Diachronic vocabulary adaptation for broadcast news transcription

Alexandre Allauzen and Jean-Luc Gauvain

Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{allauzen,gauvain}@limsi.fr

10.21437/Interspeech.2005-23

Abstract

This article investigates the use of Internet news sources to automatically adapt the vocabulary of a French and an English broadcast news transcription system. A specific method is developed to gather training, development and test corpora from selected websites, normalizing them for further use. A vectorial vocabulary adaptation algorithm is described which interpolates word frequencies estimated on adaptation corpora to directly maximize lexical coverage on a development corpus. To test the generality of this approach, experiments were carried out simultaneously in French and in English (UK) on a daily basis for the month May 2004. In both languages, the OOV rate is reduced by more than a half.

1. Introduction

Most of todays methods for indexation of broadcast audio data are manual. National institutions like the INA[1] in France process thousands hours of audiovisual data on a daily basis, in order to extract semantic information, and to interpret and summarize the content of documents. The development of automatic and efficient support for these manual tasks is a great challenge and over the last decade there has been growing interest in the use of automatic speech recognition (ASR) as a tool to provide random and relevant access to large broadcast news (BN) databases [7]. The NIST evaluations on Spoken document Retrieval (SDR) showed in 2000 that the quality of ASR transcripts is good enough to enable a variety of applications such as content-based document retrieval [6]. The performance of BN transcription system has been improved since, with reported word error rates in several languages between 10 and 15%.

Even though the linguistic properties of BN data continually change over time, most ASR systems use static language models (LM). In particular, the vocabulary (the set of words that can be recognized by the system) is usually selected using a large, fixed training corpus. There is often a substantial gap between the epoch of the LM training corpus and the audio data to process (two years is not unusual), because of the high cost of collecting and processing texts. In contrast, the news domain is characterized by rapid changes in topic, with corresponding changes in vocabulary items and linguistic content. The longer the gap between the training data epoch and the processed data, the higher the expected proportion of words that did not appear in the training data. The recognition vocabulary usually consists of the most frequent words in the training corpus. When large amounts of training data are available, words in recent data can be favored. Out of vocabulary words (OOVs) are words not included in the recognition vocabulary, therefore cannot be hypothesized by the system. These words often cause additional recognition errors in their immediate context and are mainly named entities. Furthermore, named entities are an important lexical class for which recognition errors have a significant impact on the indexation accuracy. One of the conclusions of the SDR track highlighted by the NIST was the need to develop new methods for linguistic model adaptation over time [6].

In this article, method for adapting the vocabulary of a speech recognizer to the news content is investigated. We propose to make use of texts collected from websites to be able to model the lexical content of the news on a daily basis. The next section describes the baseline vocabulary, followed by the automatic algorithm for vocabulary adaptation. Then the use of the Web as a mean to gather corpora from a variety of news sources is described. Finally experiments are reported where the recognition vocabulary is updated on a daily basis over a month-long test period in both languages.

2. The baseline vocabulary

For both languages, the baseline vocabulary used in this work contains 65k words selected from the same training texts that are used for language modeling. The most appropriate texts for modeling the linguistic content of BN data are accurate manual transcripts of BN shows. Unfortunately, this type of data does not exist in a sufficient quantities especially for languages other than English, and additional types of texts are often used such as newspapers and newswires. To deal with different sizes of text from the various sources, the data are usually grouped into subcorpora depending on their types. For each subcorpus a set of words is selected using thresholds on the word frequencies, and the word lists are merged to create the baseline vocabulary. For both languages, the training corpora mainly date from the Nineties and are fully described in in [2] for French, and in [7] for the English system.

A typology of OOV words was carried out on different test sets of French BN transcripts dating from 2000, 2002 (both are described in [2]) and 2004. The last set is the one used in this work, and is further described in section 4. Globally the OOV rate increases over time with 0.9% in 2000, 1.4% in 2002 and 2.1% in 2004. More than the half (51% in 2000 and 55% in 2002) of OOV forms are named entities and 14% are common names. On the test set of 2004, the trend is more pronounced with a proportion of 65% of the OOV words which are proper names. Moreover, on the corpus of 2002 which spans a two month period, 34 % of OOV forms occur on only one day during the period. This study shows that the OOV words mainly concern word classes which are important in an indexing task, and most of them change over time.

[1] Institut National de l'Audiovisuel: http://www.ina.fr

3. Vocabulary adaptation

To adapt the recognizer vocabulary, the first need is up-to-date data called adaptation data. Then one option is to use the same methodology as was used for the baseline vocabulary. That is to add the counts of the words in the adaptation data to those in the other sources. However, the new words (the ones we want to get) will necessarily be relatively infrequent in the combined data. Alternatively the counts can serve as an additional subcorpora, for which appropriate thresholds would need to be selected before merging the word list with the most frequent words in the other subcorpora. In either case this method requires manual intervention to tune the word frequency thresholds. An other solution is to develop an automatic adaptation method which makes use of contemporaneous texts to model the lexical and linguistic content of the news [8, 5]. The method used in this work is a vectorial-based adaptation method which efficiently combines several adaptation corpora by optimizing a vectorial criterion on a development corpus, thus eliminating the need for human intervention. This algorithm was introduced in [2].

As in information retrieval [9], a document, a corpus or more generally a distribution of words can be represented by a vector in the word space. In this article, the term frequency is chosen to be the quantifier. The classical *idf* term (inverse document frequency) is not used, because when building a vocabulary, all words have to be considered, not only the "informative" ones. The K adaptation corpora (X_1, X_2, ... , X_K) and the development corpus Y are represented as vectors ($\vec{X_1}$, $\vec{X_2}$,... , $\vec{X_K}$) and $\vec{Y}$. If the words of $\vec{Y}$ are sorted by their frequencies, and if the N most frequent words are selected, we obviously obtain the N words vocabulary with the lowest OOV rate estimated on the development corpus. The vectorial adaptation therefore aims to estimate $\vec{Y}$ by a linear combination of the adaptation vectors.

If $\mathcal{F} = (\vec{X_1}, \vec{X_2}, ..., \vec{X_K})$ is a linearly independent family, $\mathcal{F}$ forms a basis of the subspace $\mathcal{S}$ and $\dim(\mathcal{F}) = K$. There are two possibilities, depending whether $\vec{Y}$ is a member of $\mathcal{S}$ or not. The first case means that $\vec{Y}$ can be generated by a linear combination of vectors of $\mathcal{F}$, but in practice it never happens. In the second case, $\vec{Y}$ is not directly reachable, and its best approximation is the orthogonal projection $\vec{Y_p}$ in $\mathcal{S}$. Therefore the vectorial algorithm aims to compute the coordinates of $\vec{Y_p}$ in $\mathcal{F}$. A sufficient condition for linear independence of $(\vec{Y_p}, \mathcal{F})$ is that each vector contains at least one word which does not occur in the others. The algorithm can be described as follow:

1. The linear independence of $(\vec{Y}, \mathcal{F})$ is checked with the sufficient condition previously given. If $\vec{Y}$ is not linearly independent, $\vec{Y} = \vec{Y_p}$, and go to step 2. If one vector of $\mathcal{F}$ is not linearly independent, $\mathcal{F}$ is not a basis. Then stop and remove one or more vectors.
2. $(\vec{x_1}, ..., \vec{x_K})$ an orthogonal basis of $\mathcal{S}$ is built.
3. $\vec{Y}$ is projected on this basis using the usual formula:

$$\vec{Y_p} = \sum_{i=1}^{K} \frac{\vec{Y}^t \vec{x_i}}{\| \vec{Y} \| \cdot \| \vec{x_i} \|} \vec{x_i} \ . \qquad (1)$$

4. The reverse change of basis is performed ($x \rightarrow X$) to obtain the coefficients $(\alpha_i)_{i=1}^{K}$ defined by:

$$\vec{Y_p} = \sum_{i=1}^{K} \alpha_i \vec{X_i} \ . \qquad (2)$$

5. The N words with the highest interpolated frequencies are selected from $\vec{Y_p}$ to constitute the adapted vocabulary $V_{vect}(d)$.

This algorithm does not require any *a priori* knowledge about the language, the adaptation data, their sizes, natures or relations. It can deal with corpora with very different sizes without one being masked by the others.

4. The Web: a diachronic adaptation corpus

Several web sites provide up-to-date news reports. Some provide selected articles or abstracts from national newspapers whereas others publish dispatches from newswire press agencies. By judiciously selecting web sites, it is possible to daily collect texts which reflect the lexical and linguistic content of current news events. However, the texts available on the Web come from a variety of heterogeneous sources. Thus the text data need to be collected, cleaned and normalized after having been downloaded from web sites before they can be combined and used to adapt the recognition vocabulary.

4.1. Text collection and normalization

HTML files were collected on a daily basis from selected websites. Scripts were developed to clean the downloaded web pages in order to extract useful information items (date, title and text) and to convert the pages to a common format for normalization. The coding of accents and other diacritic signs was also standardized and HTML markers removed.

The text normalization process defines what is considered to be a word. It also aims to transform written texts in order to create LM training data [1]. A specific normalization aspect appeared in the following English experiments, since most of the available textual resources are suited for American English and not British English BN data. So we may ensure the mapping between this two kind of English, and British texts are converted in an "American style". The converting rules are inspired by the freely available package VarCon[2].

In addition to the problems encountered in the processing heterogeneous texts, there also are typographical errors. It is not possible to use a finite and fixed word list to filter typographic errors in texts extracted from the web [4]. The normalization process is therefore divided in three steps. The first step processes ambiguous punctuation marks (such as hyphen and apostrophe). The second step uses the baseline Vocabulary to process compound words and sentence initial capitalization. In the final step, number are expanded to word sequences.

4.2. Corpus selection

Websites are obviously selected for their content and reliability. Besides, one of the main constraints is to keep following experiments as comparable as possible in both languages. Therefore we focus on sites which have both French and English versions and provide quite similar data.

On the European channel *Euronews* website[3], videos of the same BN shows are available in seven European languages: English and French, as well as German, Italian, Spanish, Portuguese and Polish. Approximate transcripts are associated with these videos and supply test and development corpora. The behavior of each text source is summarized in the upper part of

[2] http://wordlist.sourceforge.net

[3] http://www.euronews.net

Source	*Mean*	*Min*	*Max*
E^{UK}	1.8k	0.5k	2.6k
Y^{UK}	96.3k	64.8k	130.6k
E^{FR}	2.0k	0.5k	2.9k
Y^{FR}	64.7k	38.4k	86.4k
A_{28}^{UK}	2,268.9k	1,4071.4k	2,752.2k
B_7^{UK}	11.0k	10.0k	12.2k
A_{28}^{FR}	1,786.6k	1,753.3k	1,815.5k
B_7^{FR}	11.9k	7.3k	13.7k

Table 1: Mean, minimum and maximum of number of kilo words per day and per source on May 2004.

Table 1 in terms of the mean, minimum and maximum of number of words per day observed on May 2004. Sources are named E^{FR} for French and E^{UK} for English. They respectively provide a mean of 2.0k and 1.8k words a day. And whereas only a small amount of textual data are available, this websites leads to a real experimental opportunity, with BN shows in the both languages and well suited associated texts.

To increase the amount of data, the second selected web site is the news portal of Yahoo[4] which publish dispatches from newswire press agencies. This additional source enables to collect a mean of 64k words a day in French (denoted Y^{FR} in Table 1) and a mean 96k words a day in English (denoted Y^{UK}). In the following, $S(d)$ denotes all the texts downloaded from the site S dated from the day d.

In order to construct homogeneous text sets with a sufficient amount of data, it is preferable to merge the data from several consecutive days. For each day d in May 2004, the adaptation corpora called $A_k^{FR}(d)$ and $A_k^{UK}(d)$ are built by merging the texts from the current day and the $k-1$ preceding days, for example in French:

$$A_k^{FR}(d) = \bigcup [Y^{FR}(d-k+1) \cdots Y^{FR}(d)].$$

Since on consecutive days there can be a wide variation in the linguistic content of the news, a sharp but small development corpus can be obtained by gathering all texts of the day d from *Euronews* source:

$$D^{FR}(d) = E^{FR}(d) \text{ and } D^{UK}(d) = E^{UK}(d).$$

The remaining texts provide an additional adaptation set which is more related to the task than Yahoo texts. For example for French this corpus is defined by:

$$B_k^{FR}(d) = \bigcup [E^{FR}(d-k+1) \cdots E^{FR}(d-1)].$$

Constraints which lead to an selection of k are of a different type. First k must respect the week periodicity of the news websites (less news on Sunday, special sport edition on Monday, ...). This suggests using the week as a time unit for k. Then the size of the corpus has to suit with the task. However, bigger is not always better, since the corpus must remain specific to the day. Given these constraints and after development on April 2004 data, the best choice for k was found to be the same for both languages: A_{28} and B_7 are used for vocabulary adaptation. The bottom part of Table 1 shows the characteristics of the different corpora built with texts extracted from the Web.

[4]http://uk.news.yahoo.com for U.K and Ireland and http://fr.news.yahoo.com for France

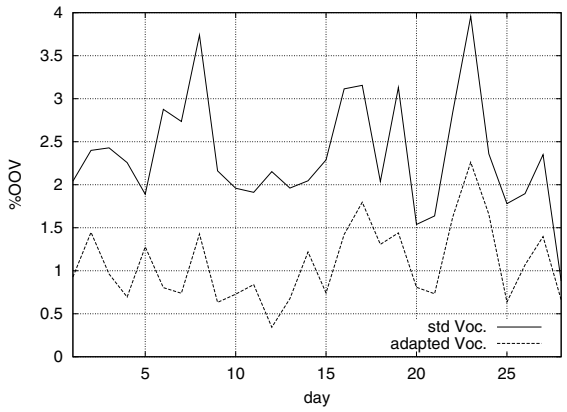

Figure 1: French experiment : OOV rate estimated for each day d on the test set $D^{FR}(d)$ with the baseline vocabulary and the French adapted vocabulary.

5. Experiments

For each day d, test sets are $D^{FR}(d)$ and $D^{UK}(d)$. To adapt vocabularies to these targets, the word frequencies of corpora $D^{FR}(d-1)$ and $D^{UK}(d-1)$ constitute the development vectors $\vec{Y}$ for the algorithm. That amounts to adapting the vocabulary to the $d-1$ day and to test this adaptation for the day d. Two kinds of training vector are needed: training vectors whose represent the punctual lexical content of todays news, and a training vector which aims to model the general part of language. Corpora $A_{28}^{FR}(d-1)$ and $B_7^{FR}(d-1)$ for French, and $A_{28}^{UK}(d-1)$ and $B_7^{UK}(d-1)$ for English are specifically designed for the first goal and are used to build training vectors $\vec{X_1}$ and $\vec{X_2}$.

The general French vector $\vec{X_3}$ is derived from the word frequency list estimated on a fixed corpus comprised of 332 M words of newspaper texts from *Le Monde* and *Le Monde diplomatique*. To filter out typographical errors, words occurring less than 10 times are excluded. For English, the general vector $\vec{X_3}$ is also derived from the word frequency list estimated on newspapers texts used to train the baseline American LM [7].

Adaptation experiments are carried out for the four early weeks of May 2004. In both languages, the baseline vocabulary is updated using the vectorial algorithm. Figures 1 and 2 shows the variation of the OOV rate with both the baseline and adapted vocabularies respectively in French and English. In French the OOV rates obtained with the adapted vocabulary are significantly lower those for the baseline vocabulary for all days. The absolute gain ranges from 0.22% to 2.31%, with a mean of 1.44%. The average relative reduction in OOV rate is 61%. In English, the tendency is the same except for the two first days, when both OOV rates are nearly equal. This is mainly due to the very low amount of training data for these two days. Over the four weeks, the absolute gain ranges from 0% to 1.97%, with a mean of 0.85%. This yields to an average relative reduction of the OOV rate of 56%.

The lexical rank of a word in the adapted vocabulary is obtained by sorting words according to their interpolated fre-

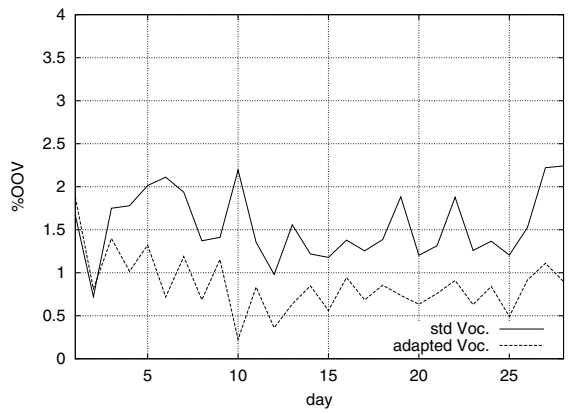

Figure 2: English experiment : OOV rate observed for each day d on the test set $D^{UK}(d)$ with the baseline vocabulary and the English adapted vocabulary.

quency in the vector $\vec{Y_p}$. The mean lexical rank for new words is 42004 in French and 18892 for English. However some of the the new words are ranked among the most frequent words in $\vec{Y_p}$. The 20 words with the highest lexical rank observed for the 14th May are listed in Table 2. In both languages, almost all forms are proper names (*Kadyrov, Ghraïb, Moqtada, ...*) or derived from a proper name (*Sadr's, Kadyrow's*). In French, some words were already included in the baseline vocabulary in an uncapitalized form. These words can be considered as named entities because they refer to specific contexts or entities, for example "*Premier* Ministre" (Prime minister) or "Congrès" (for the American Congress).

Different values of k were experimented for the adaptation corpora. When k varying between 1 an 28 for A_k and B_k, the OOV rates observed with the adapted vocabulary increase slightly in both languages. For vocabulary selection, the linear interpolation of unigram distributions was proposed in [3] to maximize the likelihood of a development set. This technique was experimented in our evaluation framework and yields to a OOV reduction which is twice lower. An other solution is just to add the word counts observed on the adaptation corpus to the baseline counts and to re-estimate the vocabulary. This solution only yields to smaller OOV reduction of 9%.

6. Conclusions

This article has investigated the use of Internet sources for the daily adaptation of the vocabulary of a broadcast news transcriptions system in two languages: French and British English. Specific methods are used to create text resources from news websites, including processing steps for information extraction from the HTML files, text normalization to be closer to spoken language, and corpus gathering. For each day of May 2004, three corpora are built for adaptation, development and test.

A vectorial algorithm for vocabulary adaptation is used which combines word frequencies vectors estimated on adaptation corpora to directly maximize lexical coverage on a development corpus. To assess the language portability of this algorithm, a symmetrical framework is proposed to evaluate the approach for French and for British English on a daily basis. Experiments show a significant reduction of the OOV rate compared with the baseline vocabulary: a relative decrease of 61 % in French and 56% in English.

French		*English*	
Premier	93	Ghraib	81
Kadyrov	100	Sadr	119
Ghraïb	109	Kadyrov	165
Moqtada	122	Hoon	214
Rumsfeld	139	Moqtada	265
Poutine	184	Battisti	273
Sasser	322	Kerbala	347
Armée	326	Taguda	349
Akhmad	330	Sadr's	422
Hoon	331	Fallujah	580
Falloujah	414	Euronews	697
Battisti	435	Najaf	744
Zeitoun	436	Musab	1121
Ramzan	437	Ramzan	1127
Congrès	539	Kadyrow's	1150
Comité	580	Kirkuk	1155
Najaf	592	Latif	1158
Donaldson	608	Ghraib's	1186
Taguba	610	infernal	1191
EADS	611	Falluja	1467

Table 2: The 20 words with highest lexical rank obtained with $V_{vect}(d)$ for May 2004 the 14th

7. References

[1] G. Adda, M. Adda-Decker, J.L. Gauvain, and L. Lamel. Text normalization and speech recognition in French. In *Proc. Eurospeech*, volume 5, pages 2711–2714, Rhodes, September 1997.

[2] A. Allauzen and J.L Gauvain. Adaptation automatique du modèle de langage d'un système de transcription de journaux parlés. *Traitement Automatique des langues*, 44(1): 11–31, 2003.

[3] Wen Wang Anand Venkataraman. Techniques for effective vocabulary selection. In *Proceedings of Eurospeech 2003*, pages 245–248, September 2003.

[4] C. Barras, A. Allauzen, L. Lamel, and Jean-Luc Gauvain. Transcribing audio-video archives. In *Proc. ICASSP*, volume 1, pages 13–16, Orlando, Florida, May 2002.

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

[6] 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*, pages 107–130, Gaithersburg, Maryland, November 1999.

[7] J.-L. Gauvain, L. Lamel, and G. Adda. Audio partitioning and transcription for broadcast data indexation. *MTAP Journal*, 14(2):187–200, 2001.

[8] T. Kemp and A. Waibel. Reducing the oov rate in broadcast news speech recognition. In *Proc. ICSLP*, volume 5, pages 1839–1842, Sydney, Australia, 1998.

[9] G. Salton and M. J. McGill. Smart and sire experimental retrieval system. In K. Sparck-Jones and P. Willet, editors, *Readings in Information Retrieval*, pages 381–399. Morgan-Kaufmann, 1997.