HAL
open science

First Broadcast News Transcription System for Khmer Language

Sopheap Seng, Sethserey Sam, Laurent Besacier, Brigitte Bigi, Eric Castelli '

To cite this version:

Sopheap Seng, Sethserey Sam, Laurent Besacier, Brigitte Bigi, Eric Castelli '. First Broadcast News Transcription System for Khmer Language. Sixth International Conference on Language Resources and Evaluation, 2008, Marrakech, Morocco. pp.2658-2661. hal-01392538

HAL Id: hal-01392538

https://hal.science/hal-01392538

Submitted on 13 Dec 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

First Broadcast News Transcription System for Khmer Language

Sopheap Seng*, Sethserey Sam*', Laurent Besacier*, Brigitte Bigi*, Eric Castelli'
* LIG 220, rue de la chimie, B.P. 53 38041 Grenoble Cedex 9
' MICA 1 Dai Co Viet, Hanoi, Vietnam
E-mail: Sopheap.Seng@imag.fr

Abstract

In this paper we present an overview on the development of a large vocabulary continuous speech recognition (LVCSR) system for Khmer, the official language of Cambodia, spoken by more than 15 million people. As an under-resourced language, develop a LVCSR system for Khmer is a challenging task. We describe our methodologies for quick language data collection and processing for language modeling and acoustic modeling. For language modeling, we investigate the use of word and sub-word as basic modeling unit in order to see the potential of sub-word units in the case of unsegmented language like Khmer. Grapheme-based acoustic modeling is used to quickly build our Khmer language acoustic model. Furthermore, the approaches and tools used for the development of our system are documented and made publicly available on the web. We hope this will contribute to accelerate the development of LVCSR system for a new language, especially for under-resource languages of developing countries where resources and expertise are limited.

1. Introduction

With respect to speech recognition, the Khmer language bears challenging characteristics: (1) the lack of language resources (text and speech corpora) in digital form, (2) the writing system without explicit word boundary, which calls for automatic segmentation approaches to make statistical language modeling feasible and (3) the acoustic and phonologic characteristics that are not yet well studied. The statistical nature of the approaches used in automatic speech recognition requires a great quantity of language resources in order to perform well. For under-resourced languages which are mostly from developing countries, those resources are available in a very limited quantity because of its economic interest and the lack of standardized automatic processing tools (standard character encoding, word processing software). In this situation, language data collection is a challenging task and requires innovative approaches and tools. Similar to Chinese and Thai, Khmer is written without spaces between words. A sentence in Khmer ពណ៌សម្តេចថាខ្មៅ could be segmented into ពណ៌|ស|ម្តេច|ថា|ខ្មៅ (color|white |why|say|black) or ពណ៌|សម្តេច|ថា| ខ្មៅ (color|king|say |black). A correct segmentation of a sentence into words requires the full knowledge of the vocabulary and of the semantics of the sentence. The automatic segmentation method which is generally based on a vocabulary can not give 100% of correct segmentation because of the ambiguities and performs worst when the out-of-vocabulary rate is high. This makes text data processing for word n-gram language modeling complicated and other modeling units must be investigated. Note that a text in Khmer could be segmented into words, syllables and character-cluster (group of inseparable characters). The character-cluster could be a potential modeling unit as its segmentation is trivial because of its non-ambiguities structure.

In this paper we present an overview on the development of a large vocabulary continuous speech recognition (LVCSR) system for Khmer. We first describe our methodologies and tools for data collection and quick development of a new ASR system for under-resourced language. In section 3, we discuss the use of word and sub-word units in statistical language models for Khmer. Our objective is to investigate the potential of sub-word units for unsegmented language. We present in section 4 the process of automatic generation of grapheme based and phoneme based Khmer pronunciation dictionaries for acoustic modeling. The experimental framework and the recognition results are presented in section 5. Section 6 concludes the work and gives some future perspectives.

2. Language data acquisition

2.1. Text Corpus

Creating a statistical language model consists in estimating from a text corpus the probability of word n-gram. A large amount of in domain text data (several hundred millions words) is needed in order to obtain accurate probability estimation. As our system is targeted to automatic broadcast news transcription, the classic way to get in domain text data is to take content from newspapers. Method for text collection from the web is becoming more and more popular as the web allows obtaining freely and quickly a large quantity of text. Recently, several research works proposed techniques to exploit the resources from the web for natural language processing. In [1], a web robot that retrieves text from the Internet to build a text corpus is proposed. From some given starting points on the web, the robot can reach and retrieve recursively text documents and html pages. However, we must control the robot in order to get only the text in the target domain and language. Another approach in [2] consists in estimating words n-gram probabilities using the World Wide Web. The probabilities are estimated from the number of pages found using a given search engine. Those kinds of methods applied well to languages which have already a significant coverage on the Internet. For an

under-resourced language like Khmer, the number of websites and the speed of Internet connections are often limited. There are only around 340.000 websites registered in Cambodian domain name *.kh* (results from Google) and most of them propose contents in English instead of Khmer.

In our case, retrieving the Khmer pages from some well selected news websites allows us to get big quantity of text more rapidly than using a robot to crawl many sites on the net as proposed in [1]. Once html pages are retrieved, further processing is needed in order to build a text corpus:

- Filtering in order to extract only text from html pages
- Converting legacy character encoding to standard Unicode encoding
- Segmenting text into sentences and word or sub-word units using automatic segmentation tools
- Converting special signs and numbers to text
- Normalizing the words

By using the *ClipsTextTK* [3], this process could be done rapidly by adapting the language dependent part of the toolkit for Khmer language. We developed tools for the conversion of encoding (from legacy ad-hoc code to Unicode), the conversion of special characters and numbers to text. For automatic segmentation, we developed segmentation tools to segment text into sentences, words and character-cluster (see example in table 1). The word segmentation tool is developed using an algorithm which segments a text into words based on a list of vocabulary of 18000 words obtained from the official Khmer dictionary (*Chhoun Nat* dictionary) with an optimization criteria: *longest matching*. This word segmentation tool, estimated on some held-out data, gives 95% of correct word segmentation. Syllable and character cluster segmentation is done using rules created with linguistic knowledge. The syllable segmentation is not trivial and gives only 85% of performance while the character-cluster segmentation is 100% correct because of its non-ambiguities atomic structure. A complete *ClipsTextTK-Khmer* version is added to the toolkit.

Sentence	ព្រះពុទ្ធជាព្រះបរមគ្រូនៃយើង											
Word	ព្រះពុទ្ធ			ជា	ព្រះបរម				គ្រូ	នៃ	យើង	
Syllable	ព្រះ		ពុទ្ធ	ជា	ព្រះ	ប		រម	គ្រូ	នៃ	យើ	ង
CC	ព្រ	ះ	ពុ ទ	ជ	ា	ព្	ร	ម	គ្	ូ	យ	ើ
Translation	Buddha is our supreme teacher											

Table 1: Example of Segmentation of Khmer text

The collection of Khmer text from the Internet during 3 months (from December 2006 to February 2007) allows us to get 25130 (448Mb) *html* pages from 5 selected daily news news websites. A text corpus of 15,5 millions words (249Mb) is obtained after applying the *ClipsTextTK* toolkit. In these 15.5 millions words, 15 % of OOV words are found compared to the original 18,000 vocabulary. A non negligible part of the OOV words is probably due to the word segmentation errors, while another part corresponds to real OOV words.

2.2. Speech Corpus

To train the acoustic models for our system, a speech corpus is needed. Speech corpus can be created by recording a well prepared text read by professional readers in a studio. The recording task is however very time and resource consuming as we need to prepare the text data and scenarios and run the recording process. To obtain speech signal quickly and freely, we tried several techniques. The first consists in searching the websites that propose the radio broadcast news in Khmer language. Many organizations such as Voice of America, Radio Australia and Radio FreeAsia have broadcast program in Khmer language and put on their website the entire broadcast news for public download. Most of the time, the transcripts are also available. From those sites, we can retrieve quickly a big quantity of speech signal but with a poor quality (narrowband) because the signal is compressed. In order to obtain a good quality speech signal, with help from our partner, Institut de Technologie du Cambodge (ITC) in Cambodia, we built a recording system from basic equipments: a computer with a radio receiver card installed, a recording program that we scheduled to record several hours of broadcast news of different radio stations in Phnom Penh, Cambodia. From this operation, we got recordings of 30h of good quality speech signal of radio broadcast news in Khmer language. A manual transcription campaign of the recording speech signals was organized at ITC. Twenty volunteers (students at ITC) who were motivated to contribute to the development of the language resources for Khmer were recruited and trained to do the manual transcription. By using *Transcriber* [4], 6h30mn of speech signal were manually transcribed in Unicode Khmer script (only speech read in the studio was transcribed and without extra detailed information).

This 6h30mn of transcription contains 3200 phrases of 45200 words pronounced by 8 different speakers (3 women). 172 phrases (25mn of speech signal) are then extracted to serve as test corpus during the evaluation of our ASR system.

3. Language modeling

The statistical nature of the approaches used in language modeling requires a large quantity of text corpus in order to make accurate word probability estimation. Word, which is often defined as a sequence of characters separated by space, is traditionally the basic unit of modeling and work well for languages like English and French. For languages which have a very rich morphology where prefixes and suffixes augment word stems to form words and for the languages without explicit word boundary, traditional word definition is not appropriate and leads to a high out-of-vocabulary (OOV)

rate. In this case, language models must be estimated from error-prone word segmentation. In addition, when the text data available is limited, it will lead to poor estimates of the language model probabilities, and hence may hurt ASR performance. This is typically the case for languages like Khmer. Alternatively, we can make language model estimation at sub-word level (syllable or character). This has potentially bad consequences on the word coverage of the n-gram models but it allows more accurate probability estimation because sub-word vocabulary is smaller than word vocabulary.

Some previous works using sub-word units for language modeling have recently been published for Arabic, Turkish (morphological analysis). Data-driven or fully unsupervised [5] word decomposition algorithms were used like in [6, 7] as well as working on the character level for unsegmented languages like in [8]. For character-based language like Chinese, mixed vocabularies containing both characters and a set of frequent words (mostly 2 characters words) are used in language modeling [9].

Note that a text in Khmer could be segmented into word, syllables or character-cluster. The character-cluster could be a good modeling unit as its segmentation is trivial because of its non-ambiguities atomic structure. We proposed to use these word and sub-word units in language modeling for Khmer. We will discuss the performance of these different LMs in the experimentation section.

4. Acoustic modeling

4.1. Automatic pronunciation generation

The pronunciation dictionary provides the link between sequences of acoustic units and words as represented in the language model. Whereas text and speech corpora can be collected, pronunciation dictionary is generally not directly available. While a manually generated pronunciation dictionary gives a good quality ASR, this task is time consuming and requires extended knowledge on the acoustic and phonology systems of the language in question. There were several techniques found in the literature for generating a pronunciation dictionary. Among them we can mention [10] which proposed a modeling technique based on pronunciation rules. This method requires knowledge on the target language and also of its phonetic rules. Grapheme based modeling has been successfully addressed for different languages [11, 12]. It has the advantage of being straightforward and fully automatic. For Khmer language, a grapheme based dictionary is generated by converting the Khmer character in its Unicode representation to its Unicode name in Roman representation. There are totally 77 graphemes in modern Khmer alphabet: 33 consonants symbols, 16 dependents vowels symbols, 16 independent vowels and 12 diacritics and signs. In Unicode Standard 5.0 code charts, each Khmer code has a name in Roman. The following table shows the correspondence between a Khmer word and its grapheme based pronunciation model.

Khmer word	Grapheme based pronunciation
ចចក	Ca Ca Ka
ចតុមុខ	Ca Ta U Mo U Kha
ក្រាមដី	Ka COENG Ro OO Mo Da II

Table 2: Example of Khmer grapheme based dictionary

4.2. Acoustic modeling

The most basic acoustic modeling for Khmer language is grapheme based modeling. We used our grapheme based dictionary which has 77 modeling units. We used SphinxTrain [13] toolkit from Sphinx project for building Hidden Markov Models (HMMs) acoustic models. With our speech training database described previously, we train acoustic models based on grapheme. Context-independent (CI) and context-dependent (CD with 1000 tied states) models are both considered. We obtained 2 acoustic models : Grapheme_CI and Grapheme_CD.

5. Experiments

5.1. ASR System

Sphinx3 decoder [13] is used to in our experiments. The model topology is a HMM with 8 Gaussian mixtures per state. The pre-processing of the system consists of extracting a 39 dimensional feature vector of 13 MFCCs, the first and second derivatives.

Our text corpus collected from the web is first segmented into words and we extract 20,000 most frequent words to be used as the test vocabulary. This word vocabulary and the corresponding LM training corpus are also segmented in syllables (8,800 syllables vocabulary) and character-cluster (3500 character-clusters vocabulary). The transcript from speech corpus is also used for language models as it is from the same broadcast news source as the test corpus. The language models used in our experiments are obtained by linear interpolation of web corpus LM and the speech corpus transcript LM. A development corpus is used to tune the interpolation parameters.

In addition to Word Error Rate (WER) measure, we use Syllable Error Rate (SER) and Character Cluster Error Rate (CCER) for evaluation. Since Khmer word and syllable segmentation is not a trivial task and segmentation errors may prevent a fair comparison of different ASR hypotheses, we believe that comparing the ASR hypotheses at the character-cluster level gives a more accurate idea of the relative performance of different systems. The test is run on our test corpus which contains 172 utterances (around 20mn of speech signal).

5.2. Word and Sub-word language models

In this experiment, we train 3 trigram LMs by using respectively word, syllable and character-cluster as modeling units.

LM	Acoustic model	WER	SER	CCER
LMword	Grapheme_CI	54.1	37.4	31.8
LMsyl	Grapheme_CI	-	46.1	39.2
LMcc	Grapheme_CI	-	-	45.9
LMword	Grapheme_CD	47.8	26.9	**20.8**
LMsyl	Grapheme_CD	-	44.3	25.0
LMcc	Grapheme_CD	-	-	32.3

Table 3: Test results of word and sub-word units LMs

From the results in table 6, we can see that word unit is still the best unit of modeling. A Khmer word is in average composed of 3.2 syllables and 4.3 characters clusters. Thus, the effective span of a trigram model of syllables or character-cluster is shorter than that of a word trigram model, which leads to less accurate prediction. The advantage of these sub-word units may remain when there are high OOV rates in the test corpus.

The results show the potential of grapheme based acoustic modeling when a phoneme-based pronunciation dictionary is not available. We can see that the context-dependent acoustic model performs better than the context-independent model in our case.

The big difference between WER and CCER is partly due to the word segmentation errors that occur in output hypotheses and in the reference. This suggests that CCER gives a more accurate evaluation and should be used if we want to compare different systems using different segmentation units.

6. Conclusion

We presented in this paper an overview of the development of a new large vocabulary continuous speech recognition system for Khmer language. As an under-resourced language, building a LVCSR system for Khmer has to deal with several challenging tasks such as the lack of language resources and the text segmentation problem. We proposed methodologies for rapid language data collection and processing based mainly on *Opensource* tools to build a speech recognition system for Khmer language. Our second contribution is a publicly available manual of development of a speech recognition system for a new under-resource language [14]. Our current work consists in improving our ASR system for Khmer by trying to exploit the combination of word and sub-word units in statistical language modeling.

7. References

[1] Vaufreydaz, D. (2002) Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue. *Thèse de doctorat de l'Université J. Fourier - Grenoble I*, France.

[2] Zhu, X., Rosenfeld, R. (2001) Improving Trigram Language Modelling with the World Wide Web. In *Proc. ICASSP*, pages 533-536, Salt Lake, USA.

[3] www-clips.imag.fr/geod/User/brigitte.bigi/

[4] Barras, C., et al. (2000) Transcriber: development and use of a tool for assisting speech corpora production. In *Speech Communication* Vol 33, No 1-2.

[5] Kurimo, M., et al. (2006) Unsupervised segmentation of words into morphemes - Morpho Challenge 2005: Application to Automatic Speech Recognition. In *Proc. Interspeech*, pages 1021-1024, Pittsburgh, PA.

[6] Abdillahi, N., et al. (2006) Automatic transcription of Somali language. In *Proc. Interspeech*, pages 289-292, Pittsburgh.

[7] Afify, M., et al. (2006) On the use of morphological analysis for dialectal Arabic Speech Recognition. In *Proc. Interspeech* pages 277-280, Pittsburgh, PA.

[8] Denoual, E., Lepage, Y. (2006) The character as an appropriate unit of processing for non-segmenting languages. *NLP Annual Meeting*, pages 731-734, Tokyo, Japan.

[9] Chen, L., et al. (2000) "Broadcast News Transcription in Mandarin," Proc. ICSLP'2000, Beijing, China.

[10] Huang, X., et al. (2001) Spoken Language Processing – A Guide to Theory, Algorithm, and System Development, *Prentice Hall*.

[11] Billa, J., et al. (2002) Audio indexing of Arabic broadcast news. In *Proc. IEEE International Conference on Acoustique, Speech and Signal Processing*. Pages 5-8, Orlando.

[12] Bisani, M., Ney., H. (2003) Multigram-based grapheme-to-phoneme conversion for LVCSR. In *Proc. EUROSPEECH*. Pages 933-936 Geneva, Switzerland.

[13] http://cmusphinx.sourceforge.net/html/cmusphinx.php

[14] http://www-clips.imag.fr/geod/User/sopheap.seng/asr/