OPEN VOCABULARY ASR FOR AUDIOVISUAL DOCUMENT INDEXATION

Alexandre Allauzen and Jean-Luc Gauvain*

Spoken Language Processing Group (http://www.limsi.fr/tlp)
LIMSI-CNRS, Univ. Paris XI, bat. 508, 91403 Orsay, France

{allauzen,gauvain}@limsi.fr

ABSTRACT

This paper reports on an investigation of an open vocabulary recognizer that allows new words to be introduce in the recognition vocabulary, without the need to retrain or adapt the language model. This method uses special word classes, whose n-gram probabilities are estimated during the training process by discounting a mass of probability from the out of vocabulary words. A part of speech tagger is used to determine the word classes during language model training and for vocabulary adaptation. Metadata information provided by the the French audiovisual archive institute are used to identify important document-specific missing words which are added to appropriate word class in the system vocabulary. Pronunciations for the new words are derived by grapheme-to-phoneme conversion. On over 3 hours of broadcast news data, this approach leads to a reduction of 0.35% in the OOV rate, of 0.6% of the word error rate, with 80% of the occurrences of the newly introduced being correctly recognized.

1. INTRODUCTION

National archive institutions like INA in France, or the BBC in England, index over a hundred thousand hours of audiovisual data on a yearly basis. Most of their methods are manual and aim to extract semantic information from the document and summarize its content. The recent transition at such institutes from analog to digital storage media, has started a change in the indexing chain for audiovisual documents. With digital storage, the same medium can be used for the audiovisual data and its associated metadata but at the same time introduces the challenge of developing automatic methods and efficient support for these manual tasks.

Over the last decade there has been growing interest in the use of automatic speech recognition (ASR) as a tool to provide relevant access to large broadcast news (BN) archives [1]. The NIST evaluations on Spoken document Retrieval (SDR) showed that ASR systems can transcribe contemporary BN data with a sufficient quality to enable a variety of applications such as content-based document retrieval [2]. Moreover, the SDR evaluations highlight the need for vocabulary and LM adaptation techniques to adjust to linguistic changes in audiovisual collection over time. In addition, there is often a substantial gap between the epoch of the LM training corpus and the audio data to process (over a year is not unusual), because of the high cost of collecting and processing training texts. This gap tends to increase the expected proportion of out of vocabulary (OOV) words, which are mainly named entities, and often cause recognition errors in their immediate context. Furthermore, named entities are an important lexical class for which recognition errors have a significant impact on the indexation accuracy [2].

Previous work investigating automatic LM adaptation methods to transcribe BN shows on a daily basis, such as the rolling LM [3, 4], have highlighted the difficulties in collecting a sufficient amount of well suited data for recent news or for archived documents [5, 6]. However, when indexing a collection of audiovisual documents, prior knowledge sources specify certain words, such as the speaker's name or terms specific to the show which are likely to have appeared in the broadcast. In this case, documentalists may want to manually include these words in the ASR vocabulary, even if appropriate training texts are not available.

This practical need sets the stage for this work. Not matter how large a recognition vocabulary is, given that it is selected from a training corpus, it will never cover all words which can possibly occur in a document. We investigate an open vocabulary recognizer that makes use of a static LM incorporating lexical back-off (LBO), thus allowing the addition of new words without the need for LM retraining or adaptation. The basic idea is depicted in Figure 1, where the metadata of the audiovisual document is used to provide a set of new words, which are automatically added to the lexical back-off in the language model. In the next section the application which guides this research is presented. A brief overview of the LIMSI BN transcription system is provided is Section 3. Section 4 describes the lexical back-off method, and experimental results are given in Section 5.

2. CONTEXT

The problem investigated in this work stems from an application proposed by the French national archive institute. The goal is to complement the manually specified metadata of an audiovisual document with the ASR output, using the metadata provided by the INA archivist as a prior information source to adapt the LM to the audiovisual document without additional training data. The metadata are derived from documents which may come

*This work was carried out during A. Allauzen's thesis in collaboration with the Research and Experiment Direction of the French National Audiovisual Institute (INA, http://www.ina.fr/recherche/index.en.html)

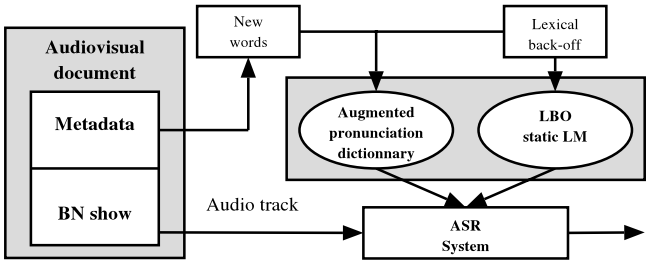

Figure 1: Use of document metadata to select new words to be included in the open-vocabulary LM using the LBO.

from the producers or the broadcast channel, and which include various kinds of information such as the date and time of diffusion, technical features (storage medium or video materials used to film or record the show), diffusion rights (which summarize the speakers' names including the anchor, reporters and guests), and a summary. All of the relevant information is compiled by an archivist in a controlled format (called a document summary) which is the only form accessible to search engines. The document summary is a set of field/value pairs, and except for the summary field, information is expressed in a controlled language as lists of proper names, keywords, and technical attributes.

For the prime time BN shows used in our experiments (described in Section 5), metadata contain only an average of 350 words of texts. This quantity of texts is not sufficient for adapting a language model. A complementary solution is to use the document summaries to collect a language model adaptation corpus [7]. While this approach is applicable for recent BN shows, but not for older archive documents since it is very difficult to find related electronic texts. The average OOV rate of the text summaries estimated with the baseline vocabulary is 1.6% and the OOV words are mainly named entities (70%). Our objective is to allow the recognition of the OOV words observed in the document summaries.

3. BASELINE TRANSCRIPTION SYSTEM

The LIMSI broadcast news transcription system has 2 main components, an audio partitioner and a word recognizer [1]. Data partitioning divides the continuous audio stream into homogeneous segments, associating cluster, gender and bandwidth labels with each segment. The speech recognizer uses continuous density (CD) HMMs with Gaussian mixture for acoustic modeling and n-gram statistics estimated on large text corpora for language modeling. Each context-dependent phone model is a tied-state, left-to-right CDHMM with Gaussian mixture observation densities.

Word recognition is performed in three steps: 1) initial hypothesis generation, 2) lattice generation, 3) final hypothesis generation. A 3-gram language model is used in the first two decoding steps. The final hypotheses are generated by rescoring the lattice with a 4-gram LM and acoustic models adapted with the hypotheses of step 2. The acoustic models were trained on about 100 hours of recent French broadcast news data. The phone models are position-dependent triphones, with about 10k tied-states for the largest model set. The state-tying is obtained via a divisive, decision tree based clustering algorithm.

The baseline language models were obtained by interpolation of n-gram back-off language models trained on 4 data sets: 73 M words of press service transcripts; 274 M words of *Le Monde* and *Le Monde Diplomatique* newspaper texts; 67 M words from *Agence France Presse* newswire texts; and 1.6 M words corresponding to the transcriptions of the acoustic training data. The interpolation coefficients of these LMs were chosen so as to minimize the perplexity on a set of development texts.

The recognition lexicon contains 65333 words, and has a lexical coverage of 98.8% of the same development set. Each lexical entry is described as a sequence of elementary units, taken from a 33 phone set, with 3 additional units to represent silence, filler words, and breath noises. The lexical pronunciations were initially derived from the grapheme-to-phoneme rules and manually verified, adding alternate and contextual pronunciations.

4. LEXICAL BACK-OFF

Three steps are necessary to add a word in an ASR system. First the word must be added to the vocabulary, then a phonetic transcription must be associated with it, and thirdly, it must be included in the n-gram distributions of the LM with non-zero probabilities. To introduce a word in the vocabulary without retraining the LM, we propose to use special forms called back-off word classes (BOW). During LM training one of these forms replaces one or more words which are not yet known, by discounting a mass of probability from the OOV words. Then, prior to decoding, new words can be added as alternate orthographic forms of these special classes, thereby allowing the LM to assign a probability to sentences including these words.

Back-off words

The easiest solution is to choose only one BOW. This form already exists in the standard LM, usually denoted as `<UNK>`, the form reserved for OOV words. But this solution does not model the type or the linguistic role of words. For the INA application, it is important to differentiate content words such as proper names from rare forms of conjugated verbs. Therefore a part-of-speech (POS) tagger is used to map a word to its BOW. The Brill tagger[1] trained for French by INALF[2] [8] is used to tag the training corpus described in section 3. The lexicon and other resources of the tagger were adapted to match the text normalization used for the LM training corpus [1].

A subcorpus of 103 M of words of recent data was extracted from the training data and used to determine the set of BOWs. The OOV rate of this subset is 1.54% with respect to the baseline vocabulary. This subcorpus was POS-tagged. The proportions of the most frequent tags for OOV words are summarized in Table 1. It can be seen that more than half of the OOV word occurrences are proper names, and that the second category, nouns, represent less than 20%. Most of the others categories are verb forms. The less frequent tags were discarded because they mainly correspond to typographic errors, abbreviations and neologisms. Moreover the sum of their contribution to the OOV words is less than 0.2%.0.05%.

Our final tag set for lexical back-off consists of 20 tags. These tags were derived from the tags listed in Table 1, by adding number distinction (singular vs plural) where appropriate. For proper

[1] Available at http://www.cs.jhu.edu/ brill

[2] Available at http://jupiter.inalf.cnrs.fr/WinBrill

POS	Occurrences	% of OOV
Proper name	854043	53.7
Nouns	266062	16.7
Conjugated verbs	157980	9.9
Adjectives	133363	8.4
Participle past as adjective	91599	5.8
Infinitive verb	34365	2.2
Adverb	13486	0.8
Participle past	9212	0.6
Foreign nouns	4028	0.2
Participle present	2167	0.1
Cardinal	1886	0.1

Table 1: Occurrences of POS tags for OOV words and their relative proportions, estimated on the sub-corpus with the standard vocabulary.

Original	Ozon et Brahem	sympathisent	à Bromley
Mapped	Ozon et <SBP>	<VCJ:pl>	à <GEO>

Figure 2: Example of mapping of a text containing OOVs: *Brahem* is an unknown proper name (*SBP*), *sympathisent* is a plural verb (*VCJ:pl*), and *Bromley* is an unknown geographic name (*GEO*).

names, a post processing is performed using lists of geographic names and first names collected on the Web to divide these into three classes (Proper name, Geographic name and First name).

Phonetic transcription of new words

New words are introduced in the vocabulary as alternate orthographic forms of their BOWs. A pronunciation must be determined for words that are not already in the LIMSI reference pronunciation dictionary. For this, the LIMSI Text-To-Speech system described and evaluated in [9] is used to automatically generate the missing phonetic transcription.

Building the language model

The new LM is built using the same text corpus and method described in section 3. The only differences are the vocabulary and the pre-processing step. The vocabulary is derived from the standard vocabulary by adding the twenty BOWs and the text data are pre-processed so as to be coherent with this new vocabulary: an OOV word is replaced by its associated BOW if the BOW belongs to the vocabulary, otherwise it is mapped as usual on the unknown symbol <UNK>. An example mapping for a text containing OOVs is shown in Figure 4. BOWs are considered like the other words during the training process. The addition of the 20 BOWs to the vocabulary, increases the 4-gram LM size by 11.5% resulting in an LM with over 15 millions of bigrams, 17 millions of trigrams and 15 millions of 4-grams.

The probability of a new word w given its (n-1)-gram history h is estimated as follows:

$$P(w|h) = P(B_w|h)P(w|B_w),$$

where B_w is the BOW corresponding to the word w, and $P(B_w|h)$ is the n-gram probability assigned to the BOW. Table 2 gives the lexical ranks of the ten most frequent BOWs, all of which appear in the top 300 vocabulary items. Estimating

BOW	*lexical rank*
Proper name	25
Singular noun	84
Plural noun	111
Conjugated verb (sg)	129
Conjugated verb (pl)	141
Singular adjective	151
Geographic name	159
Plural adjective	214
Infinitive	266
First name	297

Table 2: Lexical ranks of the top ten lexical back-off words.

$P(w|B_w)$ is problematic, since relevant training data is usually not available. In the following experiments a unigram distribution estimated on the associated document summary is used.

5. EXPERIMENTS

The recognition experiments were carried out using three 45-minute news shows and the first 15 minutes of seven news shows broadcast in January 2002. The data were collected from the WEB site of one the major French television channels, and the audio quality is low, being transmitted at 16kbps. There are on average 7930 words per complete broadcast, and 2850 words for each 15 minute excerpt. The 3h15 of data was manually transcribed using the Transcriber tool.[3] and scoring is performed using the BN scoring scripts provided by the NIST.[4] The OOV rate using the baseline vocabulary is 1.1%, which causes on the order of 2%-2.5% word errors. In order to verify that the BOW model is appropriate for this data, the manual transcriptions were POS tagged. The resulting tag distribution is very similar to that of the training texts, the main sources of OOV words being proper names (58.3% of the OOV words), conjugated verbs (15.8%) and nouns (11.2%).

The goal of this experiment is to simulate the use of the open vocabulary LM, adapting the LM using the document's metadata as a prior information source. For each BN show, the associated document summary is cleaned and normalized as for LM training [4]. The summaries are then POS tagged, and the OOV words are added to the vocabulary as a member of their respective BOWs. Their frequency counts are used to estimate the term $P(w|B_w)$ for each new word w. An average of 109 words are automatically introduced for each show, of which 70% are assigned to the BOW for proper names (a larger percentage than in the development data), and with nouns and conjugated verb forms being the next most represented categories.

The ASR results are summarized in Table 3. The baseline results are somewhat higher than previously reported results [1], due to the audio quality of the compressed data. The addition of the new words from the document summaries results in a relative OOV rate reduction of 31%, from 1.10% to 0.75%. Most of the remaining OOV words are verb and adjective inflexions. About 25% are named entities, but were not considered relevant or overlooked by documentalist since they are not in the document summary and are mainly the names of small towns or interviewed

[3] Available at www.sourceforge.net
[4] http://www.nist.gov/speech/tools/

Vocabulary	*%OOV*	*%WER*	$\Delta WER/\Delta OOV$
Baseline	1.10	25.5	
Summary	0.75	24.9	1.7
Oracle	0.00	22.6	2.5

Table 3: ASR results obtained with the *Baseline* vocabulary and with lexical back-off for the *automatic* and the *oracle* experiments. The last column measures the ratio of the absolute error reduction in WER and OOV rate.

passers-by. The average absolute WER reduction is 0.6%, with a ratio between absolute improvement in WER and the OOV rate of about 1.7. This conforms with the empirical evidence that there are on average 1.5-2 errors per OOV. However these results do not imply that all words introduced via BOW are correctly recognized. About 80% of these words are correctly recognized, with slightly better recognition (84%) for named entities.

An oracle experiment was performed to estimate an upper-bound on the gain that can be obtained with this method. This experiment is carried out by adding all the OOV words in the manual transcripts to the recognition vocabulary via their associated BOWs. There are an average of 48 words per show added to the vocabulary. Using the transcripts rather than the document summaries results in fewer words being added to the vocabulary, since the summaries may contain descriptive information which does not appear explictly in the broadcast. As shown in Table 3, adding all the OOV words in the vocabulary leads to an absolute WER reduction of 2.9%. The ratio of the WER reduction to the OOV reduction is 2.5, and as before, about 80% of the new words are correctly recognized. The larger improvement results from a lower confusion rate within a BOW, since there are about half the number of alternate orthographic forms per BOW in the oracle experiment.

An error analysis by BOW showed that for proper names, geographic locations, and first names, one-third of the errors are due to problems in the the phonetic transcription of the word. These errors mainly concern foreign proper names that are not detected as foreign by the TTS system such as *Prso* or *Vanbergue*.

6. CONCLUSIONS

This article reports on investigations using lexical back-off to allow new words to be introduced in the vocabulary of an ASR system without the need to retrain or adapt the language model. This method makes use of special word classes (BOW) to assign probabilities to events not explicitly represented in the baseline LM, and to introduce new words as alternate orthographic forms of the BOWs during the recognition process. Words are linked with their lexical back-off classes via their POS tag.

A recognition experiment was carried out using prime time broadcast news shows collected from the Web site of a major television channel along with associated manually created document summaries provided by INA. The document summaries were automatically processed to obtain the words which were added to the recognizer vocabulary via lexical back-off, and their pronuciations were generated using grapheme-to-phoneme conversion. This approach yields a 30% reduction in the OOV rate, with 80% of the occurrences of the newly introduced words being correctly recognized. Moreever, about 84% of the introduced named entities, the most important category for an indexing task, are correctly recognized. In order to assess the maximum gain that can be obtained with the proposed method, an oracle experiment was carried out. An error analysis by BOW showed that for proper names, geographic locations, and first names, the main source of error is the automatic phonetic transcription.

ACKNOWLEDGMENTS

The authors wish to thank Laurent Vinet from INA for introducing them to the documentary context and for providing the metadata used in this work, and Patrick Paroubek from LIMSI for his support on POS taggers.

REFERENCES

[1] J.-L. Gauvain, L. Lamel, and G. Adda, "Audio partitioning and transcription for broadcast data indexation," *MTAP Journal*, vol. 14, no. 2, pp. 187–200, 2001.

[2] J. Garofolo, G. Auzanne, and E. Voorhees, "The trec spoken document retrieval track: A success story," in *Proceedings of the 8th Text Retrieval Conference TREC-8*, Gaithersburg, Maryland, November 1999, pp. 107–130.

[3] Cedric Auzanne, John S. Garofolo, Jonathan G. Fiscus, and William M. Fisher, "Automatic language model adaptation for spoken document retrieval," in *SDR 2000*. TREC 9, 2000.

[4] A. Allauzen and J.L Gauvain, "Adaptation automatique du modèle de langage d'un système de transcription de journaux parlés," *Traitement Automatique des langues*, vol. 44, no. 1, pp. 11–31, 2003.

[5] Marcello Federico and Nicola Bertoldi, "Broadcast news adaptation using contemporary texts," in *Proc. Eurospeech*, Aalborg, Denmark, September 2001, pp. 239–242.

[6] C. Barras, A. Allauzen, L. Lamel, and Jean-Luc Gauvain, "Transcribing audio-video archives," in *Proc. ICASSP*, Orlando, Florida, May 2002, vol. 1, pp. 13–16.

[7] L. Chen, J.-L. Gauvain, L. Lamel, G. Adda, and M. Adda-Decker, "Language Model Adaptation for Broadcast News Transcription," in *Proc. ISCA ITRW 2001 Adaptation Methods for Speech Recognition*, Sophia-Antipolis, August 2001.

[8] E. Brill, "Some advances in rule based part-of-speech tagging," in *Proceedings of the Twelfth National Conference on Artificial Intelligence*, AAAI, Ed., Seattle, WA, 1994, pp. 722–727.

[9] F. Yvon, P. Boula de Mareüil, C. d'Alessandro, V. Aubergé., M. Bagein, G. Bailly, F. Béchet, S. Foukia, J.-P. Goldman, E. Keller, D. O'Shaughnessy, V. Pagel, F. Sannier, J. Véronis, and B. Zellner, "Objective evaluation of grapheme-to-phoneme conversion for text-to-speech synthesis in french," *Computer Speech and Language*, vol. 12, no. 3, 1998.