The Automated Translation of Chinese-to-English Creative Literature

Carl Hoover (Computer Science) & Heather Sommer (English)
Mentor: Dr. Michael Wick (Computer Science)
University of Wisconsin-Eau Claire

Overview

* While the statistical approach to translation is the most popular, it might not be the best option for culturally-sensitive translations that would be prevalent in creative literature, such as:

  * idiomatic expressions

  * folk-lore references

  * culturally-based metaphor

* A statistical translator created at John Hopkins University (2003) performed impressively when given a document-specific corpus - a drastic limitation (Byrne 1).

* A translator created by researchers at Brown University and University of South California using a syntax-based approach performed significantly higher than IBM Model 4 (Charniak 1).

* A study out of the Center for Spoken Language Research at University of Colorado at Boulder applies semantic/thematic roles to the parts of a sentence in an attempt to garner a shallow semantic translation that can be expanded upon (Jurafsky "Semantic" 1).

* Limitations of current studies is partially because Chinese lacks the specific word-boundaries of English/Romantic languages and thus is more difficult to parse (Luo 1).

Abstract

The purpose of this project is to develop translation software that will assist translators, students, and scholars in translating those passages which may not see publication otherwise. Translation is a difficult process, at best, and one which is often shied away from in favor of studying language in spoken or theoretical applications. By compiling a database to allow for the corpora analysis of creative works previously translated and employing an example-based approach to translation (in addition to the currently popular statistical approach), this daunting task can be made significantly easier. This, in turn, allows the translation to focus on bringing to light the author's style and true intent in writing.

This project will require students to continue to research the current state of Chinese-to-English translation software and applications in both the United States and China, and to draw on any open-source code that is available for later comparison.

Background

Preliminary research reveals that current work in the machine translation of (Traditional) Chinese-to-English focuses on a statistical approach. This approach measures the probability of a word appearing next to the previously translated word based on studies of corpora (collections of texts) on the same subject matter stored in databases. This approach does not allow for a culturally-sensitive translation with regards to the source language, as meaning is often lost through metaphor or folk-lore references.

Because databases for a statistical approach are based on a specific, industry or field-related corpora to improve its efficacy, the likelihood that a creative work translated in this manner would be even grammatically accurate is highly unlikely. A statistical translation cannot account for stylistic choices made by the author, literary allusions, or original metaphors because a statistical approach to translation does not account for the spontaneity of language. As databases drawn from the appropriate corpora are compiled, this problem is lessened - but statistics-based translations still fails to account for the creativity of language, a principle that allows humans to express an unlimited number of ideas never before expressed.

Specific Tasks

For the first phase, students conducted background research and setup for the development portion of the project. This portion will require students to continue to extensively research machine translation theory and the current state of Chinese-to-English translation software and applications in both the United States and China. Then students will draw on any open-source code that is available from these translators for later comparison.

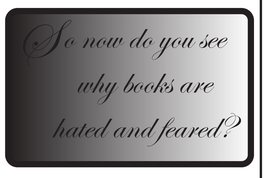
The second phase of the project will require students to find an accurate translation of an elementary Chinese text. Students will then run the Chinese version through the current machine translators (located during phase one) and gauge the accuracy of each based on an instrument developed by the students for this purpose. At this time, students will analyze the open-source code obtained for reasons why the translations might be inaccurate and seek ways to combat this during the final phase.

Phase three will require students to either locate corpora or, if necessary, create a corpus of creative works and compile these into a larger, more comprehensive database to be used during development. These corpora will be analyzed for cultural context and stylistics.

Finally, students will work to build a program that is able to handle all the nuances of translation in creative works. Previous research will help the students to discern what methods will be implemented.

Project Significance

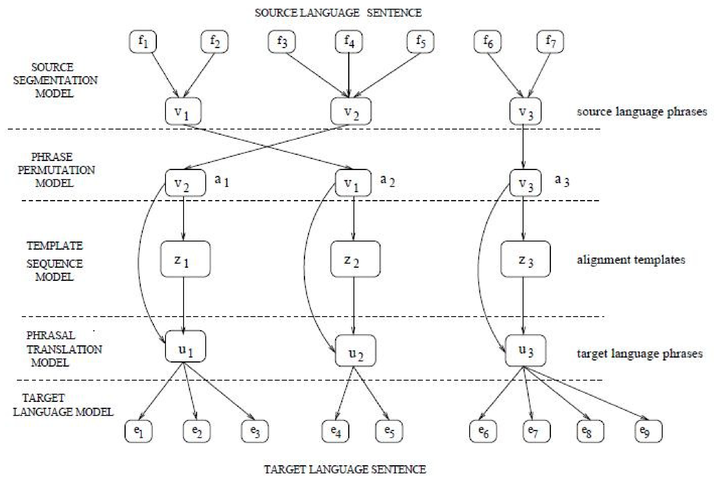
The "firewall" of accidental censorship by way of incorrect translations supplements the intentional censorship of creative works by oppressive Chinese government regulations. It is uncertain how many books never reach publication in China due to censors, but it is not difficult to find a text in the United States that found international acclaim despite never seeing publication in the author's native country. In "Fahrenheit 451," Ray Bradbury writes "So now do you see why books are hated and feared? They show the pores in the face of life" [1]. These "pores" are key to looking further into a culture by exploring those facets otherwise hidden. As long as culturally-inaccurate translations are published, the struggle of revolutionary authors to be heard remains and the likelihood of provocative works slipping through the cracks of social awareness because of their inaccessibility remains a constant threat. This research holds the potential of broadening the cultural awareness of English-speaking nations to include the creative works of Chinese authors.



Prior Studies

John Hopkins University, 2003

One translator, developed and piloted by a group at John Hopkins University in 2003, used this approach in 100,000 sentence pairs to an overall impressive effective rate; however, it is worth noting that each new document introduced to the system for translation also required a specific corpus document with translations needed specifically for that document (Byrne 1). This breakthrough - though significant for the study - proves to be *a very limiting factor* when those using the translator would be translating the document for perhaps the first time and likely without significant knowledge of the Chinese language.



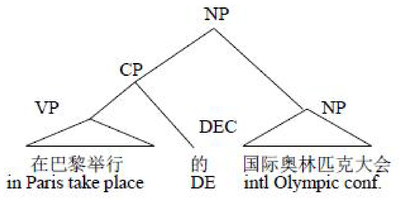
***Alignment Template Translation Model Architecture**:*
"Phrase-level alignment between source and target phrases and a word-level alignment between words in these phrases implemented via individual alignment templates" (Byrne 1).

Brown University / University of South California, 2003

Another group of researchers from Brown University and University of South California used a syntax-based approach to statistical translation, with a base of 347 sentences and an overall effective rate that was *70% higher* than IBM Model 4 when both the syntax was good and the meaning preserved (Charniak 1). The high success rate of this particular model seems to stem from translation at the node-level rather than at the sentence-level, which is an excellent starting point.

Center for Spoken Language Research at University of Colorado at Boulder, 2004

A study out of the Center for Spoken Language Research at University of Colorado at Boulder takes a different approach to the translation of Chinese-to-English, instead applying semantic/thematic roles to the parts of a sentence in an attempt to garner a shallow semantic translation that can be expanded upon (Jurafsky "Semantic" 1) and would account for particles in Chinese that do not directly translate to English easily (such as particles 的 and 了). This approach is especially interesting, as Chinese is a character-based language and the pictorial role of these characters often contributes to a seemingly shallow and simple translation in a case such as 出⿰, which translates in this way as "enter now" but translates better into English as "to emerge." These differences are not usually such a hindrance in human translation, but in MT which relies on algorithms that can generate all possible translations, someone is still needed that understands both the origin and target languages so that the correct translation is chosen. Currently, this is where the statistical approach comes into play: by weighing 'what usually follows what', the interloper (in this case, the bilingual person needed to complete the selection process) is cut out. By potentially combining statistical and semantic approaches, a more accurate translation could be reached. Because of the strict constraints on adjunct ordering in Chinese, a semantic approach will be very beneficial to this project, as will access to the Collins (1999) parser mentioned in this article (Jurafsky "Semantic" 8).



Example of DE Construction:
Particle "de" does not have an English equivalent and relies instead on semantic translation - which classical statistical translation software did not account for. Graphic from Jurafsky.

Difficulties in Translation of Chinese-to-English

One problem in the parsing of Chinese sentences is that Chinese lacks syntactic word-boundaries (Luo 1). Previous knowledge of Chinese is required if someone is to try to translate a document, insofar as he or she must be aware of what characters constitute a word, and opinions regarding this, particularly in proper names, measure words, and compound nouns, are debated even among native speakers (Luo 1). These word breaks are less important, however, in a parser that looks at the character-level and seems to then simultaneously provide word-segmentation, POS (part-of-speech) tagging, and constituent labeling (Luo 2):

1. Word-level POS tags become labels in character trees.

2. Character-level tags are inherited from word-level POS tags after appending a positional tag;

3. For single-character words, the positional tag is "s"; for multiple-character words, the first character is appended with a positional tag "b", last character with a positional tag "e", and all middle characters with a positional tag "m" (Luo 2).

"(IP (NP (NP 天 津 港/NR) (NP 扩 建/NN 工程/NN)) (VP 开工/VV) 。/PU)"
would become
"(IP (NP (NP (NR 天/nrb 津/nrm 港/nre)) (NP (NN 扩/nnb 建/nne) (NN 工/nnb 程/nne))) (VP (VV 开/vvb 工/vve)) (PU 。/pus))."

Example of Luo's word-segmentation, POS tagging, and constituent labeling

Future of This Study

It is possible that a character-based approach, modified, and compounded with both a modified statistical approach (which may or may not get thrown out at the constituent-level) and a semantic/thematic approach will prove beneficial to our project.

Works Cited

Bradbury, Ray. Fahrenheit 451. New York: Simon and Schuster, 1993.

Byrne, W. 'and' Khudanpur, S. 'and' Kim, W. 'and' Kumar, S. 'and' Pecina, P. 'and' Virga, P. 'and' Xu, P. 'and' Yarowsky, D.. "The Johns Hopkins University 2003 Chinese-English Machine Translation System." (2003) 15 Oct 2008 <http://www.clsp.jhu.edu/people/xp/publications/mtsummit03.pdf>.

Charniak, Eugene, 'and' Knight, Kevin, 'and' Yamada, Kenji. "Syntax-based Language Models for Statistical Machine Translation." (2003) 15 Oct 2008 <http://www.isi.edu/natural-language/projects/rewrite/mtsummit03.pdf>.

Hill. "Who Is Ms. Hill?" 02 April 2010 <http://studentweb.cortland.edu/alicia.hill/mypage/mypage/Who_is_Ms._Hill.html>.

Jurafsky, Daniel, 'and' Sun, Honglin. "Shallow Semantic Parsing of Chinese." (2004) 15 Oct 2008 <http://www.stanford.edu/~jurafsky/Sun-Jurafsky-HLT-NAACL04.pdf>.

Luo, Xiaoqiang. "A Maximum Entropy Chinese Character-Based Parser." 15 Oct 2008 <http://www.aclweb.org/anthology-new/W/W03/W03-1025.pdf>.

Acknowledgments

* UWEC Differential Tuition

* UW-Eau Claire Center of Excellence for Faculty and Undergraduate Student Research Collaboration