

A Unified Approach to Transliteration-based Text Input with Online Spelling Correction

Hisami Suzuki

Jianfeng Gao

Microsoft Research

One Microsoft Way, Redmond WA 98052 USA

{hisamis, jfgao}@microsoft.com

Abstract

This paper presents an integrated, end-to-end approach to online spelling correction for text input. Online spelling correction refers to the spelling correction as you type, as opposed to post-editing. The online scenario is particularly important for languages that routinely use transliteration-based text input methods, such as Chinese and Japanese, because the desired target characters cannot be input at all unless they are in the list of candidates provided by an input method, and spelling errors prevent them from appearing in the list. For example, a user might type *suesheng* by mistake to mean *xuesheng* 学生 'student' in Chinese; existing input methods fail to convert this misspelled input to the desired target Chinese characters. In this paper, we propose a unified approach to the problem of spelling correction and transliteration-based character conversion using an approach inspired by the phrase-based statistical machine translation framework. At the phrase (substring) level, *k* most probable pinyin (Romanized Chinese) corrections are generated using a monotone decoder; at the sentence level, input pinyin strings are directly transliterated into target Chinese characters by a decoder using a log-linear model that refer to the features of both levels. A new method of automatically deriving parallel training data from user keystroke logs is also presented. Experiments on Chinese pinyin conversion show that our integrated method reduces the character error rate by 20% (from 8.9% to 7.12%) over the previous state-of-the-art based on a noisy channel model.

1 Introduction

This paper addresses the problem of online spelling correction, which tries to correct users' misspellings as they type, rather than post-editing them after they have already been input. This online scenario is particularly important for languages that routinely use transliteration-based text input methods, including Chinese and Japanese: in these languages, characters (called *hanzi* in Chinese and *kanji/kana* in Japanese) are typically input by typing how they are pronounced in Roman alphabet (called *pinyin* in Chinese, *romaji* in Japanese), and selecting a conversion candidate among those that are offered by an input method system, often referred to as IMEs or input method editors. One big challenge posed by spelling mistakes is that they prevent the desired candidates from appearing as conversion candidates, as in Figure 1: *suesheng* is likely to be a spelling error of *xuesheng* 学生 'student', but it is not included as one of the candidates.

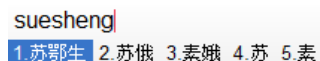


Figure 1: Spelling mistake prevents the desired output (学生) from appearing in the list of candidates

This severely limits the utility of an IME, as spelling errors are extremely common. Speakers of a non-standard dialect and non-native speakers have a particularly hard time, because they may not know the standard pronunciation of the word to begin with, preventing them from inputting the word altogether. Error-tolerant word completion and next word prediction are also highly desirable features for text input on software (onscreen) keyboards for any language, making the current work relevant beyond Chinese and Japanese.

In this paper, we propose a novel, unified system of text input with spelling correction, using