Adaptation of speech recognition vocabularies for improved transcription of YouTube videos

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

Adaptation of speech recognition vocabularies for improved transcription of YouTube videos

DENIS JOUVET[1,2,3], **DAVID LANGLOIS**[1,2], **MOHAMED AMINE MENACER**[1], **DOMINIQUE FOHR**[1,2,3], **ODILE MELLA**[1,2,3], **AND KAMEL 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*
[1] *{firstname.lastname}@loria.fr*

Compiled February 14, 2018

This paper discusses the adaptation of speech recognition vocabularies for automatic speech transcription. The context is the transcription of YouTube videos in French, English and Arabic. Baseline automatic speech recognition systems have been developed using previously available data. However, the available text data, including the GigaWord corpora from LDC, are getting quite old with respect to recent YouTube videos that are to be transcribed. After a discussion on the performance of the ASR baseline systems, the paper presents the collection of recent textual data from internet for updating the speech recognition vocabularies and for 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 the update of the speech recognition vocabularies, for the three languages. Moreover, some speech recognition evaluation results are provided and analyzed.

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 an empirical or an automatic selection process applied to relevant adaptation textual data. 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. In the project, the way to do that, is to generate a summary of the video for having access to its 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.

The baseline ASR systems used at the beginning of the project have been developed from previously available corpora. For what concerns the linguistic part, that means that the vocabularies and the associated language models have been elaborated

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

from quite old text data. Consequently, the vocabularies are somewhat outdated, and they are not relevant for a proper processing of person names and location names 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], they also impact on information retrieval performance [2] (even if, in such a case, strategies can be used in order to compensate for the resulting speech recognition errors). A large part of out-of-vocabulary words are names of persons or names of locations, and they convey a very important and useful information for information retrieval and for understanding the content of 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.

Unknown 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 [3], including the use of a probabilistic model for guessing base forms [4] in English and Finish, and a morphological guesser for lemmatization in Arabic [5]. However, such approaches for dealing with written texts are not applicable to speech recognition for large vocabulary.

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 text 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, etc.), time period, and size (from a few million words up to more than several hundred million words). In such a case, it is not suitable to concatenate all the text corpora and just select the most frequent words. Indeed, 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. The most trivial method consists in selecting the most frequent words in each sub-corpus, and then putting together all the selected words to obtain the target vocabulary. However such an ad hoc process requires tuning selection thresholds. To avoid that, various approaches 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 unigram distributions associated to each sub-corpus, that matches the best with the unigram distribution of some development set [6, 7]; 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 [8]. 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.

As an alternative to unigram-based vocabulary selection, learning-based approaches have been developed to select the most interesting words that should be inserted in a speech recognition vocabulary to handle out-of-vocabulary words for a given target evaluation corpus. In [9], both semantic and acoustic scores are combined for selecting the most interesting OOV word candidates. Relevant textual documents are used to get a list of possible OOV word candidates. The semantic score of an OOV word candidate is based on the comparison of the tf-idf (term frequency-inverse document frequency) values of vocabulary words occurring in the neighborhood of the OOV word candidate in relevant documents, with those of vocabulary words occurring in spoken documents of the target corpus. For computing acoustic scores, OOV word candidates are temporarily added to an ASR lexicon to recognize the spoken documents. In [10], potential OOV words are proposed based on different types of relatedness between in-vocabulary words and words retrieved from web-based resources. One approach exploits the semantic relations of WordNet, whereas the other one relies on word-embedding representations.

In some languages, as for example Chinese, computing the OOV rate is complex as it needs a word-based segmentation of the text. Often, Chinese OOV words get segmented into characters that are present in large generic Chinese lexicons, and thus do not end up as actual OOV words; [11] proposes an approach to deal with this problem. However, such phenomenon does not occur for the languages we are dealing with (French, English and Arabic) as words are separated by spaces in textual documents.

Here we are concerned by the transcription of videos in AMIS project. Videos have been collected from YouTube, in several languages: French, English and Arabic. They correspond to various TV channels, as for example: Alarabiya, Alquds, Euronews, France 24, BBC, etc.

Hence, this paper investigates the selection of speech recognition vocabularies in French, English and Arabic, for the automatic transcription of YouTube videos. It is organized as follows. Section 2 presents the baseline speech recognition systems, and discusses speech recognition performance. Section 3 describes the collection of the textual data over internet, and presents an analysis of the collected data, with a comparison to the Gigaword data sets. Finally, Section 4 details the selection of speech recognition vocabularies and discusses some evaluation results, with a detailed focus on French data for which more evaluation results are available.

2. BASELINE ASR SYSTEMS

Baseline ASR systems have been developed at the beginning of the AMIS project using previously available data: audio data for estimating the parameters of the acoustic models and text data for estimating the lexicons and the associated language model. After a short presentation of the different models, this section presents how some reference data is obtained for performance evaluation, and then discusses the resulting evaluation.

A. Modeling

The speech recognition systems are based on KALDI speech recognition toolkit [12].

Acoustic modeling relies on Deep Neural Networks (DNN), as such modeling provides the best performance [13]. For each language, 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, which corresponds to the number of shared densities of the initial GMM-based speech recognition system.

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

The classical n-gram approach is used for language modeling. Table 1 presents some characteristics of the baseline ASR systems. Vocabulary sizes vary from about 100 k words (for French and Arabic) to 150 k words (for English). The average number of pronunciations variants per word 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 per word is due to the absence of diacritic marks, which indicate short vowels, in the spelling of the vocabulary words.

Acoustic models for French and English have been trained using transcribed speech corpora of more than 200 hours of speech signal. The Arabic acoustic model has been trained using a smaller speech corpora, of about 50 hours. The largest part of those speech corpora corresponds to radio broadcast news. On the corresponding evaluation data sets, the word error rates were around 13% to 14% for the three languages (i.e., French, English and Arabic); i.e., for matching conditions.

B. Performance evaluation

As mentioned before, we are dealing with YouTube videos in French, English and Arabic, that corresponds to various TV channels. As illustrated in Figure 1, it is possible to get on the web some text data associated to YouTube videos. The YouTube web page usually provides only a few lines of text (for example four lines only for this video). However, for most of the videos corresponding to Euronews channel, a longer text can be found on the Euronews web page (an example of an Euronews description is provided in Figure 1).

It was observed that these texts often match approximately with portions of the audio of the corresponding videos. Hence we decided to extract some reference sentences from these texts and to use them for evaluating speech recognition performance.

B.1. Obtaining some reference data

Listening to several videos, it was observed that many sentences (from the YouTube web pages or from the Euronews web pages) matched quite well with the audio; whereas some other sentences did not match at all, or were not at the right place. Thus our goal here is to extract whole sentences matching reasonably well with the audio signal and to use them as approximate transcriptions for performance evaluation. To extract such sentences, the text from the web pages is aligned with the speech recognition output of the baseline systems. The alignment relies on an optimized set of costs. Substitution costs between words are dependent on the amount of letters that differ between these words. This leads to small substitution costs when the two words are similar. Insertion and deletion costs are dependent on the length of the words (small cost for short words, larger costs for longer words).

The main punctuation (i.e., dot, question mark, etc.) in the textual data is used to detect the beginning and end of the sentences. Also, as a reasonably correct timing of the beginning and end of the utterances is necessary for speech recognition evaluation purposes, we pay more attention to the matching costs at the beginning and end of sentences. So, sentences for which the alignment cost on the first three words and on the last three words is below a given threshold are selected as reference sentences. With such a criteria we assume that the timing obtained for the beginning and end of the sentences are quite correct. Note that we ignore sentences or segments for which the insertion rate or the deletion rate is greater than 30% (i.e., a clear missmatch between text and audio).

Table 2. Amount of (approximate) reference data obtained for each language.

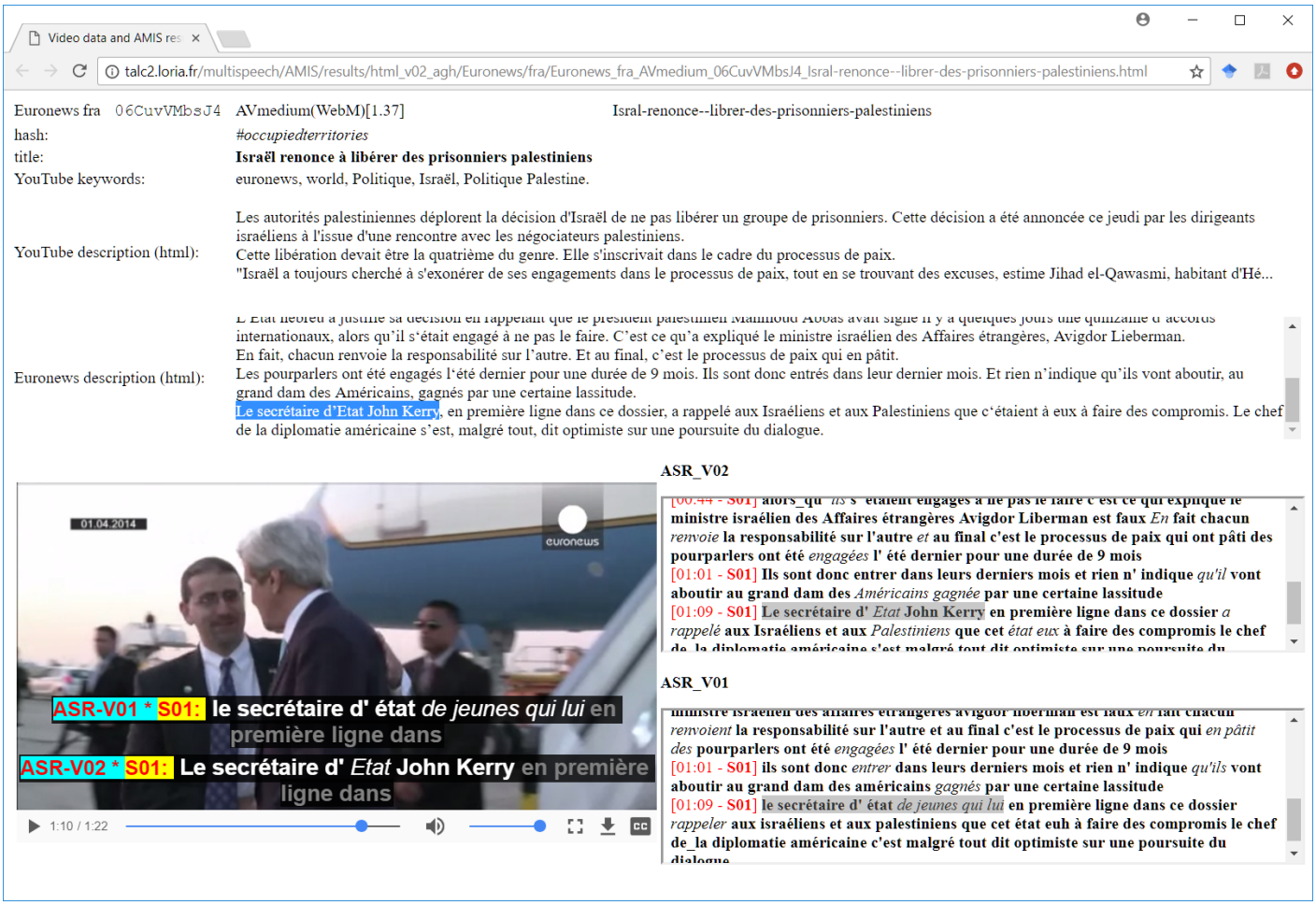
Language	YouTube text		Euronews text	
	Sentences	Words	Sentences	Words
French	5.6 k	119 k	7.3 k	146 k
English	2.6 k	51 k	4.0 k	77 k
Arabic	0.8 k	21 k	0.5 k	11 k

Table 2 displays the amount of reference data that has been obtained by this process for each language, using YouTube text data (on the left) or Euronews text data (on the right). More reference data have been obtained for French, than for English and Arabic. Almost 50% of the text from the web pages was kept as reference data for French, whereas this goes down to 30% for English and Arabic. An interesting point to note is that the extracted sentences have a rather similar length across web text sources (YouTube web page vs. Euronews web page) and languages (French, English and Arabic): on average around 20 words per sentence.

B.2. ASR performance evaluation

The approximate reference data have been used to get an estimate of the word error rates (WER) for the baseline systems. It should be noted that this is just a rough estimate, and that this estimated WER is likely to be higher than the actual WER as the text used

Fig. 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, the YouTube and Euronews text description, when available, are also displayed (middle part).



as reference is not the exact transcription of what is said in the videos.

For French, the estimated WER is 17.8% on YouTube-based references, and 17.2% on Euronews-based references. In both cases, the deletion and insertion rates are balanced (around 3.5 to 4.0%). On English data, similar WERs are observed: 17.4% on YouTube-based references, and 17.8% on Euronews-based references. Again the deletion and insertion rates are balanced (around 3.5 to 4.2%). With respect to Arabic data, the first estimate is higher, however several normalization steps are not yet included in the evaluation process, and the text data and ASR outputs are not consistent with respect to diacritics and some prefixes.

On French data, a more detailed analysis of the speech recognition results has been conducted with respect to part-of-speech (POS) tagging. Treetagger [14] has been used to annotate the tokens of the reference sentences in terms of POS. An analysis of the performance on the tokens associated to the NAM label (i.e., names of persons, names of locations, etc.) shows a 39% word error rate on those tokens (to compare to the global WER which is less than 18%). It was also observed that 14% of these name tokens are out-of-vocabulary, with respect to the baseline system.

3. WEB TEXTUAL DATA

The above evaluation confirms that the vocabularies in the baseline ASR systems, which have been defined according to previously available text corpora, 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 have been collected over the internet, in a time period matching that of the videos. This section also describes the elaboration of textual test and development sets that are latter used to select the vocabularies and to evaluate its

coverage.

A. 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 have been collected, is 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", etc.), long non-Arabic text in Arabic web pages, etc. 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 3. Note that during this preprocessing, all capital letters have been kept.

Table 3. 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

B. 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, etc. For most of the videos, YouTube provides short descriptions which gives an idea of the content of the video (an example of a YouTube description is available in the top part of Figure 1), and thus contain names of persons and of locations occurring in the videos. Such data have been collected for all AMIS videos (a few thousand videos per language). These data are used as a test data set for evaluating the percentage of occurrences of out-of-vocabulary words.

Table 4 indicates the amount of word occurrences in the tests sets and in the development sets, for each language. It should be noted that the (approximate) reference sentences extracted from YouTube web pages in Section 2-B.1 come from this textual test corpus.

C. Development corpus

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

Table 4. Amount of word occurrences per language in development and test sets.

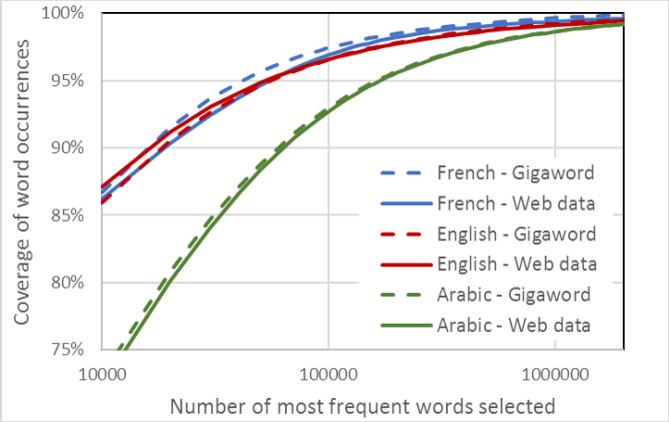
	French	English	Arabic
Development set	1500 k	1720 k	1240 k
Test set	250 k	280 k	70 k

D. Analysis of the collected web data

An analysis of the data collected from the web (cf. Section 3-A) has been carried out. For each data set collected from internet (i.e., French, English and Arabic), the frequency of occurrences of the word tokens has been analyzed. The same analysis was applied on the GigaWord corpora available from the LDC (French [15], English [16], and Arabic [17]).

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 (solid red curve) 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 green curve). Solid lines correspond to the data collected on internet. Dotted lines correspond to the GigaWord corpora.

Fig. 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, this is probably due to the morphological richness of Arabic.

4. UPDATE OF VOCABULARIES

The text data collected over internet (cf. Section 3-A) are 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. A more detailed analysis is provided for French, including speech recognition performance.

A. Selection of vocabulary words

The vocabulary selection process relies on a 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 sites, is done also for English and Arabic data, leading to 22 data subsets for English and 29 data subsets for Arabic.

Once unigrams models are estimated on each data subset, they are linearly combined to make a global unigram. The weights of the linear combination are estimated with an Estimation-Maximization (E.M.) 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 associated to the unigram corresponding to the Euronews data may be due to the fact that the development set is also 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.

B. Analysis of word coverage

To analyze the potential benefit of the new selected lexicons, we have used the text data corresponding to the YouTube descriptions (cf. Section 3-B) as test sets. On these test sets, we have evaluated the amount of out-of-vocabulary words for the various vocabularies: baseline ASR vocabulary, and new vocabularies of various sizes (100 k, 200 k, 400 k, and 800 k words). Results for the 3 languages are reported in Table 6. For comparison purpose, Table 5 reports the percentages of out-of-vocabulary words on the development sets (cf. Section 3-C).

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 sizes of the baseline lexicons are given in Table 1). For each language, a similar 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 in the development and test sets. 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 between the three languages, the OOV rates are smaller for the French data than for the English data and the largest OOV rates are observed for the Arabic language. Note that a large OOV rate on Arabic data was also observed in other studies related to statistical modeling of Arabic [18, 19].

Table 5. 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 6. 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%

C. ASR evaluation on French

To check the benefit of the new vocabularies, they have been used for a new transcription of a random subset of AMIS videos. As an example on French data, 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 vocabulary, 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.

Table 7. Performance with old and new lexicons on a subset of French videos.

	Baseline	New 100 k word lexicon
Global WER	17.4%	14.9%
Global OOV	1.3%	0.8%
WER on names	40%	27%
OOV on names	12%	5%

Table 7 details on a subset of French videos the benefits of the updated lexicons. Although the evaluation is restricted to a subset of French videos corresponding to about 6900 words, the global word error rate of the baseline system is in line with the WER estimated on the whole set of French videos (cf. Section 2-B.2), and the OOV rate is also in line with the OOV rate estimated on the textual test set for French (cf. Table 6). With the 100 k word new lexicon, the WER is significantly reduced (by 2.5% absolute), and the OOV is also reduced (down to 0.8%, which is similar to the OOV rate estimated on the textual test set for that lexicon). The second part of the table focuses on the names (names of persons, names of locations, ...) according to the POS tagging provided by TreeTagger. The WER and OOV rate reductions observed on the names are quite large: WER reduced by one third, and OOV rate reduced by more than 50%.

5. CONCLUSION

This paper has investigated the problem of out-of-vocabulary words in the transcription of videos in French, English and Arabic. A large part of out-of-vocabulary words concerns names of persons and of 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 have been collected over internet in a time period matching that of the videos. These data collected over internet have 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 words are needed in Arabic than in French or English to achieve a similar coverage of the word occurrences.

The collected data have 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.

Some speech recognition performance are also given and discussed. On French data, a detailed analysis is also provided with respect to names (of persons, of locations, etc.). This detailed analysis shows the poor coverage of the names by the baseline lexicons, and also demonstrates the benefits of the updated lexicons, both in term of WER reduction and OOV rate reduction.

ACKNOWLEDGMENTS

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).

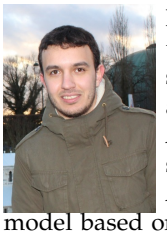
REFERENCES

1. R. Rosenfeld, "Optimizing lexical and ngram coverage via judicious use of linguistic data," in "Proceedings of EUROSPEECH 1995, 4th European Conference on Speech Communication and Technology," (Madrid, Spain, 1995), pp. 1763–1766.
2. P. C. Woodland, S. E. Johnson, P. Jourlin, and K. S. Jones, "Effects of out of vocabulary words in spoken document retrieval (poster session)," in "Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval," (ACM, 2000), pp. 372–374.
3. M. Attia, J. Foster, D. Hogan, J. Le Roux, L. Tounsi, and J. Van Genabith, "Handling unknown words in statistical latent-variable parsing models for Arabic, English and French," in "Proceedings of NAACL HLT 2010, First Workshop on Statistical Parsing of Morphologically-Rich Languages," (Association for Computational Linguistics, Los Angeles, California, 2010), pp. 67–75.
4. K. Lindén, "A probabilistic model for guessing base forms of new words by analogy," in "Proceedings CICLing 2008, 9th International Conference on Intelligent Text Processing and Computational Linguistics," (Haifa, Israel, 2008), pp. 106–116.
5. M. Attia, Y. Samih, K. Shaalan, and J. Van Genabith, "The floating Arabic dictionary: an automatic method for updating a lexical database through the detection and lemmatization of unknown words," in "Proceedings of COLING 2012," (Mumbai, India, 2012), pp. 83–96.
6. A. Venkataraman and W. Wang, "Techniques for effective vocabulary selection," in "Proceedings of EUROSPEECH 2003, 8th European Conference on Speech Communication and Technology," (Geneva, Switzerland, 2003), pp. 245–248.
7. A. Allauzen and J. L. Gauvain, "Automatic building of the vocabulary of a speech transcription system, in French: Construction automatique du vocabulaire d'un système de transcription," in "Proceedings JEP 2004, Journées d'Etudes sur la Parole," (Fès, Morocco, 2004).
8. D. Jouvet and D. Langlois, "A machine learning based approach for vocabulary selection for speech transcription," in "Proceedings TSD 2013, International Conference on Text,

Speech and Dialogue," (Pilsen, Czech Republic, 2013), pp. 60–67.
9. S. Yamahata, Y. Yamaguchi, A. Ogawa, H. Masataki, O. Yoshioka, and S. Takahashi, "Automatic vocabulary adaptation based on semantic and acoustic similarities," IEICE TRANSACTIONS on Inf. Syst. **97**, 1488–1496 (2014).
10. M. Sun, Y.-N. Chen, and A. I. Rudnicky, "Learning oov through semantic relatedness in spoken dialog systems," in "Sixteenth Annual Conference of the International Speech Communication Association," (2015).
11. Y. Zhang, P. Zhang, T. Li, and Y. Yan, "An unsupervised vocabulary selection technique for chinese automatic speech recognition," in "Spoken Language Technology Workshop (SLT), 2016 IEEE," (IEEE, 2016), pp. 420–425.
12. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in "Proceedings ASRU 2011, IEEE Workshop on Automatic Speech Recognition and Understanding," (Waikoloa, HI, USA, 2011).
13. G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep Neural Networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Process. Mag. **29**, 82–97 (2012).
14. H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in "Proceedings of International Conference on New Methods in Language Processing," (Manchester, UK, 1994), p. 154.
15. D. Graff, A. Mendonça, and D. DiPersio, "French Gigaword, third edition, LDC2011T10," (2011).
16. D. Graff, , and C. Cieri, "English Gigaword, LDC2003T05," (2003).
17. R. Parker *et al.*, "Arabic Gigaword, fifth edition, LDC2011T11," (2011).
18. K. Meftouh, K. Smaïli, and M. T. Laskri, "Comparative study of Arabic and French statistical language models," in "Proceedings ICAART 2009, International Conference On Agents and Artificial Intelligence," (Porto, Portugal, 2009).
19. K. Meftouh, M. T. Laskri, and K. Smaïli, "Modeling Arabic language using statistical methods," Arab. J. for Sci. Eng. **35**, 69–82 (2010).

Denis Jouvet graduated from Ecole Polytechnique (1979) and received the Engineer degree (1981) and Ph.D. degree in signal processing (1988) both from Ecole Nationale Supérieure des Télécommunications, Paris, and the habilitation degree from University of Avignon (France) in 2008. He was with France Télécom RD labs, in Lannion, from 1981 to 2009, where he was involved in, and then also managing, automatic speech recognition studies and the development of automatic speech recognition technology for interactive vocal services. He participated in the European projects SMADA on directory assistance, and DIVINES on intrinsic speech variabilities. In 2009, he joined the PAROLE team in the LORIA laboratory in Nancy, and he is currently the team leader of the MULTISPEECH team. He was and is still involved into collaborative projects (e.g., Interreg Allegro, Eurostar Emospeech, ANR+DFG IFCASL, Chist-ERA AMIS, . . .). He is author or co-author of more than 150 publications. His current main interests include automatic speech recognition, automatic speech synthesis, stochastic modeling, deep learning, signal processing, prosodic features, non-native automatic speech recognition and computer assisted foreign language learning.

David Langlois defended his PhD thesis in Computer Science in July 2002 on "Distant Events and Impossible Events in Statistical Language Modeling" at Henri Poincaré University (Nancy, France). Since September 2003 he is an assistant professor in Computer Science at University of Lorraine. His research interests are language modeling, automatic speech recognition and machine translation. He published more than 30 articles in international conferences (Interspeech, ICSLP, LREC,...). He was the co-supervisor of 2 theses and is co-supervising a third one in the scope of Machine Translation. He has been involved in MIAMM (an European project http://miamm.loria.fr/) and he participated to ESTER evaluation campaign (broadcast news transcription). Today, he is involved in the AMIS project (Access Multilingual Information opinionS, http://deustotechlife.deusto.es/amis/project/). Moreover, he participated three times to the part "Quality Estimation" of the World Machine Translation Evaluation Campaign.



Mohamed Amine Menacer is a Phd student at the University of Lorraine since 2016. His research interests centre around speech recognition and machine translation. In 2015, Mohamed Amine received his Master's degree in computer science at the high school of computer science in Algeria. His work was untitled "A new language model based on possibility theory", where he worked on the integration of a new language model into a statistical machine translation system. He occupied an engineer position during one year in a start up working on management software. His doctoral works are a part of the AMIS project where the aim is to help people to understand the main idea of videos in foreign languages.

Dominique Fohr graduated from ENSEM (engineering school) in 1983 and received his PhD in 1986 at the University Henri Poincare (Nancy, France). Since 1986, he is a researcher at CNRS and works in the MultiSpeech team of the Loria-Inria laboratory. He was involved in European projects SAM (Speech Assessment Methodologies), AITRAS (An Intelligent Real Time System for Signal Understanding), ROARS (RObust Analytical speech Recognition System), HIWIRE (Human Input that Works In Real Environments) and AMIS (Access Multilingual Information opinionS). His research interests include automatic text-to-speech alignment, computer aided foreign language learning, robust speech recognition and broadcast news transcription. He is author or co-author of more than 100 publications. He is co-author of ANTS and KATS (Automatic News Transcription Systems), Starap (toolkit to help the making of sub-titles for TV shows), and CoALT (software for Comparing Automatic Labelling Tools).

Odile Mella defended her PhD thesis in Computer Science in 1993 on "analysis of acoustic and phonetic features for speaker characterization" at the University of Nancy (France). Since 1982, she is assistant professor at University of Lorraine and teaches to both undergraduate and graduate students. She was head of the master's degree program in Computer Networking during several years. She is member of the Multispeech team at INRIA/LORIA. Her research fields are large vocabulary automatic speech recognition, speech-text alignment and non-native speech. She participated to several projects: AMIS, ContNomina, IFCASL, ORFEO, EVAL-ESTER,... She is co-author of ANTS (Automatic News Transcription System), KATS (Kaldi-based Automatic Transcription system) and ASTALI (Automatic Speech-Text Alignment Software) used in these projects.

Kamel Smaïli is Professor at the university of Lorraine since 2002, he obtained a PhD from the university of Nancy 1 in 1991 on automatic speech recognition. He defended his HDR in 2001 (Statistical language modelling: from speech recognition to machine translation). His research interest since more than 20 years concerns statistical language modelling for automatic speech recognition. Since 2000 he oriented his research towards speech to speech translation. He participated to several European and national projects concerning automatic speech recognition: COCOS, MULTIWORKS, COST, MIAMM, IVOMOB (RNRT project) and CMCU. He advised 14 PhD and HDR students and he participated to more than 35 PhD committees in France, Germany, Spain and Algeria. He took part to several program committees: Interspeech, Eurospeech, ICSLP, ICASSP, TALN, ICWMI, SIIE, TAIMA, Machine Translation, Computer speech and language, Speech communication, Journal of Natural Language Engineering,... He has been invited several times to give talks as invited speakers in Japan, France, Tunisia, Algeria and Morocco. He published 90 papers in international conferences and journals and 20 papers in francophone conferences. Furthermore, Mr Smaili was the head of MIAGE (Master and Bachelor) department for 7 years and the head of UFR (equivalent to a faculty) Mathematics and Informatics for 5 years where he managed more than 30 permanent people and 120 temporary positions, 500 students.