In Proceedings of the Tenth European Conference on Speech Communication and Technology - EuroSpeech / Eighth INTERSPEECH, Antwerp, Belgium, August 2007

Web-Based Language Modelling for Automatic Lecture Transcription

Cosmin Munteanu[1], *Gerald Penn*[1,2], *Ron Baecker*[1,2]

[1]Department of Computer Science, University of Toronto, Toronto, M5S 3G4, Canada
[2]Knowledge Media Design Institute, University of Toronto, Toronto, M5S 2E4, Canada
mcosmin@cs.toronto.edu, gpenn@cs.toronto.edu, rmb@kmdi.toronto.edu

Abstract

Universities have long relied on written text to share knowledge. As more lectures are made available on-line, these must be accompanied by textual transcripts in order to provide the same access to information as textbooks. While Automatic Speech Recognition (ASR) is a cost-effective method to deliver transcriptions, its accuracy for lectures is not yet satisfactory. One approach for improving lecture ASR is to build smaller, topic-dependent Language Models (LMs) and combine them (through LM interpolation or hypothesis space combination) with general-purpose, large-vocabulary LMs. In this paper, we propose a simple solution for lecture ASR with similar or better Word Error Rate reductions (as well as topic-specific keyword identification accuracies) than combination-based approaches. Our method eliminates the need for two types of LMs by exploiting the lecture slides to collect a web corpus appropriate for modelling both the conversational and the topic-specific styles of lectures.

Index Terms: speech recognition, language modelling, corpus building, topic dependent, lecture transcription.

1. Introduction

Internet broadcasting (webcasting) is becoming an increasingly popular method of delivering lectures and academic presentations over the Internet. At the same time, more of these media are being archived and accessed by users through interactive systems such as ePresence (`http://epresence.tv/`). However, without transcripts, users of webcast lectures are faced with far greater difficulty in performing tasks that are easily achieved with archives of text documents, such as retrieval, browsing, or skimming.

Currently, due to adverse acoustic and linguistic characteristics, ASR systems do not perform satisfactorily in domains such as lectures or conference presentations. Most lecture recognition systems achieve Word Error Rates (WERs) of about 40-45% [1, 2], quite far from the minimum WER of 25% for a transcript to be useful and accepted by users as determined in another study [3] (some reports suggest a 20-30% WER for lectures recorded in more artificial and better controlled conditions [4, 5]).

Significant research efforts are dedicated to the improvement of LMs for lecture transcription, the main goal being finding appropriate methods of modelling the dual nature of lecture speech: it is characterized as large-vocabulary, continuous-speech, speaker-independent and as topic- and domain-specific. Typically, solutions for this problem were sought by building separate LMs targeting the reduction of WER independently for each trait, while the recognition was performed using either interpolated LMs or separate models followed by a combination of the resulting hypothesis spaces.

One of the disadvantages of previous approaches is the more complicated process of determining and acquiring the most appropriate corpora for building two or more separate models. Other shortcomings include the need to fine-tune the interpolation or hypothesis space combination parameters and the difficulties in (automatically or manually) extracting reliable topic-specific keywords from additional knowledge sources (e.g., lecture slides). In our research, we propose an approach that eliminates the need for multiple models, yet achieves similar or better WER reductions. Our method exploits the slides used in the lecture or presentation to be transcribed; by using the entire content of the slides as web-search queries, it retrieves web corpora that can be used to directly train a single LM suitable for both the conversational and the topic-specific styles of lectures. In this paper, we will present the implementation details of our method, discuss several web retrieval and training alternatives, and compare the best solution with current interpolation-based LM adaptation methods.

2. Related Work

Automatic lecture transcription is one of the most challenging areas of ASR research. As mentioned in Section 1, the WER of current systems is still higher than the minimum level for which transcripts are accepted by humans. During recent years, several approaches were proposed to improve the ASR systems used in lecture transcription. Although some significant improvements can be achieved through acoustic model adaptation if manual transcripts of the same lecturer are available ([6, 7]), most research on speaker-independent lecture ASR have turned to the LM as the focus of their efforts ([1, 5]).

In order to improve the LMs for lecture recognition, some of the most recent work have tried to exploit the World Wide Web as a source of corpora for training topic-specific LMs used to adapt general-purpose LMs by extracting keywords from the slides presented in lectures and using such keywords as queries to retrieve relevant web documents (an approach resulting in an average of 14% WER reduction relative to a baseline with a high out-of-vocabulary rate for lectures in controlled conditions [4]). Web-based corpora building is used not only in lecture ASR, but in other topic-dependent recognition tasks, such as call routing applications (by extracting relevant semantic information from web data, as in [8]), telephone and meeting transcription (by using N-gram statistics to retrieve web data that better matches the conversational style of the recordings [9]), or transcription of financial transactions (by manually identifying topic-specific keywords and building web queries of keyword-centered N-grams extracted from existing transcriptions, as in [10]). However, such approaches require the availability of either manual transcripts (costly or impractical) or of an automatic keyword extraction procedure (not necessarily guaranteed to yield the most relevant keywords to a specific topic).

The World Wide Web is not the only source of building topic-specific language models. As part of the larger research area of LM adaptation to a specific topic, various external knowledge sources were considered, ranging from metadata (e.g. the record of a customer calling a company's customer service [11]) to the output from a first pass of the ASR system ([12]). Such external sources of knowledge can also be used for the domain of lecture transcription, for example the textbook on which the lecture is based, if such material is available in a machine-usable format [2, 6] (which yields a relative WER reduction of 5 to 7%).

As previous approaches to topic-specific modelling rely on multiple LMs, the drawbacks of such solutions stem from the need to accurately extract keywords, to determine and acquire appropriate corpora for training the LMs, and to properly adjust the interpolation parameters (making such approaches less suitable for integration into webcast systems). In the following section, we will describe our approach that overcomes such drawbacks, while achieving similar or better WER reductions.

3. Web-Based LM for Lectures

The fundamental principle of our web-based language modelling method consists in treating the entire content of the lecture/presentation slides as the source of external knowledge used in building the dedicated LMs. As it will be shown in this Section, through this approach we eliminate the need to build both topic-dependent and general-purpose LMs.

3.1. General Algorithm

Figure 1 describes the algorithm used to collect the web-based corpora used in training lecture-specific LMs. It assumes every lecture is accompanied by slides, which are mostly organized in bullet form (one idea constitutes a line on the slide), however, every line on the slides is treated as a separate web query (even if part of a larger text). Figure 2 shows several examples of typical queries based on lines from slides. There is no pre-processing of the slides, each web query being an exact copy of a slide line. Not all lines consist exclusively of topic-specific keywords (since many slide bullets do not contain any keywords, while some lines are artifacts of the slide conversion process), which ensures that the corpora retrieved using such queries appropriately matches both the topic-specific and conversation style of a lecture.

```
◇ for every slide $S_i$ in a lecture
  ◇ for every line $L_{i,j}$ on slide $S_i$
    ◇ define $T_{i,j}$ as the text of line $L_{i,j}$
    ◇ run web search with query $Q = T_{i,j}$
    ◇ retrieve most relevant first $N$ documents in PDF format
    ◇ convert retrieved documents to text corpora $C_{i,j}$
```

Figure 1: Web-based LM building using lecture slides.

3.2. Corpora Adjustment

Several parameters can be adjusted both during corpora retrieval and language modelling:

Number of documents to be retrieved for each slide line (N in Figure 1), which will be the main factor influencing the size of the final corpus.

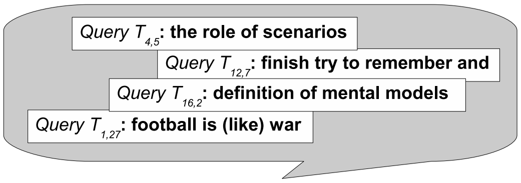
Query $T_{4,5}$: the role of scenarios
Query $T_{12,7}$: finish try to remember and
Query $T_{16,2}$: definition of mental models
Query $T_{1,27}$: football is (like) war

Figure 2: Examples of slide bullets used as web queries. These apparently un-related examples are from the same lecture ("Conceptual Design") of the third year Computer Science undergraduate course "The Design of Interactive Computational Media".

Percentage of non-dictionary words permitted (corpus filtering). For each retrieved document $C_{i,j}$, sentences (or lines) in $C_{i,j}$ for which the number of words not found in an existing initial dictionary exceeds a desired threshold are removed from the corpus. This is useful for preserving the integrity of the LM with respect to the pronunciation dictionary (Section 4 describes how the dictionary is defined).

3.3. LM and ASR Scope Alternatives

Once all corpora $C_{i,j}$ are collected and filtered, LMs can be built using the entire collection or a slide-specific selection. Three alternatives are proposed: one LM for the entire lecture, one LM for each slide, and one LM for each cluster of slides. For the latter two, the slides must be time-indexed (by recording the time of each change of slides).

One LM per lecture: in order to obtain a single model M for the entire lecture, all collected corpora will be joined in a single corpus $C = \sum_{i,j} C_{i,j}$ on which M will be trained.

One LM per slide: for every slide S_i of a lecture, a corresponding LM M_i will be trained for every $C_i = \sum_j C_{i,j}$. During the recognition process, a separate model M_i will be used for the audio segment of the lecture corresponding to the time span of slide S_i.

One LM per cluster of slides: assuming slides are numbered chronologically[1], for each slide S_i, a cluster of slides of range r is defined as $\overline{S}_i(r) = \{\bigcap_k S_k | i - r \leq k \leq i + r\}$. The related corpora are also clustered in the same manner: $\overline{C}_i(r) = \{\bigcap_k C_k | i - r \leq k \leq i + r\}$, where $C_k = \sum_j C_{k,j}$. Thus, individual LMs $\overline{M}_i(r)$ are separately trained on corpora clusters $\overline{C}_i(r)$ and subsequently used during recognition for the audio segments associated with the time span of $\overline{S}_i(r)$.

4. Experiments and Evaluation

We carried out an extensive evaluation of the approach proposed in Section 3, in which several combinations of corpora and LM scope parameters were tested (as the proposed method was not refined during the experiments, no developmental iteration was performed). The experiments were conducted using the SONIC toolkit [13]. We used the acoustic model that is part of the toolkit (built on 30 hours of data from 283 speakers from the

[1] If the same slide is displayed more than once during a lecture (e.g. it is re-visited by the lecturer), the multiple occurrences of that slide are treated as separate slides and numbered accordingly.

WSJ0 and WSJ1 subsets of the 1992 development set of the Wall Street Journal (WSJ) Dictation Corpus [14]).

For all the LMs used (web-based as well as baseline models), a pronunciation dictionary was custom-built to include all words appearing in the corpus on which the LM was trained. The pronunciations were extracted from existing initial dictionaries (the 5K-word WSJ dictionary included with the SONIC toolkit and the 100K-word CMU pronunciation dictionary [15]). For all models, we allowed one non-dictionary word per line of corpus (only for lines longer than four words) – for non-dictionary words that remained in each corpus the SONIC's `sspell` lexicon access tool was used to generate pronunciations using letter-to-sound predictions. The 3-gram LMs were trained using the CMU-CAM Language Modelling Toolkit [16], with a training vocabulary size of 40K words (the out-of-vocabulary rate was low for all models – averaging 0.3% for the baseline and below 0.1% for all other models).

4.1. Test Data

The test data consist of four lectures of approximately 50 minutes each, recorded in different weeks of the same course. The recordings were collected in a large, amphitheatre-style, lecture hall (200 seats), using the AKG C420 head-mounted directional microphone. The lecturer is male, early 60s, and a native speaker of English. The recordings were not intrusive, and no alterations to the lecture environment or proceeding were made. The mono recordings were digitized using the TASCAM US-122 interface as uncompressed audio files with 16KHz sampling rate and 16-bit samples. The audio recordings were manually segmented at pauses longer than 200ms.

4.2. Web-Based Lecture Models

For each of the four lectures all LM training options described in Section 3.3 were considered (with a range $r = 1$ for the cluster option). In terms of the number of retrieved documents (Section 3.2), for each LM training option we allowed three values for N: 10, 20, and 30 documents per bullet (the actual number was in some cases slightly lower than N due to web retrieval and PDF conversion errors). The Google APIs (`http://code.google.com/`) were used for returning the URLs of the web documents relevant to each query (the search was limited to documents in English).

4.3. Baseline Models

The transcripts of the Switchboard (SWI) corpus [17] were used for training the baseline model (SWI LM). The SWI corpus is a large collection of about 2500 scripted telephone conversations between approximately 500 English-native speakers, suitable for the conversational style of lectures (as also suggested in [6]).

In order to compare our web-based modelling with interpolation-based optimizations, two more baseline LMs were built for each of the four lectures. For these, a set of keywords relevant to each lecture was manually extracted from the slides by the teaching assistant associated with the course. A query was constructed with the selected keywords, and 200 relevant web documents were retrieved on which a language model (KEYW LM) was trained[2]. Finally, the KEYW LM was statically interpolated (with the default interpolation weight $\lambda = 0.5$) with the SWI LM to generate the third baseline LM.

LM scope	Docs per slide bullet	Lecture			
		1	2	3	4
Lecture	10	42.34	41.38	44.19	48.35
	20	**41.71**	**40.70**	44.18	47.56
	30	42.03	41.01	**43.73**	**46.95**
Slide	10	51.02	51.25	50.63	57.29
	20	48.37	49.06	49.79	55.26
	30	47.38	47.63	49.04	54.60
Cluster	10	46.94	46.99	49.02	54.41
	20	46.17	46.24	49.43	52.89
	30	45.21	46.31	48.36	52.65
SWI		47.26	48.08	48.71	50.48
KEYW		44.04	45.39	46.39	50.39
SWI+KEYW		41.11	43.43	42.64	46.39

Table 1: The WERs corresponding to the web-based, baseline, and interpolated LMs over the four lecture recognition tasks. The best scores for the web-based LMs are highlighted.

4.4. Results: WER Reduction

Table 1 presents the WER for each of the four lectures on ASR runs using the LMs described in this section. The lowest WER among the web-based LMs is achieved for training over the corpus relevant to the entire lecture, where the number of documents retrieved for each slide bullet ranges from 20 to 30. Comparatively, the baseline model (SWI LM) yields WERs higher on average by relatively 11%, while the average difference between the web-based LMs and the best model trained with manual supervision (SWI+KEYW) is less than 1%.

4.5. Results: Precision & Recall of Keywords

Text transcripts are often used in automatic information retrieval tasks (e.g. using queries to search a webcast lecture repository for a particular topic). However, one of the challenges associated with such retrievals is the accuracy of keyword transcription during the ASR process. For this, we have decided to evaluate our proposed web-based modelling for lectures through the Precision and Recall of the transcribed keywords (measured against the manual transcripts). Since no general agreement exists on how to automatically identify such keywords, we used a manually-generated list of keywords (an approach similar to that taken in [6] where the course textbook's index was used). For each lecture, the list was set to an arbitrary length equal to 1% of the number of words in the manual transcript of each lecture, and words on the list were selected by the teaching assistant associated with the course from all words appearing on slides. Table 2 compares the Precision and Recall scores of the two best web-based models with that of the baseline and the keyword-based models. As can be observed, Precision scores are higher for the web-based models than for the Switchboard models and similar to those for the keyword-interpolated Switchboard model, while Recall scores are higher for web-based models even than those for the manually-supervised SWI+KEYW model.

5. Conclusions and Future Work

In this paper we proposed an algorithm for corpora building and language modelling aimed at improving the accuracy of automatic lecture transcription. Our approach eliminates the need for multiple models, while achieving similar or better WER reductions than existing methods based on

[2] The training corpora varied in size from 1.1M words (KEYW LM) to 3.1M (SWI LM) to 26.3M words (average of best-scoring web-based LMs).

LM scope	Docs per slide bullet	Lecture			
		1	2	3	4
		Precision			
Lecture	20	**93.41**	**91.64**	92.02	87.70
Lecture	30	92.79	91.33	**93.01**	**87.90**
SWI		91.39	87.22	90.20	83.24
KEYW		93.01	89.88	91.26	85.71
SWI+KEYW		93.99	90.60	92.78	85.52
		Recall			
Lecture	20	**81.04**	**81.94**	71.49	67.72
Lecture	30	80.26	81.44	**71.49**	**68.99**
SWI		57.92	51.55	57.02	48.73
KEYW		79.48	75.52	69.01	64.56
SWI+KEYW		77.14	74.48	69.01	59.81

Table 2: Precision and recall scores for keyword detection of the two best web-based models compared to the baseline and interpolated models.

interpolating general-purpose with topic-specific models (as well as improved Precision and Recall scores for keyword identification). By using the entire content of the presented slides as web queries, only a single web-based corpus needs to be collected (on which a language model is trained). Beside allowing for an entirely automated lecture-specific modelling, our approach does not rely on identifying keywords for each lecture topic. Therefore, this unsupervised method is suitable for integration into webcast archive systems such as ePresence.

Since one of the typical sources of recognition errors is the large vocabulary (and LM) size, future work will look at improving the web-based retrieval to better maximize the match between the retrieved corpus and the conversational style of a lecture and of a particular lecturer. For this, we are considering exploiting the manually-corrected partial transcripts of the first lectures in a course (corrections that are facilitated by our collaborative transcript editing interface described in [18]).

6. Acknowledgements

This research was funded by the NSERC Canada Network for Effective Collaboration Technologies through Advanced Research (NECTAR). The authors wish to also thank our undergraduate research collaborator Yuecheng Zhang for the implementation of the Google API scripts.

7. References

[1] E. Leeuwis, M. Federico, and M. Cettolo, “Language modeling and transcription of the TED Corpus lectures,” in *Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing – ICASSP*, vol. 1, 2003, pp. 232–235.

[2] B.-J. Hsu and J. Glass, “Style & topic language model adaptation using HMM-LDA,” in *Proc. ACL Conf. on Empirical Methods in Natural Language Processing – EMNLP*, 2006, pp. 373–381.

[3] C. Munteanu, R. Baecker, G. Penn, E. Toms, and D. James, “The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives,” in *Proc. ACM Conf. on Human Factors in Computing Systems – CHI*, 2006, pp. 493–502.

[4] I. Rogina and T. Schaaf, “Lecture and presentation tracking in an intelligent meeting room,” in *Proc. IEEE Conf. on Multimodal Interfaces – ICMI*, 2000, pp. 47–52.

[5] K. Kato, H. Nanjo, and T. Kawahara, “Automatic transcription of lecture speech using topic-independent language modeling,” in *Proc. ISCA Conf. on Spoken Language Processing – ICSLP-Interspeech*, vol. 1, 2000, pp. 162–165.

[6] A. Park, T. Hazen, and J. Glass, “Automatic processing of audio lectures for information retrieval,” in *Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing – ICASSP*, vol. 1, 2005, pp. 497–500.

[7] C. Fügen, M. Kolss, D. Bernreuther, M. Paulik, S. Stüker, S. Vogel, and A. Waibel, “Open domain speech recognition & translation: Lectures and speeches,” in *Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing – ICASSP*, vol. 1, 2006, pp. 569–572.

[8] D. Hakkani-Tür and M. Rahim, “Bootstrapping language models for spoken dialog systems from the World Wide Web,” in *Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing – ICASSP*, vol. 1, 2006, pp. 1065–1068.

[9] I. Bulyko, M. Ostendorf, and A. Stolcke, “Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures,” in *Proc. HLT-NAACL*, 2003, pp. 7–9.

[10] R. Sarikaya, A. Gravano, and Y. Gao, “Rapid language model development using external resources for new spoken dialog domains,” in *Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing – ICASSP*, vol. 1, 2005, pp. 573–576.

[11] M. Bacchiani and B. Roark, “Meta-data conditional language modeling,” in *Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing – ICASSP*, vol. 1, 2004, pp. 241–244.

[12] K. Seymore and R. Rosenfeld, “Large-scale topic detection and language model adaptation,” School of Computer Science, Carnegie Mellon University, Tech. Rep. CMU-CS-97-152, 1997.

[13] B. L. Pellom, “SONIC: The University of Colorado continuous speech recognizer,” University of Colorado, Boulder, Colorado, Tech. Rep. TR-CSLR-2001-01, 2001.

[14] “The Wall Street Journal Dictation Corpus (DARPA-CSR),” The Linguistic Data Consortium, LDC94S13, 1992.

[15] “The CMU Pronouncing Dictionary v. 0.6,” http://www.speech.cs.cmu.edu/, 1998.

[16] P. Clarkson and R. Rosenfeld, “Statistical language modeling using the CMU-Cambridge Toolkit,” in *Proc. ISCA European Conf. on Speech Communication and Technology– Eurospeech*, vol. 1, 1997, pp. 2707–2710.

[17] J. Godfrey, E. Holliman, and J. McDaniel, “SWITCHBOARD: telephone speech corpus for research and development,” in *Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing – ICASSP*, vol. 1, 1992, pp. 517–520.

[18] C. Munteanu, Y. Zhang, R. Baecker, and G. Penn, “Wiki-like editing of imperfect computer-generated webcast transcripts,” in *Proc. Demo track of ACM Conf. on Computer Supported Cooperative Work – CSCW*, 2006, pp. 83–84.