Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 28 (2014) 256–277

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Normalization of informal text

Deana L. Pennell *, Yang Liu

The University of Texas at Dallas, 800 West Campbell Road, Richardson, TX 75080, United States

Received 17 August 2011; received in revised form 2 March 2013; accepted 12 July 2013
Available online 23 July 2013

Abstract

This paper describes a noisy-channel approach for the normalization of informal text, such as that found in emails, chat rooms, and SMS messages. In particular, we introduce two character-level methods for the abbreviation modeling aspect of the noisy channel model: a statistical classifier using language-based features to decide whether a character is likely to be removed from a word, and a character-level machine translation model. A two-phase approach is used; in the first stage the possible candidates are generated using the selected abbreviation model and in the second stage we choose the best candidate by decoding using a language model. Overall we find that this approach works well and is on par with current research in the field.
Published by Elsevier Ltd.

Keywords: Text normalization; Noisy text; NLP applications

1. Introduction

Text messaging is a rapidly growing form of alternative communication for cell phones. This popularity has caused safety concerns leading many US states to pass laws prohibiting texting while driving. The technology is also difficult for users with visual impairments or physical handicaps to use. We believe a text-to-speech (TTS) system for cell phones can decrease these problems to promote safe travel and ease of use for all. Normalization is the usual first step for TTS.

SMS lingo is similar to the chatspeak that is prolific on forums, blogs and chatrooms. Screen readers will thus benefit from such technology, enabling visually impaired users to take part in internet culture. In addition, normalizing informal text is important for tasks such as information retrieval, summarization, and keyword, topic, sentiment and emotion detection, which are currently receiving a lot of attention for informal domains.

Normalization of informal text is complicated by the large number of abbreviations used. Some previous work on this problem used phrase-based machine translation (MT) for abbreviation normalization; however, a large annotated corpus is required for such a method since the learning is performed at the word level. By definition, this method cannot make a hypothesis for an abbreviation it did not see in training. This is a serious limitation in a domain where new words are created frequently and irregularly.

This work is an extension of our work in Pennell and Liu (2010, 2011, 2011). In this paper, we establish two sets of baseline results for this problem on our data set. The first uses a language model for decoding without use of an

* Corresponding author. Tel.: +1 469 585 5182.
E-mail addresses: deana@hlt.utdallas.edu (D.L. Pennell), yangl@hlt.utdallas.edu (Y. Liu).

0885-2308/$ – see front matter. Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.csl.2013.07.001

Table 1
Methods for processing unseen tokens during normalization.

Method	Formal example	Informal example
as chars	RSVP	"cu" (see you)
as word	NATO	"l8r" (later)
expand	Corp.	"prof" (professor)
combine	WinNT	"neway" (anyway)

abbreviation model, while the second utilizes a state-of-the art spell checking module, Jazzy Idzelis (2005). We then compare the use of our two abbreviation models for decoding informal text sentences. We also determine the effects on decoding accuracy when more or less context is available. Finally, we combine the two systems in various ways and demonstrate that a combined model performs better than both systems individually.

2. Related work

This section briefly describes relevant work in fields directly related to our research, though not always directly applied to informal text. We describe the tasks of modeling and expanding abbreviations in text as well as research on normalization of text in both formal and informal domains.

2.1. Abbreviation modeling and expansion

Abbreviation expansion is a common problem when processing text from any domain. Willis et al. (2002) studied abbreviation generation in the hopes of lowering the effort of text entry for people with motor disabilities; the user could enter abbreviated text that would be expanded by a system to be read by his or her conversation partner. They asked a group of young people to abbreviate a text of 500 characters to progressively smaller lengths, assuming they are charged per letter but there is a hefty fee for every error in decoding by another person. Although they do not attempt to expand the abbreviations automatically, they produce a set of rules by which participants produced abbreviations, using both deletion and substitution.

Early work by Pakhomov (2002) showed that medical abbreviations can be modeled and expanded using maximum entropy modeling. He used contextual information to help disambiguate medical terms assuming that an abbreviation and its correct expansion will be found in similar contexts.

Wong et al. (2006) introduced their ISSAC system that works on top of a spell checker (Aspell) to simultaneously perform spelling correction, abbreviation expansion and capitalization restoration. Their model gives weight to the normalized edit distance, domain significance, number of hits for a word on Google, appearance of the word/abbreviation pair (WAP) in an online abbreviation dictionary and the original weight given to a suggested correction by Aspell. In addition, a word is given more weight if it has been seen paired with the current abbreviation earlier in the document. Their reranking scheme provides a significant increase in accuracy over Aspell alone.

Yang et al. (2009) work with abbreviations for spoken Chinese rather than for English text messages, but their process is quite similar to our CRF system. They first perform an abbreviation generation task for words and then reverse the mapping in a look-up table. They use conditional random fields as a binary classifier to determine the probability of removing a Chinese character to form an abbreviation. They rerank the resulting abbreviations by using a length prior modeled from their training data and co-occurrence of the original word and generated abbreviation using web search.

2.2. Text normalization

Abbreviation expansion is just one of many techniques needed for the task of text normalization. Text normalization is an important first step for any text-to-speech (TTS) system. Regardless of the size of the training corpus, there will always be tokens that do not appear and have unknown pronunciations. Text normalization has been widely studied in many formal domains. Sproat et al. (2001) provides a good resource for text normalization and its associated problems.