PENS: A Machine-aided English Writing System for Chinese Users

Ting Liu[1] Ming Zhou Jianfeng Gao Endong Xun Changning Huang
Natural Language Computing Group, Microsoft Research China, Microsoft Corporation
5F, Beijing Sigma Center
100080 Beijing, P.R.C.
{ i-liutin, mingzhou, jfgao, i-edxun, cnhuang@microsoft.com }

Abstract

Writing English is a big barrier for most Chinese users. To build a computer-aided system that helps Chinese users not only on spelling checking and grammar checking but also on writing in the way of native-English is a challenging task. Although machine translation is widely used for this purpose, how to find an efficient way in which human collaborates with computers remains an open issue. In this paper, based on the comprehensive study of Chinese users requirements, we propose an approach to machine aided English writing system, which consists of two components: 1) a statistical approach to word spelling help, and 2) an information retrieval based approach to intelligent recommendation by providing suggestive example sentences. Both components work together in a unified way, and highly improve the productivity of English writing. We also developed a pilot system, namely PENS (Perfect ENglish System). Preliminary experiments show very promising results.

Introduction

With the rapid development of the Internet, writing English becomes daily work for computer users all over the world. However, for Chinese users who have significantly different culture and writing style, English writing is a big barrier. Therefore, building a machine-aided English writing system, which helps Chinese users not only on spelling checking and grammar checking but also on writing in the way of native-English, is a very promising task.

Statistics shows that almost all Chinese users who need to write in English[1] have enough knowledge of English that they can easily tell the difference between two sentences written in Chinese-English and native-English, respectively. Thus, the machine-aided English writing system should act as a consultant that provide various kinds of help whenever necessary, and let users play the major role during writing. These helps include:

1) Spelling help: help users input hard-to-spell words, and check the usage in a certain context simultaneously;
2) Example sentence help: help users refine the writing by providing perfect example sentences.

Several machine-aided approaches have been proposed recently. They basically fall into two categories, 1) automatic translation, and 2) translation memory. Both work at the sentence level. While in the former, the translation is not readable even after a lot of manually editing. The latter works like a case-based system, in that, given a sentence, the system retrieve similar sentences from translation example database, the user then translates his sentences by analogy. To build a computer-aided English writing system that helps Chinese users on writing in the way of native-English is a challenging task. Machine translation is widely used for this purpose, but how to find an efficient way in which human collaborates well with computers remains an open issue. Although the quality of fully automatic machine translation at the sentence level is by no means satisfied, it is hopeful to

1 Now Ting Liu is an associate professor in Harbin Institute of Technology, P.R.C.

provide relatively acceptable quality translations at the word or short phrase level. Therefore, we can expect that combining word/phrase level automatic translation with translation memory will achieve a better solution to machine-aided English writing system [Zhou, 95].

In this paper, we propose an approach to machine aided English writing system, which consists of two components: 1) a statistical approach to word spelling help, and 2) an information retrieval based approach to intelligent recommendation by providing suggestive example sentences. Both components work together in a unified way, and highly improve the productivity of English writing. We also develop a pilot system, namely PENS. Preliminary experiments show very promising results.

The rest of this paper is structured as follows. In section 2 we give an overview of the system, introduce the components of the system, and describe the resources needed. In section 3, we discuss the word spelling help, and focus the discussion on Chinese pinyin to English word translation. In addition, we describe various kinds of word level help functions, such as automatic translation of Chinese word in the form of either pinyin or Chinese characters, and synonym suggestion, etc. We also describe the user interface briefly. In section 4, an effective retrieval algorithm is proposed to implement the so-called intelligent recommendation function. In section 5, we present preliminary experimental results. Finally, concluding remarks is given in section 6.

1 System Overview

1.1 System Architecture

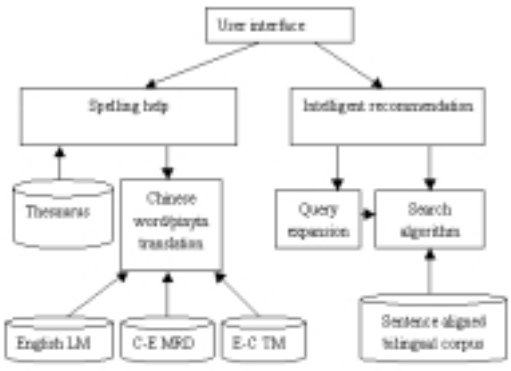

Figure 1 System Architecture

There are two modules in PENS. The first is called the spelling help. Given an English word, the spelling help performs two functions, 1) retrieving its synonym, antonym, and thesaurus; or 2) automatically giving the corresponding translation of Chinese words in the form of Chinese characters or pinyin. Statistical machine translation techniques are used for this translation, and therefore a Chinese-English bilingual dictionary (MRD), an English language model, and an English-Chinese word- translation model (TM) are needed. The English language model is a word trigram model, which consists of 247,238,396 trigrams, and the vocabulary used contains 58541 words. The MRD dictionary contains 115,200 Chinese entries as well as their corresponding English translations, and other information, such as part-of-speech, semantic classification, etc. The TM is trained from a word-aligned bilingual corpus, which occupies approximately 96,362 bilingual sentence pairs.

The second module is an intelligent recommendation system. It employs an effective sentence retrieval algorithm on a large bilingual corpus. The input is a sequence of keywords or a short phrase given by users, and the output is limited pairs bilingual sentences expressing relevant meaning with users' query, or just a few pairs of bilingual sentences with syntactical relevance.

1.2 Bilingual Corpus Construction

We have collected bilingual texts extracted from World Wide Web bilingual sites, dictionaries, books, bilingual news and magazines, and product manuals. The size of the corpus is 96,362 sentence pairs. The corpus is used in the following three cases:

1) Act as translation memory to support the Intelligent Recommendation Function;
2) To be used to acquire English-Chinese translation model to support translation at word and phrase level;
3) To be used to extract bilingual terms to enrich the Chinese-English MRD;

To construct a sentence aligned bilingual corpus, we first use an alignment algorithm doing the automatic alignment and then the alignment result are corrected.

There have been quite a number of recent papers on parallel text alignment. Lexically based techniques use extensive online bilingual lexicons to match sentences [Chen 93]. In contrast, statistical techniques require almost no prior knowledge and are based solely on the lengths of sentences, i.e. length-based alignment method. We use a novel method to incorporate both approaches [Liu, 95]. First, the rough result is obtained by using the length-based method. Then anchors are identified in the text to reduce the complexity. An anchor is defined as a block that consists of *n* successive sentences. Our experiments show best performance when *n*=3. Finally, a small, restricted set of lexical cues is applied to obtain for further improvement.

1.3 Translation Model Training

Chinese sentences must be segmented before word translation training, because written Chinese consists of a character stream without space between words. Therefore, we use a wordlist, which consists of 65502 words, in conjunction with an optimization procedure described in [Gao, 2000]. The bilingual training process employs a variant of the model in [Brown, 1993] and as such is based on an iterative EM (expectation-maximization) procedure for maximizing the likelihood of generating the English given the Chinese portion. The output of the training process is a set of potential English translations for each Chinese word, together with the probability estimate for each translation.

1.4 Extraction of Bilingual Domain-specific Terms

A domain-specific term is defined as a string that consists of more than one successive word and has certain occurrences in a text collection within a specific domain. Such a string has a complete meaning and lexical boundaries in semantics; it might be a compound word, phrase or linguistic template. We use two steps to extract bilingual terms from sentence aligned corpus. First we extract Chinese monolingual terms from Chinese part of the corpus by a similar method described in [Chien, 1998], then we extract the English corresponding part by using the word alignment information. A candidate list of the Chinese-English bilingual terms can be obtained as the result. Then we will check the list and add the terms into the dictionary.

2 Spelling Help

The spelling help works on the word or phrase level. Given an English word or phrase, it performs two functions, 1) retrieving corresponding synonyms, antonyms, and thesaurus; and 2) automatically giving the corresponding translation of Chinese words in the form of Chinese characters or pinyin. We will focus our discussion on the latter function in the section.

To use the latter function, the user may input Chinese characters or just input pinyin. It is not very convenient for Chinese users to input Chinese characters by an English keyboard. Furthermore the user must switch between English input model and Chinese input model time and again. These operations will interrupt his train of thought. To avoid this shortcoming, our system allows the user to input pinyin instead of Chinese characters. The pinyin can be translated into English word directly.

Let us take a user scenario for an example to show how the spelling help works. Suppose that a user input a Chinese word “完成” in the form of pinyin, say “wancheng”, as shown in figure1-1.

Figure 1-1

PENS is able to detect whether a string is a pinyin string or an English string automatically. For a pinyin string, PENS tries to translate it into the corresponding English word or phrase directly. The mapping from pinyin to Chinese word is one-to-many, so does the mapping from Chinese word to English words. Therefore, for each pinyin string, there are alternative translations. PENS employs a statistical approach to determine the correct translation. PENS also displays the corresponding Chinese word or phrase for confirmation, as shown in figure 1-2.

Figure 1-2

If the user is not satisfied with the English word determined by PENS, he can browse other candidates as well as their bilingual example sentences, and select a better one, as shown in figure 1-3.

Figure 1-3

2.1 Word Translation Algorithm based on Statistical LM and TM

Suppose that a user input two English words, say EW_1 and EW_2, and then a pinyin string, say *PY*. For *PY*, all candidate Chinese words are determined by looking up a Pinyin-Chinese dictionary. Then, a list of candidate English translations is obtained according to a MRD. These English translations are English words of their original form, while they should be of different forms in different contexts. We exploit morphology for this purpose, and expand each word to all possible forms. For instance, inflections of "go" may be "went", and "gone". In what follows, we will describe how to determine the proper translation among the candidate list.

Figure 2-1: Word-level Pinyin-English Translation

As shown in Figure 2-1, we assume that the most proper translation of *PY* is the English word with the highest conditional probability among all leaf nodes, that is
According to Bayes' law, the conditional probability is estimated by

$$P(EW_{ij} \mid PY, EW_1, EW_2) = \frac{P(PY \mid EW_{ij}, EW_1, EW_2) \times P(EW_{ij} \mid EW_1, EW_2)}{P(PY \mid EW_1, EW_2)} \quad (2\text{-}1)$$

Since the denominator is independent of EW_{ij}, we rewrite (2-1) as

$$P(EW_{ij} \mid PY, EW_1, EW_2) \propto P(PY \mid EW_{ij}, EW_1, EW_2) \times P(EW_{ij} \mid EW_1, EW_2) \quad (2\text{-}2)$$

Since CW_i is a bridge which connect the pinyin and the English translation, we introduce Chinese word CW_i into

$$P(PY \mid EW_{ij}, EW_1, EW_2)$$

We get

$$P(PY \mid EW_{ij}, EW_1, EW_2) = \frac{P(CW_i \mid EW_{ij}, EW_1, EW_2) \times P(PY \mid CW_i, EW_{ij}, EW_1, EW_2)}{P(CW_i \mid PY, EW_{ij}, EW_1, EW_2)} \quad (2\text{-}3)$$

For simplicity, we assume that a Chinese word doesn't depends on the translation context, so we can get the following approximate equation:

$$P(CW_i \mid EW_{ij}, EW_1, EW_2) \approx P(CW_i \mid EW_{ij})$$

We can also assume that the pinyin of a Chinese word is not concerned in the corresponding English translation, namely:

$$P(PY \mid CW_i, EW_{ij}, EW_1, EW_2) \approx P(PY \mid CW_i)$$

It is almost impossible that two Chinese words correspond to the same pinyin and the same English translation, so we can suppose that:

$$P(CW_i \mid PY, EW_{ij}, EW_1, EW_2) \approx 1$$

Therefore, we get the approximation of (2-3) as follows:

$$P(PY \mid EW_{ij}, EW_1, EW_2) = P(CW_i \mid EW_{ij}) \times P(PY \mid CW_i) \quad (2\text{-}4)$$

According to formula (2-2) and (2-4), we get:

$$P(EW_{ij} \mid PY, EW_1, EW_2) = P(CW_i \mid EW_{ij}) \times P(PY \mid CW_i) \times P(EW_{ij} \mid EW_1, EW_2) \quad (2\text{-}5)$$

where $P(CW_i \mid EW_{ij})$ is the translation model, and can be got from bilingual corpus, and $P(PY \mid CW_i)$

is the polyphone model, here we suppose $P(PY|CW_i) = 1$, and $P(EW_{ij} \mid EW_1, EW_2)$ is the English trigram language model.
To sum up, as indicated in (2-6), the spelling help find the most proper translation of *PY* by retrieving the English word with the highest conditional probability.

$$\arg\max_{EW_{ij}} P(EW \mid PY, EW_1, EW_2) = \arg\max_{EW_{ij}} P(CW_i \mid EW_{ij}) \times P(EW_{ij} \mid EW_1, EW_2) \quad (2\text{-}6)$$

3 Intelligent Recommendation

The intelligent recommendation works on the sentence level. When a user input a sequence of Chinese characters, the character string will be firstly segmented into one or more words. The segmented word string acts as the user query in IR. After query expansion, the intelligent recommendation employs an effective sentence retrieval algorithm on a large bilingual corpus, and retrieves a pair (or a set of pairs) of bilingual sentences related to the query. All the retrieved sentence pairs are ranked based on a scoring strategy.

3.1 Query Expansion

Suppose that a user query is of the form CW_1, CW_2, … , CW_m. We then list all synonyms for each word of the queries based on a Chinese thesaurus, as shown below.

$$\begin{array}{cccc} CW_{11} & CW_{21} & \cdots & CW_{m1} \\ CW_{12} & CW_{22} & \cdots & CW_{m2} \\ \cdots & \cdots & \cdots & \cdots \\ CW_{1n_1} & CW_{2n_2} & \cdots & CW_{mn_m} \end{array}$$

We can obtain an expanded query by substituting a word in the query with its synonym. To avoid over-generation, we restrict that only one word is substituted at each time.
Let us take the query “声音 效果” for an example. The synonyms list is as follows:
声音 =〉 声，音，音响，声响，响声，超声波，……
效果 =〉 作用，功效，实效…….
The query consists of two words. By substituting the first word, we get expanded queries, such as “声 效果”, “音 效果”, “音响 效果”, etc, and by substituting the second word, we get other expanded queries, such as “声音 作用”, “声音 功效”, “声音 实效”, etc.

Then we select the expanded query, which is used for retrieving example sentence pairs, by estimating the mutual information of words with the query. It is indicated as follows

$$\arg\max_{i,j} \sum_{\substack{k=1 \\ k \neq i}}^{m} MI\ (CW_k, CW_{ij})$$

where CW_k is a the *k*th Chinese word in the query, and CW_{ij} is the jth synonym of the *i*-th Chinese word. In the above example, “音响 效果” is selected. The selection well meets the common sense. Therefore, bilingual example sentences containing “音响 效果” will be retrieved as well.

3.2 Ranking Algorithm

The input of the ranking algorithm is a query *Q*, as described above, *Q* is a Chinese word string, as shown below

$Q = T_1, T_2, T_3, \ldots T_k$

The output is a set of relevant bilingual example sentence pairs in the form of,
$S=\{(Chinsent, Engsent) \mid Relevance(Q,Chinsent) > \delta \ or \ Relevance(Q,Engsent) > \delta\}$
where *Chinsent* is a Chinese sentence, and *Engsent* is an English sentence, and δ is a threshold.

For each sentence, the relevance score is computed in two parts, 1) the *bonus* which represents the similarity of input query and the target sentence, and 2) the *penalty*, which represents the dissimilarity of input query and the target sentence.
The bonus is computed by the following formula:
Where：

$$Bonus_i = \sum_{j=1}^{m} \log(W_j \times tf_{ij}) \times \log(n / df_j) / L_i$$

W_j is the weight of the jth word in query Q, which will be described later, tf_{ij} is the number of the *j*th word occurring in sentence *i*, *n* is the number of the sentences in corpus, df_j is the number of

sentence which contains *Wj*, and L_i is the number of word in the *i*th sentence.

The above formula contains only the algebraic similarities. To take the geometry similarity into consideration, we designed a penalty formula. The idea is that we use the editing distance to compute that geometry similarity.

$$R_i = Bonus_i - Penalty_i$$

Suppose the matched word list between query Q and a sentence are represented as A and B respectively：

$A_1, A_2, A_3, \ldots, A_l$
$B_1, B_2, B_3, \ldots, B_m$

The editing distance is defined as the number of editing operation to revise B to A. The penalty will increase for each editing operation, but the score is different for different word category. For example, the penalty will be serious when operating a verb than operating a noun

$$Penalty_i = \sum_{j=1}^{h} \log(W_j' \times E_j) \times \log(n / df_j) / L_i$$

where：

W_j' is the penalty of the jth word;

E_j the editing distance;

We define the score and penalty for each kind of part-or-speech

POS	Score	Penalty
Noun	6	6
Verb	10	10
Adjective	8	8
Adverb	8	8
Preposition	8	8
Conjuction	4	4
Digit	4	4
Digit-classifer	4	4
Classifer	4	4
Exclamation	4	4
Pronoun	4	4
Auxilary	6	6
Post-preposition	6	6
Idioms	6	6

We then select the first Φ sentences and output.

4 Experimental Results & Evaluation

In this section, we will report the primary experimental results on 1) word-level pinyin-English translation, and 2) example sentences retrieval.

4.1 Word-level Pinyin-English Translation

Firstly, we built a testing set based on the word aligned bilingual corpus automatically.
Suppose that there is a word-aligned bilingual sentence pair, and every Chinese word is labelled with Pinyin. See Figure 4-1.

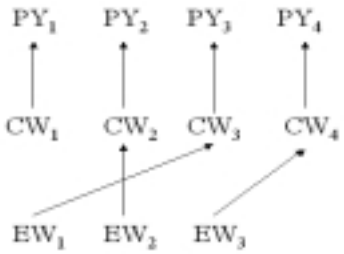

Figure 5-1: An example of aligned bilingual sentence

If we substitute an English word with the piny
Figure 4-1: An example of aligned bilingual sentence

If we substitute an English word with the pinyin of the Chinese word which the English word is aligned to, we can get a testing example for word-level Pinyin-English translation. Since the user only cares about how to write content words, rather than function words, we should skip function words in the English sentence. In this example, suppose EW_1 is a function word, EW_2 and EW_3 are content words, thus the extracted testing examples are:

$EW_1\ PY_2\ (CW_2, EW_2)$
$EW_1\ EW_2\ PY_4\ (CW_4, EW_3)$

The Chinese words and English words in brackets are standard answers to the pinyin. We can get the precision of translation by comparing the standard answers with the answers obtained by the Pinyin-English translation module.

The standard testing set includes 1198 testing sentences, and all the pinyins are polysyllabic. The experimental result is shown in Figure 4-2.

	Shoot Rate
Chinese Word	0.964942
English Top 1	0.794658
English Top 5	0.932387
English Top 1 (Considering morphology)	0.606845
English Top 5 (Considering morphology)	0.834725

Figure 4-2: Testing of Pinyin-English Word-level Translation

4.2 Example Sentence Retrieval

We built a standard example sentences set which consists of 964 bilingual example sentence pairs. We also created 50 Chinese-phrase queries manually based on the set. Then we labelled every sentence with the 50 queries. For instance, let's say that the example sentence is

他根据自己的调查研究作出了结论。(He drew the conclusion by building on his own investigation.)

After labelling, the corresponding queries are "调查 研究", and "作出 结论", that is, when a user input these queries, the above example sentence should be picked out.

After we labelled all 964 sentences, we performed the sentence retrieval module on the sentence set, that is, PENS retrieved example sentences for each of the 50 queries. Therefore, for each query, we compared the sentence set retrieved by PENS with the sentence labelled manually, and evaluate the performance by estimating the precision and the recall.

Let *A* denotes the number of sentences which is selected by both human and the machine, *B* denotes the number of sentences which is selected only by the machine, and *C* denotes the number of sentences which is selected only by human.

The precision of the retrieval to query *i*, say *Pi*, is estimated by $Pi = A / B$ and the recall *Ri*, is estimated by $Ri = A/C$. The average precision is $P = \frac{\sum_{i=1}^{50} P_i}{50}$, and the average recall is $R = \frac{\sum_{i=1}^{50} R_i}{50}$.

The experimental results are $P = 83.3\%$, and $R = 55.7\%$. The user only cares if he could obtain a useful example sentence, and it is unnecessary for the system to find out all the relevant sentences in the bilingual sentence corpus. Therefore, example sentence retrieval in PENS is different from conventional text retrieval at this point.

Conclusion

In this paper, based on the comprehensive study of Chinese users requirements, we propose a unified approach to machine aided English writing system, which consists of two components: 1) a statistical approach to word spelling help, and 2) an information retrieval based approach to intelligent recommendation by providing suggestive example sentences. While the former works at the word or phrase level, the latter works at the sentence level. Both components work together in a unified way, and highly improve the productivity of English writing.

We also develop a pilot system, namely PENS, where we try to find an efficient way in which human collaborate with computers. Although many components of PENS are under development, primary experiments on two standard testing sets have already shown very promising results.

References

Ming Zhou, Sheng Li, Tiejun Zhao, Min Zhang, Xiaohu Liu, Meng Cai （1995）. DEAR: A translator's workstation. In *Proceedings of NLPRS'95, Dec. 5-7, Seoul.*

Xin Liu, Ming Zhou, Shenghuo Zhu, Changning Huang (1998), Aligning sentences in parallel corpora using self-extracted lexical information, *Chinese Journal of Computers (in Chinese), 1998, Vol. 21 (Supplement):151-158.*

Chen, Stanley F.(1993). Aligning sentences in bilingual corpora using lexical infromation. In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics, 9-16, Columbus, OH.*

Brown. P.F., Jennifer C. Lai, and R.L. Merce. (1991). Aligning sentences in parallel corpora.In *Proceedings of the 29th Annual Conference of the Association for Computational Linguistics, 169-176,Berkeley.*

Dekai Wu, Xuanyin Xia (1995). Large-scale automatic extraction of an English-Chinese translation lexicon. *Machine Translation, 9:3-4, 285-313 (1995)*

Church, K.W.(1993), *Char-align.* A program for aligning parallel texts at the character level. In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics, 1-8, Columbus, OH.*

Dagan, I., K.W. Church, and W.A. Gale (1993) Robust bilingual word alignment for machine aided translation. In *Proceedings of the workshop on Very Large Corpora, 69-85, Kyoto, Auguest.*

Jianfeng Gao, Han-Feng Wang, Mingjing Li, and Kai-Fu Lee, 2000. A Unified Approach to Statistical Language Modeling for Chinese. In *IEEE, ICASPP2000.*

Brown, P. F., S. A. DellaPietra, V.J. Dellapietra, and R.L.Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2): 263-311

Lee-Feng Chien, 1998. PAT-Tree-Based Adaptive Keyphrase Extraction for Intelligent Chinese Information Retrieval. Special issue on "*Information Retrieval with Asian Language*" *Information Processing and Management*, 1998.