Overcoming Language Barriers in the Internet Era

- A Foreign Language Reading Assistance system

Hang Li， Yunbo Cao, and Cong Li

2002-9-12

Technical Report
MSR-TR-2002-91

Microsoft Research, Asia

5F Sigma Center, No.49 Zhichun Road, Haidian District
Beijing, China, 100080

Overcoming Language Barriers in the Internet Era

- A Foreign Language Reading Assistance System

Hang Li, Yunbo Cao, and Cong Li
Microsoft Research Asia
5F Sigma Center, No.49 Zhichun Road, Haidian District, Beijing China
{hangli, i-yuncao, i-congl}@microsoft.com

The traditional problem that exists from the time of Tower of Babel becomes more serious in the internet era. That is how we read and write foreign languages. Figure 1 shows an estimate of the distribution of languages on the web. We see that about three fourths of the web pages are in English and non-English speakers have needs to read the documents. On the other hand, English speakers cannot read roughly one fourth of the web pages in other languages. We are more challenged than people in any other era by the language barriers that stand in our way.

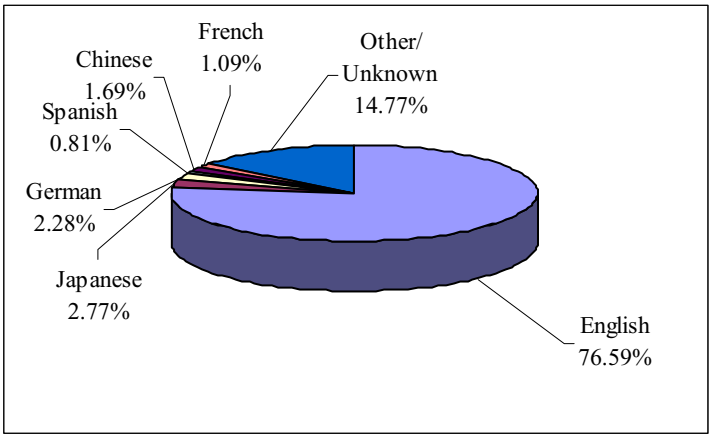

Figure 1: Language Shares on the Web
(http://www.statistics.com/content/datapages/data5.html)

Our proposal here is to use multi-lingual data on the web to help overcome the difficulties in reading foreign languages. Specifically, we describe how we use statistical machine learning techniques to perform *intelligent* foreign language reading assistance.

Starting from 'Reading Assistance'

Figure 2 shows a sample of Chinese-English translation results output by a machine translation system. English speakers can get a rough sense of what the original Chinese text is describing, but they may have difficulties in understanding the details. This example indicates that full machine translation has made substantial achievements, but its quality has not yet reached a satisfactory level.

素闻京城为蛮荒之地，几近于沙漠，不降甘霖，飞沙走石，夏日如炙，冬寒侵人，恐怖异常。不料北京地界内，竟有一处龙庆峡，山青水秀，景色宜人。近日众人于小雨之中游览，更觉气息清新凉爽，山水交映若画，如临仙境。

The element hears the national capital place for 蛮 the uncultivated land, several nearly to the desert, does not fall the timely rain, the blown sand walks the stone, the summer day like 炙, the winter cold invades the person, the terror is unusual. But unexpectedly Beijing 地界 inside, unexpectedly has one dragon celebrates the canyon, mountain blue water Xiu, the scenery is pleasant. Recently the numerous people toured inside drizzle, sense the breath fresh was cool, the scenery junction reflected if picture, like near fairyland.

Figure 2: Example Translation Results

In this report, we consider an alternative approach to full machine translation, i.e., foreign language reading assistance. The basic feature of such a system is translation dictionary consultation (i.e., translation at the word or phrase level).

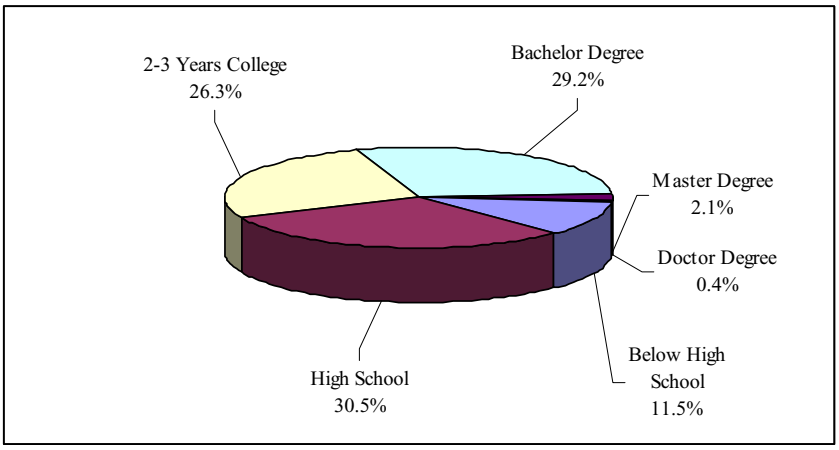

Figure 3: Educational Degrees of Internet Users in China (http://www.cnnic.net.cn/)

Figure 3 shows a survey result on educational degrees of internet users in China. Nearly ninety percent of the users have education backgrounds beyond high school and they can read English, although their English reading abilities may vary. Therefore, at least for many Chinese internet users to read English, a reading assistance tool would be more helpful, given the fact that full machine translation still needs significant improvements. The situations in other Asian countries such as Japan and Korea are very similar.

Our English reading assistance system, which we call English Reading Wizard (ERW), provides two basic features: dictionary consultation by mouse hovering and dictionary consultation in a reference window. Figure 4 displays the working of the two features. When a user puts the cursor on a word (or a phrase), ERW shows in a pop-up menu the translations of the word. When a user searches a word (or a phrase) in the reference window, ERW shows the detailed explanations on the translations. Currently ERW supports English-Chinese and English-Japanese translations.

In order to make ERW more useful, we have developed two new features which utilize multi-lingual data on the web and employ statistical machine learning techniques.

The first new feature is to automatically extract translations of words or phrases from the web, especially when no translation can be found in the ‘local’ dictionary. That is to say, this feature is designed to deal with the ‘out of vocabulary’ problem which often plagues ERW.

The second new feature is to rank translations of words or phrases based on contexts. As in many cases there are ambiguities in translations, and thus putting the correct translations on the top will save users time in dictionary consultation. We use data on the web to train classifiers for translation ranking.

There are several commercial products for foreign language reading assistance (e.g., Ciba http://www.iciba.net/). None of them, however, has the intelligent features which ERW offers. There are also some studies on extracting translation knowledge from web (e.g., Nagata et al. 2001). They are, however, pursued for constructing a lexicon, not for creating a reading assistance system.

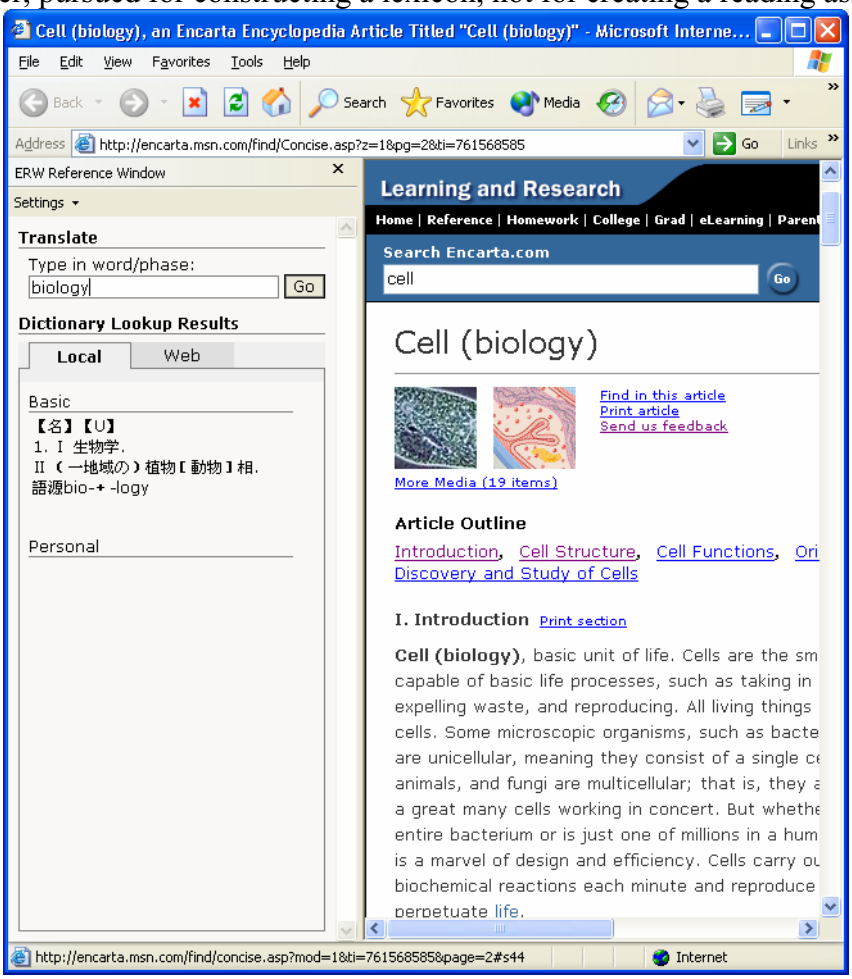

Figure 4: Dictionary Consultation

Extracting Translations from Web

ERW can extract translations from web under the help of a search engine. Figure 5 shows the result when a user looks for translations of the phrase 'dendritic cell' (from English to Japanese). It also gives the links of example web pages, which contain both the original phrase and the translations. By looking at the pages, the user can understand how the phrase and its translations are used.

This feature can help find translations which do not exist in the local dictionary according to our experiments. In one experiment, we used ERW to extract translations for 1000 noun-noun pairs, and found that for 72.9% of the noun-noun pairs, ERW was able to find *correct* translations. Among the noun-noun pairs for which correct translations have been found, 11% of them (e.g., 'opera buffa') were those whose translations cannot be obtained by just looking up the translations of each of its nouns in the local dictionary (i.e., compositionally creating translations).

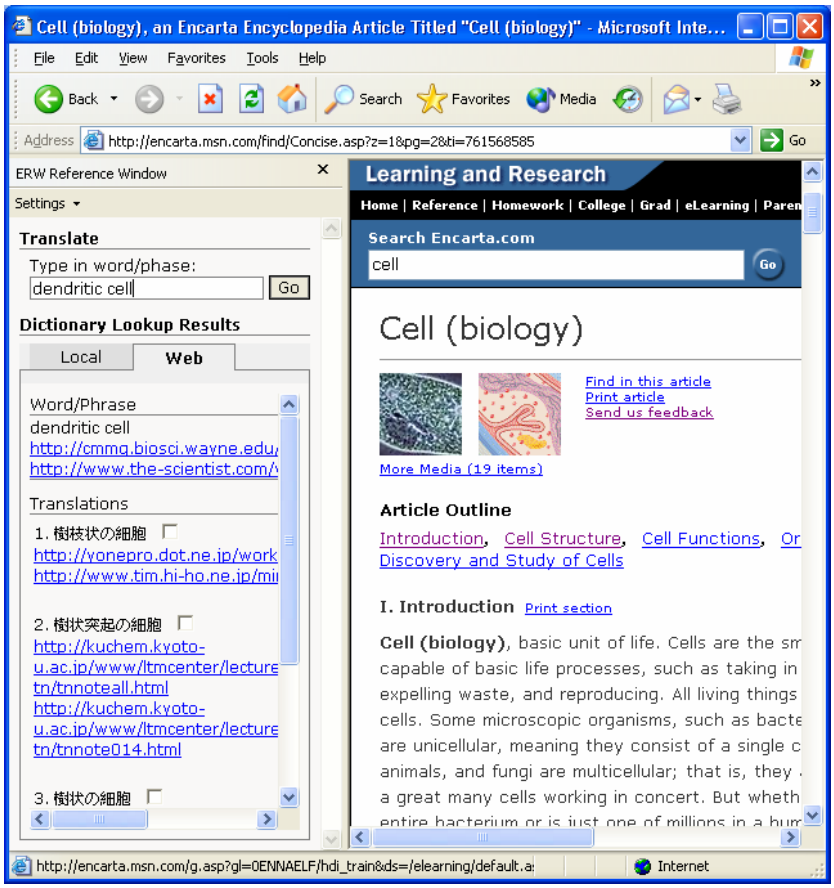

Figure 5: Translation Extraction from Web

ERW relies on two methods for translation extraction: the partial parallel method and the compositional method.

Nagata et al. (2001) observed that there are many *partial parallel corpora* between English and Japanese (or other Asian languages) on the web, and most typically English translations of Japanese terms are parenthesized and inserted immediately after the Japanese terms in Japanese documents.) The partial parallel method in ERW is based on Nagata et al's observation. Figure 6 illustrates the process of extracting Chinese translations for an English phrase 'information asymmetry' with this method.

1. Input 'information asymmetry';
2. Search the English Base NP on web sites in *Chinese* and obtain documents as follows (i.e., using partial parallel corpora):

公司的控制者和管理者通常掌握着许多外部投资者所不了解的信息，即在内部人与外部人之间存在信息不对称（information asymmetry）。

3. Find the most frequently occurring Chinese phrases immediately before the brackets containing the English Base NP, using a suffix tree;
4. Output the Chinese phrases and their document frequencies:

信息不对称 5
信息失衡 5

Figure 6: Partial Parallel Method

The compositional method, which we have developed (Cao and Li 2002), comprises of two major steps: translation candidate collection and translation selection. In translation candidate collection, it searches translation candidates of a given phrase/word on the web. In translation selection, it finds out translation(s) from the translation candidates on the basis of context similarities.

Figure 5 illustrates the process of collecting Chinese translation candidates for an English phrase 'information age' in translation candidate collection.

1. Input 'information age';
2. Consult English-Chinese word translation dictionary:
   information -> 信息
   age -> 年龄 (how old somebody is)
   时代 (historical era)
   成年 (legal adult hood)
3. *Compositionally* create translation candidates in Chinese:
   信息年龄
   信息时代
   信息成年
4. Search the candidates on web sites in Chinese and obtain the document frequencies of them (i.e., numbers of documents containing them):
   信息时代 10000
   信息年龄 10
   信息成年 0
5. Output candidates having non-zero document frequencies and the document frequencies:
   信息时代 10000
   信息年龄 10

Figure 7: Translation Candidate Collection

Translation selection is based on the following observation: the contexts of a translation tend to be similar to the contexts of the original word or phrase. If the contexts of a candidate are similar enough to those of its original word or phrase, we view the candidate as a possible translation. Our method uses the frequencies of surrounding words as contexts. Details of translation selection can be found in (Cao and Li 2002); we describe here only the basic idea.

Assume that we are to judge if the Chinese phrase '信息时代' is a translation of the English phrase 'information age'. We can obtain, as in Figures 8 and 9, frequency vectors of the two phrases at the same time when we perform translation candidate collection. We next see if the vectors are similar enough. It is not possible, however, to straightforwardly calculate the similarity between the two vectors because they belong to different languages. Our method employs the Expectation and Maximization (EM) algorithm (Dempster et al. 1977) and a translation dictionary (as in Figure 10) to transform a vector from one language into the other (e.g., from Figure 8 to Figure 11).

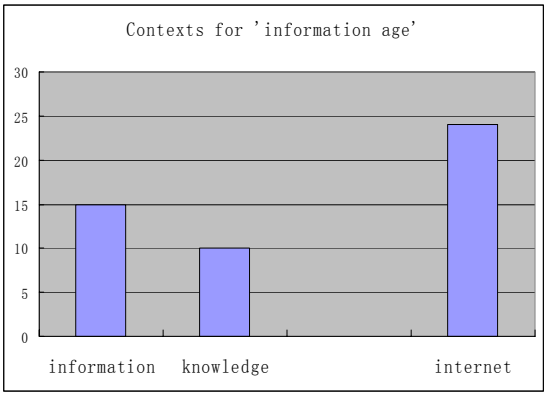

Figure 8: Frequency Vector

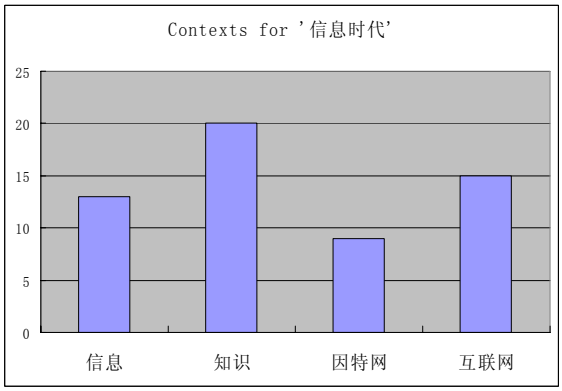

Figure 9: Frequency Vector

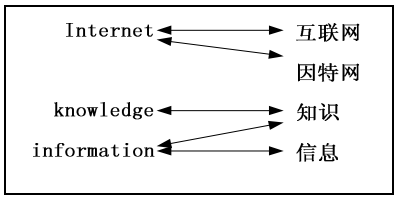

Figure 10: Translation Dictionary

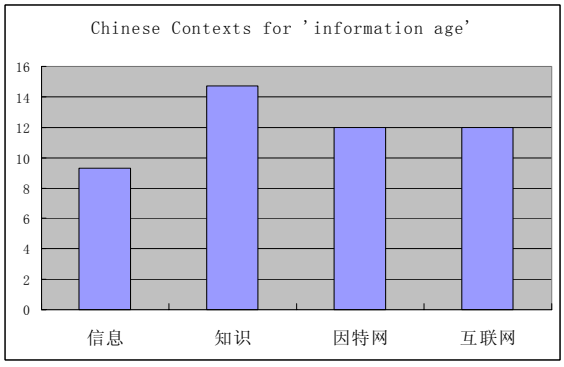

Figure 11: Frequency Vector

Let us denote the contexts of a word or phrase in English as $(f(e_1),f(e_2),..,f(e_m))$, where $f(e_i)$ represents the frequency of context word e_i. We assume the data is generated by the following mixture model:

$$P(e)=\sum_{c\in C}P(c)P(e\,|\,c)$$

where $P(e)$ and $P(c)$ represent the distributions for generating English words e and Chinese words c respectively, and $P(e|c)$ the conditional distribution of generating English words e given Chinese words c. If there is no translation relationship between e and c in the dictionary, we assume that $P(e|c)$ is zero.

We can employ the EM algorithm to iteratively estimate the parameters of the above mixture model. It turns out to utilize the many-to-many mapping relationship between the words in English and Chinese to distribute the frequencies of words from English into Chinese (from Figure 8 to Figure 11). After the transformation, we can conduct similarity calculation between frequency vectors (Figure 9 and Figure 11).

Fung and Yee (1998) also proposed to conduct translation selection on the basis of similarities between context vectors. However, they assumed that there is only one-to-one mapping relationship between the words in a dictionary. That is too strict an assumption in practice. Experimental results indicate that our method performs *significantly* better than their method (see Cao and Li 2002).

Translation Ranking based on Context

ERW can rank translations based on contexts (also under the help of a search engine). For example, the word 'plant' has translations '工厂 (gongchang)' (factory) and '植物 (zhiwu)' (vegetation) in Chinese. When the sentence is "a new automated manufacturing plant will be built in the city", the former translation will be ranked on the top, and when the sentence is "there are lots of plant and animal species in this area", the latter translation will be ranked on the top.

This feature can significantly reduce human efforts in dictionary consultation according to our experiments. 'Effort' here is defined as average number of translations which have to be read until the correct translation is found. Table 1 presents the evaluation results in terms of effort with respect to the ambiguous words 'interest' and 'line'. 2291 sentences containing 'interest' and 4419 sentences containing 'line' were used for the evaluation. We take as baseline the method of ranking translations in descending order of their frequencies. From the table, we see that our ranking method significantly improves upon the baseline method.

Table 1: Efforts in Dictionary Consultation

Word	Number of translations	Efforts		Efforts reduction
		Ranking based on translation frequency	Ranking based on context	
interest	4	1.81	1.37	24.4%
line	6	2.35	1.83	22.1%
Average		2.16	1.67	22.7%

The translation ranking task can be regarded as a classification problem in which English sentences are examples, and the correct translations of the target word (e.g., 'plant') in the respective sentences are classification decisions. A supervised learning method can be used in advance to construct classifiers for translation disambiguation (cf., Mitchell 1997). The classifiers treat context words of the target word as features and assign probabilities to its translations. Since supervised learning methods need labeled data which is expensive to create, we have developed a new *unsupervised* method which effectively uses a small number of labeled data and a large number of unlabeled data. We call this method Bilingual Bootstrapping. Since the data on the web are by nature *unlabeled*, we can use them to perform Bilingual Bootstrapping. This is exactly what we do

in ERW for translation ranking. We refer to (Li and Li 2002) for details of Bilingual Bootstrapping, and introduce here only the basic idea.

Figure 13 shows the process of creating a classifier for 'plant' with Bilingual Bootstrapping. There are also two other classifiers involved on the Chinese side, which correspond to 'gongchang' and 'zhiwu' respectively. This is because the two words in Chinese also have *translation* ambiguities. Figure 12 shows the translation relationship between the words.

As shown in Figure 13, Bilingual Bootstrapping makes use of a small number of labeled sentences collected from a dictionary and a large number of unlabeled sentences collected from the web. It first constructs three classifiers by using labeled data in both languages. This is possible because it can automatically transform labeled data from one language into the other using a translation dictionary and the EM algorithm.

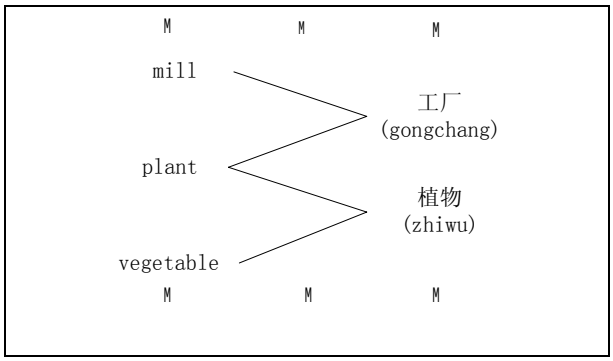

Figure 12: Translation Relationship between Words

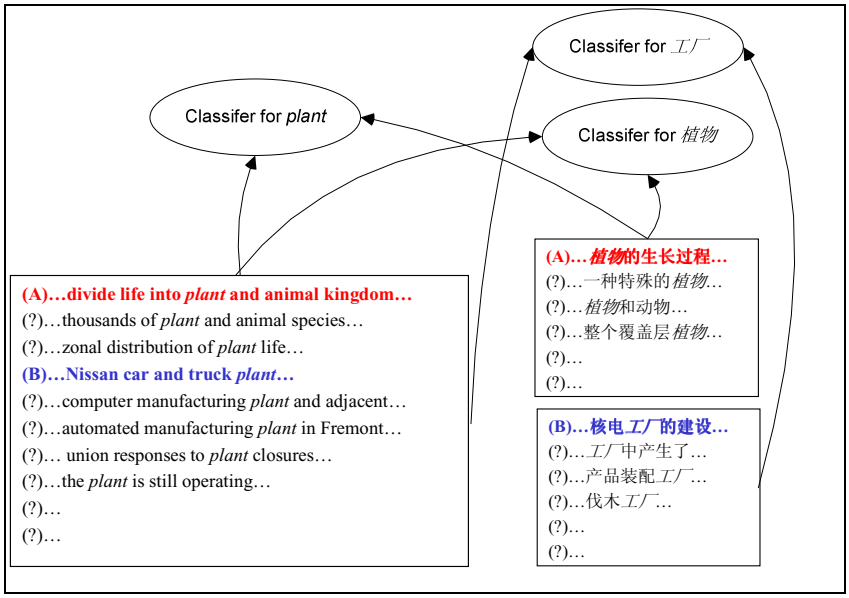

Figure 13: Bilingual Bootstrapping (1)

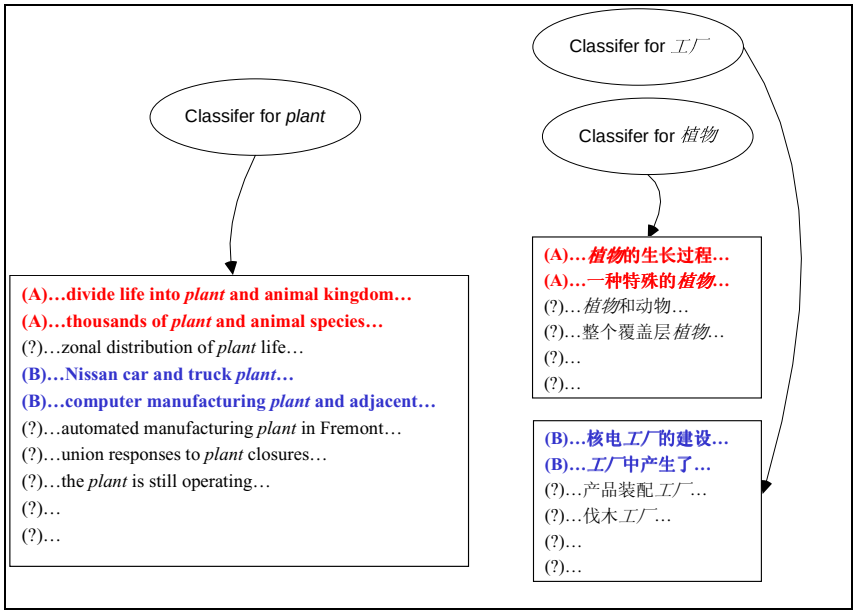

Figure 14: Bilingual Bootstrapping (2)

As shown in Figure 14, Bilingual Bootstrapping next uses the constructed classifiers to further label some unlabeled sentences. It repeats the procedures in Figures 13 and 14, until no sentence can be labeled.

Yarowsky (1995) proposed a bootstrapping method (i.e., an unsupervised method) for translation disambiguation. Because it is one conducted only in a single language (here in English), we refer to it as Monolingual Bootstrapping. Experimental results indicate that Bilingual Bootstrapping can *significantly* outperform Monolingual Bootstrapping. The better performance of Bilingual Bootstrapping can be attributed to its successful use of data in *two languages* (see Li and Li 2002 for detail).

We note that Bilingual Bootstrapping fits well with the internet environment. First, since we perform Bilingual Bootstrapping in a word-by-word fashion, it is easy to employ a web search engine to collect data. For example, when we construct a classifier for 'plant' we only need to obtain from web those sentences which contain the word. Second, because we conduct Bilingual Bootstrapping using bilingual data, it is beneficial to use the large amount of multi-lingual data available on the web.

Using Web More Effectively

Machine translation is still a challenging task for humans. Among the problems we need to resolve, words or phrases translation is no doubt on the top priority, because words and phrases are 'atoms and molecules' of languages.

In this report, we have described two new technologies which effectively use multi-lingual data on the web to conduct intelligent word or phrase level translation. We note that web offers many opportunities for realizing high-quality translation. We conclude this report by giving an example.

Suppose that we have a server and clients (Figure 15). Users can send word or phrase translation requests from the clients and the server can extract translations from web as we do in ERW. After that, users choose the translations they consider correct and send them back to the server. The

server views the results as those of a voting on correct translations and updates its translation dictionary accordingly. The dictionary on the server thus becomes a *live* translation dictionary jointly created by internet users.

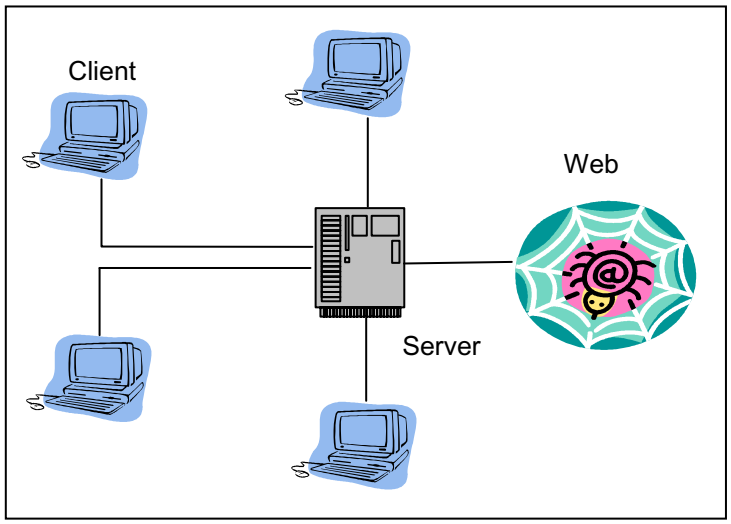

Figure 15: Translation Voting on Web

Acknowledgements

We are very grateful to Ming Zhou, Chang-Ning Huang, Jianfeng Gao, Chris Pratley, and Eric Brill for their many valuable suggestions on this work. We thank Yuan-Yuan Zhang and Zhanyi Liu for their comments on this report.

References

Y. Cao and H. Li, 2002. Base Noun Phrase Translation Using Web Data and the EM Algorithm. In *Proceedings of the 19th international Conference on Computational Linguistics*.

A. P. Dempster, N. M. Laird, and D. B. Rubin, 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society B*, vol. 39, pp. 1-38

P. Fung and L.Y.Yee, 1998. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In *Proceedings of the 18th International Conference on Computational Linguistics and the 36th Annual Meeting of Association for Computational Linguistics*, pp. 414-420.

C. Li and H. Li, 2002. Word Translation Disambiguation Using Bilingual Bootstrapping. In *Proceedings of the 40th Annual Meeting for Computational Linguistics*, pp. 343-351.

T. Mitchell, 1997. *Machine Learning*. McGraw Hill.

M. Nagata, T. Saito, and K Suzuki, 2001. Using the Web as a Bilingual Dictionary. In *Proceedings of ACL 2001 Workshop on Data-Driven Methods in Machine Translation*, pp. 95-102.

D. Yarowsky, 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33rd Annual Meeting for Computational Linguistics*, pp.189-196.