About vocabulary adaptation for automatic speech recognition of video data

Denis Jouvet, David Langlois, Mohamed Amine Menacer, Dominique Fohr, Odile Mella, Kamel Smaïli

About vocabulary adaptation for automatic speech recognition of video data

D. Jouvet[1,2,3], *D. Langlois*[1,2], *M.A. Menacer*[1], *D. Fohr*[1,2,3], *O. Mella*[1,2,3], *K. Smaïli*[1,2]

[1] Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
[2] CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
[3] Inria, Villers-lès-Nancy, F-54600, France

{ denis.jouvet, david.langlois, mohamed-amine.menacer, dominique.fohr, odile.mella, kamel.smaili }@loria.fr

Abstract

This paper discusses the adaptation of vocabularies for automatic speech recognition. The context is the transcriptions of videos in French, English and Arabic. Baseline automatic speech recognition systems have been developed using available data. However, the available text data, including the GigaWord corpora from LDC, are getting quite old with respect to recent videos that are to be transcribed. The paper presents the collection of recent textual data from internet for updating the speech recognition vocabularies and training the language models, as well as the elaboration of development data sets necessary for the vocabulary selection process. The paper also compares the coverage of the training data collected from internet, and of the GigaWord data, with finite size vocabularies made of the most frequent words. Finally, the paper presents and discusses the amount of out-of-vocabulary word occurrences, before and after update of the vocabularies, for the three languages.

Index Terms: Speech recognition, vocabulary, vocabulary adaptation, vocabulary selection.

1. Introduction

The vocabulary is one of the key components of an automatic speech recognition (ASR) system. It needs to be adequate with respect to the considered speech recognition task, and this is usually achieved through a training or adaptation process. That is the object of this paper, which discusses the adaptation of vocabularies for automatic speech transcription of videos in French, English and Arabic, for AMIS (Access Multilingual Information opinionS) project[1]. AMIS project aims at helping users to access information from videos that are in a foreign language, that is to understand the main ideas of the video. The best way to do that, is to summarize the video for having access to the essential information. Therefore, AMIS focuses on the most relevant information in videos by summarizing and translating it to the user. Obviously, the process starts by an automatic transcription of the audio channel.

Baseline ASR systems used at the beginning of the project have been developed from available corpora. For what concerns the linguistic part, that means that the vocabularies and the associated language models have been elaborated from quite old text data. Consequently, the vocabularies are somewhat outdated, and they are not relevant for a proper processing of person names and locations that have recently emerged in the news. Besides the fact that out-of-vocabulary (OOV) words affect speech recognition performance (in average, each out-of-vocabulary word produces 1.2 errors [1]), names of persons and of locations convey a very important and useful information for understanding the content of the videos. One way to cope with this aspect is to collect large amounts of text data over the web, that correspond to about the same time period as that of the videos to be processed, and build new speech recognition vocabularies from this new text data.

Unknows words are also problematic in natural language processing, for example for syntactic parsing and for machine translation. Several papers have investigated the handling of unknown words [2], including the use of a probabilistic model for guessing base forms [3] in English and Finish, and a morphological guesser for lemmatization in Arabic [4]. However, such approaches for dealing with written texts are not applicable to speech recognition.

With respect to speech recognition, several approaches have been developed in the past for elaborating vocabularies that are adequate for a given task. When a single text corpus is available, and when this corpus is homogeneous, the selection method is straightforward, it simply consists in selecting the most frequent words in the training corpus. However, since many years, the selection is done from numerous and heterogeneous corpora, which differ strongly in term of source or content (e.g., various radio or TV channels, journals, speech transcripts, …), time period, and size (from a few million words up to more than several hundred million words). In such case, it is not suitable to concatenate all the text corpora and just select the most frequent words. A frequent word in a small corpus, thus interesting to select, may end up with a small frequency in the concatenated data, and would thus not be selected.

When dealing with a heterogeneous set of text corpora, various selection methods have been proposed that rely on the unigram distribution of the words in each sub-corpus. A conventional approach consists in finding the linear combination of the unigrams associated to each sub-corpus, that matches the best with the unigram distribution of some development set [5],and [6]; then the words having the largest unigram values (according to the combined unigram distribution) are selected. The combination parameters are obtained through an expectation-maximization process. Selection approaches based on neural networks have also been investigated [7]. It should be noted that all these techniques require the availability of a development set, representative of the task, for optimizing the unigram combination weights.

This paper investigates the selection of speech recognition vocabularies in French, English and Arabic, for the automatic transcription of videos in AMIS project. It is organized as follows. Section 2 presents the baseline speech recognition systems. Section 3 describes the collection of the textual data over internet. Section 4 presents an analysis of the collected data, with a comparison to the GigaWord data sets. Finally, Section 5 details the selection of speech recognition vocabularies and discusses some evaluation results.

[1] http://deustotechlife.deusto.es/amis/project/

2. Baseline ASR systems

The speech recognition systems are based on the KALDI speech recognition toolkit [8].

Acoustic modeling is based on Deep Neural Networks (DNN), as such modeling provides the best performance [9]. The DNN has an input layer of 440 neurons (11 frames of 40 coefficients each), 6 hidden layers of 2048 neurons each, and the output layer has about 4000 neurons, corresponding to the number of shared densities of the initial GMM-based speech recognition system. The classical n-gram approach is used for language modeling.

Table 1. *Some characteristics related to linguistics aspects of the baseline ASR systems.*

	French	English	Arabic
Text training data (number of word occurrences)	1,620 M	155 M	1,000 M
Vocabulary size (number of words)	97 k	150 k	95 k
Number of pronunciation variants per word.	2.1	1.1	5.1

Table 1 presents some characteristics of the baseline ASR systems, with respect to some linguistic aspects. Vocabulary sizes vary from about 100 k words (for French and Arabic) to 150 k words (for English). The average number of pronunciations variants vary from 1.1 for English, to 2.1 for French, and 5.1 for Arabic. In French, most of the pronunciation variants are due to the optional mute-e at the end of many words, and to possible liaison consonants with following words starting by a vowel. In Arabic, the larger number of pronunciation variants is due to the absence of diacritic marks, which indicate short vowels, in the spelling of the vocabulary words.

3. Web textual data

As the vocabularies in the baseline ASR systems haven been defined according to available text corpora, that are rather old, the vocabularies are somewhat outdated, and they do not properly reflect the names of persons and locations observed in the recently collected videos of AMIS project. To update the vocabularies, new text data has been collected over the internet, in a period matching the period of the videos. This section also describes the elaboration of the test and development sets.

3.1. Training corpus

A few newspaper, radio and TV web sites in French, English and Arabic have been selected for collecting text data. A script was used to crawl web pages from the given sites over several months. The period over which text data was collected, was the same for the three languages.

A preprocessing has been applied on the raw text data collected from the various web sites. It mainly consists in removing useless data (e.g., date tags, hour tags, some keywords such as "view image", "download", ...), long non-Arabic text in Arabic web pages, ... Moreover, all duplicated sentences were also removed. About 80% of the amount of collected data is thus ignored. The amount of word occurrences available per language, after this preprocessing, is reported in Table 2. Note that during this preprocessing, all capital letters have been kept.

Table 2. *Amount of word occurrences per language for the web training data, and for the GigaWord data.*

Language	Web data	GigaWord
French	1.9 G	0.8 G
English	2.9 G	4.1 G
Arabic	0.7 G	1.1 G

3.2. Test corpus

The videos processed in AMIS project have been collected from Youtube. They correspond to various channels such as Alarabiya, Alquds, BBC, EnnaharTV, Euronews, France24, RT, SkynewsArabia… For most of the videos, Youtube provides short descriptions which correspond to the content of the video, and thus contain names of persons and locations occurring in the videos. Such data has been collected for all AMIS videos (a few thousand videos per language). This data is used as a test data set for evaluating the percentage of occurrences of out-of-vocabulary words. Table 3 indicates the amount of word occurrences in the development sets, for each language. An example of YouTube description is available in the top part of Figure 1.

3.3. Development corpus

On Euronews web site, one can find descriptions of Euronews videos. Such descriptions generally provide detailed information on the content of the video, which in some cases, is rather similar to a transcription of its content.

Independently of the collection of videos for AMIS project, another set of about 8000 videos, in Arabic language, were collected from Euronews web site. Cross language links available in the Euronews descriptions make possible to collect also the descriptions in French and in English for those videos. This led to about 8000 text descriptions in French, in English and in Arabic. This data set, which is not associated to AMIS videos, but comes from a similar period was used as a development set for the selection of the new vocabularies.

Among the videos collected in AMIS project, a part of them corresponds to Euronews. Hence, you can find in the top part of Figure 1 an example of a Euronews description (long description), along with the YouTube description (which is much shorter, four lines only).

Table 3. *Amount of word occurrences per language in development and test sets.*

	French	English	Arabic
Development set	1500 k	1720 k	1240 k
Test set	250 k	280 k	70 k

4. Analysis of the collected web data

An analysis of the data was carried out. For each data set collected from internet (i.e., French, English and Arabic), the frequency of occurrences of the words has been analyzed. The same analysis was applied on the GigaWord corpora available from the LDC (French [10], English [11], and Arabic [12]).

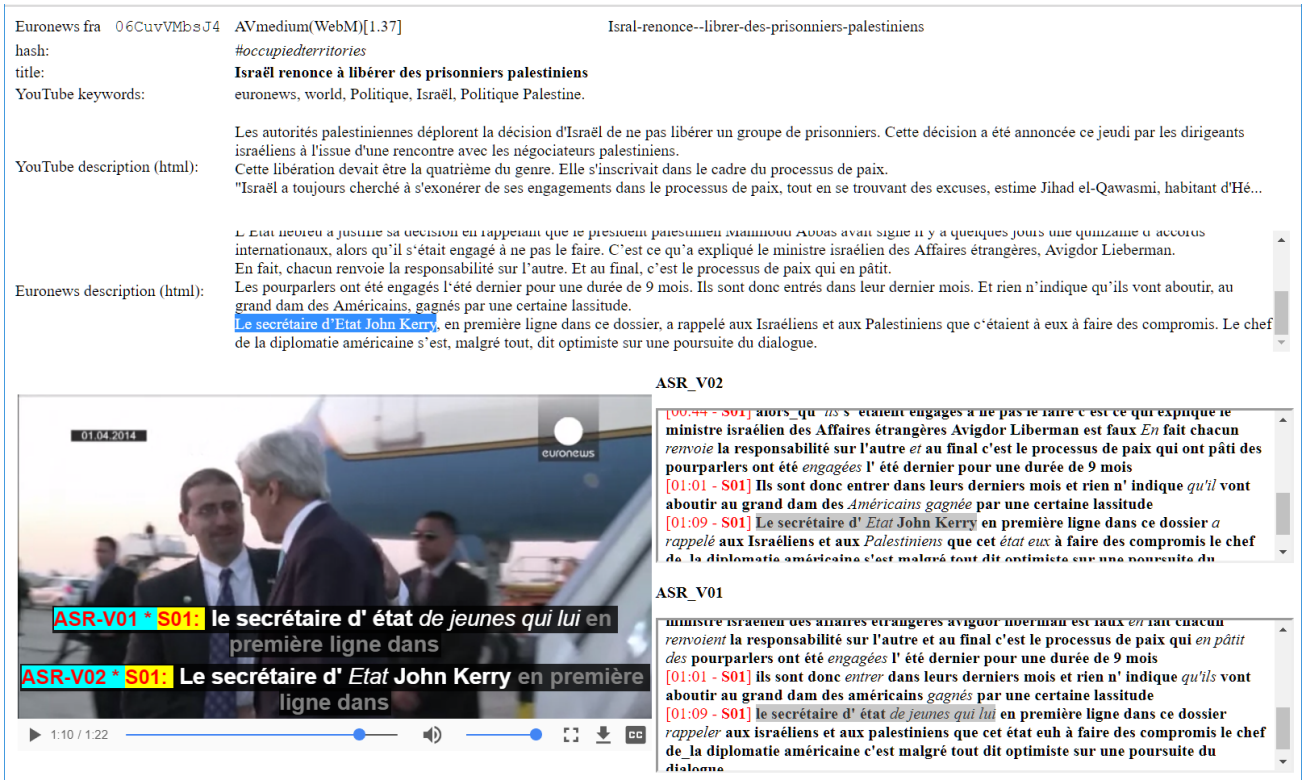

Figure 1: *Display of speech recognition results achieved with the old and new vocabularies. Both speech recognition results (ASR-V01 corresponding to the baseline ASR, and ASR-V02 corresponding to ASR with the updated vocabulary) are displayed as synchronized subtitles (bottom-left) and in separate frames (bottom-right). For helping checking recognition performance, when available, the YouTube and Euronews description are also displayed (middle part).*

As a result, Figure 2 displays, for each corpus, the coverage of the word occurrences with respect to the most frequent word tokens of the corresponding corpus. For example, for English with data collected over internet the 100,000 most frequent words cover about 96.6% of the 2,880 million word occurrences of the English data; whereas for Arabic, the 100,000 most frequent words cover only 92.7% of the 690 million word occurrences of the Arabic data. Solid lines correspond to the data collected on internet. Dotted lines correspond to the GigaWord corpora.

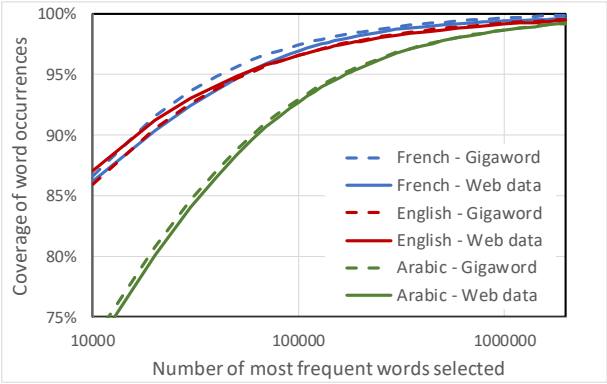

Figure 2: *Coverage of text data with respect to the most frequent words (of each data set).*

On the figure, the 4 curves corresponding to “French (gigaword corpus)”, “French (internet data)”, “English (gigaword corpus)”, and “English (internet data)” are very similar. For each language, internet data and GigaWord data leads to very similar results. The figure also shows that to reach a given coverage, much more words are needed in Arabic than in French and English languages, probably due to the morphological richness of Arabic.

5. Updated vocabularies

The new text data collected over internet is used here as text material from which new ASR vocabularies are selected for transcribing AMIS videos. Results are then analyzed mainly in terms of percentage of out-of-vocabulary words in the test sets for the different languages and different vocabulary sizes.

5.1. Selection of vocabulary words

The selection process relies on the conventional approach. First a unigram is estimated on each subset of the training corpus, corresponding to a radio channel, a TV channel, a journal, etc. For example, for the French language, the various subsets correspond to Euronews, France 24, France Inter, Le Monde, Le Figaro, L’Humanité, and so on. Overall, there are about 30 subsets for the French language. A similar splitting, according to web site, is done also for English and Arabic data, leading to 22 subsets for English and 29 subsets for Arabic.

Once unigrams models are trained on each subset, they are linearly combined to make a global unigram. The weights of the linear combination are estimated with an Estimation-Maximization (EM) algorithm to match as best as possible the unigram estimated on the development data. The objective

function that the E.M. algorithm optimizes is the Kullback-Leibler distance between the unigram distribution corresponding to the linear combination of the unigrams estimated on each sub-corpus, and the unigram distribution estimated on the development corpus.

On the French data, the two largest combination weights and the associated sub-corpus are the following: 0.876 for Euronews, and 0.106 for France24. All the other weights are below 0.01. A similar behavior is observed for the other languages. The large weight obtained for the Euronews channel may be due to the fact that the development set is made of descriptions of Euronews videos.

Finally, the selected vocabulary corresponds to the words that have the largest probability in the combined unigram. For each language, four vocabularies have been extracted corresponding respectively to the 100 k, 200 k, 400 k and 800 k most probable words.

5.2. Analysis of results

The best analysis that could be carried out requires a manual transcription of a large subset of AMIS videos. As such transcription is not available for AMIS videos, we have used the text data corresponding to the Youtube descriptions as test sets. On the test sets, we evaluated the amount of out-of-vocabulary words for the various vocabularies: baseline ASR vocabulary, and new vocabularies (100 k, 200 k, 400 k, and 800 k words). Results for the 3 languages are reported in Table 5. For comparison purpose, Table 4 reports the percentages of out-of-vocabulary words on the development sets.

Table 4. *Percentage of out-of-vocabulary words in the development sets for each language and vocabulary. Sizes of baseline vocabularies are specified in Table 1.*

	French	English	Arabic
Nb. words	51 k	64 k	129 k
Nb. occurrences	1500 k	1720 k	1240 k
Baseline (95 to 150 k)	1.8%	7.2%	17.4%
New 100 k	0.4%	1.1%	5.5%
New 200 k	0.1%	0.4%	3.1%
New 400 k	0.1%	0.3%	1.5%
New 800 k	0.1%	0.3%	0.2%

Table 5. *Percentage of out-of-vocabulary words in the test sets for each language and vocabulary. . Sizes of baseline vocabularies are specified in Table 1.*

	French	English	Arabic
Nb. words	20 k	21 k	20 k
Nb. occurrences	250 k	280 k	70 k
Baseline (95 to 150 k)	1.8%	5.5%	16.4%
New 100 k	0.8%	3.3%	6.8%
New 200 k	0.4%	2.7%	4.5%
New 400 k	0.2%	1.9%	3.1%
New 800 k	0.2%	1.5%	2.0%

As can be seen on these tables, the percentage of out-of-vocabulary words is much lower with the new vocabularies than with the old ones. The same behavior is observed on the development and on the test sets. In all cases, increasing the size of the vocabularies significantly reduces the percentage of out-of-vocabulary words. For example, for the English data, on the test set, the OOV rate was reduced from 5.5% with the baseline vocabulary (150 k words) to 3.3% with the new 100 k word vocabulary, and then to 2.7%, 1.9% and 1.5% respectively with the 200 k, 400 k and 800 k vocabularies.

Comparing the languages, the OOV rates are smaller for the French data than for the English data. The largest OOV rates are observed for the Arabic language. The large OOV rate on Arabic data was also observed in other studies related to statistical modeling of Arabic [13] and [14].

To check the benefit of the new vocabularies, they have been used for a new transcription of AMIS videos. The two speech recognition results obtained with the old vocabulary, and with the new vocabulary (100 k words), are displayed as simultaneous subtitles. Figure 1 provides a typical example of the recovery of names of persons thanks to the new vocabularies. “John Kerry” was not present in the old French vocabulary, and thus the corresponding occurrence was replaced by a sequence of short words which are acoustically close. As the person name is missing in the old vocabualry, the corresponding transcription (cf. line ASR-V01 in Figure 1) gets difficult to understand, and an important information (the name “John Kerry”) is missing; such behavior will also impact the machine translation process. With the new vocabularies, this problem is overcome.

6. Conclusion

This paper has investigated the problem of out-of-vocabulary word in the transcription of videos in French, English and Arabic. A large part of out-of-vocabulary words concerns names of persons and locations, which convey an important information for understanding the content of videos. To elaborate speech recognition vocabularies that are adequate for the transcription of the videos, large amount of data has been collected over internet in a period matching the period of the videos. This data collected over internet has been compared to the well-known GigaWord corpora, available from LDC. The behavior (coverage) of the frequent words of each corpus, is similar between the data collected over the web and the GigaWord data. Nevertheless, the comparison shows that much more (frequent) words are needed in Arabic than in French or English to achieve a similar coverage of the word occurrences.

The collected data has been used to elaborate updated vocabularies in French, English and Arabic. Different sizes have been considered from 100 k words up to 800 k words. Noticeable reductions in the OOV rates are observed when the vocabulary size increases. The smallest OOV rates are observed on French data, and the largest ones on Arabic data.

7. Acknowledgements

Part of this work was supported by the Chist-Era AMIS (Access Multilingual Information opinionS) project. Also, some experiments presented in this paper have been carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https: //www.grid5000.fr).

8. References

[1] Rosenfeld, R., "Optimizing lexical and ngram coverage via judicious use of linguistic data", *EUROSPEECH'95, 4th European Conf. on Speech Communication and Technology*, pp. 1763-1766, Madrid, Spain, 1995.

[2] Attia, M., Foster, J., Hogan, D., Roux, J. L., Tounsi, L., & Van Genabith, J., "Handling unknown words in statistical latent-variable parsing models for Arabic, English and French", *NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pp. 67-75, 2010.

[3] Lindén, K., "A probabilistic model for guessing base forms of new words by analogy", *Computational Linguistics and Intelligent Text Processing*, pp. 106-116, 2008.

[4] Attia, M., Samih, Y., Shaalan, K. F., & Van Genabith, J., "The Floating Arabic Dictionary: An Automatic Method for Updating a Lexical Database through the Detection and Lemmatization of Unknown Words". *COLING*, pp. 83-96, 2012.

[5] Venkataraman, A., and Wang, W., "Techniques for effective vocabulary selection", *INTERSPEECH'2003, 8th European Conf. on Speech Communication and Technology*, pp. 245-248, Geneva, Switzerland, 2003.

[6] Allauzen, A., and Gauvain, J.-L., "Automatic building of the vocabulary of a speech transcription system", In French "Construction automatique du vocabulaire d'un système de transcription", *JEP'2004, Journées d'Etudes sur la Parole*, Fès, Maroc, 2004.

[7] Jouvet, D., and Langlois, D,. "A machine learning based approach for vocabulary selection for speech transcription". *TSD'2013, Int. Conf. on Text, Speech and Dialogue*, Pilsen, Czech Republic, 2013.

[8] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al., "The kaldi speech recognition toolkit", *ASRU'2011, IEEE Workshop on Automatic Speech Recognition and Understanding*, Waikoloa, HI, USA, 2011.

[9] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R.., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups", *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[10] Graff, D., Mendonça, A, and DiPersio, D.. "French Gigaword Third Edition LDC2011T10". DVD. Philadelphia: Linguistic Data Consortium, 2011.

[11] Graff, D., and Cieri, C., "English Gigaword LDC2003T05". Web Download. Philadelphia: Linguistic Data Consortium, 2003.

[12] Parker, R., et al., "Arabic Gigaword Fifth Edition LDC2011T11". Web Download. Philadelphia: Linguistic Data Consortium, 2011.

[13] Meftouh, K., Smaili, K., & Laskri, M. T., "Comparative Study of Arabic and French Statistical Language Models". *ICAART*, pp. 156-160, 2009.

[14] Meftouh, K., Tayeb Laskri, M., & Smaili, K., "Modeling Arabic Language using statistical methods". *Arabian Journal for Science and Engineering*, 35(2), 69, 2010.