A finite-state approach to Chinese word segmentation for rule-based machine translation

Blake Perry Smith · perry.smithb@gmail.com · perry952 · Robert Reynolds · Digital Humanities

Project Purpose

Chinese in its written form, whether typed or penned, does not separate its characters by spaces. Imagine if this were the case with English, and a sign for a job fair were to display "opportunityisnowhere." Regardless of their intent being to announce that "opportunity is now here," that can easily be read pessimistically. Our purpose is to advance computer effectiveness in automatically parsing unrestricted Chinese text and then, considering all possible parses, correctly selecting the true word boundaries as instructed by a system of Mandarin-specific rules.

Project Importance

Mandarin Chinese is indisputably the most widely used language in the world with nearly 900 million native speakers and over a billion speakers when including those who speak it as a second language.[1] If languages that share the same writing system are included under the broader label "Chinese", then the number of speakers accounts for about 18% of the world's population. Accurately segmenting Chinese text is therefore vital to enabling uninhibited communication between East and West as it is a fundamental task for automatic machine translation. Because Mandarin is a Lingua Franca in its region, it can serve as a pivot language[2] in that process, e.g. a stepping stone between a language like English and a language like Cantonese which uses Chinese characters.

The best Chinese word segmentation systems are reaching accuracy rates of around 90% (Duan, Sui, Ge, 2014). This means that one in ten words is mis-parsed, which is far worse than human performance. Many attempts to solve the segmentation problem are proprietary systems. We believe that for our contribution to have a real chance of being incorporated in the collaborative attempt at advancing this technology, we need to produce work under a free-use, open-source license. In our view, this is imperative when seeking to make a lasting contribution in the field of computational linguistics where each new breakthrough is built on recent progress of other scholars. Proprietary work often excludes itself from this process. Therefore, we intend to contribute to the Apertium[3] project, which is an open-source machine translation platform supported by an international community of computational linguists.

Project Profile Body

When you type in a word that your word processor does not recognize, your entry is underlined in red. With a right click, you then have the option to "add to dictionary." Similarly, when a word-parsing program—called a segmenter—comes upon an unknown character, it is excluded in the program's output unless we specifically equip it with a vocabulary large enough to process it.

We plan to compile thousands of terms into a digital lexicon so that the associated segmenter can achieve complete coverage when parsing our gold-standard corpus and produce a lineup of all possible parses. Additionally, we aim to formulate context-dependent rules in the form of a constraint grammar.[4] Our final outcome then, is the result of these two systems (segmenter and grammar) working in concert. The constraint grammar considers every possible parse as provided by the segmenter, and then removes options which are not possible given the surrounding context. We will evaluate the performance of our work by comparing our system's output with the gold-standard parses of human annotators done as a part of the Universal Dependencies[5] project.

Dr. Reynolds' own work as a computational linguist with Russian includes formulating rules within the constraint grammar paradigm and actively contributing within the Apertium platform. His expertise in these spheres and their associated processes, coupled with my familiarity with Mandarin and concomitant competence in computer science tasks will allow us to begin our work immediately.

We will first divide our procured gold-standard corpus of 500 hand-segmented sentences containing over 12,500 words into two subcorpora. A large portion, likely ninety percent, will be used as a

---

[1] https://www.ethnologue.com/language/cmn
[2] https://en.wikipedia.org/wiki/Pivot_language
[3] https://www.apertium.org
[4] https://visl.sdu.dk/constraint_grammar.html
[5] https://github.com/UniversalDependencies/UD_Chinese/blob/master/zh-ud-dev.conllu

development corpus with the aim of achieving full coverage by our segmenter before December 1st. The remainder will be reserved as an untouched *test corpus* to be used for evaluation at the conclusion of our project. Updating the dictionary in order to reach that goal entails scouring the internet for existing character banks and appending those words to our program's lexicon.

Next, we will formulate the rules of the constraint grammar. Dr. Reynolds' expertise handling the technical aspects of rule development will be crucial at this stage. Each rule expresses a constraint of what is possible in a language. For example, an English constraint grammar rule could easily handle a phrase-pattern like "the push" where the second word can technically be either a noun or a verb. Though the correct reading seems obvious to a human, computers require instruction. In this example, the rule might say that if a word following "the" can be either a verb or a noun, eliminate the verb option.

In order to account for the nuances of Mandarin in our rule writing, we plan to collaborate with Mandarin linguists who specialize in the study of the language's unique properties. At this stage, we will constantly be testing our rules and the order in which they are being applied on our *development corpus*. Doing so enables the flexibility of making slight adjustments to our grammar as needed. The work of formulating rules for a language-specific constraint grammar is never done. When it comes time to draft a report of our work and we are forced to a stopping point, we will assess the precision of the current rule set by comparing their output to that of the manual segmentations in our *test corpus*.

Anticipated Academic Outcome

In the form of a nine-page paper, we plan to submit a summary of our methods and results to either the Student Research Workshop of the North American Chapter of the Association for Computational Linguistics (NAACL 2018) or the annual conference of the International Committee on Computational Linguistics (COLING 2018).

Qualifications

Dr. Reynolds is an expert in rule-based approaches to natural language processing. His dissertation research produced multiple state-of-the-art technologies for the Russian language, including automatic stress placement, and automatic language-learning exercise generation from online texts. He has extensive experience mentoring students as part of the Google Summer of Code and Google Code-In initiatives, having volunteered with the Apertium project every year since 2013. He is currently serving on thesis committees for two Linguistics MA students, and supervising multiple undergraduate students on a variety of projects sponsored by the Office of Digital Humanities.

I feel that I am uniquely qualified to approach this research project and the real-world problems it addresses because of my combined competence in programming, Chinese, and linguistics. Since taking an introductory computer science course in 2015, I have continued to self-educate by completing online courses in programming and expanding my scope to include more programming languages. Currently, I am learning Python in Dr. Reynolds' Text Processing & Analysis course. It is likely that I will use Python to power much of the computation involved in this endeavor. My formal study of Mandarin here at BYU has given me the foundation with the language necessary for the project at hand. As a student of linguistics, I have gained the skills to discuss a language in formal terms, think critically about its linguistic properties, and appreciate its unique complexities.

Project Timetable

October: Compile terms into the open-source Apertium dictionary
November: Dictionary compilation continued with 100% coverage of development corpus by Dec 1
December: Begin work on constraint grammar (CG)
January: Consult Mandarin linguistics specialists to fine-tune CG
February: Conclude work on CG; begin composition of 9-page paper describing methods and results
March: Submit paper to either NAACL (March 2) or COLING (March 16)
April/May: Notifications
June-August: Main Conference: NAACL June 2-4, 2018 or COLING August 22-25

Scholarly Sources

Duan, Huiming, Zhifang Sui, and Tao Ge. "The CIPS-SIGHAN CLP 2014 Chinese Word Segmentation Bake-off." Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, January 2014. doi:10.3115/v1/w14-6814.