Chapter 6 Conclusion and Future Work

6.1 Our Contributions to Written Taiwanese Resources and Processing

We have described the tasks we have performed in written Taiwanese related research.

For digital written Taiwanese resources, the important infrastructure, we have established:

(a) A 22,000 entry online Taiwanese syllable dictionary (OTSD). This had a total of more than 290,000 searches from more than 32,000 different IP addresses (as of January 2003), with more than 250 searches per day for the past year (as of December 30, 2008) (Iunn, 2003c, 2003f);

(b) A 62,000 entry online Taiwanese-Mandarin dictionary (OTMD). This had a total of more than 2.4 million searches from more than 125,000 different IP addresses (as of December 2002), with more than 2,700 searches per day for the past year (as of December 30, 2008); we also developed a Google gadget interface for the OTMD (Iunn, 2000, 2002, 2003g, 2007c);

(c) A Taiwanese corpus with 5,800,000 syllables in HR mixed script and 3,400,000 syllables in POJ script, and the Online Taiwanese Concordancer System (OTCS) based on this corpus. This had a total of nearly 1,900,000 searches from more than 56,400 different IP addresses (as of January 2003), with about 1,630 searches per day for the past year (as of December 30, 2008) (C.-C. Cheng et al., 2007; Iunn, 2003b, 2003e; Iunn & Lau, 2007);

(d) A Preliminary Taiwanese Word Frequency Report for the Taiwanese POJ and HR mixed scripts based on the above Taiwanese corpus (Iunn, 2005b, 2005c);

(e) A 2,580,000-word Digital Archive Database for Written Taiwanese (2nd stage) (DADWT), which contains literature data with POJ and HR mixed scripts paragraph alignment. This had a total of more than 1,320,000 page visits (as of December 2006), with 1,672 page visits per day on average. We also developed a Google gadget interface for the DADWT, which can randomly select an article (Iunn, 2006a, 2007b, 2007d);

For the coding and I/O of POJ, we proposed a two-stage search method via string matching and a filter program. We also proposed a query expansion scheme for toneless, glottal stop, checked syllable, and vowel searches, and described a display method. The problems mentioned above are quite different from other languages, such as English and Mandarin. The Taiwanese syllable query expansion is an important achievement since no other systems fully provide these functions. We also provide the first online Taiwanese word

segmentation system.

In relation to the processing techniques, we translated every word into Mandarin via the OTMD, obtained the POS information from the CED made by the CKIP group, proposed a rule-based tone sandhi algorithm to solve the Taiwanese tone sandhi problem, and implemented an online text-to-speech system to read out the Taiwanese literature data for users (Iunn, 2006a). We achieved accuracy rates of 97.4% and 89.0% for the training and test data, respectively. These accuracy rates are higher than other research results so far.

We also proposed a statistics-based POS tagging method using the OTMD and 10-million-word Mandarin training database to tag the Taiwanese. We followed the tagset drawn up by CKIP, did the POJ script and HR mixed script word alignment work, searched the OTMD to find corresponding Mandarin candidate words, selected the most adequate Mandarin word using an HMM probabilistic model from the Mandarin training data, and tagged the word using an MEMM classifier. We achieved an accuracy rate of 91.5% in this work. It is difficult to make a comparison with other research since the tagsets are different.

Since we have been performing this pioneering work in written Taiwanese related research, other researchers and graduate/PhD students have contacted us to get Taiwanese texts in order to do related search. Most of them have used the Taiwanese corpus and DADWT, including (K.-i. Chan, 2008; Niu, 2004). Yi-fen Huang, a PhD student at CMU[1], put part of the DADWT data into the CMU

[1] Carnegie Mellon University.

SPICE[2] system to include Taiwanese text-to-speech and speech recognition. Hong-tin Teng, an assistant professor in the Department of Taiwanese Languages and Literature of National Taichung University, utilized the contents of our website to teach her students. In addition, the SMHLA project intends to use the method we have proposed to perform the POS tagging task for Hakka.

Other researchers have not contacted us, but did their research using the OTCS or DADWT, including (Chang, 2007; S.-L. Chen, 2006; Y.-F. Cheng, 2007; Liao, 2008). In the "Workshop on The Use of Language Corpora of Taiwan '台灣語言語料庫使用工作坊,' " held on February 2 and 3, 2007, at National United University, Ying Cheng introduced Taiwan Southern Min research using the OTCS (MOE Advisory Office, 2007)

At times we have received email from strangers asking us to fix our website when it had problems. Additionally, an undergraduate studying Taiwanese languages once told us that they could not finish their homework when our website was out of service[3].

6.2 Future Work and Prospects for Written Taiwanese Processing Research

It might be possible to improve the Taiwanese tone sandhi problem in the following ways:

[2] Speech Processing - Interactive Creation and Evaluation.

[3] Personal communication in July 2008.

(a) Solicit assistance from linguists. It is hoped that linguistics will define a standard for part-of-speech analysis and word segmentation, and that a dictionary conforming to such a standard will be built.

(b) Improve word segmentation, especially the processing of morphology, quantitative words, and proper nouns.

(c) Improve the processing of POS tags to account for ambiguity.

(d) Change the dictionary's POS tags, such as by making use of Embree's POS analysis (Embree, 1984).

(e) Improve the sandhi rules.

(f) Find alternative ways of modeling sandhi processing such as template theory or optimality theory.

(g) Use a machine learning method to model tone sandhi processing if we can construct a corpus with tone sandhi markers.

In relation to the Taiwanese POS tagging, if we could construct a Taiwanese Mandarin parallel corpus, we could then use other methods, like the Coerced Markov Models proposed by (Fung & Wu, 1995), to do the Taiwanese POS tagging task.

Based on the results mentioned above, we will try to describe the blueprint for our future written Taiwanese processing research.

First, in relation to the infrastructure, we need to amend the original data and prepare the following items:

(a) A suitable tagset for the Taiwanese language;

(b) An electronic word dictionary based on a word segmentation standard.

Second, we need to establish a Taiwanese corpus with syntactic tags. We also need to establish a Taiwanese corpus with semantic tags, and the discussion of the semantic role of Taiwanese is necessary. This suggested corpus should provide both POJ and HR mixed script transcription and tone sandhi markers.

Finally, we can construct a Taiwanese treebank.

If we submit the Taiwanese corpus to the Linguistic Data Consortium (LDC), the status of this language will be promoted (UPenn, 1992).

In addition, speech and text conversion techniques, and an OCR technique (from images to text, like a Google Book search (Google Inc., 2007)) for Taiwanese are also important.

On the other hand, if we have already established the Taiwanese corpora, at least at the 10-million-word level for example, it will be possible for us to develop the field of Taiwanese applied linguistics via computational research. We think that some of the issues worth investigating include:

(a) Zipf's law:

Zipf's law states that, given a large corpus of natural language, where the words are listed in descending order of frequency, with *f* the frequency of a word and *r* its rank, then $f \propto \frac{1}{r}$. Mandelbrot derived a more general relationship between rank and frequency: $f = P(r+\rho)^{-B}$, where P, B, and ρ are the parameters of a text. These parameters are different for different languages. What are the parameters of Taiwanese (Manning & Schütze,

1999)?

(b) Lexicography:

"Collins COBUILD Learner's Dictionary" is the first dictionary to use the computational corpus-based approach. They use word frequency data to select words to include in the dictionary, with contextualized examples from the corpus. There are also other corpus-based dictionaries, like "The New Oxford Dictionary of English," "The Oxford-Hachette French Dictionary," *etc*. Can we establish a Taiwanese dictionary via corpus?

(c) Lexical change:

Every language changes day by day, but many people believe that the Taiwanese language is changing more rapidly than other languages, mainly under the influence of Japanese and Mandarin, because of political or historical factors. If we attain a diachronic Taiwanese corpus with data from different time periods, we can get more precise quantitative data to describe this phenomenon (Iunn & Kao, 2004; Khu, 2008; Li, 2000; McEnery, Xiao, & Tono, 2006).

(d) Script selection for HR mixed script:

Though the written Taiwanese orthography has not yet been standardized, the specific written form is gradually being accepted by some through common practice. We think the mainstream written Taiwanese orthography is HR mixed script. What percentage of the POJ in the HR mixed script comes from the word types/tokens' point of view? Why does a writer select POJ or a Han character? Are there different selection attitudes in different

genres? The Taiwanese corpus may give us more satisfactory answers (K.-i. Chan, 2008).

(e) Co-occurrence of words:

Word usage is an important factor in language learning. It is necessary for us to establish Taiwanese collocation data via the Taiwanese corpus.

(f) Machine translation:

Taiwan is a multi-ethnic and multi-lingual society, whose languages interact with each other frequently. It is necessary to develop language translation systems, such as Taiwanese/Mandarin, Taiwanese/Hakka, Taiwanese/Aboriginal languages, *etc*.

On the other hand, Taiwanese/English and Taiwanese/Japanese translations are also important when we want to communicate with the international community.

Housewives are also a societal reality in Taiwan, and translation between Taiwanese and Southeast Asian languages is becoming increasingly important. However, we think that this will be difficult to realize in the near future due to a lack of resources.

Written Taiwanese processing and Taiwanese computational linguistics are nearly uncultivated fields, and need many researchers.