SkELL: A Discovery-Based Chinese Learning Platform

Simon Smith (Coventry University, UK)

This research will:

- Extend a corpus-based language learning platform to Chinese.
- Enable an exploratory/discovery-based learning approach for Chinese learning
- Conduct grammatical analysis of Chinese for natural language processing

Mandarin Chinese has the largest number of native speakers of all languages. Many people are starting to learn Chinese, and more resources enabling its learning are needed. Already, there is a wide range of apps and web learning platforms, targeting in particular the mastery of Chinese characters (Shen, 2005). Other software packages focus on the acquisition of vocabulary, for example through the use of online flashcards (Edge, Searle, Chiu, Zhao, & Landay, 2011). Various online dictionaries are available, some offering advanced features.

What we have not seen so far is applications which make use of or are based on authentic text data, even in online dictionaries. As was once more often the case with TESOL materials, dictionaries give example sentences which are either invented, or drawn from literary sources that do not reflect modern usage. Corpora are not widely used by Chinese lexicographers and materials developers (Li & Smith, 2015).

One particularly powerful corpus-based tool, currently available for English and Russian learning, is SkELL (Sketch Engine for Language Learning; Kilgarriff, Marcowitz, Smith, & Thomas, 2015). SkELL is a user-friendly system which allows learners to explore the behaviour of words in context, presenting a variety of example sentences, and showing how words participate in collocations and interact grammatically with other words. We will extend SkELL to the Chinese language.

With SkELL, language learners are able to consult three types of word usage display, based on large corpora. They are Example Sentences, which uses the GDEX algorithm (Kilgarriff, Husák, McAdam, Rundell, & Rychlý, 2008) to find the best dictionary examples from a corpus; Wordsketch, which gives a single-screen summary of a word’s usage, showing which other words it typically collocates with, and in what grammatical relations; finally there is Similar Words, which is a distributional thesaurus. SkELL is based on the architecture of Sketch Engine, a general purpose corpus analysis tool which is not specifically configured for language learning, but which does offer access to several large Chinese corpora. Most of these, in common with corpora of other languages available on Sketch Engine, have been segmented (divided into words) and POS-tagged. In order to provide the Wordsketch and Similar Words features, a sentence parsing function is also required, and this takes the form of a set of language-specific grammatical relations rules which are executed on the fly (when the user requests the feature). The rules are specified as regular expressions over parts of speech.

For our implementation of SkELL, we will evaluate a range of segmenting and POS tagging options, including the tools of Academia Sinica (Ma & Chen, 2005; Tseng & Chen, 2002), City University Hong Kong (Sun, Shen, & Tsou, 1998) and Stanford University (Toutanova, Klein, Manning, & Singer, 2003; Tseng, Chang, Andrew, Jurafsky, & Manning, 2005). We will also optimize the grammatical relations rules, to take better account of the relatively free word order of Chinese (including for example topicalization and object-fronting) than does the ruleset currently implemented on Sketch Engine.

The poster will set out plans for the SkELL implementation, and will incorporate a small-scale evaluation of the currently available Chinese Wordsketch function, exploring ways in which the grammatical relation ruleset might be optimized.

References

Edge, D., Searle, E., Chiu, K., Zhao, J., & Landay, J. A. (2011). MicroMandarin: mobile language learning in context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3169-3178). ACM.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings Euralex*.

Kilgarriff, A., Marcowitz, F., Smith, S., & Thomas, J. (2015). Corpora and language learning with the Sketch Engine and SKELL. *Revue française de linguistique appliquée*, *20*(1), 61-80.

Li, W. & Smith, S. (2015). Introduction. In Zou, B., S. Smith & M. Hoey (Eds.), *Corpus Linguistics in Chinese Contexts*. Basingstoke: Palgrave.

Ma, W. Y. & Chen, K. J. (2005). Design of CKIP Chinese word segmentation system. *Chinese and Oriental Languages Information Processing Society*, *14*(3), 235-249.

Shen, H. H. (2005). An investigation of Chinese-character learning strategies among non-native speakers of Chinese. *System*, *33*(1), 49-68.

Sun, M., Shen, D., & Tsou, B. K. (1998). Chinese word segmentation without using lexicon and hand-crafted training data. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2* (pp. 1265-1271). Association for Computational Linguistics.

Toutanova, K., Klein, D., Manning, C. & Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003* (pp. 252-259).

Tseng, H., & Chen, K. J. (2002). Design of Chinese morphological analyzer. In *Proceedings of the first SIGHAN workshop on Chinese language processing* (pp. 1-7). Association for Computational Linguistics.

Tseng, H., Chang, P., Andrew, G., Jurafsky, D., & Manning, C. (2005). A conditional random field word segmenter. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing* (pp. 168-171).