Test Bed for Building Resources for Multi-Lingual Processing

Carol Nichols
DMP Research Participant, University of Pittsburgh
cln23@pitt.edu

Abstract

Annotated examples are extremely valuable to the improvement of natural language processing tools, but are expensive to obtain from human annotators. We seek to make annotating an easier and more intuitive process for two natural language processing applications: machine translation and part-of-speech tagging. Our test bed allows the annotator to provide more information than has previously been collected from similar interfaces. This program is also ready to be integrated with an active learning framework in order to ask annotators for information which would provide the machine attempting to learn with the most useful examples.

1 Introduction

Natural language processing encompasses many problems in computer science relating to making computers able to work with human language. One subtopic is machine translation, which seeks to have a computer translate from one language to another. This work is difficult for computers because of many issues such as the ambiguity of a word's meaning and the differences in structure between languages. The best way to improve machine translation accuracy is to have the machine learn in a supervised way by providing the model with annotated examples of alignments between words of a sentence in one language and its translation in another. The machine can learn important information from these examples such as which words translate to which words in the other language and where in the translated sentence these words are placed. This information can then be applied to the translation of previously unseen sentences.

In addition to improving machine translation, we are also interested in improving natural language processing tools for languages other than English. Much of the research done in natural language processing has been done in English, and there are many annotated examples in English for many different tools. If we wanted to improve these tools for languages other than English (especially those languages which are very dissimilar to English, such as Chinese or Arabic) we would need annotated examples in that language, but word alignments between that language and English could also be useful since so much information is already available for English. Exploiting the relationship between English and another language could possibly improve the tools when only a few annotated examples in that language are available. One tool we are interested in is a part-of-speech tagger, and my program can gather manually annotated part-of-speech tags along with manually annotated word alignments.

When working to improve a tool such as a part-of-speech tagger, we are also interested in using an active learning framework in order to get the most information out of the least number of annotated examples. In this framework, a tagger would attempt to tag words in sentences with their part-of-speech. Those sentences which the tagger has the lowest accuracy would then be sent to my test bed program for a human to correct. This reduces the number of sentences a human has to annotate since we are only asking them to do sentences the machine is not able to on its own. Asking a human to annotate thousands of sentences that a machine can already tag with accuracy is wasteful.

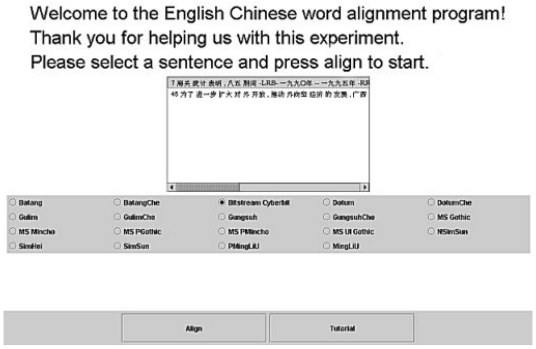
In addition to making annotation easier for the people doing it by reducing the number of sentences we need them to annotate, we want to make the interface they will use as intuitive as possible. We did not want to make any assumptions about the computer ability of the people doing the alignments. We did not want them to have to install any programs or change any complicated settings, so my program is a JAVA applet which is web accessible. The annotators only need to open a JAVA-enabled web browser to my program and they can start annotating. A useful side effect of the program's web accessibility is that the annotators do not need to be physically present at the research site. All the information is handled over the internet.

The general layout of the program is based on previously available annotation tools, most notably the interface used to create the gold standard in The Blinker Project (Melamed, 1998) and one created by my mentor, Rebecca Hwa. A sentence in English and a sentence in Chinese are displayed and the user is able to click on words from both sentences and mark them as aligned to each other graphically with a line between them. My program has additional features which will be explained in section 2. Information about the internal representation of the data collected and how it can be used is in section 3. Section 4 contains the biggest problems we ran into while developing this test bed. Section 5 is about future work we are proposing to do that involves this test bed program, and section 6 contains conclusions.

2 Features

Upon loading the test bed, the annotator must provide a username. This allows the annotator to save their work and return to it at a later date. New users are forced to view a tutorial detailing how to use the program. The tutorial is accessible from the main part of the program as well, in case the user needs help.

The main menu screen provides font display choices since some fonts which seem to be able to display Chinese are actually Japanese fonts which do not display all the characters correctly. The user can select the one which best displays the Chinese sentences in the menu. Next the user must select a sentence and click Align to begin annotating.



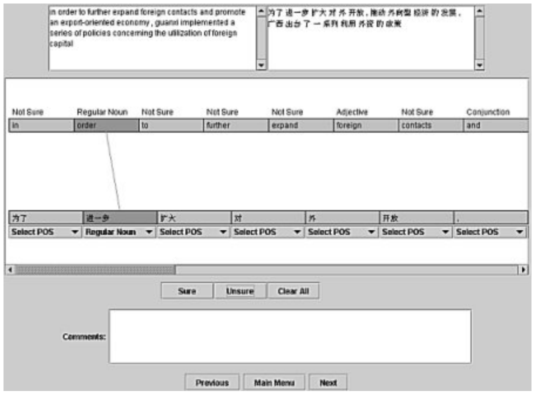
Welcome to the English Chinese word alignment program!
Thank you for helping us with this experiment.
Please select a sentence and press align to start.

7 海关 统计 表明，八五 期间 -LRB- 一九九O年 -- 一九九五年 -RR
46 为了 进一步 扩大 对 外 开放，推动 外向型 经济 的 发展，广西

○ Batang ○ BatangChe ● Bitstream Cyberbit ○ Dotum ○ DotumChe
○ Gulim ○ GulimChe ○ Gungsuh ○ GungsuhChe ○ MS Gothic
○ MS Mincho ○ MS PGothic ○ MS PMincho ○ MS UI Gothic ○ NSimSun
○ SimHei ○ SimSun ○ PMingLiU ○ MingLiU

The English and Chinese sentences are displayed above the actual alignment panel for easy reading. Then the English sentence and the Chinese sentence are displayed in horizontal rows of rectangles which contain the presegmented words. These rectangles are initially gray. When the user clicks on a rectangle, it becomes selected in purple. When a word from the other sentence is also selected, a purple line is drawn between the two rectangles. Now the user must click a button marked “Sure” or a button marked “Unsure” to make the alignment permanent. More than one word from each sentence may be selected in the case of a phrase which cannot be logically separated, such as idioms or the French “ne… pas” construction that typically aligns to the English “not”, and all words in a phrase are automatically aligned to all other words in the phrase.

If a word appears in one sentence but has no translation in the other sentence, the user can right click on the word to select it in red. “Sure” or “Unsure” must still be clicked to mark this word permanently as not translated.



in order to further expand foreign contacts and promote an export-oriented economy , guanxi implemented a series of policies concerning the utilization of foreign capital

为了 进一步 扩大 对 外 开放，推动 外向型 经济 的 发展，广西 出台 了 一 系列 利用 外资 的 政策

Not Sure	Regular Noun	Not Sure	Not Sure	Not Sure	Adjective	Not Sure	Conjunction
in	order	to	further	expand	foreign	contacts	and

为了	进一步	扩大	对	外	开放	，
Select POS	Regular Noun	Select POS	Select POS	Select POS	Select POS	Select POS

Part-of-speech tags for the English words are given and appear above the English sentence. Below the Chinese sentence are menus which contain part-of-speech tags. When an English word is aligned with a Chinese word, the Chinese word automatically receives the part-of-speech of the English word, but the user can correct this if it is wrong.

The time a user spends on a particular sentence is recorded along with their alignments and part-of-speech choices. This feature serves to let the researcher know exactly how much time they are asking of their annotators to see if the amount of work is reasonable. The timer is not visible to the user.

An input text area for comments is underneath the alignment panel. This allows users to provide information to the researcher about why they aligned the sentences the way they did, problems they had, or assumptions they made. These comments are stored with the other information from the alignment.

When all the words in a sentence have been marked as aligned with another word or not translated, and all the part-of-speech tags in the second sentence have been marked, an asterisk appears next to that sentence in the main menu. Once all sentences currently available to the user have been completed, a new button appears beneath the main menu that says “Get New Session.” This button gets a new group of sentences for the user. Currently the user is not able to return to a previous group of sentences, and a warning stating this is in the tutorial. This may be changed in the future.

Upon exit from the program, the server program goes through all the information from all the sentences the user has aligned and creates two aggregate data files. One contains the word alignment information, the other contains part-of-speech information, and both are in the format typically needed by machine translation models that use the alignment and part-of-speech taggers that use part-of-speech tags.

3 Internal Representation

The main directory of the program contains the program and its configuration files, including the files of the English and Chinese sentences, the English part-of-speech tags that go with the English sentences, and the part-of-speech tags to display in the menus below the Chinese words. Subdirectories are created for each user, named by their username. Within each user subdirectory are more subdirectories by session number. These session subdirectories contain local copies of the English sentence file, Chinese sentence file, and English part-of-speech tags – but these only contain the sentences in this session so that they do not have to be extracted from the master file at run time.

Each sentence that the user has worked on has its own file in the session subdirectory titled by its master sentence ID number. This file contains the English part-of-speech tags, English sentence, Chinese sentence, user created Chinese part-of-speech tags, user created alignment, time spent by the user and comments left. The alignment is represented by a sparse matrix that uses the English words as row headings and Chinese words as column headings. At the intersection of an English word and Chinese word, a 0 indicates that these two words were not aligned to each other. A 1 indicates a “Sure” link while 2 means “Unsure”. The first row and column in the matrix represent not translated, so a 1 or 2 in the first row or column means that word was marked as not translated.

Also in the main user directory are the aggregate data files previously mentioned for the word alignment and part-of-speech data, and a file containing the IDs of all the sentences in all the sessions of this user.

4 Problems and Solutions

The two biggest problems I encountered while developing this program were having Chinese characters display in a JAVA applet and working around applet security restrictions. Researching the solution to the character display was difficult as many web pages and tutorials offered conflicting advice. The correct solution turned out to be having a font file of a font that could display Chinese with the JAVA compiler and encoding all input and output streams.

JAVA applets are not permitted to write files to protect computers from viruses. They also are only allowed to make network connections with their host. One of our computers was set up as a webserver for us, and I wrote a server program that runs on the webserver. The applet connects with this server program and sends it the user data. The server program is a JAVA application which is allowed to write files.

5 Future Work

We are interested in combining this applet with a part-of-speech tagger in order to create an active learning framework. The tagger would attempt to choose part-of-speech tags for words in the sentences, and would be better at tagging some sentences than it would at others. In order to improve the accuracy of the tagger, it would send the sentences it is the worst at tagging to my program for a human to annotate. This would improve the tagger the most in the least amount of time with the least amount of human effort, since the tagger is only asking to see examples that it especially needs to see, and not more sentences similar to those it is already capable of tagging. Currently my program can take a list of sentences with scores and first ask a user to annotate those sentences with the lowest scores. We also have part-of-speech taggers; future work will include making a program which scores the sentences.

Other improvements we are interested in making include a program to gather statistics about multiple users' annotations on the same sentences. This would be helpful if we were interested in creating a gold standard or deciding which annotators were the most consistent. Also, there are other graphical representations of the sentence alignments we could have used, such as a grid representation where one sentence is the row headings and one sentence is the column headings, and words are aligned to each other by marking their intersecting cell. Some users may find this easier to work with, and improving ease of use for different people is one of our future goals.

6 Conclusions

This program will make it easier for annotators to provide us with more valuable data to use to improve machine translation, part-of-speech taggers, and other Natural Language Processing tools for languages other than English. We will also have gathered other useful information such as how sure the annotator is of their alignments and how long it took them that other programs previously used have not been able to provide. The web accessibility of this program makes it easier to use and easier to obtain more data from more people. This program will also be part of an active learning framework to further improve the efficiency of a machine learning model.

References

I. Dan Melamed (1998). Manual Annotation of Translational Equivalence: The Blinker Project. In *IRCS Technical Report #98-07*.