Unsupervised Language Model Adaptation using Utterance-based Web Search for Clinical Speech Recognition

Robert Herms[1], Daniel Richter[2], Maximilian Eibl[1], and Marc Ritter[2]

[1]Chair Media Informatics, [2]Junior Professorship Media Computing,
Technische Universität Chemnitz, 09107 Chemnitz, Germany
{robert.herms,daniel.richter,maximilian.eibl,marc.ritter}
@cs.tu-chemnitz.de

Abstract. In this working notes paper we present our methodology in clinical speech recognition for the Task 1.a.1 of the CLEF eHealth Evaluation Lab 2015. The goal of this task is to minimize the word-detection errors. Our approach is based on the assumption that each spoken clinical document has its own context. Hence, the recognition system is adapted for each document separately. The proposed method performs two-pass decoding whereas the first transcript is processed to queries which are used for retrieving web resources as adaptation data to build a document-specific dictionary and language model. The second pass decodes the same document using the adapted dictionary and language model. The experimental results show a reduction of the insertion errors in comparison to the baseline system, but no improvement of the overall incorrectness percentage across all spoken documents.

Keywords: Speech recognition, Language modeling, Unsupervised adaptation, Information retrieval, Clinical texts

1 Introduction

In general, the creation of acceptable transcripts of spoken language requires high human intervention and remains time- as well as cost-intensive. Since manual generated transcriptions are a challenging task, especially for large and heterogeneous datasets, it is more appropriate to apply automatic speech recognition (ASR). In the medical domain, ASR supports a typical handover workflow as a first step by transforming verbal clinical information into electronic structured records. The CLEF eHealth Evaluation Lab 2015 [1] aims to ease patients and nurses in understanding and accessing eHealth information. The goal of Task 1.a [2] is to convert verbal nursing handover to free-form text documents, whereas the challenge of Task 1.a.1 is to minimize word-detection errors by addressing the correctness of the speech recognition engine itself.

In this connection, out-of-vocabulary (OOV) has a serious impact on ASR results. It necessarily requires the utilization of domain-specific language models (LMs) in order to cope with the huge amount of data and different topics. For this purpose, the adaptation of a generic LM with a more specific LM using weighted linear interpolation is a common way. Supervised LM adaptation is very costly for huge amount of data and different topics, since the generation of specific corpora takes a lot of time. A conclusive way is an unsupervised method, which takes the context of a situation into account. As described in [3], it is not suitable for unsupervised adaptation to use the hypothesis of an ASR system as adaptation data. This is due to the fact, that automatic generated transcripts contain recognition errors and do not counteract the OOV problem. However, transcripts can be processed to queries and used in an information retrieval system, e.g., [3, 4]. Resources such as specific corpora or the web with HTML pages (e.g., [4–6]), RSS Feeds and Twitter (e.g., [7]) are very useful in order to obtain further textual data for the LM adaptation. Moreover, this enables to get new specific vocabulary for covering the OOV (names, brands, technical terms, etc.). Additional data especially from out-of-domain does not always lead to improvements (e.g., [8]). In contrast, domain-specific data is helpful to address certain topics. Hence, the work [5] proposed a complete unsupervised technique based on information retrieval methods to build a thematically coherent adaptation corpus using the web. However, in [4] was clarified that the application of topic specific LMs is not easy to handle for an out-of-the-box ASR system, especially, if the topic is very heterogeneous or the contents change dynamically.

In this working notes paper we present our methodology and the results we obtained in Task 1.a.1 of the CLEF eHealth Evaluation Lab 2015. Our approach is based on the assumption, that each spoken clinical document has its own context. Therefore, we suggest adapting ASR for each document separately. The proposed method uses a two-pass decoding strategy. First, the transcript of a document is generated by an ASR system. Keywords of the utterances are extracted and used as queries in order to retrieve web resources as adaptation data to build a document-specific dictionary and LM. Finally, re-decoding of the same document is performed using the adapted dictionary and LM. The developed system was already applied in the previous works [9] and [10].

This Paper is organized as follows: In the next section we present our method for unsupervised language model adaptation in clinical speech recognition. In Section 3 we describe the applied dataset, the experimental setup, and the evaluation results. Finally, we conclude this paper in Section 4 and give some future directions.

2 Adaptation Method

Our method works out-of-the-box with a two-pass decoding strategy. First, a transcript of utterances in the spoken document is generated by ASR. The segmentation of the transcript into several units is performed by the recognizer itself using long silences. Each segment ranges from a short statement to a whole sentence. The segments are processed and used as queries for retrieving adaptation data to build a document-specific dictionary and LM. The second pass of the recognizer decodes the same document using the adapted dictionary and LM.

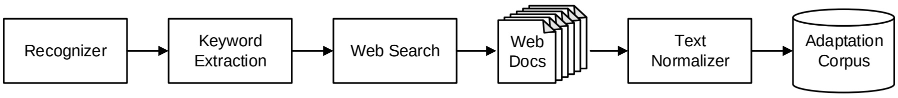

Fig. 1. Process chain for the retrieval of web-based adaptation data using the transcript of recognized speech.

2.1 Retrieval of Adaptation Data

As shown in Fig. 1, a transcribed segment generated by the ASR system is used for building a query in order to perform a web search. Since a segment often contain more words than useful for a web search query, especially for retrieving documents in a close context, the following steps are performed to limit their number:

1. Nouns, plural nouns and the corresponding adjectives are extracted to obtain the most meaningful words.
2. A pre-defined stop-word list is applied which is derived from the training data and contains unnecessary as well as recurring vocabulary (e.g., date and time specification)
3. If 2. yields more keywords than a predefined threshold, the sequence of words is split into several parts with almost the same number of words fulfilling the requirements and each of these parts is considered to be a separate query. Otherwise there is only one query.

For each resulting query a web search is conducted. The amount of the retrieved web documents is combined and normalized before adding to the adaptation corpus. In detail, the pure articles of the retrieved web documents are extracted and special characters, acronyms and numbers are converted in order to be conform to the conventions of the pending adaptation process and the ASR system. These steps are performed for all transcribed segments of one spoken document and their normalized texts are accumulated to one corresponding adaptation corpus.

2.2 Dictionary and LM Adaptation

The accumulated adaptation corpus is used for modifying the base dictionary and the base LM as illustrated in Fig. 2. The pronunciation dictionary adaptation aims to enrich a base dictionary by new vocabulary coming from the adaptation corpus. For this purpose, the vocabulary of the corpus is extracted and compared to the base dictionary. The additional vocabulary is phonetically transcribed by a grapheme-to-phoneme (G2P) decoder and combined into a temporary dictionary. Finally, the temporary and the base dictionary are merged to an adapted dictionary.

The LM adaptation is performed by a weighted linear interpolation of the temporary and the base LM. The temporary LM is trained by means of the

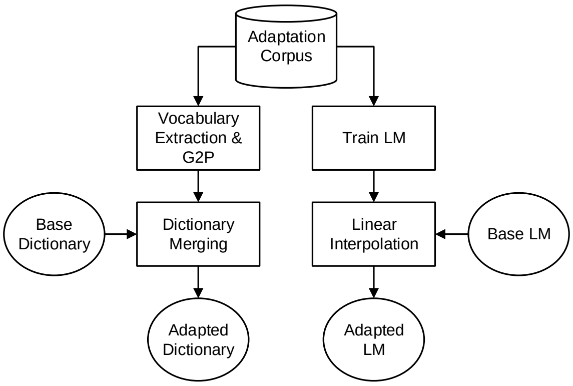

Fig. 2. Procedure of the pronunciation dictionary and LM adaptation based on the accumulated adaptation corpus.

adaptation corpus, whereas the base LM is a more general model trained on topic-independent data collections. Finally, the vocabulary of the resulting model is a superset of the vocabulary of both, the temporary and the base LM.

3 Experiments and Results

Before describing the details of the experimental setup, the used dataset is introduced and some observations on conducted preliminary experiments are stated. Afterwards we discuss our experimental intermediate as well as final results of the evaluation.

3.1 Dataset

In this work the NICTA Synthetic Nursing Handover Data dataset [11] is used which was created at NICTA in 2012-2014 for clinical speech recognition and information extraction related to nursing shift-change handover. The training set as well as the test set consist of 100 written, free-form text documents and the corresponding recorded audio files spoken by an Australian registered nurse with over twelve years of working experience. The text documents of the test set were not released for evaluation purposes. Furthermore, this dataset includes recordings lasting about half an hour of her reading an excerpt of “The Final Odyssey” as initialization data for speech recognition engines.

In a preliminary experiment the ASR output of the training set was compared to the written, free-form text documents. This exposed some common errors, which are partially already mentioned in the task description. Further investigation revealed, that there are some abbreviations used instead of the correctly spelled words. However, these words are verbalised correct by the nurse

Table 1. List of typical abbreviations in free-form text documents of the training set which were predominantly used instead of the correct spelling.

Correct spelling	Used abbreviation
years	yrs
hours	hrs
doctor	dr
antibiotics	abs
arteriovenous fistula	avf
blood pressure	bp
blood pressures	bps
hypertension	hpn
level of consciousness	loc
prednisolone	pred

in the provided audio files. Therefore, a list of usual misspellings was created, which should be used instead of the correct words generated by the standard configuration of the ASR system. For instance, only 26 appearances of the correctly spelled word "years" were counted, but 53 appearances of the abbreviation "yrs". Using the abbreviation should lead to less substitutions counted for the final evaluation metric. More detected substitutions are shown in Table 1. Beside these misspellings in the text documents, there are some misspeaks and following corrections by the nurse as well as some filler words like "ehm". Another observation is the usage of numerals and numbers in the written, free-form text documents. Only 22 numerals from "one" to "eight" are used but 266 numbers up to three digits (more than half of this with only one digit) were found. Hence, always using numbers instead of numerals seems promising.

3.2 Experimental Setup

ASR was performed using the engine of the open-source framework sphinx-4 [12]. As a basic configuration, we used already existing components. We applied the acoustic model "HUB4" (http://www.speech.cs.cmu.edu/sphinx/models/), which has been trained using 140 hours of 1996 and 1997 hub4 training data. It includes 3-state within-word and cross-word triphone Hidden-Markov-Models with 8 Gaussian mixture models. We performed maximum a posteriori (MAP) adaptation by using the initialization data of the training set to update the parameters of the acoustic model in order to better match the observed data. Next, we used the pronunciation dictionary "cmudict.0.7a_SPHINX_40" (https://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/sphinxdict/), which comprises 133k words and the corresponding phonetic transcription. We modified the dictionary concerning the notation of the clinical reports, for instance, adding new vocabulary or replacing words with abbreviations. Moreover, we assigned some vocabulary to filler words to avoid misinterpretation caused by fillers (e.g., "ah" or "ehm"). As a generic LM we used the "US English Generic

Language Model" (http://sourceforge.net/projects/cmusphinx/files/) comprising about 3.2M trigrams. The utilization of the generic LM on the free-form text documents of the training set results in a perplexity of 324.8 and 1.1k OOV words. In our experiments language modeling was conducted with the SRILM toolkit [13]. We performed LM adaptation using the generic LM and the free-form text documents of the training set in order to obtain our base LM for the proposed method. Our goal was to generate a background model which has the properties of clinical documents as well as an appropriate generalization. Hence, we assigned equal interpolation weights.

Concerning the proposed adaptation method, the temporary LMs were constructed as trigram models using Kneser-Ney smoothing. These models were combined with the base LM by means of a weighted linear interpolation in order to perform the LM adaptation. We used the WFST-driven G2P framework Phonetisaurus [14] to phonetically transcribe temporary dictionaries. For this purpose, we trained a G2P model based on 133k words from the applied pronunciation dictionary, which works stable for typical English words.

The accumulation of the adaptation corpus was achieved by parsing the website of the Journal of Postgraduate Medicine (http://www.jpgmonline.com). To accomplish this, a segment of the first-pass transcript is processed by the Stanford Lexical Parser [15] to extract keywords. We assigned a threshold of 5 for the keyword extraction, i.e., if there are more than five words left, the sequence of words is split into separate parts with an almost equal number of words by trying to keep adjectives and their corresponding nouns together. Considering the next noun belonging to an adjective results in parts of up to seven words. Each part is considered to be a separate query which is utilized in the search function of the web portal. The resulting list of full-text articles was prioritized concerning a relevance of 50% and higher. We limited the maximum number of retrieved articles to 100.

3.3 Results

The intermediate results of our adaptation method are relevant for further processing steps and consequently for the final results. Table 2 gives an overview of the mean values concerning the retrieval process across all spoken documents. Comparing the numbers of segments per spoken documents leads to the conclusion, that the speech recognizer was able to detect much more pauses in the training set than in the test set. Fewer segments in the test set also lead to fewer but longer queries built by the system, as seen in line two and three of Table 2. As the number of web documents retrieved per query is almost equal for both datasets, the number of web documents per spoken document is also much higher for the training set than for the test set.

The number of tokens per web document (after normalization) are quite similar. Hence, the differences in the resulting adaptation corpora for the training and the independent test set, as shown in Table 3, can only be traced back to the differences in the number of segments per spoken document. The statistics in Table 3 indicate that more tokens in the adaptation corpus lead to more types

Table 2. Mean values concerning the retrieval process of the adaptation corpus across all spoken documents in the training set and the independent test set.

	Train set	Test set
# Segments per spoken document	6.1	2.3
# Queries per spoken document	7.0	4.9
# Tokens per query	2.8	3.6
# Web documents per query	47.9	49.2
# Web documents per spoken document	337.2	238.6
# Tokens per web document	1,311.0	1,324.0

Table 3. Statistics of the resulting adaptation corpus across all spoken documents in the training set and the independent test set.

	Train set			Test set		
	Tokens	Types	New Types	Tokens	Types	New Types
Min	18.0k	3.9k	0.3k	31.6k	5.1k	1.0k
Max	1,074.7k	34.7k	14.9k	629.6k	25.0k	9.8k
Mean	441.9k	20.4k	7.5k	315.9k	17.4k	5.9k
Median	394.1k	20.0k	7.4k	299.3k	17.3k	5.8k
SD	248.7k	6.1k	3.1k	155.8k	4.9k	2.3k

and also to more new types, which were used to extend the dictionary of the ASR system for the second pass.

A series of speech recognition experiments was carried out using the two different interpolation weights λ=0.8 and λ=0.9 for the proposed adaptation method on the training set and the independent test set. These weights imply a higher preference of the base LM which was designed for the clinical free-form text documents. The results are illustrated in Table 4. The baseline results were achieved by Dragon Medical 11.0 which was trained on the initialization data and applied with the vocabulary of nursing. In general, it can be seen that for both datasets our system had more substitution errors and deleted some more words than the baseline system. We considered the specific notation of the written, free-form text documents such as abbreviations or fillers, which leads to less inserted words than the baseline system. Compared to the baseline, the mean value of the incorrectness percentage across all documents in the training set was improved by our system with 2.1% (λ=0.9). The reason for that is the consideration of the specific notation of the written, free-form text documents and the adaptation of the generic LM using the free-form text documents of the training set which increases the probability of recognizing the correct words.

The performance measurements of our system on the test set in comparison to the training set concerning the mean of the incorrectness percentage across all documents show similar results with a decrease of 3.6% (λ=0.9). Our system generated many substitution errors on the test set with 36.6% and a difference

Table 4. Performance measurement on the training set and the independent test set using different interpolation weights (λ) of the proposed method for unsupervised language model adaptation.

	Train set			Test set		
	Baseline	λ=0.8	λ=0.9	Baseline	λ=0.8	λ=0.9
% Correct words	72.3	64.6	65.6	73.1	53.7	54.3
% Substituted words	24.1	28.6	27.9	22.6	36.7	36.6
% Deleted words	3.6	6.8	6.6	4.3	9.6	9.1
% Inserted words	28.2	19.5	19.6	11.6	6.5	6.6
% Incorrect words	55.9	54.9	54.1	38.5	52.8	52.3
Incorrectness percentage across all documents						
Min	30.2	22.0	22.0	20.7	26.2	26.2
Max	137.5	142.5	142.5	59.1	92.0	92.0
Mean	57.8	56.5	55.7	39.5	52.6	52.1
Median	55.5	52.9	53.0	39.1	51.7	51.6
SD	17.0	18.6	18.4	9.8	13.1	12.9

of 14.0% to the baseline that is crucial for the overall incorrectness percentage. We achieved a mean value of 52.1% and consequently 12.6% over the baseline. However, we could achieve a reduction of the insertion errors with 5.1% (λ=0.8) and 5.0% (λ=0.9). All in all, the evaluation on the test set shows that our system did not improve the baseline mean incorrectness percentage. The configuration with the interpolation weight λ=0.9 was just slightly better than the lower one.

4 Conclusions

We presented a method for unsupervised language model adaptation in automatic speech recognition for the Task 1.a.1 of the CLEF eHealth Evaluation Lab 2015. Our approach is based on the assumption, that each spoken clinical document has its own context. Hence, the recognition system is adapted for each document separately. The method uses a two-pass decoding strategy, whereas the first transcript is processed to queries, which are used for retrieving web resources as adaptation data to build a document-specific dictionary and language model. The second pass of the speech recognizer decodes the same document using the adapted dictionary and language model. The experimental results on the test set showed a reduction of the insertion errors in comparison to the baseline system. We achieved a mean value of 52.1% incorrectness across all documents. All in all, we did not improve the baseline incorrectness percentage, since our system produces more substitution errors and deleted some more words. The configuration of our method with the interpolation weight λ=0.9 was just slightly better than λ=0.8.

However, further improvements could be achieved by a more sophisticated selection of the retrieved adaptation data. For instance, a model for disease classification in text corpora could be helpful to obtain only specific adaptation data for the corresponding spoken document. Moreover, it would be interesting to use further resources from web, like Twitter and RSS Feeds. For future work, the investigation of phonetics for accented speech and consequently the application of pronunciation modeling should enhance the performance of the recognition system.

Acknowledgments. This work was partially funded by the German Federal Ministry of Education and Research within the project MACeLot (funding code 16SV7260) and the program of Entrepreneurial Regions InnoProfile-Transfer in the project group localizeIT (funding code 03IP608X).

References

1. Goeuriot, L., Kelly, L., Suominen , H., Hanlen, L., Névéol, L., Grouin, C., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth Evaluation Lab 2015. CLEF 2015 - 6th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS). Springer (2015)
2. Suominen, H., Hanlen, L., Goeuriot, L., Kelly, L., J F Jones, G.: Task 1a of the CLEF eHealth Evaluation Lab 2015: Clinical speech recognition. Working Notes of the CLEF 2015 - 6th Conference and Labs of the Evaluation Forum. (2015)
3. Chen, L., Lamel, L., Gauvain, J.-L., Adda, G.: Dynamic language modeling for broadcast news. In: 8th International Conference on Spoken Language Processing, pp. 997–1000. INTERSPEECH, Jeju Island, Korea (2004)
4. Meng, S., Thambiratnam, K., Lin, Y., Wang, L., Li, G., Seide, F.: Vocabulary and language model adaptation using just one speech file. In: IEEE International Conference on Acoustics Speech and Signal Processing, pp. 5410–5413. ICASSP (2010)
5. Lecorvé, G., Gravier, G., Sebillot, P.: An unsupervised web-based topic language model adaptation method. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5081–5084. ICASSP (2008)
6. Tsiartas, A., Georgiou, P., Narayanan, S.: Language model adaptation using www documents obtained by utterance-based queries. In: IEEE International Conference on Acoustics Speech and Signal Processing, pp. 5406–5409. ICASSP (2010)
7. 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, pp. 2698–2702. INTERSPEECH, Lyon, France, (2013)
8. Iyer, R., Ostendorf, M.: Relevance weighting for combining multi-domain data for n-gram language modeling. Computer Speech & Language, vol. 13, no. 3, pp. 267–282. (1999)
9. Herms, R., Ritter, M., Wilhelm-Stein, T., Eibl, M.: Improving Spoken Document Retrieval by Unsupervised Language Model Adaptation Using Utterance-Based Web Search. In: 15th Annual Conference of the International Speech Communication Association, pp. 1430–1433. INTERSPEECH, Singapore (2014)
10. Wilhelm-Stein, T., Herms, R., Ritter, M., Eibl, M.: Improving Transcript-Based Video Retrieval Using Unsupervised Language Model Adaptation. In: Information Access Evaluation. Multilinguality, Multimodality, and Interaction, pp.110–115. Springer (2014)

11. Suominen, H., Zhou, L., Hanlen, L., Ferraro, G.: Benchmarking clinical speech recognition and information extraction: New data, methods, and evaluations. JMIR Medical Informatics. (2015)
12. Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., Woelfel, J.: Sphinx-4: A Flexible Open Source Framework for Speech Recognition. Technical Report. Sun Microsystems, Inc., Mountain View, CA, USA. (2004)
13. Stolcke, A., Zheng, J., Wang, W., Abrash, V.: SRILM at sixteen: Update and outlook. In: Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop, p.5. (2011)
14. Novak, J. R.: Phonetisaurus: A wfst-driven phoneticizer. The University of Tokyo, Tokyo Institute of Technology, pp.221–222. (2011)
15. Klein, D., Manning, C. D.: Fast Exact Inference with a Factored Model for Natural Language Parsing. In: Advances in Neural Information Processing Systems 15 (NIPS 2002), Cambridge, MA: MIT Press, pp. 3–10. (2003)