1998 International Computer Symposium
Workshop on Artificial Intelligence
December 17-19, 1998, N.C.K.U., Tainan, Taiwan, R.O.C.

A Portuguese-Chinese Corpus-Based Machine Translation System

Chi-Man Pun *Yi-Ping Li*

Faculty of Science and Technology
University of Macau
P.O. Box 3001, Macau
E-Mail: {fstpcm, fstypl}@sftw.umac.mo

Abstract

This paper describes an approach to build a practical and useful corpus-based machine translation (CBMT) system which employs a statistical approach with automatic bilingual alignment support. Our improved algorithm for aligning bilingual parallel texts can achieve 97% of accuracy, which acts as the strong back-end of our CBMT system to build up the Portuguese-Chinese corpus-base.

Keywords: machine translation, corpus, alignment, nearest example retrieval.

1 Introduction

Our approach to build a CBMT system is corpus-based, supports automatic sentence alignment, employs statistical methodology, is a semi-automatic machine translation system, and makes use of both machine aided human translation (MAHT) and fully automated machine translation (FAMT) theories. The conventional rule-based machine translation (RBMT) fails due to the lack of formal and complete descriptions of the human languages. On other hand, the availability of a large amount of bilingual Portuguese-Chinese parallel texts in Macau nowadays, solves the one of major problems in CBMT, that is, finding the data source of parallel texts for the Portuguese-Chinese corpus-base. Finding the suitable data source for parallel texts is the first important step to make the implementation of our CBMT system to become a reality.

2 Background and Previous Work

The research on machine translation (MT) had been started since 1950s. However, there was no breakthrough for a satisfactory MT system. This is due to the fact that a perfect fully automatic MT system depends on exact knowledge of how human languages work. However, even now there is no complete and formal descriptions on human languages to be available. Why, then, human can easily handle the language problem? The answer is that the human brain is specially structured for language. The evidence comes from two main sources. The first is the study of brain injuries. There are specific areas of the brain which, when injured, impair a person's use of language in specific way. The second kind of evidence comes from language acquisition. All children learn the language of the people around them, whether or not anyone makes any effort to teach them how to talk. In addition, the human brain has a very hung memory capacity to store lots of things.

However, today's computer architecture is totally different from human brain, that is, they are not specially structured for languages. On the other hand, the memory capacity and performance of nowadays' computers is comparative to human brain. Hence, instead of studying on the pure linguistic MT systems, we decided to focus on the study of statistical-based MT systems, which could be more practical and useful. Recently, the ideas of translator's workstation [Zhou95] and large corpus-base building from aligning a large amount of bilingual parallel texts [Gale91] give us some insights on how to make a MT system more practical and useful.

3 Corpus-Based Machine Translation

Definition:

> Corpus-based machine translation (CBMT) is a statistical-based approach to perform automatic machine translation by means of retrieving translation units, such as words, phrases, or sentences, from a large bilingual corpus-base with possibly regular and irregular transformations.

The idea of CBMT is actually deduced from the idea of example-based machine translation (EBMT). The core part of CBMT is, of course, the bilingual corpus-base , like the bilingual examples in EBMT. In fact, both CBMT and EBMT share almost the same theory, except that their basic translation units are different. In CBMT, basic translation units for the bilingual corpus-base can be sentences, phrases or words.

There are also mainly three key issues in CBMT (Fig. 3-3):

1. Building the bilingual corpus-base at sentence, phrases and word level from parallel texts.
2. A mechanism for retrieving from the corpus-base the example that best matches the input sentence.

3. Exploiting the retrieved translation units to produce the actual translation of the input sentence.

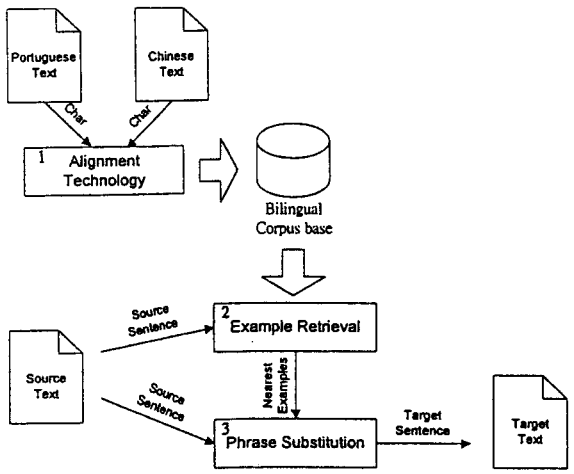

Figure 3-3 Process of CBMT

3.1 Building Portuguese-Chinese bilingual corpus-base by alignment

Alignment of two parallel texts is to show or mark which parts of one text are translated by which parts of second text, or formally,

Definition:

An alignment is a parallel segmentation of the two texts, typically into small logical units such as sentences and words, such that the n^{th} segment of the first text and the n^{th} segment of the second are mutual translations.

Approaches to sentence alignment basically fall into two main classes: lexical and statistical. In lexically-based methods, they use a lot of online bilingual lexicons to match sentences and heuristic linguistic knowledge. In statistical-based methods, they usually require no or very little linguistic knowledge and are based solely on the lengths of sentences. The main advantage of this approach is that they can be both accurate and producing very good performance.

3.1.1 Statistical-based alignment

Given a pair of parallel texts (passages), choose the alignment that maximizes the probability over all possible alignments. Formally,

$$\arg\max_{A} \text{Prob}(A \mid T_1, T_2) \qquad (1)$$

where A is an alignment, and T_1 and T_2 are the Portuguese and Chinese texts, respectively. An alignment A is a set consisting of $L_1 \leftrightarrow L_2$ pairs where each L_1 or L_2 is a Portuguese or Chinese passage.

We can now do some approximations so that the formulation is not too general. Assume that the probabilities of the individual aligned pairs within an alignment are independent, that is, A_i does not depend on A_j for any $i \neq j$, then the first approximation is as follow:

$$\text{Prob}(A \mid T_1, T_2) \approx \prod_{(L_1 \leftrightarrow L_2) \in A} \text{Prob}(L_1 \leftrightarrow L_2 \mid T_1, T_2) \qquad (2)$$

Usually each $\text{Prob}(L \leftrightarrow L_2 \mid T_1, T_2)$ depends not on the entire texts, but only on the contents of the specific passages within the alignment, then we have the second approximation as follow:

$$\text{Prob}(A \mid T_1, T_2) \approx \prod_{(L_1 \leftrightarrow L_2) \in A} \text{Prob}(L_1 \leftrightarrow L_2 \mid L_1, L_2) \qquad (3)$$

From (1), the maximization of the approximation (3) to the alignment probabilities is easily converted into a minimum-sum problem:

$$\begin{aligned} &\arg\max_{A} \text{Prob}(A \mid T_1, T_2) \\ &\approx \arg\max_{A} \prod_{(L_1 \leftrightarrow L_2) \in A} \text{Prob}(L_1 \leftrightarrow L_2 \mid L_1, L_2) \\ &= \arg\min_{A} \sum_{(L_1 \leftrightarrow L_2) \in A} -\log \text{Prob}(L_1 \leftrightarrow L_2 \mid L_1, L_2) \end{aligned} \qquad (4)$$

The *log* is introduced here so that adding the probabilities will produce desirable results.

3.1.2 Incorporation of Lexical cues in Pure Length-based method

Many alignment methods employed solely statistical criteria, like Gale and Church's pure length-based alignment method [Gale91]. Some other alignment methods employ solely lexical criteria. Only few, like Wu's, combine both statistical and lexical criteria in sentence alignment [Wu95]. However, there will have performance problems when the lexical database becomes larger. Our purpose is to incorporate lexical criteria without giving up the high performance statistical approach.

Pure length-based method is based on the fact that longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences. From (3), we can rewrite the probability function as follow:

$$\text{Prob}(L_1 \leftrightarrow L_2 \mid L_1, L_2) \approx \text{Prob}(L_1 \leftrightarrow L_2 \mid l_1, l_2) \qquad (5)$$

In order to incorporate lexical criteria into the pure length-based method, we include some lexical parameters in (5).

$$\text{Prob}(L_1 \leftrightarrow L_2 \mid L_1, L_2) \approx \text{Prob}(L_1 \leftrightarrow L_2 \mid l_1, l_2, v_1, w_2, \ldots, v_n, w_n) \quad (6)$$

where l_1 = the length of L_1 and l_2 = the length of L_2, measured in number of characters, v_i = the number of Portuguese cue of $type_i$ in $L1$ and w_i = the number of Chinese cue of $type_i$ in $L2$ (see Table 3-3).

The following is the criteria for choosing the lexical cues:

1. should be highly reliable and occur frequently so that violations, which wastes the additional computation, happen only rarely;
2. domain specific lexical cues are useful for building domain specific corpus-base;
3. both Portuguese and Chinese fields of a lexical cue should be unique.

In similar way of [Gale91], the dependence can be encapsulated by different parameter δ_i, and the $L1 \leftrightarrow L2$ pairs in alignment set can also be restricted into six matches: 1-1, 0-1 or 1-0, 2-1 or 1-2, 2-2 mappings. The approximation is estimated as follow:

$$\text{Prob}(L_1 \leftrightarrow L_2 \mid l_1, l_2, v_1, w_2, \ldots, v_n, w_n) \approx \text{Prob}(\text{match} \mid \delta_0(l_1, l_2), \delta_1(v_1, w_1), \ldots, \delta_n(v_n, w_n)) \quad (7)$$

Cue No.	Type	Portuguese	Chinese
1	adj.	1	一
2	adj.	2	二
3	adj.	3	三
...		...	...
n	noun	Agosto	八月

Table 3-3 Some examples of Portuguese-Chinese lexical cues.

By Bayes' Theorem, we have

$$\text{Prob}(\text{match} \mid \delta_0, \delta_1, \ldots, \delta_n) = \frac{\text{Prob}(\delta_0, \delta_1, \ldots, \delta_n \mid \text{match})\,\text{Prob}(\text{match})}{\text{Prob}(\delta_0, \delta_1, \ldots, \delta_n)} \approx \text{Prob}(\delta_0, \delta_1, \ldots, \delta_n \mid \text{match})\,\text{Prob}(\text{match}) \quad (8)$$

where $Prob(\delta_0, \delta_1, \ldots, \delta_n)$ is a normalizing constant which can be also ignored. It will not affect the result of the minimization because $\delta_0, \delta_1, \ldots, \delta_n$ does not depend on *match*.

Assuming all δ_i values are approximately independent, we have,

$$\text{Prob}(\delta_0, \delta_1, \ldots, \delta_n \mid \text{match}) \approx \prod_{i=0}^{n} \text{Prob}(\delta_i \mid \text{match}) \quad (9)$$

Similarly, we follow the definition of the function δ in [Gale91], since it has a approximately normal distribution with mean zero and variance one:

$$\delta_0 = \frac{l_2 - l_1 c}{\sqrt{l_1 s^2}} \quad (10)$$

where l_1 = the length of $L1$ (Portuguese Sentence) and l_2 = the length of $L2$ (Chinese Sentence), c is the expected number of characters in $L2$ per character in $L1$, and s^2 is the variance of the number of characters in $L2$ per character in $L1$. Here, the mean, c, and variance, s^2, have exactly the same values for paragraph and sentence alignment.

and,

$$\delta_i = \frac{w_i - v_i m}{\sqrt{v_i v^2}}, \quad i = 1..n \quad (11)$$

where v_i = the number of Portuguese cue of $type_i$ in $L1$ (Portuguese Sentence) and w_i = the number of Chinese cue of $type_i$ in $L2$ (Chinese Sentence), m is the mean of the number of Chinese cue of $type_i$ in $L2$ per number of Portuguese cue of $type_i$ in $L1$, and v^2 is the variance of the number of Chinese cue of $type_i$ in $L2$ per number of Portuguese cue of $type_i$ in $L1$.

The mean and variance calculated in [Gale91] are primarily for English-French and English-German alignment. We have to work out our own mean and variance for Portuguese-Chinese alignment. Based on the data of our own Portuguese-Chinese parallel texts in [6], we define our own mean and variance as follow:

c (mean) = 0.694
s^2 (variance) = 0.204

In addition, the number of Portuguese cue of $type_i$ and the number of Chinese cue of $type_i$ are most probably one-to-one mapping, that is, if there is only one occurrence of Portuguese cue in a particular Portuguese sentence, then most probably there should have only one occurrence of Chinese cue in the corresponding Chinese sentence. Hence, we define the nearly standard values for the mean and variance as :

m (mean) = 0.9
v^2 (variance) = 0.01

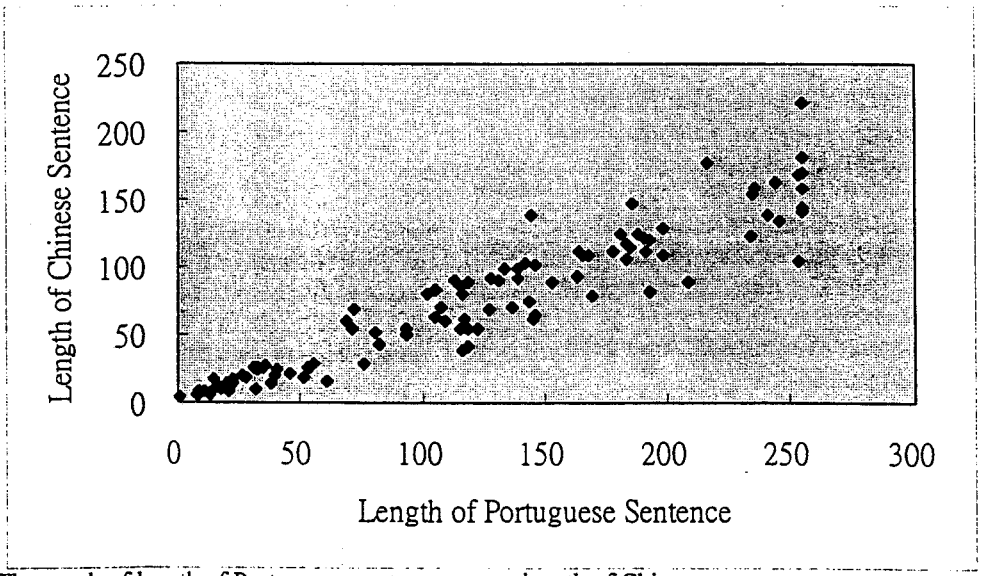

Figure 3-5 The graph of length of Portuguese sentence versus length of Chinese sentence.

Category	Prob(match)
1-1	0.899
1-0 or 0-1	0.0099
2-1 or 1-2	0.089
2-2	0.011

Table 3-2 The probability of the six matches in sentences

Similarly to[Gale91], we use the same probability table for *Prob(match)* as described in previous section. Hence, the conditional probability $Prob(\delta_i \mid match)$ can also be estimated in the same way by:

$$Prob(\delta_i \mid match) = 2(1 - Prob(|\delta_i|)) \qquad (12)$$

where $Prob(|\delta_i|)$ is the probability that a random variable, z, with a standardized normal distribution, has magnitude at least as large as $|\delta_i|$. That is,

$$\text{Prob}(\delta_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\delta_i} e^{-z^2/2}\, dz\,, i = 0..n \qquad (13)$$

Finally, the completed formula to select suitable alignment pairs for length-base sentence alignment with the incorporation of lexical cues can be obtained by (4), (9), (12), (13) as follow:

$$\arg\min_A \sum_{(L_1 \leftrightarrow L_2) \in A} -\log \text{Prob}\,(L_1 \leftrightarrow L_2 \mid L_1, L_2)$$

$$\approx \arg\min_A \sum_{match \in A} -\log \text{Prob}\,(\delta_0, \delta_1, ..., \delta_n \mid \text{match})\ \text{Prob (match)}$$

$$\approx \arg\min_A \sum_{match \in A} -\log \prod_{i=0}^{n} (2\,(1 - \text{Prob}(|\delta_i|))\ \text{Prob (match)}$$

$$\approx \arg\min_A \sum_{match \in A} -\log \prod_{i=0}^{n} \left(2\left(1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\delta_i} e^{-z^2/2} dz\right)\right) \text{Prob (match)} \qquad (14)$$

where δ_i and *Prob(match)* are defined in (10, 11) and Table 3-2, respectively.

3.2 Nearest Examples Retrieval

Definition:

A **nearest example** of an input sentence is the one whose has the highest score on the metric of similarity (as defined below) after the search in the bilingual corpus-base with regular and irregular transformations.

In establishing a mechanism for the nearest examples retrieval, the most important task is to define the metric of similarity between two sentences.

Definition of Similarity Metric

- the comparison units can be sentences, phrases or words, giving higher score to longer unit;
- if the functional words in both sentences fall into the same classes, that is, they have same pure form, after the regular and irregular transformation, then this pair has higher score than different classes.
- the minimum difference in the length of two sentences will have higher score.

Our nearest mechanism support two kinds of searching methods: fully automatic and semi-automatic search. In later case, the user (human translator) has to input a search expression, which support fuzzy, wildcard and Boolean operators. Our MT system will find out all examples in

bilingual corpus-base that satisfy the conditions of the given search expression. In fully automatic search, our MT system solely depends on the similarity metric to work out a suitable search expression , then performs the examples retrieval task.

3.3 Phrase Substitution Technique in Translation

In our approach, we prefer the simple pure phrase-substitution technique and concentrate on domain specific translations, that is, we tolerate such low quality translation as the front-end of our MT system in general case. Then we provide an user-friendly environment and large corpus-base building capability for human translator's post-editing work as the strong system back-end.

Our phrase substitution technique works as follow:

1. Extract all possible phrases in the source sentence if no maximum phrase-length restriction is given, or extract only those phrases whose length is less than or equal to the maximum phrase-length in the source sentence.
2. Search the bilingual corpus-base for each of extracted phrases, giving higher preferences to longer phrases. Mark the phrase with highest preference.
3. Select a suitable explanation if multiple explanations is available in the marked phrase.
4. Substitute the marked phrase with the selected explanation.
5. Go to step 2, repeat the process for the remaining non-substituted phrases until all words in source sentence are substituted.

4 System Design

Our system design has two main parts: Translator's Workstation and CBMT Central. In the top level design, firstly we base on the three visual forms in our user-interface to break the large Translator's Workstation module as follow:

1. *MDI form* - manages the functions of menus, dialogs, toolbar, printing, file open / save, child windows arrangements (tile, cascade or sizing).
2. *PC edit form* - manage the functions of Portuguese, Chinese or mixed editing and viewing, including font selection, size, color, word wrap.
3. *CBMT main form* - manage the interface for CBMT central to provide searching, options selection, display of query results, list-box, checkbox, radiobox, buttons and scroll box controls.

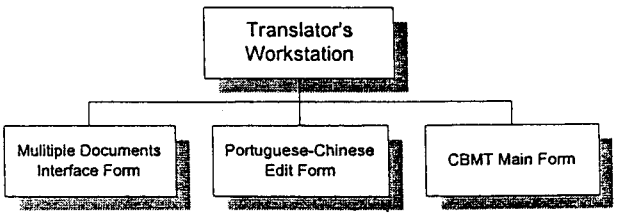

Figure 4-3 Top Level Design of Translator's Workstation

Secondly, we break another large module, CBMT Central, into four non-visual parts as follow:

1. *Lexical and syntax analyzer* - manages the functions of loading Portuguese-Chinese parallel texts, produce tokens (sentences) to alignment and CBMT modules, and perform syntax checking on search expression.
2. *Alignment module* - manages the functions of probability calculations, paragraph, sentence alignments.
3. *CBMT module* - manages the functions of phrase substitutions, nearest examples retrieval, order of corpus-base searching.
4. *Database management module* - manages all low-level database functions of data retrieval and updating of the bilingual corpus-base.

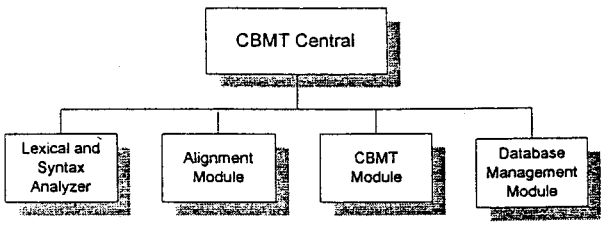

Figure 4-4 Top Level Design of CBMT Central

5 Evaluation

There are three main algorithms in CBMT Central: the alignment algorithm, the nearest examples retrieval algorithm, and the phrase substitution algorithm. Among of them, only the alignment algorithm can easily produce a lots of results for evaluation. In addition, the correctness of its results is easy to justify manually. However, the linguistic correctness criteria of nearest examples retrieval algorithm and phrase substitution algorithm is difficult to define due to the fact that there are many ambiguous cases in different levels of linguistic analysis (including syntax, semantics and pragmatics). Without such correctness criteria, we can only give some typical examples for evaluation of the nearest examples retrieval algorithm and phrase substitution algorithm.

The alignment algorithm is a length-based method with incorporation of lexical cues, which has two steps: aligning the paragraphs first and then sentences. Although these two steps employ the same algorithm except the

parameters: the variance and mean are different, they can produce very different results because of the differences in distributions between paragraphs and sentences (as shown Table 5-1 & 5-2). The aligned results are compared with a human alignment to decide whether a particular pair of alignment is correct or not. The task of human alignment is very simple, which is mostly to identify the wrong sentence segmentation and minor linguistic ambiguity.

Some of typical examples from outputs of the nearest examples retrieval algorithm and phrase substitution algorithm are shown in Table 5-3 & 5-4.

	1-1	1-0	0-1	2-1	1-2	2-2
Total	384	0	0	35	3	2
Correct	380	0	0	33	1	2
Incorrect	4	0	0	2	2	0
% Correct	0.99			0.94	0.33	

Table 5-1 Results of sentence alignment

	1-1	1-0	0-1	2-1	1-2	2-2
Total	411	0	0	3	3	0
Correct	411	0	0	2	2	0
Incorrect	0	0	0	1	1	0
% Correct	100			66.7	66.7	

Table 5-2 Results of paragraph alignment

	Portuguese Sentence / Matching Expression	Nearest example	Chinese Translation
1	7 de Julho	de 7 de Agosto	八月七日
2	O livro e o caderno são novos	Tendo em atenção o proposto pelo Governador de Macau	鑑於澳門總督之建議
3	*artigo*	Cumpridas as formalidades previstas na alínea a) do artigo 48.° do Estatuto Orgânico de Macau	經遵守《澳門組織章程》第四十八條第二款 a)項所規定的程序

Table 5-3 Some examples of nearest examples retrieval algorithm.

Our experiments of CBMT central have produced encouraging results, especially in alignment algorithm, which can have 97% of accuracy in sentence alignment and 99% of accuracy in paragraph alignment. From the result of sentence alignment, most of aligned pairs are 1-1 mapping and its error rate is only about 1%. The main reason for these results is the high frequency of 1-1 alignments in Portuguese-Chinese translations. There is no examples in 1-0 and 0-1 mappings because of the extremely low probability in these two cases and in fact there are no such sentences in the original texts. The result of high error rate in the 2-1 mappings is due to the specially structure in original texts and different punctuation between them (sometimes the Portuguese text uses full stops to separate numbering header, sometimes uses parentheses). Generally, the error rate depends mostly on following four factors:

1. Sentence Length
2. Paragraph Length
3. Category Type
4. Probability Measure

	Portuguese Sentence	Translated Chinese Sentence
1	O livro e o caderno são novos	這書和這本子是新的
2	Ela comprou uma saia preta e um chapéu preto	她買一裙子黑色的和一帽黑色的
3	Ele emprestou me dois livros	他出借我二書
4	Ele é estudante esperançoso	他是有前途的學生
5	Gosto de ouvir críticas tuas	味的聽評論你的

Table 5-4 Some examples of phrase substitution algorithm.

From the typical examples of nearest examples retrieval algorithm, it is useful if the input sentence is similar to some examples (or in the same domain) in the corpus-base (like example 1). Since most of our examples in the corpus-base are domain specific to laws, it seems useless for some dissimilar or different domain input sentence (like example 2). However, the users can retrieve the examples they want by means of using matching expressions (like example 3).

From the typical examples of phrase substitution algorithm, the translated Chinese sentence is good in the first example due to the same word ordering in both sentences. In examples 2 & 3, the Chinese translations are not good but still readable. However, in example 5, the Chinese translation is wrong because our phrase substitution algorithm automatically selects the most popular one if more than one explanations are available, in this case it selected the wrong explanation "味" for "Gosto", it should be "喜歡". One solution is add more phrases to the corpus-base. Like example 4, the algorithm makes use of the phrase substitution of "estudante esperançoso" to "有前途的學生", so that the quality of the Chinese translation is better.

6 Conclusion

In 1997, a Portuguese-Chinese CBMT system with built in translation editor has been implemented as a research project in University of Macau. Several improvements have been made on the pure length-based algorithm proposed by Gale and Church: 1. Based on our own Portuguese-Chinese data, recalculate two languages

dependent parameters: variance and mean; 2. Incorporation of lexical cues in their pure length-based algorithm. In addition, a domain specific Portuguese-Chinese corpus-base (400 entries) and a general Portuguese-Chinese dictionary (30000 entries) have been built up for both searching and translation.

Although Microsoft Windows is still very popular nowadays, it would be useful to re-implement our system into a machine independent version, for example, using JAVA as the programming language and one SQL server as middle layer for remote data retrieval. In addition, the feature of multimedia capabilities, such as, digitized speech and pictorial dictionary, is very useful for the learning of languages. Furthermore, on-line dictionary and translation web pages can be setup up to extend the usage of our system since more and more WWW users have been increased.

Although our alignment algorithm is very accurate and fast, the aligned results are not very useful in automatic translation, because most of sentences are not short and concise enough for phrase substitution algorithm. One obvious improvement is based on the results of sentence alignment to do word alignment. However, only partial word alignment (that is, only parts of sentences, for example, proper nouns and technical terms, are aligned) could have high accuracy, because it often difficult to decide just which words in a sentence are responsible for a given one in the translated sentence and some words apparently translate only morphological or syntactic phenomena rather than other words.

Our phrase substitution algorithm is corpus-based with simple regular and irregular transformations. Therefore, even very obvious mistakes could be made if there are multiple explanations existed or different positioning scheme in different languages. Although some of the cases can be avoided by adding more phrases and special cases in the corpus-base, another problems like the downgrade of performance, finding the source and input of those phrases, will then happen. However, the number of phrases and sentences in a natural language is not merely arbitrarily large. It is potentially infinite. This is because of the large number of choices, in both lexical and syntactic level, which can be made in the production of a sentence. Also, sentences can be recursive, like the construct of noun clauses in Portuguese. Therefore, a more reasonable approach is to add some linguistic rules to the phrase substitution algorithm, so that most of the obvious mistakes can be avoided and those linguistic rules could guide phrase substitution algorithm to substitute the right words in the right places. With the foundation of CBMT system, a more powerful and reliable fully automatic machine translation system could be implemented by hybrid approach of corpus-based and conventional rule based.

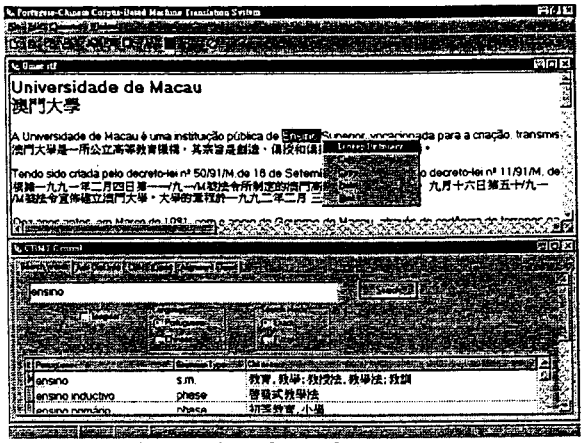

Figure 6-1 The user interface of CBMT system.

References and Bibliography

[1] Church, Kenneth W. and Robert L. Mercer (1993), "Introduction to the Special Issue on Computational Linguistics Using Large Corpora", *Using Large Corpora*, The MIT Press, 1994, p.1-23.

[2] Collins, Bróna, Pádraig Cunningham and Tony Veale (1996), "An Example-Based Approach to Machine Translation", *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (Expanding MT Horizons)*, p.1-13, Canada.

[3] Covington, Michael A. (1993), *Natural Language Processing for Prolog Programmers*, Prentice Hall.

[4] Cranias, Lambros, Harris Papageorgiou and Stelios Piperidis (1995), "A Matching Technique in Example-Based Machine Translation", *Procedings of Natural Language Processing Pacific Rim Symposium*, p.100-104, Seoul.

[5] Gale, William. A. and Kenneth W. Church (1993), "A Program for Aligning Sentences in Bilingual Corpora", *Using Large Corpora*, The MIT Press, 1994, p.75-102.

[6] Imprensa Oficial de Macau (1995), *Código Penal de Macau*, Imprensa Oficial de Macau.

[7] Kay, Martin and Martin Röscheisen (1993), "Text-Translation Alignment", *Using Large Corpora*, The MIT Press, 1994, p.121-142.

[8] Klavans, Judith and Evelyne Tzoukermann (1995), "Combining Corpus and Machine-Readable Dictionary Data for Building Bilingual Lexicons", *Machine Translation*, p.185-218, Netherlands.

[9] Kumano, Akira and Hideki Hirakawa (1995), "Building an MT dictionary from parallel texts based on linguistic and statistical information", *Procedings of Natural Language Processing Pacific Rim Symposium*, p.76-81, Seoul.

[10] Macklovitch, Elliott and Marie-Louise Hannan (1996), "Line 'Em up: Advances in Alignment Technology and Their Impact on Translation Support Tools", *Proceedings of the Second Conference of the Association for Machine*

Translation in the Americas (Expanding MT Horizons), p.145-156, Canada.

[11] Picchi, Eugenio, Carol Peters and Elisabeth Marina (1992), "A Translator's Workstation", *Third International Conference on Computational Linguistics (COLING-92)*.

[12] Simard, Michel and Pierre Plamondon (1996), "Bilingual Sentence Alignment: Balancing Robustness and Accuracy", *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (Expanding MT Horizons)*, p.135-144, Canada.

[13] Smith, Gary (1988), *Statistical Reasoning*, Allyn and Bacon, Inc., Second Edition.

[14] Wang, Sou Ying (1992), *A falar é que a gente entende... a Gramática*, Institvto Portvgvês do Oriente, Macau.

[15] Wu, Dekai and XuanYin Xia (1995), "Large-Scale Automatic Extraction of an English-Chinese Translation Lexicon", *Machine Translation*, 9:3-4, p.285-313.

[16] Zhou, Ming, Sheng Li, Changning Huang, Chuanbao Li, Tiejun Zhao, Min Zhang, Xiaohu Liu and Meng Cai (1995), "DEAR: A Translator's Workstation", *Procedings of Natural Language Processing Pacific Rim Symposium*, p.388-393, Seoul.

[17] 周明 (1996), *達雅翻譯工作站技術報告*, 清華大學.

[18] 周漢軍, 王增揚, 趙鴻玲, 崔維孝 (1994), *簡明葡漢詞典 DICIONÁRIO CONCISO PORTUGUÊS-CHINÊS*, 商務印書館, 北京.