A novel Post-editing algorithm to optimize Amharic speech recognition for speech translation

Michael Melese Woldeyohannis[1], Laurent Besacier[2], Million Meshesha[1]
[1]Addis Ababa University, Addis Ababa, Ethiopia
[2]LIG Laboratory, UJF, BP53, 38041 Grenoble Cedex 9, France
[michael.melese@aau.edu.et, laurent.besacier@imag.fr, million.meshesha@aau.edu.et]

Abstract

In this study an attempt is made to optimize the output of Amharic ASR using n-gram based post-edit algorithm without human intervention. Experimental result shows that the integration of post-editing algorithm with the ASR improves the accuracy of recognition. The post-edit algorithm on Amharic ASR using corpus based n-gram approach resulted an absolute improvement by 1.42% from the word based recognition accuracy. Now we are working towards further improving propagated errors and connecting the optimized ASR to Amharic-English speech translation.

1 Introduction

Computers that can understand natural language can facilitate communication between the people who speak different languages (Honda, 2003). Among this communication, speech is one of the fundamental for humankind. Speech translation is the process by which spoken source phrases are translated to a target language using a computer (Gao et al., 2006). The state-of-the-art of speech translation system can be seen as the integration of three major cascading components (Gao et al., 2006; Jurafsky and Martin, 2008); ASR, MT and TTS. Speech translation research for major and technological supported languages has been conducted since the 1983s by NEC Corporation (Kurematsu, 1996). On the contrary, attempts for under-resourced languages, like Amharic, not yet particularly started due to unavailability of speech and text corpora (Melese et al., 2016). However, a number of attempts have been made for Amharic ASR using different techniques, data and tool which is now inaccessibility. Thus, the main aim of this study is to investigate the possibility to design Amharic-English speech translation system that controls ASR errors propagation using post-edit algorithm.

2 Background

Amharic is a Semitic language derived from Ge'ez with the second largest speaker in the world next to Arabic (Thompson, 2016). The language is used as official working language of government of Ethiopia. Unlike other Semitic languages, such as Arabic and Hebrew, modern Amharic script has inherited its writing system from Ge'ez using a grapheme based writing system called fidel(Yimam, 2000). Tourism become a pleasing sustainable economic development serving as an alternative source of foreign exchange for the counties like Ethiopia (UNWTO, 2016). Majority of the tourists can speak and communicate in English to exchange information about tourist attractions. Due to this, language barriers are a major problem for today's global communication. As a result, they look for an alternate option that lets them communicate with others using speech translation technology (Nakamura, 2009). This is especially true for under-resourced language like Amharic.

The development of an accurate, efficient and robust speech translation system has a lot of challenges beside the challenges of cascading components; such as ASR, SMT and TTS (Nakamura, 2009). These include, conversational speaking style, openness of domain, scarcity of data and morphological variation which results in less recognition accuracy as part of the Amharic-English speech translation. Moreover, the error generated in ASR further propagate to the succeeding machine translation component which results in low performance beside the component integration and lack of standard speech and text corpora. On the other hand, a morpheme LM with a phonemic AM based recognition leads to better recognition accuracy for the Amharic language (Melese et al., 2016). Thus, we need an optimized recognition accuracy of recognition using post-edit algorithm.

Speech and text corpora are the fundamental resources for speech translation; Collecting and preparing corpora is a challenging and expensive task which require innovative data collection methodologies (Simons and Fennig, 2017). Due to this, parallel English-Arabic text was acquired from BTEC 2009 which is made available through IWSLT (Kessler, 2010). Once the Amharic-English BTEC corpus is prepared, it is divided into training, tuning and testing set with a proportion of 69.33% (19472 sentences), 1.78%(500 sentences) and 28.88%(8112 sentences), respectively. Then, LIG-Aikuma (Blachon et al., 2016), a smartphone application was used to record a 7.43 hours of Amharic speech from 8112 sentences. While for

training, a 20 hours of read speech corpus prepared by Abate et al. (2005) from 10,875 sentences. Table 1 presents the training, development and language model data used for Amharic speech recognition.

	Train	Test	Language Model	
			Word	Morpheme
Sentence	10,875	8,112	261,620	261,620
Token	145,404	50,906	4,223,835	5,773,282
Type	24,653	4,035	328,615	141,851

Table 1: Distribution of Amharic speech data.

	Unit		Train	Dev	Test
Amharic	Word	Sentence	19,472	500	8,112
		Token	107,049	2,795	37,288
		Type	18,650	1,470	4,168
	Morpheme	Sentence	19,472	500	8,112
		Token	145,419	3,828	50,906
		Type	15,679	1,621	4,035
English	Word	Sentence	19,472	500	8,112
		Token	157,550	4,024	55,062
		Type	10,544	1,227	3,775

Table 2: Distribution of Amharic-English SMT data.

In addition to this, the Amharic language model (LM) data collected for Google project (Tachbelie and Abate, 2015) have been used beside the separate BTEC data excluding the test set. Consequently, corpus based and language independent segmentation have been applied on a training, development and test set of Amharic SMT data. Morfessor is used to segment words to a sub word units to solve the unknown word problems to via morphological segmentation. Table 2 presents summary of the corpus used for Amharic-English SMT using word and morpheme as a unit. Similarly, a corpus containing 681,910 sentences (11,514,557 tokens of 582,150 types) data crawled from web including news and magazine for a corpus based n-gram post-edit algorithm. From this data, a total of 5,057,112 bigram, 8,341,966 trigram, 9,276,600 quadrigram and 9,242,670 pentagram word sequences have been extracted after preprocessing.

3 Methodology

In this study we propose a post-editing module for Amharic ASR that can detect the kinds of errors described in the next paragraph, identify plausible alternatives to the errors, and finally, choose the most plausible to correct the output. The correction is made using n-gram data store using minimum edit distance and perplexity before the error heads to SMT. The first phase of post editing is to detect the error from ASR recognition output. Basically, to detect an error, recognized morpheme units are concatenated to form a word and its existence is checked in unigram Amharic dictionary. Thus, a morpheme-based ASR outputs are concatenated to form a word based sentence. Then the sentence are tokenized in to sequence of words. If the word is not in the unigram Amharic dictionary, then the "word" is considered as an error and concatenated to the remaining words. If the error is detected during the first phase, then the correction proposal phase takes the sentence with error mark and creates (w-n+1)[1] n-grams after adding start and end symbol. Subsequently, we select the n-grams with error marks and search in n-gram data store to select possible candidates for correction after removing the error mark. If there is no candidate in n-gram, then go for (n-1)-gram order until bigram.

Once the candidate identified, the suggestion is made taking the minimum edit distance between the error detected and suggestion selected. In this phase, the sum of maximum edit distance has been set experimentally to 16. The edit distance 16 was selected to provide at least one suggestion per sentence and minimize the computation of perplexity. Finally, the suggestion is made primarily using minimum edit distance then by calculating the perplexity. The minimal edit distance is computed between the word considered as an error and n-gram based possible suggested word in a sentence. If the edit distance is the same for different suggestion, then the decision is made by selecting the good perplexity.

4 Result and discussion

In ASR experiments, Kaldi (Povey et al., 2011), SRILM (Stolcke et al., 2002) and Morfessor (Smit et al., 2014) have been used for ASR, LM and unsupervised segmentation, respectively. The entire ASR experiment is conducted using a morpheme-based LM with phoneme-based AM. Accordingly, the Amharic ASR experiment shows a 76.4% accuracy for the morpheme-based. Then, after the concatenation of morphemes to words, a 77.4% word-based recognition accuracy have been achieved.The morpheme based recognition followed by post-edit resulted in 78.5% WRA. The result obtained from the n-gram post-edit experiment shows an absolute advance by 1.42% from word recognition accuracy of 77.4% by concatenating a 76.4% morpheme based recognition. The current study proves the possibility of enhancing the performance of ASR by controlling speech recognition error using post-editing algorithm. Further works need to be done to integrate post-editing based ASR with SMT as part of Amharic-English speech translation beside the preparation of linguistic motivated pronunciation dictionary for speech translation.

[1] w is number of token in sentence and n specifies n-grams. Otherwise, the sentence is considered as correct.

References

Solomon Teferra Abate, Wolfgang Menzel, Bairu Tafila, et al. 2005. An amharic speech corpus for large vocabulary continuous speech recognition. In *INTERSPEECH*, pages 1601–1604.

Tadesse Anberbir and Tomio Takara. 2009. Development of an amharic text-to-speech system using cepstral method. In *Proceedings of the First Workshop on Language Technologies for African Languages*, pages 46–52. Association for Computational Linguistics.

Laurent Besacier, V-B Le, Christian Boitet, and Vincent Berment. 2006. Asr and translation for under-resourced languages. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 5, pages V–V. IEEE.

David Blachon, Elodie Gauthier, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker, and Annie Rialland. 2016. Parallel speech collection for under-resourced language studies using the lig-aikuma mobile device app. *Procedia Computer Science*, 81:61–66.

Yuqing Gao, Liang Gu, Bowen Zhou, Ruhi Sarikaya, Mohamed Afify, Hong-Kwang Kuo, Wei-zhong Zhu, Yonggang Deng, Charles Prosser, Wei Zhang, et al. 2006. Ibm mastor system: Multilingual automatic speech-to-speech translator. In *Proceedings of the Workshop on Medical Speech Translation*, pages 53–56. Association for Computational Linguistics.

Elodie Gauthier, Laurent Besacier, Sylvie Voisin, Michael Melese, and Uriel Pascal Elingui. 2016. Collecting resources in sub-saharan african languages for automatic speech recognition: a case study of wolof. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*

Masaaki Honda. 2003. Human speech production mechanisms. *NTT Technical Review*, 1(2):24–29.

Daniel Jurafsky and James H Martin. 2008. *Speech and language processing (prentice hall series in artificial intelligence)*. Prentice Hall.

Fondazione Bruno Kessler. 2010. A generic weaver for supporting product lines. In *International Workshop on Spoken Language Translation*, pages 11–18. ACM.

Akira Kurematsu. 1996. *Automatic Speech Translation*, volume 28. CRC Press.

Michael Melese, Laurent Besacier, and Million Meshesha. 2016. Amharic speech recognition for speech translation. In *Atelier Traitement Automatique des Langues Africaines (TALAF). JEP-TALN 2016*.

Michael Melese, Laurent Besacier, and Million Meshesha. 2017. Amharic-english speech translation in tourism domain. In *Proceedings of the Workshop on Speech-Centric Natural Language Processing, EMNLP*, pages 59–66.

Satoshi Nakamura. 2009. Overcoming the language barrier with speech translation technology. *Science & Technology Trends-Quarterly Review*, (31).

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, EPFL-CONF-192584. IEEE Signal Processing Society.

Gary F. Simons and Charles D. Fennig. 2017. *Ethnologue: Languages of the World*. SIL, Dallas, Texas.

Peter Smit, Sami Virpioja, Stig-Arne Gronroos, and Mikko Kurimo. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 21–24. European Chapter of the Association for Computational Linguistics, EACL.

Andreas Stolcke et al. 2002. Srilm-an extensible language modeling toolkit. Interspeech.

Martha Yifiru Tachbelie and Solomon Teferra Abate. 2015. Effect of language resources on automatic speech recognition for amharic. In *AFRICON, 2015*, pages 1–5. IEEE.

Mulu Gebreegziabher Teshome, Laurent Besacier, Girma Taye, and Dereje Teferi. 2015. Phoneme-based english-amharic statistical machine translation. In *AFRICON, 2015*, pages 1–5. IEEE.

Irene Thompson. 2016. About world language. Accessed: 2017-05-26.

UNWTO. 2016. World tourism organization annual report 2015. Technical report, United Nation, Madrid, Spain.

Baye Yimam. 2000. *Yeamarigna sewasew (Amharic version)*. EMPDA, Addis Ababa, Ethiopia.