Improving Transcript-Based Video Retrieval Using Unsupervised Language Model Adaptation

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

Technische Universität Chemnitz, 09107 Chemnitz, Germany
{wilt,robeh,ritm,eibl}@hrz.tu-chemnitz.de

Abstract. One challenge in automated speech recognition is to determine domain-specific vocabulary like names, brands, technical terms etc. by using generic language models. Especially in broadcast news new names occur frequently. We present an unsupervised method for a language model adaptation, which is used in automated speech recognition with a two-pass decoding strategy to improve spoken document retrieval on broadcast news. After keywords are extracted from each utterance, a web resource is queried to collect utterance-specific adaptation data. This data is used to augment the phonetic dictionary and adapt the basic language model. We evaluated this strategy on a data set of summarized German broadcast news using a basic retrieval setup.

Keywords: language modeling, out-of-vocabulary, spoken document retrieval, unsupervised adaptation.

1 Introduction

Today, there is a growing amount of videos. These videos, after being produced and published, need to be made accessible. Spoken document retrieval tries to improve the access to content, which is not properly annotated [1]. By employing automatic speech recognition (ASR) videos are transcribed so they can be processed using classical information retrieval.

In the domain of broadcast news there is a challenge, because new words, particularly named entities, are a very common phenomenon. When these words are unknown to the automatic speech recognition engine, it is unable to recognize them correctly. This problem is called out-of-vocabulary (OOV). Out of vocabulary has a serious impact on the information retrieval performance, because names, which were not recognized, cannot be searched for subsequently.

To add a word to an ASR system, which is yet unknown, three essential steps are required: At first the missing word must be identified, second a phonetic notation for the word is needed and third the language models must be adapted to include the word. Even though all three steps can be assisted manually, we want to focus on a fully automatic solution to this problem.

The required data to adapt the language model can be acquired by using the result of preliminary ASR transcripts as queries to an information retrieval

E. Kanoulas et al. (Eds.): CLEF 2014, LNCS 8685, pp. 110–115, 2014.

(IR) system [2, 3]. Web-resources like standard web pages [3, 4, 5], RSS feeds or Twitter [6] provide easy available and accessible sources for the information retrieval system.

However, the enrichment of the language model with external data, in particular with out-of-domain data, does not necessarily results in an improved recognition [7]. Saykham et al. [8] showed minor improvements by adapting a language model using recent online news texts.

2 Method

Our approach is based on the assumption that utterances in speech, which appear in temporal proximity to each other, have a common topic and share some vocabulary. Therefore we performed a block-based, unsupervised adaptation of a general language model. Additionally the base dictionary is enriched with the new phonetically transcribed vocabulary.

Our method utilizes an ASR system with a two-pass decoding strategy. First a transcript of the speech was generated by the ASR system. The segmentation of the transcript into several units was performed by the recognizer of this system. It divides segments when silences occur for an extended time. Segments ranged from short statements to whole sentences. The segments were separated into blocks of specific sizes. All segments of each block were processed and used as web queries for retrieving the adaptation data to build a block-specific dictionary and language model. Finally, the speech was transcripted again using the adapted dictionaries and language models for the corresponding blocks.

We employed a keyword extraction algorithm based on natural language processing (NLP) to reduce the word count and to narrow down the web query. The following prioritized constraints were used:

1. Nouns and named entities are directly applied to the web query.
2. If 1. returns no results, only the named entities are used as a web query.
3. If 2. returns no results, the sequence of named entities is recursively decomposed into two parts until there is any retrieved result.

The articles of the retrieved HTML pages were extracted and special characters, acronyms and numbers are converted in order to be conform to the conventions of the adaptation process and the ASR system. For example numbers are transformed to their corresponding words, because otherwise it cannot be used for the next step, where the phonetic dictionary is amended.

The phonetic dictionary adaptation aims to enrich a basic dictionary with new vocabulary of the adaptation corpus. Vocabulary from the adaptation corpus was extracted and compared to the basic dictionary. Then, the new vocabulary was phonetically transcribed using a grapheme-to-phoneme (G2P) decoder and merged into a temporary dictionary. Finally, the temporary and the basic dictionary are merged to the adapted dictionary.

Our basic language model is a general model trained on topic-independent data collection. On basis of the adaption corpus an intermediate language model was trained, which was merged with our basic language model. The vocabulary of the resulting model is finally a superset of the vocabulary of both models.

Our training set was created using short video clips of a popular German news broadcast show called “Tagesschau”. The full show is aired each day on the German television station “Das Erste”. It covers a variety of topics including politics, economy, sports and weather. There is an additional format called “Tagesschau in 100 Sekunden”, which is a summary of the most important video clips. Each show is separately produced and has a length of about 100 seconds with approximately 20 sentences and is publicly available as a webcast, which can easily be downloaded. Since 2011 we have collected various of these web clips. Our experiments were performed using a training set of the years 2011 and 2012 and a test set of 2013 and the first quarter of 2014. Therefore, we are able to investigate the out-of-vocabulary problem concerning the more recent data. The training set for acoustic as well as language modeling consists of 208 clips with a vocabulary of 8,300 words and more than 6 hours of speech. The test set is a sequence of 30 chronological combined clips with a vocabulary size of 2,500 and a total duration of 1 hour.

We used CMU Sphinx to perform acoustic modeling and train a gender-dependent triphone Hidden-Markov-Models together with eight Gaussian mixture models. The application of gender detection right before the speech recognition event allows us to apply the appropriate acoustic model. To train the basic language model, the MITLM toolkit with Kneser-Ney smoothing was used. The adaptation of the intermediate and the basic language model was performed using a linear interpolation with fixed weights.

The adaptation corpus was constructed by parsing web articles from the “Tagesschau” news portal. Even though this is the same source as for the video clips, these articles did not include transcripts of the broadcasted news and were edited independent from the video clips. Therefore they can be regarded as different. To acquire relevant articles the search function of this website was used and the results were prioritized with respect to relevance and date according to the requirements of the test set.

We tested different block sizes to determine an optimal size. For each test we increased the block size by 10 utterances. The experiments were continued until the result of the word error rate [9] (WER) was below the baseline or no further improvements were observable.

To evaluate our approach for improving spoken document retrieval we indexed all recognized statements using the Xtrieval Framework [10]. An unaltered and standard Apache Lucene version 4.3.1 was used as search and retrieval component. During the pre-processing step 231 common German stop words[1] were removed and the German Snowball stemmer was applied. Each utterance was indexed as a single document.,The resulting index contained 583 documents. For the retrieval test we created 18 topics based on knowledge of the documents. Some were especially aimed at out-of-vocabulary issues. The topics cover different scopes like political news, sports, and weather (e.g. “edathy affäre”,

[1] Retrieved from http://snowball.tartarus.org/algorithms/german/stop.txt

"olympische winterspiele in sotschi", and "regen oder schnee"). All topics were formulated into keyword queries using OR as conjunction operator.

For each block size a new index was built. 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 the mean average precision (MAP) was calculated for each block size.

3 Results and Discussion

The results in Table 1 show that the best word error rate of 38.1 percent was achieved at a block size of 30 utterances. Compared to the baseline, where no adaptation was performed, this is an improvement of 7.5 percent.

The mean average precision of the manually generated transcripts outperformed the baseline with 90.14 versus 55.66 percent. The automatically generated speech transcripts perform better than the baseline, but are still inferior to the manual transcripts. Our best result was at a block size of 50 utterances and achieved a mean average precision of 67.33 percent. This is 11.69 percent above the baseline, but 22.81 percent below the manual transcripts. The reduction of the word error rate has a positive effect on the mean average precision.

As the video clips have a size of approximately 20 utterances and are processed in chronological order, the block sizes of 30 and 50 indicate that consecutive video clips support each other. This is probably due to shared topics and thus some common vocabulary.

With increasing block size the OOV-rate is reduced. The best result is at the maximum block size of 200 with 3.7 percent and could therefore be reduced

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
BS150	121.7	4.0	43.4	0.5689
BS200	122.6	3.7	42.0	0.5843

by 9.7 percent compared to the baseline. 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 mean perplexity rose slowly for growing block sizes, probably because out-of-domain data got increasingly included in the language model. For small block sizes the perplexity on the utterance level varies more than for big block sizes. For instance, at a block size of 10 the perplexity for a certain block was as low as 56.5 but for another block it was as high as 241.0. Whereas at a block size of 200 it ranged from 117.3 to 132.1.

4 Conclusions

We presented an unsupervised method for a language model adaptation, which is used in automated speech recognition with a two-pass decoding strategy to improve spoken document retrieval on broadcast news. The experiments showed an out-of-vocabulary reduction and as a result a better performing document retrieval by applying the new models to the corresponding blocks.

The biggest improvement of mean average precision in comparison to the baseline was about 11.7 percent at a block size of 50 utterances. The best word error rate of 38.1 percent was achieved at a block size of 30. Both values were close together and show that the retrieval benefits from an improved recognition.

The remaining gap compared to the manually generated transcripts could be further reduced by using appropriate context dependent acoustic models. Further improvements may comprise an adjustment of the weights of linear interpolation of the language models. Since the test set was of German language a decomposition strategy could be beneficial towards a better mean average precision. Additionally, resources like RSS feeds or Web 2.0 might be useful as adaptation data. We intend to evaluate this method using different types of speech, e.g. a mixture of broadcast news, advertisement and talk shows. Furthermore, it would be interesting to use additional resources from the web, like Twitter and RSS Feeds as adaptation data.

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.

References

[1] Garofolo, J.S., Auzanne, C.G.P., Voorhees, E.M.: The trec spoken document retrieval track: A success story. In: Mariani, J.J., Harman, D. (eds.) RIAO, CID, pp. 1–20 (2000)

[2] Chen, L., Lamel, L., Gauvain, J.L., Adda, G.: Dynamic language modeling for broadcast news. In: 8th International Conference on Spoken Language Processing (INTERSPEECH), pp. 997–1000 (2004)

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

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

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

[6] Schlippe, T., Gren, L., Vu, N.T., Schultz, T.: 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), pp. 2698–2702 (2013)

[7] Iyer, R., Ostendorf, M.: Relevance weighting for combining multi-domain data for n-gram language modeling. Computer Speech and Language 13(3), 267–282 (1999)

[8] Saykham, K., Chotimongkol, A., Wutiwiwatchai, C.: Online temporal language model adaptation for a thai broadcast news transcription system. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) LREC. European Language Resources Association (2010)

[9] Klakow, D., Peters, J.: Testing the correlation of word error rate and perplexity. Speech Communication 38(1-2), 19–28 (2002)

[10] Kürsten, J., Wilhelm, T.: Extensible retrieval and evaluation framework: Xtrieval. In: Baumeister, J., Atzmüller, M. (eds.) LWA. Volume 448 of Technical Report, Department of Computer Science, University of Würzburg, Germany, 107–110 (2008)