Word Alignment and Cross-Lingual Resource Acquisition *

Carol Nichols and Rebecca Hwa
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260
{cln23,hwa}@cs.pitt.edu

Abstract

Annotated corpora are valuable resources for developing Natural Language Processing applications. This work focuses on acquiring annotated data for multilingual processing applications. We present an annotation environment that supports a web-based user-interface for acquiring word alignments between English and Chinese as well as a visualization tool for researchers to explore the annotated data.

1 Introduction

The performance of many Natural Language Processing (NLP) applications can be improved through supervised machine learning techniques that train systems with annotated training examples. For example, a part-of-speech (POS) tagger might be induced from words that have been annotated with the correct POS tags. A limitation to the supervised approach is that the annotation is typically performed manually. This poses as a challenge in three ways. First, researchers must develop a comprehensive annotation guideline for the annotators to follow. Guideline development is difficult because researchers must be specific enough so that different annotators' work will be comparable, but also general enough to allow the annotators to make their own linguistic judgments. Reported experiences of previous annotation projects suggest that guideline development is both an art and a science and is itself a time-consuming process (Litman and Pan, 2002; Marcus et al., 1993; Xia et al., 2000; Wiebe, 2002). Second, it is common for the annotators to make mistakes, so some form of consistency check is necessary. Third, the entire process (guideline development, annotation, and error corrections) may have to be repeated with new domains.

This work focuses on the first two challenges: helping researchers to design better guidelines and to collect a large set of consistently labeled data from human annotators. Our annotation environment consists of two pieces of software: a user interface for the annotators and a visualization tool for the researchers to examine the data. The data-collection interface asks the users to make lexical and phrasal mappings (word alignments) between the two languages. Some studies suggest that supervised word aligned data may improve machine translation performance (Callison-Burch et al., 2004). The interface can also be configured to ask the annotators to correct projected annotated resources. The idea of projecting English annotation resources across word alignments has been explored in several studies (Yarowsky and Ngai, 2001; Hwa et al., 2005; Smith and Smith, 2004). Currently, our annotation interface is configured for correcting projected POS tagging for Chinese. The visualization tool aggregates the annotators' work, takes various statistics, and visually displays the aggregate information. Our goal is to aid the researchers conducting the experiment to identify noise in the annotations as well as problematic constructs for which the guidelines should provide further clarifications.

Our longer-term plan is to use this framework to support active learning (Cohn et al., 1996), a machine learning approach that aims to reduce the number of training examples needed by the system when it is provided with more informative training exam-

*This work has been supported, in part, by CRA-W Distributed Mentor Program. We thank Karina Ivanetich, David Chiang, and the NLP group at Pitt for helpful feedbacks on the user interfaces; Wanwan Zhang and Ying-Ju Suen for testing the system; and the anonymous reviewers for their comments on the paper.

Proceedings of the ACL Interactive Poster and Demonstration Sessions,
pages 69–72, Ann Arbor, June 2005.

ples. We believe that through a combination of an intuitive annotation interface, a visualization tool that checks for style and quality consistency, and appropriate active learning techniques, we can make supervised training more effective for developing multilingual applications.

2 Annotation Interface

One way to acquire annotations quickly is to appeal to users across the Internet. First, we are more likely to find annotators with the necessary qualifications. Second, many more users can work simultaneously than would be feasible to physically host in a lab. Third, having many users annotate the same data allows us to easily identify systematic problems as well as spurious mistakes. The OpenMind Initiative (Stork, 2001) has had success collecting information that could not be obtained from data mining tools or with a local small group of annotators.

Collecting data from users over the Internet introduces complications. Since we cannot ascertain the computer skills of the annotators, the interface must be easy to use. Our interface is a JAVA applet on a webpage so that it is platform independent. An online tutorial is also provided (and required for first-time users). Another problem of soliciting unknown users for data is the possibility of receiving garbage data created by users who do not have sufficient knowledge or are maliciously entering random input. Our system minimizes this risk in several ways. First, new users are required to work through the tutorial, which also serves as a short guide to reduce stylistic differences between the annotators. Second, we require the same data to be labeled by multiple people to ensure reliability, and researchers can use the visualization tool (see Section 3) to compare the agreement rates between annotators. Finally, our program is designed with a filter for malicious users. After completing the tutorial, the user is given a randomly selected sample sentence (for which we already have verified alignments) to annotate. The user must obtain an F-measure agreement of 60% with the "correct" alignments in order to be allowed to annotate sentences.[1]

Because word alignment annotation is a useful resource for both training and testing, quite a few interfaces have already been developed. The earliest is the Blinker Project (Melamed, 1998); more recent systems have been released to support more languages and visualization features (Ahrenberg et al., 2003; Lambert and Castell, 2004). [2] Our interface does share some similarities with these systems, but it is designed with additional features to support our experimental goals of guideline development, active learning and resource projection. Following the experimental design proposed by Och and Ney (2000), we instruct the annotators to indicate their level of confidence by choosing *sure* or *unsure* for each alignment they made. This allows researchers to identify areas where the translation may be unclear or difficult. We provide a text area for comments on each sentence so that the annotator may explain any assumptions or problems. A hidden timer records how long each user spends on each sentence in order to gauge the difficulty of the sentence; this information will be a useful measurement of the effectiveness of different active learning algorithms. Finally, our interface supports cross projection annotation. As an initial study, we have focused on POS tagging, but the framework can be extended for other types of resources such as syntactic and semantic trees and can be configured for languages other than English and Chinese. When words are aligned, the known and displayed English POS tag of the last English word involved in the alignment group is automatically projected onto all Chinese words involved, but a drop-down menu allows the user to correct this if the projection is erroneous. A screenshot of the interface is provided in Figure 1a.

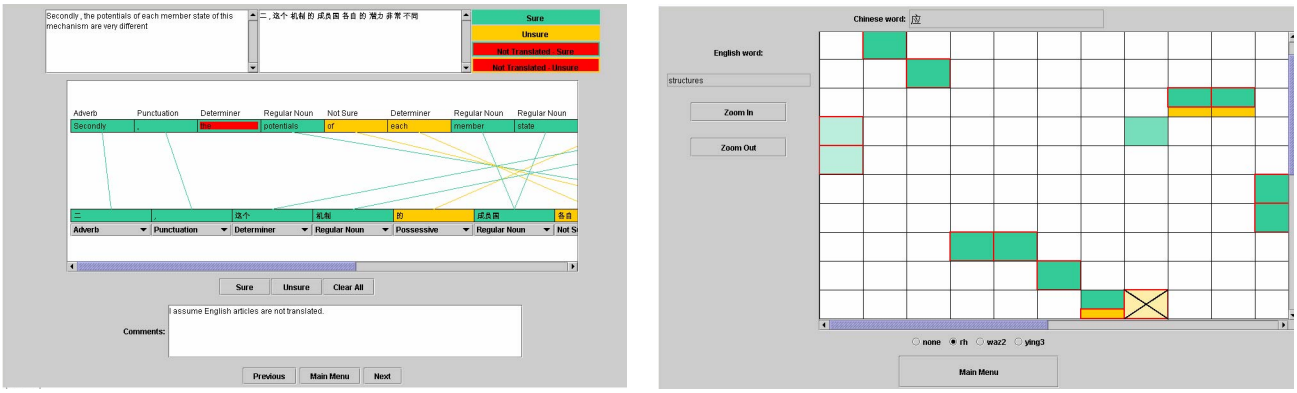
3 Tools for Researchers

Good training examples for NLP learning systems should have a high level of consistency and accuracy. We have developed a set of tools for researchers to visualize, compare, and analyze the work of the annotators. The main interface is a JAVA applet that provides a visual representation of all the alignments superimposed onto each other in a grid.

For the purposes of error detection, our system provides statistics for researchers to determine the agreement rates between the annotators. The metric we use is Cohen's K (1960), which is computed for every sentence across all users' alignments. Cohen's K is a measure of agreement that takes the total probability of agreement, subtracts the probability the agreement is due to chance, and divides by the maximum agreement possible. We use a variant of the

[1] The correct alignments were performed by two trained annotators who had an average agreement rate of about 85%. We chose 60% to be the figure of merit because this level is nearly impossible to obtain through random guessing but is lenient enough to allow for the inexperience of first time users. Automatic computer alignments average around 50%.

[2] Rada Mihalcea maintains an alignment resource repository (`http://www.cs.unt.edu/~rada/wa`) that contains other downloadable interface packages that do not have companion papers.



(a) (b)

Figure 1: (a) A screenshot of the word alignment user-interface. (b) A screenshot of the visualization tool for analyzing multiple annotators' alignments.

equation that allows for having three or more judges (Davies and Fleiss, 1982). The measurement ranges from 0 (chance agreement) to 1 (perfect agreement). For any selected sentence, we also compute for each annotator an average pair-wise Cohen's K against all other users who aligned this sentence.[3] This statistic may be useful in several ways. First, someone with a consistently low score may not have enough knowledge to perform the task (or is malicious). Second, if an annotator received an unusually low score for a particular sentence, it might indicate that the person made mistakes in that sentence. Third, if there is too much disagreement among all users, the sentence might be a poor example to be included.

In addition to catching individual annotation errors, it is also important to minimize stylistic inconsistencies. These are differences in the ways different annotators (consistently) handle the same phenomena. A common scenario is that some function words in one language do not have an equivalent counterpart in the other language. Without a precise guideline ruling, some annotators always leave the function words unaligned while others always group the function words together with nearby content words. Our tool can be useful in developing and improving style guides. It highlights the potential areas that need further clarifications in the guidelines with an at-a-glance visual summary of where and how the annotators differed in their work. Each cell in the grid represents an alignment between one particular word in the English sentence and one particular word in the Chinese sentence. A white cell means no one proposed an alignment between the words. Each colored cell has two components: an upper green portion indicating a *sure* alignment and a lower yellow portion indicating an *unsure* alignment. The proportion of these components indicates the ratio of the number of people who marked this alignment as *sure* to those who were *unsure* (thus, an all-green cell means that everyone who aligned these words together is sure). Moreover, we use different saturation in the cells to indicate the percentage of people who aligned the two words together. A cell with faint colors means that most people did not chose to align these words together. Furthermore, researchers can elect to view the annotation decisions of a particular user by clicking on the radio buttons below. Only the selected user's annotation decisions would be highlighted by red outlines (i.e., only around the green portions of those cells that the person chose *sure* and around the yellow portions of this person's *unsure* alignments).

Figure 1b displays the result of three annotators' alignments of a sample sentence pair. This sentence seems reasonably easy to annotate. Most of the colored cells have a high saturation, showing that the annotators agree on the words to be aligned. Most of the cells are only green, showing that the annotators are sure of their decisions. Three out of the four *unsure* alignments coincide with the other annotators' *sure* alignments, and even in those cases, more annotators are sure than unsure (the green areas are 2/3 of the cells while the yellow areas are 1/3). The colored cells with low saturation indicate potential outliers. Comparing individual annotator's alignments against the composite, we find that one annotator, *rh*, may be a potential outlier annotator since this person generated the most number of lightly saturated cells. The person does not appear to be malicious since the three people's overall agreements are high. To determine whether the conflict

[3]not shown in the screenshot here.

arises from stylistic differences or from careless mistakes, researchers can click on the disputed cell (a cross will appear) to see the corresponding English and Chinese words in the text boxes in the top and left margin.

Different patterns in the visualization will indicate different problems. If the visualization patterns reveal a great deal of disagreement and *unsure* alignments overall, we might conclude that the sentence pair is a bad translation; if the disagreement is localized, this may indicate the presence of an idiom or a structure that does not translate word-for-word. Repeated occurrences of a pattern may suggest a stylistic inconsistency that should be addressed in the guidelines. Ultimately, each area of wide disagreement will require further analysis in order to determine which of these problems is occurring.

4 Conclusion and Future Work

In summary, we have presented an annotation environment for acquiring word alignments between English and Chinese as well as Part-Of-Speech tags for Chinese. The system is in place and the annotation process is underway.[4]

Once we have collected a medium-sized corpus, we will begin exploring different active learning techniques. Our goal is to find the best way to assign utility scores to the as-of-yet unlabeled sentences in order to obtain the greatest improvement in word alignment accuracy. Potential information useful for this task includes various measurements of the complexity of the sentence such as the rate of (automatic) alignments that are not one-to-one, the number of low-frequency words, and the number of potential language divergences (for example, many English verbs are nominalized in Chinese), and the co-occurrence of word pairs deemed to be *unsure* by the annotators in other contexts. Furthermore, we believe that the aggregate visualization tool will also help us uncover additional characteristics of potentially informative training examples.

References

Lars Ahrenberg, Magnus Merkel, and Michael Petterstedt. 2003. Interactive word alignment for language engineering. In *Proceedings from EACL 2003*, Budapest.

Christopher Callison-Burch, David Talbot, and Miles Osborne. 2004. Statistical machine translation with word- and sentence-aligned parallel corpora. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, July.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Meas.*, 20:37–46.

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. 1996. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145.

M. Davies and J. Fleiss. 1982. Measuring agreement for multinomial data. *Biometrics*, 38:1047–1051.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Journal of Natural Language Engineering*. To appear.

Patrik Lambert and Nuria Castell. 2004. Alignment of parallel corpora exploiting asymmetrically aligned phrases. In *Proc. of the LREC 2004 Workshop on the Amazing Utility of Parallel and Comparable Corpora*, May.

Diane Litman and S. Pan. 2002. Desiging and evaluating an adaptive spoken dialogue system. *User Modeling and User-adapted Interaction*, 12(2/3):111–137.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

I. Dan Melamed. 1998. Annotation style guide for the blinker project. Technical Report IRCS 98-06, University of Pennsylvania.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447.

David A. Smith and Noah A. Smith. 2004. Bilingual parsing with factored estimation: Using english to parse korean. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

David G Stork. 2001. Toward a computational theory of data acquisition and truthing. In *Proceedings of Computational Learning Theory (COLT 01)*.

J. Wiebe. 2002. Instructions for annotating opinions in newspaper articles. Technical Report TR-02-101, University of Pittsburgh, Pittsburgh, PA.

Fei Xia, Martha Palmer, Nianwen Xue, Mary Ellen Ocurowski, John Kovarik, Fu-Dong Chiou, Shizhe Huang, Tony Kroch, and Mitch Marcus. 2000. Developing guidelines and ensuring consistency for chinese text annotation. In *Proceedings of the Second Language Resources and Evaluation Conference*, Athens, Greece, June.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *Proceedings of the Second Meeting of the North American Association for Computational Linguistics*, pages 200–207.

[4] The annotation interface is open to public. Please visit `http://flan.cs.pitt.edu/~hwa/align/align.html`