Using Bilingual Web Data to Mine and Rank Translations

Hang Li, Yunbo Cao, and Cong Li, *Microsoft Research Asia*

The English Reading Wizard uses bilingual Web and local-dictionary data to help readers understand foreign languages by translating words and phrases. Methods include the Expectation and Maximization algorithm and bilingual bootstrapping.

In the Internet era, the traditional Tower of Babel problem—how we read and write foreign languages—has become even more serious. According to research, about three-fourths of the Web pages that non-English speakers need to read are in English, while for English speakers, roughly one-fourth of the pages are in other languages (see www.statistics.com/content/datapages/data5.html).

We propose using multilingual Web data and statistical-learning methods to help readers understand foreign languages. We've created an intelligent English reading-assistance system that offers word and phrase translation with automatic mining and ranking features based on these methods.

English Reading Wizard

Full machine translation has made substantial achievements, but its quality hasn't reached a satisfactory level. Figure 1 shows such a system's Chinese-to-English translation. English speakers can get a rough sense of what the original Chinese text describes, but they'll probably have difficulties understanding the details. (For an example machine translation system, see Babelfish, http://babelfish.altavista.com.)

Nearly 90 percent of Internet users in China have educational backgrounds beyond high school, and they can read English, although their abilities vary (see www.cnnic.net.cn). For many of them, therefore, a reading-assistance tool would be more helpful than full machine translation. The situations in other Asian countries such as Japan and Korea are very similar.

Our English reading-assistance system, English Reading Wizard, provides dictionary consultation for words and phrases through two basic features: mouse hovering and searching. When a user puts the cursor on a word such as *cellular*, ERW displays the word and its translations in a pop-up menu (as shown in the lower part of Figure 2). When a user searches for a word such as *biology* by typing it in the reference window on the left, ERW displays the detailed translation under Dictionary Lookup Results. Local dictionary consultation by searching operates when the local tab is chosen in the reference window, which has both basic and personal translations. The latter is obtained from a user-compiled dictionary. ERW supports English-to-Chinese and English-to-Japanese translations.

To make ERW easier to use, we've developed two advanced features. The first, *translation mining*, automatically extracts the translations of words and phrases from the Web when no translation can be found in the local computer dictionary. This feature deals with the local "out of vocabulary" problem that often plagues a foreign language reading-assistance system.

The second advanced feature, *translation ranking,* sorts the translations of words or phrases into lists based on contexts. Because many translations contain ambiguities, putting the correct translations on the top of the translation list saves users time in dictionary consultation. This feature ranks translations existing in the local dictionary.

Several commercial products exist for foreign language reading assistance, such as Ciba (www.iciba.net), and related research has been conducted,[1] but no other product offers ERW's advanced features.

1094-7167/03/$17.00 © 2003 IEEE
Published by the IEEE Computer Society

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 1. Example Chinese-to-English results from a full machine-translation system.

Translation-mining feature

ERW extracts translations from the Web using a search engine. Figure 3 shows how translation mining works. Specifically, it shows the extracted result when a user searches for the translations of the phrase *dendritic cell* (from English to Japanese) with the Web tab selected. ERW also provides the links of example Web pages that contain both the original phrase and the translations. By looking at these pages, the user can understand how the phrase and its translations are used.

In one experiment, we used ERW to extract translations for 1,000 noun-noun pairs and found that it located correct translations for 72.9 percent of them. Of these, 11 percent (such as *opera buffa*) were translations that couldn't be obtained by just looking up the translation of each noun in the local dictionary (that is, using the compositional method to create translations).

Translation-mining method

ERW relies on both the partial parallel method and the compositional method for translation mining. Many *partial parallel corpora* exist between English and Chinese (or other Asian languages) on the Web. In these corpora, sporadically interlaced English translations appear in parentheses immediately after the terms in other languages. We based our partial parallel method on this observation. Figure 4 illustrates the process of extracting Chinese translations of the English phrase *information asymmetry* with this method.

The compositional method, which we developed,[2] has two steps: translation-candidate collection and translation selection. ERW first searches translation candidates of a given phrase on the Web and then finds the translations from the candidates.

Figure 5 illustrates the process of collecting Chinese translation candidates for the English phrase *information age*.

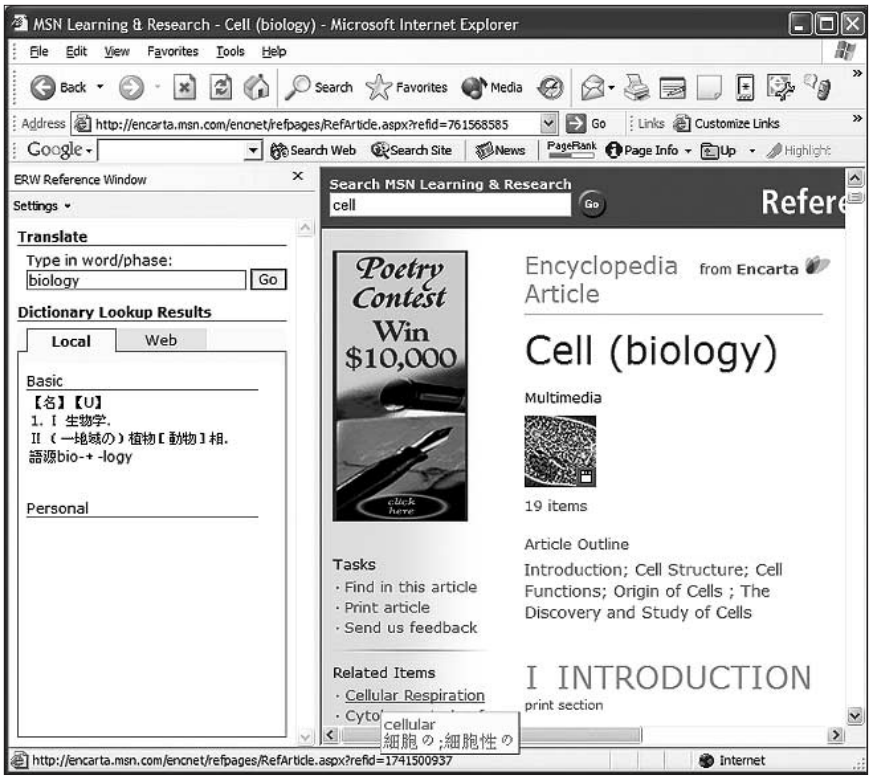

Figure 2. English Reading Wizard mouse-hovering and searching results in Japanese.

The importance of context

We based translation selection on our observation that a translation's context tends to be similar to that of the original phrase. If a candidate's context is similar enough to that of its original phrase, we view the candidate as a possible translation. Our method uses the surrounding words' frequencies to represent contexts. (Details of translation selection appear elsewhere.[2])

Say we're judging whether the Chinese phrase *xinxishidai*, obtained from translation-candidate collection, is the correct translation of the English phrase *information age* (actually, it is). The context words of *information age* in English are *Internet*, *knowledge*, and *information*. The context words of *xinxishidai* in Chinese are *hulianwang, yintewang*, *zhishi*, and *xinxi*. Figure 6 shows the translation relationship between the context words in the two languages. If two words in the two languages can be each other's translations, ERW links them in the graph. We see that a many-to-many mapping relationship exists between the context words in the two languages.

We obtain the context words' frequencies for each of the two phrases *information age* and *xinxishidai* at the same time we perform translation-candidate collection. We combine the frequencies for each phrase into a vector (see Figure 6). Vectors A and B should be the same as a result of one-to-one mapping. When vectors A and D are similar enough, we view *xinxishidai* as the correct translation

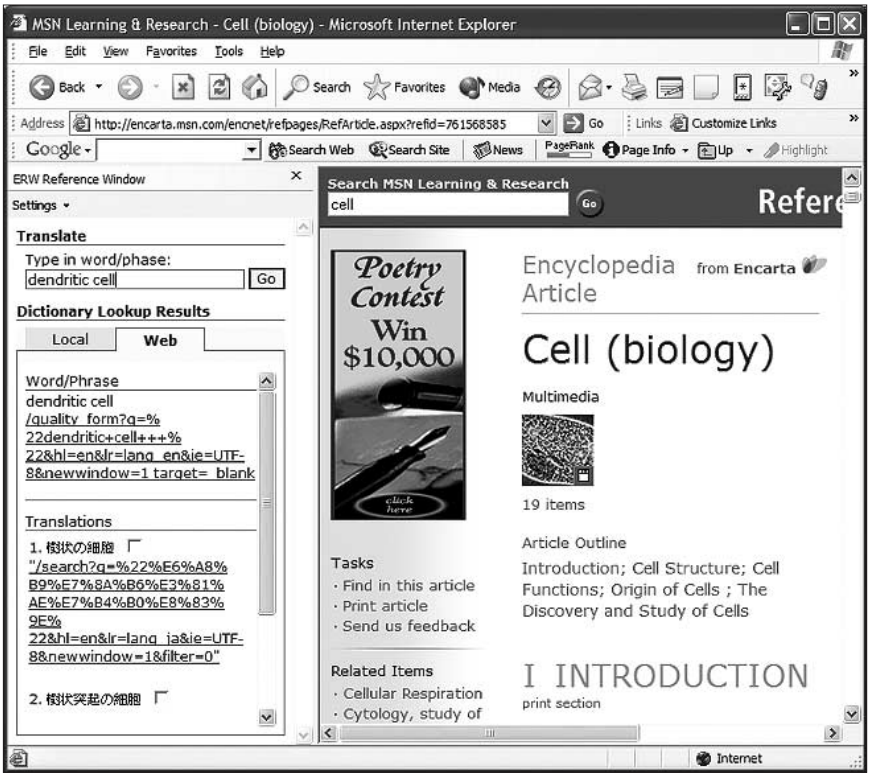

Figure 3. Translation-mining results from the Web in Japanese.

1. Input *information asymmetry*.
2. Search the English phrase on Web sites in Chinese and obtain documents as follows (that is, 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 phrase, 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 4. Partial parallel method of extracting Chinese translations.

of *information age*. Because one phrase can have several translation candidates, the problem is sorting the translation candidates in descending order of their similarity values. We evaluate the similarities through measures not discussed here.

Expectation and maximization algorithm

The fact is, you can't straightforwardly calculate the similarity between the two vectors because they don't belong to the same language (the same space). Our method employs the Expectation and Maximization (EM) algorithm[3] and a translation dictionary of context words (as Figure 6 shows) to transform a vector from one language into the other. In the example, our method transforms vector A into vector C, which is close to vector D, as expected (this transformation doesn't appear in the figure). The merit of using EM here is that even the translation relationship between the context words in the two languages is a many-to-many mapping. EM can still split the context word frequency in one language and distribute the frequencies into its translations in the other in a theoretically sound way.

Comparing methods

Many methods have been proposed for word and phrase translation mining.[4–7] ERW's partial parallel method follows Masaaki Nagata and his colleagues' similar proposal.[4] The translation-candidate collection process in the compositional method is similar to Gregory Grefenstette's proposal.[5] However, our translation-selection process is an improvement over Pascale Fung and Lo Yuen Yee's method.[6]

Fung and Yee assumed that only a many-to-one mapping (or a one-to-one mapping) relationship exists between the context words and that you can straightforwardly transform a vector from the source language into the target language. However, this approach is too strict in practice. When the relationship between the context words is a many-to-many mapping, Fung and Yee's method must cut some links (such as the dashed lines in the graph in Figure 6) to forcibly create a many-to-one mapping. Because of this, in this example, vector A will be transformed into vector B. It turns out that vectors B and D are quite different, however, although they should be similar. Because C is A's transformation and we expect A is similar to D, we also expect C is similar to D. Experimental results indicate that our method performs significantly better than Fung and Yee's.[2] If the relationship between context words is many-to-one mapping, however, our EM method equals theirs.

Translation-ranking feature

The word *plant* translates to both *zhiwu* and *gongchang* in Chinese. The former corresponds to the sense "vegetation," the latter to "factory." When the sentence is, "There are lots of plant and animal species in this area," ERW ranks the former translation higher in terms of context. When the sentence is, "A new automated manufacturing plant will be built in the city," ERW ranks the latter translation higher.

This feature can significantly reduce human effort in dictionary consultation. We define *effort* as the average number of translations that users must read until the correct translation is found in the translation list. We assume that users read the translation list from the top to the bottom when a word or phrase has several translations.

Table 1 presents the evaluation results in terms of effort with respect to the ambiguous words *interest* and *line*. *Interest* has four senses in Chinese, and *line* has six. We used 2,291 sentences containing *interest* and 4,419

sentences containing *line* for the evaluation. We took as the baseline the method of ranking translations in descending order of their frequencies. In the evaluation, we knew the correct translations for *interest* and *line* when given a sentence as context. The ranking that had the correct translation on the top is the best one. In practice, we use the correction translation's rank to measure human effort. For example, if the correct translation appears in the second position on the list, the effort is "2." We average the effort cost over all sentences in the data set. From the table, we see that our ranking method significantly improves the baseline method.

Translation-ranking method

You can regard the translation-ranking task as a classification problem in which English sentences are examples and the correct translations of the target word (such as *plant*) in the respective sentences are classification decisions. You can use a supervised learning method in advance to construct classifiers for translation disambiguation.[8] The classifiers treat the target word's context words as features and assign probabilities to the target word's translations. Because supervised learning methods need labeled data that is expensive to create, we developed a new *unsupervised* method that effectively uses a small number of labeled data and a large number of unlabeled data. We call this method *bilingual bootstrapping*. Because Web data is by nature unlabeled, we can use it to perform part of bilingual bootstrapping. This is exactly what we do in ERW for translation ranking. (The details of bilingual bootstrapping appear elsewhere.[9])

Bilingual bootstrapping's labeled and unlabeled sentences are in both English and Chinese. This method first constructs classifiers in both languages by using all labeled data. It automatically transforms labeled data from one language into the other using a translation dictionary and the EM algorithm. Bilingual bootstrapping next uses the constructed classifiers to further label some unlabeled sentences in both languages. It repeats these procedures until no further sentences can be labeled. When a classifier tries to label a sentence, its confidence level must be higher than an empirically determined threshold, say 0.9. At some time point, the two classifiers can't label any more unlabeled sentences with confidences higher than this threshold, so the iteration stops.

Figure 7 shows the translation relationship between the words *plant, gongchang,* and *zhiwu*. Here, *mill* and *plant* (*factory*) are different senses for the Chinese word *gongchang,* and *plant* (*vegetation*) and *vegetable* are different senses for *zhiwu*. That is to say, *gongchang* and *zhiwu* can also be ambiguous. Figure 8 shows how this problem is resolved.

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 adulthood)
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 their document frequencies (that is, numbers of documents containing them):
   信息时代 10,000
   信息年龄 10
   信息成年 0
5. Output candidates having nonzero document frequencies and the document frequencies:
   信息时代 10,000
   信息年龄 10

Figure 5. Translation-candidate collection.

Table 1. Translation-ranking effects.

Word	Number of translations	Efforts: Ranking based on translation frequency	Efforts: Ranking based on context	Effort reduction (%)
Interest	4	1.81	1.37	24.4
Line	6	2.35	1.83	22.1
Average	5	2.16	1.67	22.7

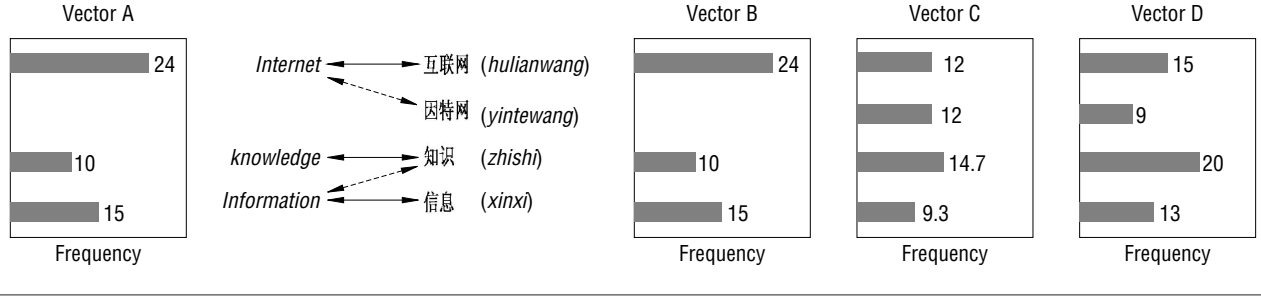

Figure 6. Translation relationship between context words and example frequency vector transformation.

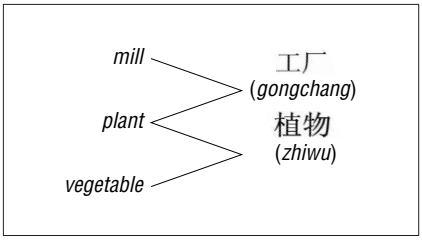

Figure 7. Translation relationship between English and Chinese words.

Figure 8a uses sentences containing *plant*, *gongchang,* and *zhiwu* to show the training procedure, constructing classifiers from labeled sentences. Figure 8b shows how the labeling procedure labels new sentences using the newly constructed classifiers.

At first, as seen on the left side of Figure 8a, sentences E1 and E4 have labels A and B, respectively. On the right side, sentences C1 and C5 receive labels A and B. Here, A represents the sense "vegetation," and B the sense "factory." Other sentences are unlabeled. Bilingual bootstrapping uses labeled sentences E1, E4, C1, and C5 to create a classifier for *plant* disambiguation (between A and B). It also uses labeled sentences E1 and C1 to create a classifier for *zhiwu* and uses labeled sentences E4 and C5 to create a classifier for *gongchang*. Bilingual bootstrapping next uses the *plant* classifier to label sentences E2 and E5 (Figure 8b). It uses the *zhiwu* classifier to label sentence C2 and the *gongchang* classifier to label sentence C6, repeating this process until no further sentences can be labeled.

Data collection

We first collected the labeled data from a dictionary. In our experiment with *plant*, we used the word *industry* as a pseudosentence, obtained from the definition of the sense "factory." We used *life* as a pseudosentence, obtained from the definition of the sense "vegetation." As we discussed earlier, our method needs a few labeled sentences for each word as a starting point and then constructs an initial classifier and a label.

Bilingual bootstrapping also fits well into the Internet environment. First, because we perform bilingual bootstrapping word by word, you can easily employ a Web search engine to collect data. When we construct a *plant* classifier, we need to obtain from the Web only those sentences that contain the word. Second, because bilingual bootstrapping uses bilingual data, it's beneficial to access the large amount of multilingual data on the Web.

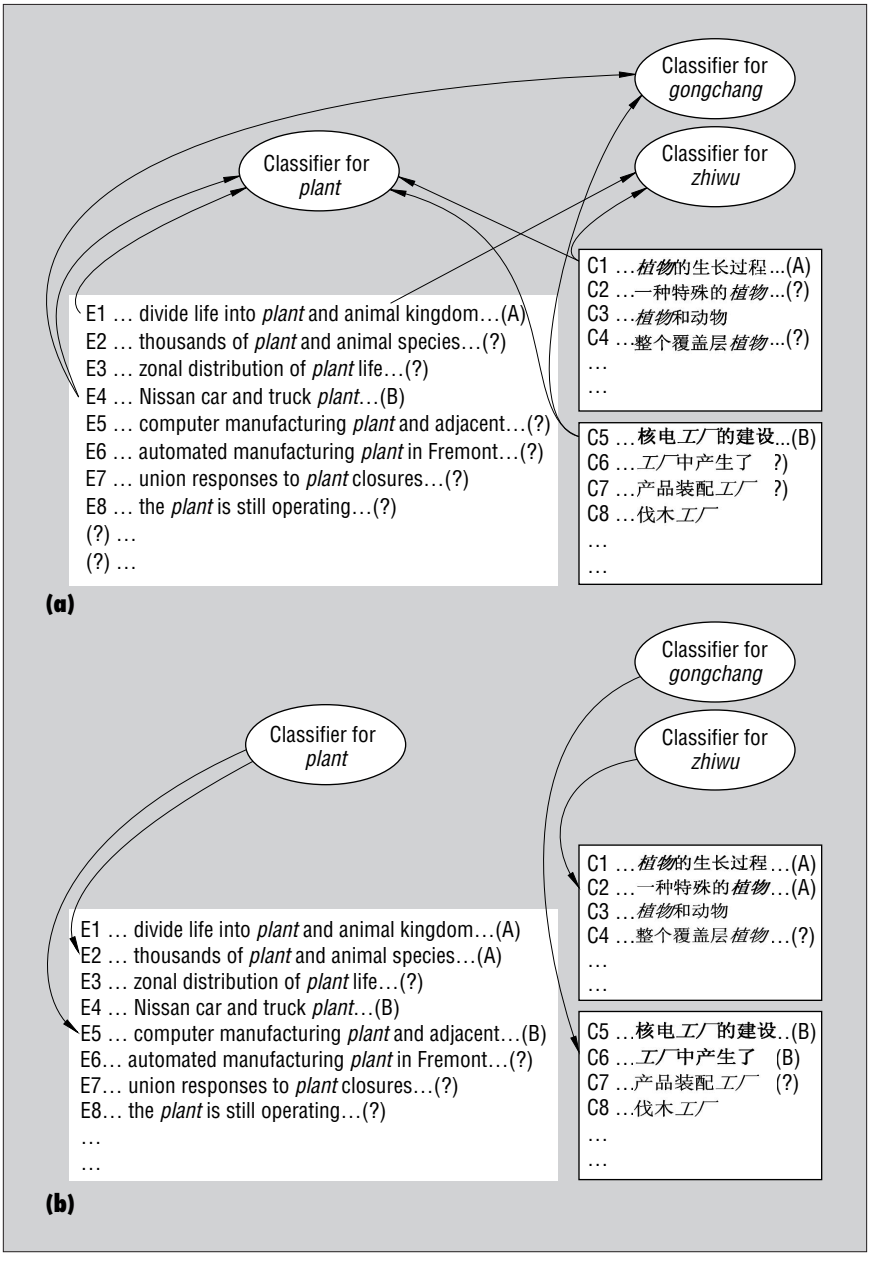

Figure 8. Bilingual Bootstrapping: (a) constructing classifiers from labeled sentences, and (b) labeling new sentences using the newly constructed classifiers.

Performance

David Yarowsky[10] proposed a bootstrapping (unsupervised) method for translation disambiguation. Because it's conducted in only one language (here, in English), we refer to it as monolingual bootstrapping. Our experimental results indicate that bilingual bootstrapping significantly outperforms monolingual bootstrapping.[9]

Bilingual bootstrapping achieves higher performance because it effectively uses the translation relationship between ambiguous words in the two languages. Most sentences containing the word *zhiwu* (such as C1 and C2) should be labeled A, and most sentences containing *gongchang* (C5 and C6) should be labeled B. That is, senses A and B are represented as two words in Chinese, and the two words aren't ambiguous. Thus, these sentences are good examples of constructing the *plant* classifier in English when they are transformed from Chinese into English.

The Internet is a rich source of data for conducting machine translation, not only in terms of data size but also because it contains parallel, link, and glossary data types. However, more efforts are needed to develop sophisticated technologies for using Web data and to introduce breakthroughs to the field. ■

Acknowledgments

We thank Ming Zhou, Chang-Ning Huang, Jianfeng Gao, Eric Brill, Jeff Reynar, Masaaki Nagata, and Chris Pratley for their many important suggestions on this work. We thank Yuan-Yuan Zhang, Zhanyi Liu, and the two anonymous reviewers for their many helpful comments on this article.

References

1. D. Bauer, F. Segond, and A. Zaenen, "LOCOLEX: The Translation Rolls off Your Tongue," *Proc. Joint Int'l Conf. Assoc. Computer and Humanities and Assoc. Literary and Linguistic Computing,* 1995, pp. 6–8.
2. Y. Cao and H. Li, "Base Noun Phrase Translation Using Web Data and the EM Algorithm," *Proc. 19th Int'l Conf. Computational Linguistics*, 2002, pp. 127–133.
3. A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Soc.: Series B*, vol. 39, 1977, pp. 1–38.
4. M. Nagata, T. Saito, and K. Suzuki, "Using the Web as a Bilingual Dictionary," *Proc. ACL 2001 Workshop Data-Driven Methods in Machine Translation,* 2001, pp. 95–102.
5. G. Grefenstette, "The WWW as a Resource for Example-Based MT Tasks," *Proc. Aslib Translating and the Computer 21 Conf.,* 1999.
6. P. Fung and L.Y. Yee, "An IR Approach for Translating New Words from Nonparallel, Comparable Texts," *Proc. 17th Int'l Conf. Computational Linguistics and 36th Ann. Meeting Assoc. Computational Linguistics*, 1998, pp. 414–420.
7. R. Rapp, "Automatic Identification of Word Translations from Unrelated English and German Corpora," *Proc. 37th Int'l Conf. Ann. Meeting Assoc. Computational Linguistics*, 1999, pp. 519–526.
8. T. Mitchell, *Machine Learning*, McGraw-Hill, 1997.
9. C. Li and H. Li, "Word Translation Disambiguation Using Bilingual Bootstrapping," *Proc. 40th Ann. Meeting Assoc. Computational Linguistics*, 2002, pp. 343–351.
10. D. Yarowsky, "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods," *Proc. 33rd Ann. Meeting Assoc. Computational Linguistics*, 1995, pp.189–196.

For more information on this or any other computing topic, please visit our Digital Library at http://computer.org/publications/dlib.

The Authors

Hang Li is a researcher at Microsoft Research Asia in Beijing. He is also an adjunct professor at Xian Jiaotong University. His research interests include statistical language learning, natural language processing, information retrieval, and data mining. He earned a PhD in computer science from the University of Tokyo. Contact him at Microsoft Research Asia, 5F Sigma Center, No. 49 Zhichun Rd., Haidian District, Beijing, China 100080; hangli@microsoft.com.

Yunbo Cao is an assistant researcher at Microsoft Research Asia in Beijing. His research interests include statistical language learning, natural language processing, and text mining. He obtained an MS in computer science from Peking University. Contact him at Microsoft Research Asia, 5F Sigma Center, No. 49 Zhichun Rd., Haidian District, Beijing, China 100080; i-yucao@microsoft.com.

Cong Li is an assistant researcher at Microsoft Research Asia in Beijing. His research interests include statistical language learning, natural language processing, and artificial intelligence. He obtained a BS in automation from Shanghai Jiaotong University. Contact him at Microsoft Research Asia, 5F Sigma Center, No. 49 Zhichun Rd., Haidian District, Beijing, China 100080; i-congl@microsoft.com.

IEEE Pervasive Computing

delivers the latest peer-reviewed developments in pervasive, mobile, and ubiquitous computing and acts as a catalyst for realizing the vision of pervasive (or ubiquitous) computing, described by Mark Weiser nearly a decade ago.

In 2003, look for articles on

- Security & Privacy
- The Human Experience
- Building Systems That Deal with Uncertainty
- Sensor and Actuator Networks

To subscribe, visit

http://computer.org/pervasive/subscribe.htm

or contact our Customer Service department:

+1 800 272 6657
toll-free in the US and Canada
+1 714 821 8380 phone
+1 714 821 4641 fax