UNSUPERVISED VOCABULARY SELECTION FOR REAL-TIME SPEECH RECOGNITION OF LECTURES

Paul Maergner[1,2], Alex Waibel[1,2], Ian Lane[1]

[1] Carnegie Mellon University, USA
[2] Karlsruhe Institute of Technology, Germany

ABSTRACT

In this work, we propose a novel method for vocabulary selection to automatically adapt automatic speech recognition systems to the diverse topics that occur in educational and scientific lectures. Utilizing materials that are available before the lecture begins, such as lecture slides, our proposed framework iteratively searches for related documents on the web and generates a lecture-specific vocabulary based on the resulting documents. In this paper, we propose a novel method for vocabulary selection where we first collect documents similar to an initial seed document and then rank the resulting vocabulary based on a score which is calculated using a combination of word features. This is a critical component for adaptation that has typically been overlooked in prior works. On the interACT German-English simultaneous lecture translation system our proposed approach significantly improved vocabulary coverage, reducing the out-of-vocabulary rate, on average by 57.0% and up to 84.9%, compared to a lecture-independent baseline. Furthermore, our approach reduced the word error rate, by 12.5% on average and up to 25.3%, compared to a lecture-independent baseline.

Index Terms— Vocabulary selection, automatic speech recognition, language model adaptation

1. INTRODUCTION

Recent advances in streaming technologies now allow research talks and lectures to be broadcasted live across educational institutes around the world. This provides students with unprecedented access to educational content no matter their physical location. However, although physical barriers are reduced, language barriers remain. Lectures may be presented in a language the student cannot understand thus limiting the usefulness of such content. Similarly, due to the lack of subtitles, live audio-video content is unsuitable for the hearing impaired. To overcome these barriers, recent works have investigated both the use of speech-translation technologies to translate lectures in real-time [1] and real-time lecture transcription for the hearing impaired [2]. Although useful, the biggest downfall of these technologies is portability since they rely on a automatic speech recognition (ASR) system which are generally optimized to a specific lecture topic. For each new topic, significant effort and cost is required to manually transcribe similar lectures, without which the system will generally perform poorly. In this work, we propose to overcome this limitation by introducing approaches to automatically adapt ASR systems to the diverse topics that occur in educational and scientific lectures.

In modern ASR systems, speech recognition is performed by applying search across three models, an acoustic model, which models the phonetic units in the input speech, a language model (LM), which models the likelihood of word sequences, and a recognition vocabulary, which models the pronunciation of individual words. The recognition vocabulary has to be small enough to allow real-time processing but remain large enough to cover the vocabulary used within the lecture. If a word is not present in the active system vocabulary it cannot be recognized and will often lead to additional errors to the surrounding content. When the mismatch between the training data used to build the ASR system and the topic of conversation is severe, vocabulary coverage is poor leading to a high number of out-of-vocabulary (OOV) words, low recognition accuracy and low intelligibility in the resulting transcript. For effective adaptation, vocabulary coverage is a key component that prior works have often overlooked.

There have been a number of recent works that have proposed methods to deal with the diversity of topics encountered in lecture speech. In [1], a topic-independent system vocabulary was selected based on word occurrence counts in both in-domain and out-of-domain corpora and lecture-independent models for speech recognition were built using these corpora. Lecture-independent models were the goal in this work and no lecture-specific adaptation was performed. In [3], an approach for language model (LM) adaptation using web data was introduced but vocabulary adaptation was not considered. An approach for joint vocabulary and LM adaptation was introduced in [4] in which words from the lecture slides were first added to the active system vocabulary and then LM adaptation, similar to [3], was performed. Within the MIT Spoken Lecture Processing Project [5] a lecture-specific vocabulary was adapted to the lecture using manually provided supplemental text, including slides, journal articles, and book chapters, which were made available prior to the lec-

ture. Although the adaptation approaches described above were shown to be effective compared to non-adapted systems, they did not significantly improve the vocabulary coverage thus limiting the usefulness of these approaches. In this work, we propose a novel approach to improve vocabulary coverage based on a feature-based vocabulary ranking scheme applied on documents automatically collected from the WWW. Our proposed approach improves vocabulary coverage, LM perplexity, and speech recognition accuracy compared to a lecture-independent system.

2. UNSUPERVISED VOCABULARY SELECTION

The vocabulary used by a presenter during a lecture can be seen as a combination of two vocabularies as described in [5]: A topic-independent lecture vocabulary, which contains vocabulary common to spontaneous speech, and a topic-dependent vocabulary. Our proposed approach for vocabulary selection uses a similar breakdown. We begin with a topic-independent lecture vocabulary, which consists of stop words and common words used in spontaneous lecture speech (in the experimental evaluation described in Section 3 our common vocabulary consisted of 1788 words). In addition to this vocabulary, we then select a topic-specific vocabulary for each lecture based on a set of initial seed documents, for example lecture-slides, handouts or book chapters. Using these seed documents, our proposed system automatically collects a large corpus of related documents from the World-Wide-Web and then an active recognition vocabulary is selected using a feature-based word ranking computed using this corpus.

2.1. Document Collection

The document collection process is performed in four steps:

1. **Word Extraction:** First, text from the lecture slides is extracted, cleaned (symbols, punctuation, and casing are removed), and split into individual words. The resulting word-list is then verified against an extremely large vocabulary[1] to remove erroneous words that were introduced during the extraction process.

2. **Query Selection:** Next, search queries are generate from the lecture slides. Here, short phrases of up to three words which do not contain any topic-independent vocabulary are selected as search queries.

3. **Web-Search:** Web-search[2] is then performed. For each search query, the 50 highest ranked documents are selected and the text from the resulting documents (web page or PDF file) are extracted.

4. **Language Verification:** For each document, language verification is performed to ensure that it is actually in the required language. When the percentage of topic-independent vocabulary in the document is below 30% the document is removed from further processing.

2.2. Vocabulary Ranking and Selection

After document collection, the resulting vocabulary is too large to be incorporated directly into an ASR system (in our work we observed vocabularies between 135k and 850k) and thus a smaller active recognition vocabulary must be selected. To select words for this smaller vocabulary, a ranking score for each word is computed. Words with the highest score are added to the vocabulary until the desired vocabulary size is reached. The *ranking score* is based on different word features. These features range from simple counts, such as word frequency ("VocCount") and number of documents the word occurs in ("DocCount"), to more powerful features which have been found to be effective in information retrieval, including cosine similarity ("DocCosineCount") and tf-idf ("tfCosineTfidf"). In total, we compared the effectiveness of 21 different features for the proposed vocabulary selection task. These are described in detail in [6]. For vocabulary ranking, we compared different scoring functions $s(w)$ to compute the ranking of each word w based on its specific word features $f_i(w)$:

1. **Single Feature Score:** The score $s_{single,i}(w)$ is based on one single feature $f_i(w)$ (e.g. DocCount).

$$s_{single,i}(w) = f_i(w) \tag{1}$$

2. **Linear Feature Combination Score:** The score $s_{linear}(w)$ is defined as a linear weighting of two or more features. For example:

$$s_{linear,i,j}(w) = \alpha \cdot f_i(w) + (1-\alpha) \cdot f_j(w) \tag{2}$$

3. **Gaussian Mixture Model Score:** The score $s_{gmm}(w)$ is based on the likelihood ratio of two Gaussian Mixture Models (GMMs). Two GMMs are trained, one on words which occur in a specific lecture and one on words which do not occur. The score $s_{gmm}(w)$ is the difference in the log-likelihood of a word feature vector for each of these GMMs. For example with the word feature vector $\mathbf{f}_{i,j}(w) = \begin{pmatrix} f_i(w) & f_j(w) \end{pmatrix}^T$:

$$s_{gmm,i,j}(w) = \log \mathrm{P}_{in}(\mathbf{f}_{i,j}(w)) - \log \mathrm{P}_{out}(\mathbf{f}_{i,j}(w)) \tag{3}$$

2.3. Lecture-specific Language Modeling

Once an active vocabulary has been selected, we adapt the language model (LM) to be applied during recognition using an approach similar to [3]. First, we train a lecture-independent LM using a large lecture-independent corpora.

[1] Unigram occurrence in the Google Book Ngrams dataset available at *http://ngrams.googlelabs.com/datasets*.

[2] Search is performed using the Microsoft Bing search engine.

Then, for each lecture we train a separate LM using the lecture slides and the resulting web documents found with our document collection approach (section 2.1). A lecture-specific LM is subsequently generated by interpolating this LM. For this interpolation, we use fixed interpolation weights of 0.5 in our experimental evaluation.

3. EXPERIMENTAL EVALUATION

We evaluated the effectiveness of the proposed method on the German speech recognition component in our German-English Simultaneous Lecture Translation system [1]. The evaluation was performed on six lectures held at Karlsruhe Institute of Technology, in 2009 and 2010. The lectures consisted of a variety of topics: Data structures (Lect1), machine translation (Lect2), mechanics (Lect3), population geography (Lect4), computer architecture (Lect5), and copyright law (Lect6). The evaluation is performed on a total of 5.7 hours of transcribed lecture audio.

3.1. Unsupervised Vocabulary Selection

First, we evaluated our proposed vocabulary selection approach in terms of the reduction in out-of-vocabulary (OOV) rate it could provide. Evaluation was performed using Lectures 1-4 as transcripts of Lectures 5 and 6 were not available when this evaluation took place.

3.1.1. Baseline System Performance

Baseline vocabularies consisting of 40k, 90k, and 300k words were selected from combined corpora of broadcast news, parliamentary debates, printed media, and university web data using the method described in [7]. Using these vocabularies, the average OOV rate across the four lectures were: 5.6% (40k), 4.0% (90k), and 3.0% (300k). Adding vocabulary that occurred in the lecture slides ("Baseline+Slides") reduced OOV rate on average by 18.2%, obtaining an OOV rate of 4.6% for the 40k case.

3.1.2. Feature-based Vocabulary Selection

First, we selected vocabularies using our single feature score (section 2.2, eq. 1). The average OOV rate using a 40k vocabularies is shown in Fig. 1. The lowest OOV rate was obtained using DocCount. The OOV rate using the DocCount feature is significantly lower than the proposed Baseline vocabularies even when slides were added. Using the DocCount feature for vocabulary selection reduced the OOV rate on average by 56.8% obtaining an OOV rate of 2.4% for the 40k case.

Next, we investigated the effectiveness of combining multiple features for vocabulary ranking. We linearly combined pairs of features using the linear feature score (section 2.2, eq. 2) evaluating across all feature combinations. We observed that combining DocCount and VocCount with $\alpha = 0.5$

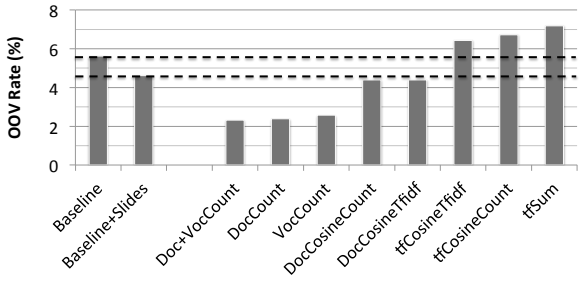

Fig. 1. Average OOV rate for different features

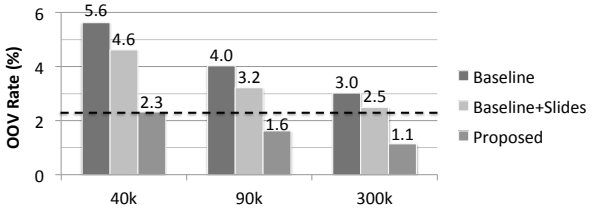

Fig. 2. Average OOV rate of vocabulary sizes.

("Doc+VocCount") obtained an average reduction of OOV rate of 1% compared to using the DocCount feature alone. GMM-based word ranking did not reduce the OOV rate compared to the linear case (section 2.2, eq. 3). We evaluated all feature-pairs and although slight improvements were gained for specific lectures no feature-pairs consistently improved performance across all lectures. Fig. 2 shows the effectiveness of our proposed linear score Doc+VocCount compared to the baseline over varying vocabulary sizes. The proposed approach reduced the OOV rate by 58.2%, 55.1%, and 57.7% for the 40k, 90k, and 300k systems. More significantly the 40k vocabulary selected with the proposed approach obtained a lower OOV rate of the 300k Baseline system, showing the effectiveness of this approach.

3.2. Lecture-dependent Language Model Adaptation

Next, lecture-specific LMs were trained using the vocabulary selected in section 3.1.2, a topic-independent corpora (1280M words), consisting of broadcast news, parliamentary debates, printed media, and web data, and a lecture-specific corpora (avg. 56M words) consisting of the slides and web documents collected using the method described in section 2.1. The resulting lecture-specific LMs obtained a significantly lower perplexity compared to the baseline lecture-independent model as shown in Table 1. On average the lecture-dependent LMs reduced perplexity by 23.2%.

3.3. Lecture-dependent Speech Recognition

Finally, we evaluated the recognition performance of our German lecture recognition system using the proposed vocabu-

	Baseline	Adapt LM
Lecture 1	344.0	261.4 (24.0%)
Lecture 2	352.0	285.7 (18.8%)
Lecture 3	325.0	199.9 (38.5%)
Lecture 4	247.1	210.0 (15.0%)
Lecture 5	274.3	170.0 (38.0%)
Lecture 6	241.3	229.9 (4.7%)
Avg. Improvement	-	**23.2%**

Table 1. Language Model Perplexity (40k Vocabulary)

Vocabulary Selection		X		X
LM Adaptation			X	X
Lecture 1	43.1	42.4	42.7	41.0 (5.0%)
Lecture 2	34.9	35.7	34.3	33.9 (2.7%)
Lecture 3	33.4	27.3	34.7	27.5 (17.6%)
Lecture 4	28.3	23.9	28.5	22.7 (20.0%)
Lecture 5	28.4	28.8	25.5	21.2 (25.3%)
Lecture 6	37.4	36.4	37.6	35.7 (4.4%)
Average Improvement	-	**5.8%**	**1.3%**	**12.5%**

Table 2. Word Error Rate (40k Vocabulary)

lary selection and LM adaptation techniques. Recognition was performed using the Janus speech recognition toolkit with speaker adapted acoustic models. The German ASR system was trained on 150 hours of audio data resulting in an acoustic model with 4000 codebooks and a maximum of 64 Gaussian mixtures. Semi-tied covariance and boosted MMI discriminative training was performed during model training.

We evaluated the speech recognition accuracy of four different systems. The lecture-independent baseline system obtained an average WER of 34.2% across the six lectures used in this evaluation. When vocabulary selection (described in section 2.2) was performed using linear feature combination score (Doc+VocCount) an average WER of 32.4% was obtained, a 5.8% relative reduction compared to the baseline system. With LM adaptation (described in section 2.3), an average WER of 33.9% was obtained, a 1.3% relative reduction compared to the baseline system. Applying both, vocabulary and LM adaptation, led to an average WER of 30.3%, a 12.5% relative reduction compared to the baseline system. On average, vocabulary selection obtained higher recognition accuracy than LM adaptation alone, but the biggest gain was obtained by combining both, vocabulary selection and language model adaptation. Although, the improvement was not equally large across all lectures the proposed approach always improved speech recognition accuracy.

4. CONCLUSION

Effective adaptation techniques are required to enable lecture transcription and lecture translation systems to perform adequately across the diverse topics that occur in educational and scientific lectures. Our proposed approach solves one of the key issues in current systems, that of selecting an appropriate topic-specific vocabulary for real-time speech recognition. Using our approach, the OOV rate was reduced by up to 84.9% (on average by 57.0%) compared to a baseline vocabulary. Furthermore by generating a lecture-specific language model incorporating the retrieved web documents, word error rate was dramatically reduced, obtaining a WER up to 25.3% lower than a lecture-independent Baseline.

Acknowledgments

We would like to thank Sebastian Stüker and Kevin Kilgour for providing the Baseline German speech recognition system used in this work as well as their technical expertise throughout the project.

5. REFERENCES

[1] Muntsin Kolss, Matthias Wölfel, Florian Kraft, Jan Niehues, Matthias Paulik, and Alex Waibel, "Simultaneous German-English Lecture Translation," in *Proc. IWSLT*, 2008, pp. 174–181.

[2] Tatsuya Kawahara, Yusuke Nemoto, and Yuya Akita, "Automatic Lecture Transcription by Exploiting Presentation Slide Information for Language Model Adaptation," in *Proc. ICASSP*, 2008, pp. 4929–4932.

[3] Cosmin Munteanu, Gerald Penn, and Ron Baecker, "Web-based Language Modelling for Automatic Lecture Transcription," in *Proc. Interspeech*, 2007, number August, pp. 2353–2356.

[4] Hiroki Yamazaki, Koji Iwano, Koichi Shinoda, Sadaoki Furui, and Haruo Yokota, "Dynamic Language Model Adaptation Using Presentation Slides for Lecture Speech Recognition," in *Proc. Interspeech*, 2007, pp. 2349–2352.

[5] James R. Glass, Timothy J. Hazen, Scott Cyphers, Igor Malioutov, David Huynh, and Regina Barzilay, "Recent Progress in the MIT Spoken Lecture Processing Project," in *Proc. Interspeech*, 2007, pp. 2553–2556.

[6] Paul Maergner, Ian Lane, and Alex Waibel, "Unsupervised Vocabulary Selection for Domain-Independent Simultaneous Lecture Translation," in *Proc. MT Summit XIII*, Xiamen, China, 2011, pp. 89–96.

[7] Sebastian Stüker, Kevin Kilgour, and Jan Niehues, "Quaero Speech-to-Text and Text Translation Evaluation Systems," in *Proc. HLRS*. 2010, pp. 529–542, Springer.