Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 56 (2014) 101–118

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Web-based tools and methods for rapid pronunciation dictionary creation

Tim Schlippe*, Sebastian Ochs, Tanja Schultz

Institute for Anthropomatics, Cognitive Systems Lab (CSL), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Available online 16 July 2013

Abstract

In this paper we study the potential as well as the challenges of using the World Wide Web as a seed for the rapid generation of pronunciation dictionaries in new languages. In particular, we describe *Wiktionary*, a community-driven resource of pronunciations in IPA notation, which is available in many different languages. First, we analyze *Wiktionary* in terms of language and vocabulary coverage and compare it in terms of quality and coverage with another source of pronunciation dictionaries in multiple languages (*GlobalPhone*). Second, we investigate the performance of statistical grapheme-to-phoneme models in ten different languages and measure the model performance for these languages over the amount of training data. The results show that for the studied languages about 15k phone tokens are sufficient to train stable grapheme-to-phoneme models. Third, we create grapheme-to-phoneme models for ten languages using both the *GlobalPhone* and the *Wiktionary* resources. The resulting pronunciation dictionaries are carefully evaluated along several quality checks, i.e. in terms of consistency, complexity, model confidence, grapheme n-gram coverage, and phoneme perplexity. Fourth, as a crucial prerequisite for a fully automated process of dictionary generation, we implement and evaluate methods to automatically remove flawed and inconsistent pronunciations from dictionaries. Last but not least, speech recognition experiments in six languages evaluate the usefulness of the dictionaries in terms of word error rates. Our results indicate that the web resources of *Wiktionary* can be successfully leveraged to fully automatically create pronunciation dictionaries in new languages.

Keywords: Web-derived pronunciations; Pronunciation modeling; Rapid bootstrapping; Multilingual speech recognition

1. Introduction

1.1. Motivation

With some 6,900 languages in the world, data resources such as text files, transcribed speech or pronunciation dictionaries are only available in the most economically viable languages. Over the past years, the World Wide Web has been increasingly used as a text data source for rapid adaptation of ASR systems to new languages and domains at low cost, e.g. websites are crawled to collect texts that are used to build language models. Moreover, prompts which might be read by native speakers to receive transcribed audio data, are extracted from the crawled text (Schultz et al., 2007). The creation of pronunciation dictionaries can be time consuming and expensive if they are manually produced by language experts. The dictionaries provide the mapping from the orthographic form of a word to its pronunciation, which is useful in both text-to-speech and automatic speech recognition (ASR) systems. They are used to train speech processing systems by describing the pronunciation of words according to manageable units, typically phonemes (Martirosian et al., 2007). Our intention is to research if web-derived pronunciations can be used to build pronunciation dictionaries from scratch or at least enhance existing ones. If pronunciations can be extracted together with the corresponding written words from the World

[*] Corresponding author. Address: Institute for Anthropomatics, Cognitive Systems Lab (CSL), Karlsruhe Institute of Technology (KIT), Adenauerring 4, 76131 Karlsruhe, Germany. Tel.: +49 721 608 4 6303; fax: +49 721 608 4 6116.

E-mail address: tim.schlippe@kit.edu (T. Schlippe).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.06.015

Wide Web, the following scenario comes within reach: Text from the ASR system's target domain is crawled and a text normalization is performed automatically. The vocabulary of the normalized crawled text is extracted, and on the basis of pronunciations from the World Wide Web, a pronunciation dictionary with that vocabulary is created. Subsequently, a language model is built on the collected text. Thereby, dictionary and language model in a new domain or language are generated without manual effort.

Our *Rapid Language Adaptation Toolkit (RLAT)* with its web-based interface is an ongoing effort towards that goal. It aims to reduce the human effort involved in building speech processing systems for new languages and domains. Innovative tools enable novice and expert users to develop speech processing models, such as acoustic models, pronunciation dictionaries, and languages models, to collect appropriate speech and text data for building these models, and to evaluate the results (Vu et al., 2010). We extended RLAT to extract pronunciations from the World Wide Web and collected pronunciations from *Wiktionary*. *Wiktionary* is a wiki-based open content dictionary, available in many languages and checked by a big community frequently and carefully. It contains pronunciations written in the International Phonetic Alphabet (IPA). The IPA, devised by the International Phonetic Association, is a standardized representation of the sounds of spoken language (IPA, 1999). Due to the fast growth in language presence on *Wiktionary* which is shown in Section 2, we see a future potential of gathering pronunciations for languages from this source that are still underrepresented.

1.2. Related work

In this section, we present methods of other researchers to generate pronunciation dictionaries, to check the quality of pronunciations and to treat erroneous dictionary entries.

In the field of speech processing, the World Wide Web has been used as a data source for improving the language model probability estimation as well as for obtaining additional training material (Zhu et al., 2001). Furthermore, several approaches to automatic dictionary generation have been introduced. Black et al. (1998) apply grapheme-to-phoneme (g2p) rules for dictionary production. Dictionaries built with these methods are often post-edited by human experts to further raise their quality: Suspicious entries are automatically or manually flagged, then examined and corrected manually. In Kominek et al. (2006), the *Lexicon Learner* asks the user, who does not have to be a language expert, to provide pronunciations for displayed words. Each word is accompanied by a suggested pronunciation, along with a synthesized audio file. The prediction is based on grapheme-to-phoneme rules that the system infers from the user's answers and which are updated after each additional word. The rules are seeded during an initialization stage in which the Lexicon Learner asks the user for the phoneme most commonly associated with each grapheme. A similar dictionary creation process that combines machine learning with minimal human intervention was proposed by Davel et al. (2004). Initial investigations to leverage off pronunciations from the World Wide Web have been described (Ghoshal et al., 2009; Can et al., 2009; Schlippe et al., 2010; Schlippe et al., 2012a; Ghoshal et al., 2009) retrieve English pronunciations in IPA and ad-hoc transcriptions from the World Wide Web and compare the pronunciations to the Pronlex dictionary[1] with phoneme error rates. To extract, validate and normalize the web-derived pronunciations, they apply English unigram grapheme-to-phoneme rules, grapheme-to-grapheme rules and phoneme-to-phoneme rules learned from the Pronlex dictionary. In Can et al. (2009), they apply their methods for English spoken term detection. Our goal in Schlippe et al. (2010), Schlippe et al. (2012a) and in this paper is to analyze a multilingual online database such as *Wiktionary* which may be a source for pronunciations of under-resourced languages as well. Our focus is also on the quantity and quality of the pronunciations which includes the impact of the web-derived pronunciations on ASR performance. To enhance quality, we analyze data-driven methods to validate and normalize them without the need of reference dictionaries in the target language. Additionally, we use the web-derived word-pronunciation pairs to investigate their suitability for g2p conversion and their generalization ability for multiple languages.

For g2p conversion, different methods have been proposed. Knowledge-based approaches with rule-based conversion systems were developed which can typically be expressed as finite-state automata (Kaplan et al., 1994; Black et al., 1998). Often, these methods require specific linguistic skills and exception rules formulated by human experts. In contrast to knowledge-based approaches, data-driven approaches are based on the idea that, given enough examples, it should be possible to predict the pronunciation of unseen words purely by analogy. The benefit of the data-driven approach is that it trades the time- and cost-consuming task of designing rules, which requires linguistic knowledge, for the much simpler one of providing example pronunciations. Besling et al. (1994) propose data-driven approaches with heuristical and statistical methods. In Kneser et al. (2000), the alignment between graphemes and phonemes is generated using a variant of the Baum–Welch expectation maximization algorithm. Chen et al. (2003), Vozila et al. (2003) and Jiampojamarn et al. (2007) use a joint-sequence model to the g2p task. Novak et al. (2011) and Novak et al. (2012) utilize weighted finite-state transducers for decoding as a representation of the joint-sequence model. Gerosa et al. (2009), Laurent et al. (2009) and Karanasou et al. (2010), apply statistical machine translation-based methods for the g2p conversion. We use Sequitur G2P, a data-driven g2p converter developed at RWTH Aachen University that works with joint-sequence models (Bisani and Ney, 2008) which we explain in detail in Section 6.2. A good overview of state-of the art g2p methods is given in (Hahn et al., 2012).

[1] CALLHOME American English Lexicon, LDC97L20.