賽德克語構詞結構之自動解析

Analyzing the Morphological Structures in Seediq Words

林川傑 Chuan-Jie Lin†

國立臺灣海洋大學資訊工程學系

Department of Computer Science and Engineering

National Taiwan Ocean University

cjlin@email.ntou.edu.tw

宋麗梅 Li-May Sung†

國立臺灣大學語言學研究所

Graduate Institute of Linguistics

National Taiwan University

limay@ntu.edu.tw

游景勝 Jing-Sheng You, 王瑋 Wei Wang, 李政勳 Cheng-Hsun Lee,

廖子權 Zih-Cyuan Liao

國立臺灣海洋大學資訊工程學系

Department of Computer Science and Engineering

National Taiwan Ocean University

{10857039, 00657120, 00657140, 00672042}@email.ntou.edu.tw

摘要

原住民族語言保存及振興的問題已日益受到重視。如果能開發出原住民族語相關自然語言處理技術，有助於原住民語資料保存及族語推廣等工作。賽德克語的詞形變化相當多樣，其中一大部分主要是為了標示動詞焦點或時貌，包括完成貌、主事焦點、受事焦點、處所焦點等等。因為這種焦點系統為南島語系所特有，若要研究台灣原住民語和中文之間的自動翻譯系統，辨別這類構詞資訊非常重要。

† 通訊作者 corresponding authors

一個賽德克詞的構詞結構會以它所參考的原形詞加上前、中、後綴的組合來呈現，然而構詞結構資訊無法由詞面直接獲得，詞典中也僅能查到各賽德克詞參考的原形詞。更特別的是，賽德克語構詞律有元音脫落規則，因此賽德克詞並非直接由原形詞接上詞綴而得。因此本論文的主要目標是自動解析賽德克語的構詞結構，在給定一個賽德克詞及其參考原形詞時，能夠解析出該賽德克詞裡所出現的前綴、中綴及後綴組合。

此外，賽德克語構詞律有元音中性化和詞尾輔音變化等等規則。在研究構詞情形的過程中我們發現，加上後綴時原形詞部份會恢復回變化之前的樣子，我們將之定義為「深層原形」。因為字典中並無此項資訊，本論文也會探討如何猜測一個賽德克詞的深層原形。實驗資料主要來自宋麗梅教授著作「賽德克語語法概論」及協助原住民族委員會開發的「賽德克語德固達雅方言」線上詞典。

首先由語法書中整理出的構詞相關知識來撰寫規則，用以偵測詞綴的出現。深層原形則是利用詞典中參考同一原形詞的不同賽德克詞來統計猜測，這些猜測結果又可再歸納出常見的變化規則用來推測新詞彙的深層原形。詞綴及深層原形解析工作在測試資料的精確率是 98.66%，而召回率是 88.29%。

至於前綴的部份，因為同一個前綴字串可能可以拆解出多種不同的前綴組合而產生歧異情形，因此改以機器學習方式進行。在測試各種方法後，效果最好的是以基本前綴為單位的二元機率模型。解決零機率的方法是降階至一元機率模型 (權重設為 α)，而一元機率模型解決零機率的方法又以 Lidstone smoothing 效能最好 (頻率增加值設為 λ)。前綴組合最佳解析正確率為 76.92%。

Abstract

The issue of preservation and revitalization of the indigenous languages is gaining attention from the public in recent days. Developing NLP techniques related to the indigenous languages will help to preserve and promote these languages. Word inflection or morphological forms in Seediq are plentiful. Major categories of the inflections are mainly for representing the focus or aspect, such as perfective aspect, active voice, patient voice, locative voice, etc. The focus system of the Austronesian languages is quite different from Chinese. It is important to identify the information of focus or aspect in words if we want to study machine translation among Taiwanese indigenous languages and other languages.

The morphological structure of a Seediq word consists of its word root, prefixes, infixes, and suffixes. This kind of information cannot be obtained directly from the surface of a Seediq

word. Dictionaries only offer the information of word roots. Furthermore, due to the rule of vowel reduction in Seediq, the surface of a Seediq word is not the same as the concatenation of affixes and word root. This paper focuses on automatically analyzing the morphological structure of a Seediq word given its word root.

Moreover, there are also rules of vowel neutralization and final consonant variation. During the research, we found that a word root would return to its original form when combining with the suffixes. We define the original form of a root word as a "deep root". Since there is no information about deep roots in the dictionary, this paper also proposes methods to predict deep roots of Seediq words. The experimental data come from the works of Prof. Li-May Sung: the grammar book "賽德克語語法概論" (An Introduction to Seediq Grammar) and the online dictionary "賽德克語德固達雅方言" (Tgdaya Seediq) from the Council of Indigenous Peoples.

First, several morphological analyzing rules were created from the knowledge provided in the grammar book. These rules were used to detect the occurrences of affixes. Deep roots were learned from the set of different words referencing to the same root words. The mapping of root words with their deep roots could be further used to derive deep-root-prediction rules for unknown words. The rule-based system successfully detected the deep root and the existence of affixes with a precision of 98.66% and a recall of 88.29% on the test data.

Because one prefix string can be divided into several different structures, we used machine learning methods to solve the ambiguity. The best system was developed by bigram model where grams were atomic prefixes. Zero probability in the bigram model was replaced by the unigram probability (weighted by α), where the unigram model was also smoothed by the Lidstone smoothing method (with an addition of λ to the frequencies). The best prefix analysis system achieved an accuracy of 76.92% on the test data.

關鍵詞：賽德克語，構詞結構自動解析，深層原形，臺灣原住民族語之自然語言處理

Keywords: Seediq, automatic analysis of morphological structures, deep root, natural language processing for Taiwanese indigenous languages

致謝

本論文之研究承蒙行政院科技部研究計畫 (MOST 109-2221-E-019 -053 -) 之經費支持，謹此致謝。