Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 35 (2016) 234–261

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Word segmentation and pronunciation extraction from phoneme sequences through cross-lingual word-to-phoneme alignment☆

Felix Stahlberg[a,*], Tim Schlippe[a], Stephan Vogel[b], Tanja Schultz[a]

[a] *Cognitive Systems Lab, Karlsruhe Institute of Technology (KIT), Germany*
[b] *Qatar Computing Research Institute, Qatar Foundation, Qatar*

Received 3 December 2013; received in revised form 5 October 2014; accepted 11 October 2014
Available online 18 October 2014

Abstract

In this paper, we study methods to discover words and extract their pronunciations from audio data for non-written and under-resourced languages. We examine the potential and the challenges of pronunciation extraction from phoneme sequences through cross-lingual word-to-phoneme alignment. In our scenario a human translator produces utterances in the (non-written) target language from prompts in a resource-rich source language. We add the resource-rich source language prompts to help the word discovery and pronunciation extraction process. By aligning the source language words to the target language phonemes, we segment the phoneme sequences into word-like chunks. The resulting chunks are interpreted as putative word pronunciations but are very prone to alignment and phoneme recognition errors. Thus we suggest our alignment model *Model 3P* that is particularly designed for cross-lingual word-to-phoneme alignment. We present two different methods (source word dependent and independent clustering) that extract word pronunciations from word-to-phoneme alignments and compare them. We show that both methods compensate for phoneme recognition and alignment errors. We also extract a parallel corpus consisting of 15 different translations in 10 languages from the Christian Bible to evaluate our alignment model and error recovery methods. For example, based on noisy target language phoneme sequences with 45.1% errors, we build a dictionary for an English Bible with a Spanish Bible translation with 4.5% OOV rate, where 64% of the extracted pronunciations contain no more than one wrong phoneme. Finally, we use the extracted pronunciations in an automatic speech recognition system for the target language and report promising word error rates – given that pronunciation dictionary and language model are learned completely unsupervised and no written form for the target language is required for our approach.

Keywords: Pronunciation dictionary; Non-written languages; Lexical language discovery; Under-resourced languages; Speech-to-speech translation; Word segmentation

☆ This paper has been recommended for acceptance by R.K. Moore.

* Corresponding author. Tel.: +49 721 608 46312.
E-mail address: fstahlberg@gmail.com (F. Stahlberg).

http://dx.doi.org/10.1016/j.csl.2014.10.001
0885-2308/

1. Introduction

1.1. Motivation

We explore cross-lingual information to discover the vocabulary of an non-written target language. We align source language words to target language phoneme sequences across languages, i.e. cross-lingually. Based on this alignment, we induce phoneme sequences forming words of the target language. The resulting pronunciations are used in automatic speech recognition (ASR) systems. Our research is inspired by the following challenges that language technology faces nowadays:

- There are over 7100 living languages and dialects in the world (Gordon and Grimes, 2014). ASR and Machine Translation (MT) systems exist only for few of them due to the large amount of monolingual or respectively bilingual data which is necessary to train such systems. For example, transcribed speech resources, large amounts of text for language modeling, pronunciation dictionaries, and parallel sentence-aligned text corpora are of great importance to create speech processing systems. Schultz and Kirchhoff (2006, p. 37) estimate that thorough transcription of 1 h conversational speech data normally takes about 20 h of effort. Undoubtedly, creating sentence-aligned parallel corpora required by MT tools is very time consuming as well. Therefore, a major challenge of language technologies in our globalized world is to overcome this dependence on large amounts of training data or to significantly reduce the costs of data collection efforts. Our approach only requires written sentences in a source language, spoken translations of them in the target language and a phoneme recognizer which can recognize the target language phonemes.
- A lot of the world's languages and dialects do not have an acknowledged written form (Schultz and Kirchhoff, 2006; Nettle and Romaine, 2000) despite their widespread use for oral communication (for example reported for Arabic dialects by Chiang et al. (2005)). Nowadays, language technology generally requires a written script. Our approach bypasses the written form, since it operates on the phoneme level on the target language side.
- Rapid language adaptation (for instance investigated by Vu et al. (2010)) deals with the challenge of rapidly porting ASR systems to new languages and domains. These techniques are suitable in situations where languages with few linguistic resources suddenly appear in the focus of interest due to various reasons – for example Haitian Creole in the scope of international relief operations for victims of the Haiti earthquake in 2010. Results of this work might be applicable to such a situation for developing ASR and MT systems in the future since the only resources required in the target language are spoken translations which can be generated in near real-time.
- Human simultaneous translations are common in the real world, for instance in the international press or multilingual parliaments. If we have access to a speech recognition system that transcribes the source language and generates the required written translations, such parallel speech (Paulik and Waibel, 2013) can be used for our approach.

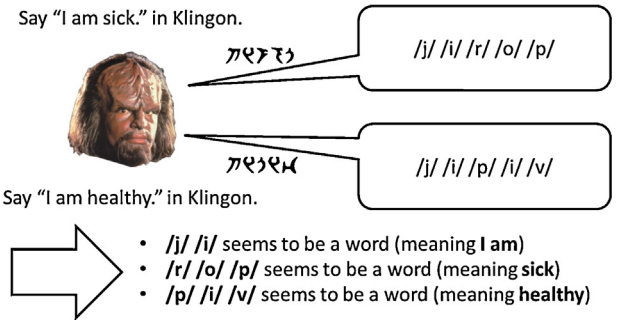

Fig. 1. Basic scenario for English-Klingon. Just by the use of English sentences in written form and their spoken translations, a human English speaker is able to locate possible word boundaries in Klingon.