Task#5: Multilingual Chinese-English Lexical Sample Task

Peng Jin,Yunfang Wu, Shiwen Yu
Institute of Computational Linguistics, Peking University
Beijing, China, 100871
{jandp, wuyf, yusw}@pku.edu.cn

1 Introduction

In this paper we put forward the multilingual Chinese-English lexical sample task. The goal of this task is to create a framework to evaluate Chinese word sense disambiguation and to stimulate relative researches.

Following the leading idea of the Senseval-3 Multilingual English-Hindi lexical sample task, we use the Chinese contexts for SemEval-2007. The "sense tags" for the ambiguous Chinese target words are given in the form of their English translations.

We will provide 40 Chinese polysemous words: 20 nouns and 20 verbs, and each sense of one word will be provided at least 15 instances, in which around 2/3 is used as the training data and 1/3 test data.. The translator comes from the Chinese Semantic Dictionary (CSD) developed by the Institute of Computational Linguistics, Peking University (ICL/PKU). The texts will be extracted from the corpus of People's Daily News, which have been word segmented and POS-tagged. The semantically ambiguous target words will be manually sense tagged with their English equivalents.

2 Sense Inventory Representation

The sense inventory used here is the English translator that comes from Chinese Semantic Dictionary (CSD) developed by ICL/PKU. The sense distinctions are made mainly according to the Contemporary Chinese Dictionary, the most widely used dictionary in mandarin Chinese, with necessary adjustment and improvement is implemented according to words usage in real texts. Word senses are described using the feature-based formalism. The features, which appear in the form "Attribute =Value", can incorporate extensive distributional information about a word sense. The feature set, constitutes the representation of a sense, while the verbal definitions of meaning serve only as references for human use. The English translation is assigned to each sense in the attribute "English translation" in CSD.

If the same English translation is assigned to several different senses, we will integrate them into only one sense. For instance, as shown in Table 1, the same English translation "enter" corresponds to both senses 1 and 3 of the verb "进", so sense 1 and 3 are merged into one sense in the multilingual Chinese-English lexical sample task.

Table 1. The English translation description of verb "进"

Senses ID	CSD Chinese sense description	English translation
1	从外面进入到里面	enter
2	吃饭	eat
3	步入先进水平（抽象事物）	enter
4	用在动词后做补语	into

Senses ID	CSD Chinese sense description	English translation
5	呈上	submit
6	比赛中升入名次	survive
7	向前移动	move forward
8	购进	buy in

3 Source Corpora

All the training and test data come from the corpus of People's Daily, which have been word-segmented and POS-tagged according to the POS tagging scheme of ICL/PKU.

4 Sense Tagged Data

The sense tagged corpus is manually constructed with the help of a word sense tagging interface developed in Java. Three annotators, two major in linguistics and one major in computer science will take part in the construction of sense-tagged corpus. A text generally is first tagged by one annotator and then verified by two checkers. Controversies over word senses will be resolved by discussion.

Checking is of course a necessary procedure to keep the consistency. Checking all the instances of a word in a specific time frame will greatly improve the precision and accelerate the speed just as in the tagging process. A software tool is designed in Java to gather all the occurrences of a word in the corpus into a checking file with the sense KWIC (Key Word in Context) format in sense tags order.

Based on this sense tagged corpus, we replace the sense id with the corresponding English translator. If different senses corresponded to the same translator, we labeled them with the same translator.

5 Evaluation Method

The training sense tagged data will be distributed to all participants. The test data will be used to evaluate the participating systems, where the target ambiguous words are explicitly marked, and the participants are required to assign one unique translator to each instance. And an answer key file will be provided as a separate one.