 42.9 | 0.5771 |
| BS120 | 120.2 | 4.3 | 41.5 | 0.6125 |
| BS130 | 122.2 | 4.4 | 41.5 | 0.6061 |
| BS140 | 121.3 | 4.5 | 42.9 | 0.5954 |
| BS150 | 121.7 | 4.0 | 43.4 | 0.5689 |
| BS160 | 122.8 | 4.0 | 44.0 | 0.5828 |
| BS170 | 125.1 | 3.9 | 43.6 | 0.6030 |
| BS180 | 122.9 | 3.9 | 40.3 | 0.5744 |
| BS190 | 123.9 | 3.8 | 41.3 | 0.5894 |
| BS200 | 122.6 | 3.7 | 42.0 | 0.5843 |

For each analysed configuration a new index was built, which contained the transcripts returned by the automatic speech recognition. All topics were searched in these indices and a list of candidate results was assembled. Each result was checked against the transcript corpus for its relevance. The result was a list of relevant and non-relevant results returned by the search. In the last step for each index the mean average precision (MAP) was calculated for all 18 topics.

### 3.3. Experimental Results

A series of experiments was carried out using different block sizes for the adaptation method to optimize a suitable block size for SDR concerning the dataset presented in this work. Each test case was incrementally increased by a block size of 10 to the value of 200, which results in a maximum of three blocks for the test set (583 hypotheses). Basically, the reduction of the WER has a positive effect on the MAP. In order to illustrate the impact of the optimization of speech recognition on the MAP using our approach, we investigate the values of WER, OOV, and the mean perplexity (MPPL) of the adapted LMs for certain block sizes.

The results are shown in Table 1. It can be seen, that using our adaptation method the best result of WER was 38.1% at a block size of 30, which reduces the rate by 7.5% relative to the baseline that was performed without any adaptation. This block size is still relatively small compared to the selected block spec-

trum of the experiments. The block size of 30 comprises more than one episode of the test set and indicates, that neighbouring episodes which are close in time result in a benefit for the adaptation. Additionally, it is obvious that with increasing block size, the OOV rate is reduced. The best result is at a maximum block size of 200 with 3.7% and could thus be reduced by 9.7%. This can be explained by the fact that adaptation data based on utterances which are more distant in time is considered in larger block sizes.

The MPPL could not be improved in any experiment relative to the baseline and increases with larger block sizes, since out-of-domain data continues to expand. Moreover, it was observed that the individual perplexities vary considerably stronger in smaller blocks allowing particular improvements. For instance, at a block size of 10, the perplexity could be reduced in a certain block to a minimum of 56.5 but ranges in other cases up to 241.0 whereas it ranges from 117.3 to 132.1 at a block size of 200. Smaller blocks imply less adaptation data, while recognition errors in the first-pass transcript may lead to inappropriate adaptation results. Therefore, the perplexities can partially be very high. However, the MPPLs in the case of smaller blocks are better concerning our test set, since the number of utterances of similar content is also relatively small. The contents of an episode extend over two to three more adjacent episodes. Larger blocks result in significantly more adaptation data and are much more unspecific concerning content. By this generalization, the individual perplexities are kept high and affect the ASR performance and thus the MAP.

The results in Table 1 show that the manual generated transcript of the test set (reference) with 90.14% outperforms the baseline with 55.66% by 34.47% in MAP. As expected, the automatically generated speech transcripts with different block sizes from our approach perform better than the baseline, but are still inferior to the reference. Our best result is located at a block size of 50 with 67.33% MAP, which is 11.69% above the baseline and 22.81% below the reference. Several factors contribute to the overall high MAP: the short document length (one utterance), the small number of documents, and the domain-specific setting.

# 4. Conclusions

We presented a method to perform a block-wise unsupervised and web-based language model adaptation for summarized sequences of spoken utterances. The main goal of this work was to improve the performance of spoken document retrieval by reducing the out-of-vocabulary rate. Experimental results showed that our method has a direct impact on document retrieval at the example of German broadcast news with 18 different topics. The best improvement of the mean average precision in comparison to the baseline was about 11.7%.

However, the remaining gap compared to the manually generated transcripts could be further reduced by a more sophisticated general language model. Further improvements may comprise an adjustment of the weights for the linear interpolation. Additionally, resources like RSS feeds or Web 2.0 might be useful as adaptation data. We are also interested in applying the automatic detection of optimal block sizes especially for dynamically changing types of reports and its contents.

# 5. Acknowledgements

This work was realized as part of the project Chrooma+ supported by the Sächsische Aufbaubank within the European Social Fund in the Free State of Saxony, Germany and the project ValidAX funded by the Federal Ministry of Education and Research, Germany.

# 6. References

[1] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "The trec spoken document retrieval track: A success story," in *Proceedings of the 6th International Conference on Computer-Assisted Information Retrieval (RIAO), College de France, France, April 12-14, 2000*, J.-J. Mariani and D. Harman (Eds.), CID, 2000, pp. 1–20.

[2] L. Chen, L. Lamel, J.-L. Gauvain, and G. Adda, "Dynamic language modeling for broadcast news," in *8th International Conference on Spoken Language Processing (INTERSPEECH — ICSLP), Jeju Island, Korea*. ISCA, 2004, pp. 997–1000.

[3] S. Meng, K. Thambiratnam, Y. Lin, L. Wang, G. Li, and F. Seide, "Vocabulary and language model adaptation using just one speech file," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 5410–5413.

[4] G. Lecorvé, G. Gravier, and P. Sebillot, "An unsupervised web-based topic language model adaptation method," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 5081–5084.

[5] A. Tsiartas, P. Georgiou, and S. Narayanan, "Language model adaptation using www documents obtained by utterance-based queries," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 5406–5409.

[6] T. Schlippe, L. Gren, N. T. Vu, and T. Schultz, "Unsupervised Language Model Adaptation for Automatic Speech Recognition of Broadcast News Using Web 2.0," in *The 14th Annual Conference of the International Speech Communication Association (Interspeech)*, Lyon, France, 2013, pp. 2698–2702.

[7] R. Iyer and M. Ostendorf, "Relevance weighting for combining multi-domain data for n-gram language modeling," *Computer Speech & Language*, vol. 13, no. 3,1999, pp. 267–282.

[8] K. Saykham, A. Chotimongkol, and C. Wutiwiwatchai, "Online temporal language model adaptation for a thai broadcast news transcription system," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias (Eds.), European Language Resources Association (ELRA), 2010, pp. 1690–1694.

[9] R. Knauf, J. Kürsten, A. Kurze, M. Ritter, A. Berger, S. Heinich, and M. Eibl, "Produce. annotate. archive. repurpose — Accelerating the composition and metadata accumulation of tv content," in *Proceedings of the 2011 ACM International Workshop on Automated Media Analysis and Production for Novel TV Services*, ser. AIEMPro. New York, NY, USA: ACM, 2011, pp. 31–36.

[10] J. Kürsten and T. Wilhelm, "Extensible retrieval and evaluation framework: Xtrieval," in *Proceedings of LWA — Lernen, Wissen, Adaption*, J. Baumeister and M. Atzmüller (Eds.), vol. 448. Department of Computer Science, University of Würzburg, Germany, 2008, pp. 107–110.