8 Chinese segmentation and collocation: a platform for blended learning

Simon Smith[1]

Abstract

Mandarin Chinese is an increasingly popular world language and object of study, and while there are numerous online character learning apps and flashcard systems, very little research has been done on inductive or autonomous learning in the realm of collocation acquisition. I propose a new Chinese implementation of a trusted corpus-based platform, currently available for English and several other languages, accompanied and enhanced by an adaptive approach to Chinese segmentation approach, whereby different ways of carving up a given sentence are selectively displayed to the learner.

Keywords: Chinese, Mandarin, CALL, blended learning, segmentation.

1. Introduction and background

Mandarin Chinese has the largest number of native speakers of all languages (Simons & Fennig, 2018). More and more people are starting to learn Chinese, in the UK and globally. Centres of Chinese cultural exchange, such as Confucius Institutes, are opening up, and increasing numbers of universities, and more recently schools as well, are starting to offer Mandarin Chinese programmes. It is fast becoming a global language, and more tools and resources enabling its learning are needed. Lo (2016), for example, shows that heritage Chinese students from the UK who are native speakers of English and Cantonese find

1. Coventry University, Coventry, United Kingdom; simon.smith@coventry.ac.uk

How to cite this chapter: Smith, S. (2018). Chinese segmentation and collocation: a platform for blended learning. In M. Orsini-Jones & S. Smith (Eds), *Flipping the blend through MOOCs, MALL and OIL – new directions in CALL* (pp. 59-65). Research-publishing.net. https://doi.org/10.14705/rpnet.2018.23.791

the mastery of Mandarin important for their careers. A renewed commitment to trade partnership with China has been noted by the UK government (Gibb & Johnson, 2015), and one likely impact of Brexit is that increased trade with China will turn command of Chinese into a yet more marketable skill.

Mandarin Chinese is popular among students of all ages, and Chen (2014) reports success in teaching the language at primary school in Britain. Secondary school uptake has increased in recent years, with General Certificate of Secondary Education (GCSE) entries up 18% in 2015, despite an overall decline in pupils sitting language exams (Guardian, 2015). Jin (2014) discusses the opportunities for enhancing students' intercultural competencies that the learning of Mandarin Chinese affords.

Modern approaches to language learning, especially English learning, emphasise the use of authentic texts (Gilmore, 2007; Nunan, 1999), so that vocabulary and patterns may be acquired by learners in genuine contexts. Such authentic texts may be conveniently gathered together in a *corpus* – that is to say, a 'body of texts', defined more explicitly by McEnery, Xiao, and Tono (2006) as "a collection of (1) *machine-readable* (2) *authentic* texts […] which is (3) *sampled* to be (4) *representative* of a particular language or language variety" (their emphasis, p. 5). The potential of *inductive* (as opposed to *deductive*) learning, where learners look at data to try to establish systematic rules, rather than being taught the rules explicitly, is now widely accepted in educational circles (Dörnyei, 2014; Larsen-Freeman & Long, 2014). Johns (1991) made what was then an innovative use of inductive learning, in an approach named Data-Driven Learning (DDL), which entails getting students to consult corpora directly. With DDL, learners can search for particular lexical and grammatical patterns that interest them. They can be trained to adopt an inductive or discovery-based approach to learning, where they work out a grammatical rule or pattern of usage from a plethora of authentic examples, as opposed to a deductive and more traditional approach where the teacher lays out rules, words, and patterns and gets the learner to practise them. There has been increasing interest in DDL in language teaching circles over the years, and the approach lends itself well to blended learning, which by definition involves autonomous study.

Almost all this work has so far focused on English learning; although Chinese is widely spoken and studied, and while there is no shortage of flashcard apps and character-practice software, there have been very few attempts to harness the power of the corpus for this language. One particularly powerful corpus-based tool, currently available for English and several other languages, is Sketch Engine for Language Learning (SkELL) (Baisa & Suchomel, 2014).

On SkELL, Chinese students can obtain three kinds of output displays about the usage of words, derived from large corpora. The first display is Example Sentences, which finds the most salient dictionary-like examples from the corpus. Then there is Word Sketch, which offers a one-page synopsis of the usage of a word, indicating for example which collocations it is most associated with, and what the grammatical relations are (e.g. what is the most salient object of this verb, or most salient modifier of this noun). The third display type is called Similar Words, a distributional thesaurus. SkELL is powered by the Sketch Engine (SkE; Baisa & Suchomel, 2014), a corpus query software suite which does not specifically target language learners, but which does allow access to a number of large Chinese corpora. Most of these have been segmented (broken up into words) and POS-tagged (Smith, 2017).

2. Methodology and approach

In this work, SkELL is extended to the Chinese language, allowing learners to view vocabulary in authentic collocational contexts, presenting a variety of example sentences, and showing how words participate in collocations and interact grammatically with other words. The implementation incorporates a standalone adaptive segmentation system using Hidden Markov Model (HMM) technology, and it will be evaluated using the training and test corpora of the first Chinese Segmentation Bakeoff of the Association for Computational Linguistics Chinese special interest group (Sproat & Emerson, 2003). A learner-friendly interface will be designed, and its use piloted with a group of intermediate Chinese learners who will be asked to evaluate its usefulness (in particular its adaptive features). The study will experiment with different ways of presenting

the varying granularity of segmentation to the learner, aiming to provide for ease of use in a blended learning context.

A particular challenge for learners of Chinese, an addition to the obvious complexity of the characters themselves, is the identification of word boundaries. The Chinese SkELL implementation will therefore incorporate a new adaptive segmentation system, which is described below.

3. Chinese vocabulary and segmentation

In the absence of clear orthographic information about word boundaries, as is available in English writing, it can be quite difficult to get even human informants to agree on where the word boundaries are in a Chinese sentence. It follows from that that it is quite difficult to write segmentation software to do the same task.

Early segmentation algorithms consisted of a dictionary search module supplemented by heuristics, typically a longest match (or maximum match) procedure (Deng & Long, 1987). This means that if several different ways of segmenting the sentence are potentially available, the way which includes the longest words will be selected.

The next phase of segmentation algorithms made use of statistical information: notably Mutual Information (MI) scores in the work of Sproat and Shih (1990), and Sun, Shen, and Tsou (1998), without the use of dictionaries. The segmenter currently in use by Baidu, the main Chinese search engine, exploits HMM technology, and offers the user of their so-called 'Jieba' segmentation software the option of adding in their own custom dictionary (Lin, 2015).

This study confronts and exploits the 'wordhood' challenge. It offers an adaptive segmentation approach, where different ways of carving up a given sentence are selectively displayed to the learner.

Wu (2003, p. 3) demonstrates how such an approach can benefit the different applications of Chinese Natural Language Processing (NLP): Machine Translation (MT), for example, generally needs the longest strings that are available in the bilingual lexicon being used, so a maximum matching algorithm is the most useful. For information retrieval, on the other hand, a user (such as a search engine user) might be interested in webpages that contain substrings of the string they entered, so a fine-grained segmentation might be more appropriate. I believe that, just as the varying granularities can be applied to different NLP applications, so too they can usefully address different language learning purposes in DDL.

For example, the string 中华人民共和国 is the official title of the People's Republic of China. This could be treated as one word, or segmented into two (中华人民 / 共和国) or three words (中华/人民/共和国). Alternatively, the learner is likely to be interested in the individual characters as morphemes, and finding out what other characters they pattern with. In a blended learning context, where guidance from the teacher is not always at hand, the student will be able to set the parameters for his or her own learning.

4. Conclusion and next steps

It was noted above that making different segmentation granularities available could benefit learners of Chinese. The adaptive segmenter described by Wu (2003) and Gao, Li, Wu, and Huang (2005) allows for several different levels of segmentation, within a "single annotated corpus that can be conveniently customised to meet different segmentation requirements" (Wu, 2003, p. 2).

Gao et al. (2005), with Wu as a co-author, implemented a similar system called MSRseg (Microsoft Research Segmenter), using transformation based learning (Brill, 1995). This is still available as a free download from Microsoft Research (although minus the adaptive component which is of particular relevance to this work). Gao et al. (2005) note that in actuality they retain only the segmentation

that involves the smallest number of words, because "we currently do not know any effective way of using multiple segmentations in [NLP] applications" (p. 541).

Adaptive segmentation does not appear to have been revisited in the literature since, and there has not been any attempt that I am aware of to integrate such a segmentation model into language learning. I therefore consider our proposal to be innovative, practical, and timely.

References

Baisa, V., & Suchomel, V. (2014). SkELL – Web interface for English language learning. In *Eighth Workshop on Recent Advances in Slavonic Natural Language Processing* (pp. 63-70). Tribun EU.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Computational Linguistics*, *21*(4), 543-565.

Chen, T. (2014). Teaching Chinese as a foreign language at primary school in England. *Quarterly Journal of Chinese Studies*, *2*(4), 67-83.

Deng, Q., & Long, Z. (1987). A microcomputer retrieval system realising automatic information indexing in Chinese. *Journal of Information Science*, *6*, 427-432 (in Chinese).

Dörnyei, Z. (2014). *The psychology of the language learner: individual differences in second language acquisition*. Routledge.

Gao, J., Li, M., Wu, A., & Huang, C. N. (2005). Chinese word segmentation and named entity recognition: a pragmatic approach. *Computational Linguistics*, *31*(4), 531-574. https://doi.org/10.1162/089120105775299177

Gibb, N., & Johnson, J. (2015). *Press release: UK-China education partnership reaches new heights*. Department for Business, Innovation & Skills https://www.gov.uk/government/news/uk-china-education-partnership-reaches-new-heights

Gilmore, A. (2007). Authentic materials and authenticity in foreign language learning. *Language Teaching*, *40*, 97-118. https://doi.org/10.1017/S0261444807004144

Guardian. (2015). *GCSE results: fall in numbers taking foreign languages 'a cause for concern'*. https://www.theguardian.com/education/2015/aug/20/gcse-results-fall-numbers-foreign-languages

Jin, T. (2014). Getting to know you: the development of intercultural competence as an essential element in learning Mandarin. *London Review of Education*, *12*(1), 20-33. https://doi.org/10.18546/LRE.12.1.04

Johns, T. (1991). Should you be persuaded: two examples of data-driven learning. In T. Johns & P. King (Eds), *Classroom concordancing* (pp. 1-16). English Language Research.

Larsen-Freeman, D., & Long, M. H. (2014). *An introduction to second language acquisition research*. Routledge.

Lin, F. (2015). JIEBA 結巴中文斷詞. https://speakerdeck.com/fukuball/jieba-jie-ba-zhong-wen-duan-ci

Lo, L. (2016). Challenges faced by Cantonese speakers in a UK university Mandarin course. In C. Goria, O. Speicher, & S. Stollhans (Eds), *Innovative language teaching and learning at university: enhancing participation and collaboration* (pp. 139-145). Research-publishing.net. https://doi.org/10.14705/rpnet.2016.000415

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: an advanced resource book*. Routledge

Nunan, D. (1999). *Second language teaching and learning*. Heinle & Heinle.

Simons, G., & Fennig, C. (Eds). (2018). *Ethnologue: languages of the world*. SIL International. http://www.ethnologue.com

Smith, S. (2017). SkELL: A discovery-based Chinese learning platform. In *Corpus Linguistics International Conference Abstracts*. http://paulslals.org.uk/ccr/CL2017ExtendedAbstracts.pdf

Sproat, R., & Emerson, T. (2003). The first international Chinese word segmentation bakeoff. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing* (pp. 133-143). SIGHAN. https://doi.org/10.3115/1119250.1119269

Sproat, R., & Shih, C. (1990). A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages, 4*(4), 336-51.

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.

Wu, A. (2003). Customizable segmentation of morphologically derived words in Chinese. *International Journal of Computational Linguistics and Chinese Language Processing, 8*(1), 1-27.

Published by Research-publishing.net, a not-for-profit association
Voillans, France, info@research-publishing.net

Flipping the blend through MOOCs, MALL and OIL – new directions in CALL
Edited by Marina Orsini-Jones and Simon Smith

Typeset by Research-publishing.net
Cover design by © Raphaël Savina (raphael@savina.net)
Cover illustration © Marina Orsini-Jones

ISBN13: 978-2-490057-16-0 (Ebook, PDF, colour)
ISBN13: 978-2-490057-17-7 (Ebook, EPUB, colour)
ISBN13: 978-2-490057-15-3 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method; with which the book is never 'out of stock' or 'out of print'.

British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, UK: British Library.
Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: juin 2018.