

English-Chinese Cross-Lingual Retrieval Using a Translation Package

K. L. Kwok

23 January, 1999

Paper ID Code: 139

Submission type: Thematic

Topic Area: I1

Word Count: 3100 (excluding refereneces & tables)

Under consideration for other conferences (specify)? will submit a 1-page summary of this content to the government sponsored Advanced Information Processing & Analysis Symposium to be held March 23-24, 1999, in D.C. If accepted, the 1-page will appear in their proceedings.

Abstract

Using a COTS English-Chinese bidirectional translation software package together with our PIRCS bilingual retrieval system, we performed English-Chinese cross-lingual retrieval experiments using the TREC Chinese collections and queries. With some simple approaches, we are able to attain effectiveness about 67% of the monolingual Chinese results.

1. Introduction

CLIR has gained importance in recent years [OaDo96,Gref98] because web browsing, accessing foreign sites, and text searching has become popular, easy and convenient. Many language pairs need to be considered, but one can fairly say that English-Chinese cross language IR would become increasingly important because of the growing significance of China in business, politics, science & technology, etc. as well as the sheer number of the Chinese population. English of course is practically the de facto world language. Thus, the ability to do effective retrieval of collections in Chinese (the target language) via queries in English (the source language) without incurring professional translation costs would be a great convenience to those users who need to search or monitor Chinese information. The scenario would be for a user, who may have little or no knowledge of the Chinese language, to be able to filter large numbers of Chinese documents by himself or herself. From a gist of the retrieved document content, the user may determine which ones to send out for professional translation.

Program for translation of web pages from English to Chinese has become common nowadays and commercial software for this purpose is readily available. It appears that for the study of the effectiveness of English-Chinese CLIR and its improvements, not much work has been done. A major difficulty may be that Chinese language is quite different from other alphabetical languages, so that certain techniques for English-European language retrieval may not apply. For example, use of cognates and stems would not be available. Upper/lower case distinction also is non-existent in Chinese and word boundaries are ambiguous, so that the reverse translation is much more problematic. Another difficulty is the scarcity of resources. For example, most cross-lingual IR studies such as [BaCr97, DaOg97, HuGr96] require a fairly large bilingual dictionary, and for English-Chinese these are not readily available. Also, large Chinese corpora and evaluated results are not easy to come by.

The last difficulty concerning corpora and evaluated results however has recently been somewhat reduced due to TREC 5 & 6 (Text REtrieval Conference) where a large collection consisting of 170MB of GB-encoded Chinese text has become available together with 54 queries and their manually judged answers [VoHa97,98]. We use this collection for our investigation into this CLIR situation.

It is also during the recent years that commercial bi-directional translation packages have become popular because many users of the web are not proficient in English. These software run on PC's and are affordable. One such is the Transperfect (<http://www.otec.com.tw>) system that has been advertised to have accuracy of over 80% from English to Chinese. Using such a software package enables us to solve three problems at once: the dictionary resource problem, the English to Chinese translation problem and its inverse, and is quite an attractive proposition. The purpose of this paper is to gain some insight into the effectiveness of such commercial software for CLIR and investigate how well it can satisfy users' needs.

Section 2 describes the functionalities of the Transperfect package. Section 3 presents some results of CLIR using this software. Section 4 presents some results of translating Chinese sub-documents retrieved into English, and in Section 5 we discuss some of the difficulties and the conclusions.

2. The Translation Software

The standard Transperfect package employs a proprietary 'Basic' English-Chinese dictionary

(some 100K entries). One can augment with 'Special' domain dictionary (such as finance, computing, etc.) at additional costs. The system can translate whole documents. It has three buttons for a users' choice to determine the output: (A) interleaved output that allows convenient comparison of the original English and the Chinese counterpart; (B) multiple meanings for a word -- words that have several senses or meanings will all be displayed (so that a user can edit and make choices); and (C) alternative translations, where some sentences may have different candidate translations and all are displayed. For IR purposes, we disregard both (A) and (C). By disabling or enabling (B) we obtain respectively a short, unique translation output (but with some wrongly translated words), and a longer, noiser output that has multiple word meanings (but with higher probability of including a correct translation to each word).

There is also the capability for a user to supply his/her own words and create a 'User' dictionary to augment the 'Basic' one. The system allows a new word to have at most three POS categories, and within each category three senses. It is in general quite adequate. One sense within each POS category is the default, and it will be used when unique output is desired. During translation, the software makes use of its own proprietary analysis of an English sentence, and choose one POS category for each word. For multiple meaning mode, all senses of the POS category will be dumped out. Examples of translation is given in the next Section. The software package was developed in Taiwan where Big5 is the default Chinese character coding. It comes with a utility to convert outputs into GB coding and vice versa.

3. Translation of the Queries

The English version of the TREC Chinese track queries were fed into Transperfect as a single file on a PC. All English words or abbreviations that are not in its dictionary and numerics are left as is in the output. The translation time averages less than 10 seconds per raw query. Conversion to GB code takes only 2 or 3 seconds for the whole file.

As an example, the translation of Query #29 'Building the information super highway' with multiple word meaning output is shown in Appendix A. The English version as well as the provided 'accurate' Chinese translation (denoted by >>>) are also displayed. It can be seen that the word 'Building' can be a noun or a verb, and the software is sufficiently smart to choose the verb POS and with the multiple translation: 建筑 [建立]. If it were a noun, it would have been: 建筑物 [建筑], the first default

entry being a 'building thing', while the 2nd entry can be noun or verb. Several words in this query are not translated appropriately such as: 'information' is mapped into 消息(news), and the alternatives 知识(knowledge), 报告(report); 'super' becomes 上等的 (upper class) 表面的 (on the surface, superficial); 'infrastructure' becomes 下部构造(lower part construction), 下部组织(lower part organization). In spite of some inappropriate conversions, this translated query actually performs well, achieving about 76% of the 'accurate' translation retrieval result.

Some glaring mistakes of translation include: human rights 'issue' 发行 (to distribute) or 流出(outflow); 'oil' 油画颜料 (oil painting material); 'concrete indicators of impact' 水泥, 混凝土, (water, mortar, sand) 冲击 [冲突/影响] (conflict, influence) 的指示器 [指示剂/指压计] (measuring meter); 'lighten' the burden, 点亮 (start a light) [变亮/减轻负担]. An example of a translated query that returns 0% of the monolingual result is #19 and is also shown in Appendix 1.

3. CLIR Retrieval: Results and Discussion

3.1 Direct Use of Translation Package

The TREC queries came with both English and Chinese versions. We assume these versions are translated pairs. This is not exactly accurate because we do find words in the Chinese version that are not in the corresponding English. On the whole they are reasonable images of each other. Our procedure is similar to what we have done in [Kwok97a]: the Chinese long queries are first used to retrieve Chinese documents using short-word segmentation with character representation. This provides the basis by which we measure our CLIR results. This is shown as result Column 1 in Table 1. The retrieval results are very good as noted before [Kwok97b,c], achieving an average non-interpolated precision of 0.535. Precision at 10, 20 documents are 0.733, 0.691 respectively. Thus, within the top 10 to 20 documents one can expect 7 to 14 of the items are relevant averaged over all queries. At 1000 retrieved the number of relevants is 4852 out of a total 5140 determined to be relevant, or nearly 94.4%.

The English queries were then translated using Transperfect into two outputs, one using the unique output mode, and the other using the multiple word meaning translation mode. The latter provides about twice as much volume. However, since Chinese retrieval seems to tolerate noise well [Kwok97c], we expect multiple output mode to be useful. Results of retrieval are tabulated in Columns 3-4 of Table 1.

Column4 shows that multiple output mode is better than unique output mode Column 3, by about 5% in top document precisions (e.g. P@10: 0.426 vs 0.406), over 10% in relevants retrieved (3333 vs. 2980), and about 1% in average precision (0.301 vs 0.298). However, compared to monolingual, multiple output translation only performs at about 56% of the basis in average non-interpolated precision (0.301 vs 0.535). Most CLIR works that are based on simple dictionary mapping produce a performance of about 50% of monolingual [xx]. Thus, this English-Chinese CLIR 56% is within the ballpark, and about 6% better. One difference worth noting is that our starting basis is very high, unlike other investigations such as English-Spanish where the basis has only about 0.2 in average precision. The number of relevants retrieved is 3533, quite a high 69% of the monolingual value of 4852. The precision at 10, 20 documents retrieved is 58-60% of monolingual. Result of this simple attempt is quite reasonably and usable.

3.2 Un-Translated Words

In this approach, English words not in the dictionary are left un-translated, so called out-of-vocabulary words or OOV. For our 54 queries, these are listed in Table 2. Most of them are proper nouns like place or organization names, acronyms, etc.. Only very few general English words are un-translated, showing that the supplied dictionary is quite good.

One could add these OOV terms to the 'User' dictionary manually, but this would only boost up the results artificially. What we have done is to delete these un-translated words from the original Chinese queries to see how the basis degrades. Name entity identification for retrieval can affect effectiveness by over 10% [ThDo97]. Our result should not be that bad because we did not remove all proper nouns, only those not translated. The retrieval results of this modified basis is tabulated in Column 2 of Table 1.

It is seen that the absence of these un-translated words in the Chinese query did lead to worse average precision result of 0.515, about 4% worse than the unmodified basis. Compared with this degraded basis, CLIR (Column 4) average precision achieves 58.4% of this result where no proper noun OOVs are present. Thus, if certain information needs can be expressed as general English without involving uncommon proper nouns, CLIR may achieve a higher percentage of monolingual retrieval.

3.3 Pre-Translation Query Expansion

In [BaCr97], it is shown for English-Spanish that if the English queries were expanded before translation, the retrieval system may bring in extra related terms and bolster the query representation. Translation of these expanded queries helps CLIR by about 10%. We also performed similar experiments for English-Chinese. Because the two are completely different, what helps between English and Spanish may not be true here, like cognates. Every new word brought in needs translation before it can be helpful. The English queries were used for normal ad-hoc retrieval using an English collection, viz. Financial Times and FBIS. The choice is deliberate because these documents may contain higher concentration of foreign stories than domestic collections like LA Times. We expanded the English queries by 15, 25 and 35 terms. On average, about 8, 17 and 28 new terms are added.

After expansion, the queries are translated as before either with or without multiple meaning. The expanded terms may be words or English phrases, and they appear on each line individually. Results are tabulated in Column 5 to 10, two columns for each expansion value. It is seen that pre-translation query expansion indeed helps, and expansion with about 25 terms appears best in terms of efficiency and effectiveness. Multiple translation output again seems to have an edge over unique output. Thus, using multiple word meaning translation, average precision, relevants retrieved and precisions @ 10, 20 documents achieve values of 0.341, 3909 and 0.524 and 0.478 respectively. These are 63.7%, 80.4% and 71.5, 69.2% of the original basis of Column 1. The average precision of 0.341 is over 10% better than not using pre-translation (0.301).

3.4 Adding an External Dictionary

As we mentioned before, most English-Chinese dictionaries are commercialized. They are designed for consultation but not for downloading into a searchable file. However, in recent months, Paul Denisowski (http://www.mindspring.com/~paul_denisowski) at his web site has accumulated a sizable Chinese-English dictionary of about 20K entries by November, 1998. It is freely available for research use. We believe that combining a basic translation package plus an external dictionary, especially one that is domain-related to a query, would help to reduce the OOV problem. In our case, the OOV words are mainly geographic place names and abbreviations and we discover that some of these are captured in the dictionary. Example entries are as follows:

东南亚国家联盟 [dong1 nan2 ya4 guo2 jia1 lian2 meng2] /ASEAN (Association of Southeast Asian Nations)/

大麻[da4 ma2] /hemp/marijuana/

新疆 [xin1 jiang1] /Xinjiang (Uygur Autonomous Region)/

乌鲁木齐 [wu1 lu3 mu4 qi2] /Urumqi (capital of Xinjiang autonomous region)/

Each entry includes a Chinese word/phrase followed by the pronunciation [...] and the English translation(s) /.../.../. The dictionary contains a number of abbreviations (like: ASEAN, NATO) and proper nouns (like: Spratly, Bosnia) that occur in our queries and left un-translated. These dictionary lookup terms are marked with an asterisk in Table 2, and affect some 14 queries. Naturally, these terms are quite specific in meaning, and having them correctly translated helps boost the average precision by nearly 5% (see Column 11-12, Table 1: 0.360 vs 0.325 and 0.357 vs 0.341). At a value of 0.357, we have achieved about 67% of the monolingual basis average precision value of 0.535, quite similar to results of other language pairs like English and Spanish. This demonstrates the importance of proper noun and abbreviation translations.

4. Translating Retrieved Documents

To complete the cross language retrieval loop, one needs to display the top retrieved documents in English for the user's perusal. One should not expect accurate translation, but a gist of the content sufficient to detect if the document is useful. In PIRCS, we store and retrieve based on sub-documents which are approximately fixed-length chunks of 550 characters ending on a paragraph. They are sufficiently small to be translated wholesale.

Transperfect has a companion product called Alexander for Chinese to English translation. It worked quite similarly to the previous package, except that instructions are all in Chinese. It operates at about the same speed as before but result is much inferior. Still, as a gloss translation utility to provide a gist of the document content, it does serve its purpose. Example of a retrieved document fragment translated into English is displayed in Appendix 2.

5. Conclusion

English-Chinese cross language IR has been considered a daunting task because the two languages are so different and inadequate translation may be quite damaging. Here, we demonstrated that using an inexpensive COTS translation package and our bilingual PIRCS retrieval system, we are able to perform quite successful CLIR by employing some simple IR methodologies and an external dictionary, achieving some 67% of monolingual effectiveness. However, we should caution that these experiments were done only with 54 queries on 170 MB of text, a scale much smaller than monolingual English retrieval. Much more experiments need to be done before one can generalize results.

Using a COTS package helps us overcome the problems of dictionary resource and 2-way translation difficulties much faster. The disadvantage is that it operates as a black box, and one cannot manipulate the inside parameters to suit our needs. Operationally, it is also inconvenient in that the translation system runs on a PC while our retrieval system is on a SUN Solaris platform. The solution is to bring our PIRCS system to the PC environment. We believe that the package does solve a large portion of the cross language problem. Additional disambiguation of the translation output could be realized by using more powerful external dictionaries and thesaurus. CLIR can also further be improved using retrieval combination techniques

Appendix 1: Examples of Translated Queries

Query #29

Building the Information Super Highway

建筑 [建立] 消息 [知识/报告] 上等的 [表面的] 公路 [大道/直截的途径]
>> 信息高速公路的建设

Information Super Highway, building

消息 [知识/报告] 上等的 [表面的] 公路 [大道/直截的途径], 建筑 [建立]
>> 信息高速公路, 建设

building the Information Super Highway, including any technical problems, problems with the information infrastructure, or plans for use of the Internet by developed or developing countries.

建筑 [建立] 消息 [知识/报告] 上等的 [表面的] 公路 [大道/直截的途径], 包括任何的技术上的 [专门的/工业的] 问题 [难题], 消息 [知识/报告] 下部构造 [下部组织] 的问题 [难题], 或为 [因为/至于] 网际网路的使用 [习惯/价值] 的计划 [策略/方法] 被 [经由/在~之前] 发展 [发达/进步] 或发展中的国家 [乡村/山地]。

>> 信息高速公路的建设, 包括任何技术上的, 或与信息基础设施有关的问题, 以及有关发达国家或发展中国家对国际网A 纒撓 τ 眉幸常®

Expanded Terms

multimedia networks, 多媒体网路 [广播网],
network project., 网路 [广播网] 计画 [放映/表]。 ,
Telecommunications, 电传视讯 [远距离通信/电磁通信],
Vice, 恶习 [恶行/虎钳],
access, 存取,
Ministry, 部 [内阁/服务],
Japanese, 日本人 [日文],
Japan, 日本.,
connected, 连接, ,
advanced, 高等的 [在前的/老的],
computer, 电脑 [电子计算机],
-speed, -加速 [超速/推进],
Gore. 缝以补裆 [刺/抵]。 ,
technology, 技术 [工业技术/科学技术]。 ,
commercial, 商业的 [商用的/商品化的],
linking, 连结 [联合/连系],
road, 道路 [公路/方法],
cables, 打海底电报,
Telephone, 打电话,
Hong Kong, 香港.

Query #19

Project "Hope"

"计画 "希望"

>>

China, Project Hope, educational level, education

中国, 计画希望, 教育的水平, 教育

>>

information on Project Hope's objectives and its results. Any document containing information on raising teachers pay, improving remote areas' education, educational reform laws, or the amount of private contributions to Project Hope is relevant.

An irrelevant document mentions Project Hope but does not provide any concrete data on the success of the project such as how each area carries out the project and how many people have benefited from it. Documents such as letters to the editor asking where to donate money for the project are irrelevant. Documents that mention educational reform but do not give concrete measures are also irrelevant.

的消息计画希望的目的和它的结果。 任何的
文件包含养育教师的消息,薪资,改良遥远的区域教育,教育的
改革法律,或私人的捐助数量为了要计画希望是有关的。
不恰当的文件提到计画希望但是不为成功提供任何的具体资料
计画例如每个区域如何实行计画,而且人们有从它来的 BENEFITED 多少。
文件例如到编者的信问该在哪里捐赠金钱因为计画是不恰当。 文件
那提到教育的改革但是不给具体的措施也是不恰当的。

>>

Expanded Terms

"Hope Project", "希望计画",
primary schools, 最主要者教,
society, 社会,
Li, 李,

primary, 最主要者,
poor, 穷人,
students, 学生,
Chinese, 中文,
rural, 乡下的,
Beijing,, 北京,,
yuan,
compulsory education, 被强制的教育 ,
century, 世纪,
compulsory, 被强制的 ,
higher education, 比较较高的教育 ,
promoting, 促进,
science, 科学 ,
Foundation, 基础,
provinces, 省,
"implementing, "实现,
quality, 品质,
vocational, 职业的,
。

Appendix 2: Fragment of a Retrieved Document (CB046010-BBW-1318-175) for Query #54

中国外交部就美国提高美台关系问题向美国政府提出强烈抗议
Chinese Foreign Ministry the United State increase the beautiful set to relate t
o the question to hand in to the American government strong protest

新华社北京九月十日电中国外交部副部长刘华秋今天约见
New 华社 Peking September 10 Chinese Foreign Ministry vice minister Lou 华秋 of
电 today invite see

美国驻华大使芮效俭, 奉命就美国
The United State 驻华 ambassador 芮效俭, receive order the United State

提高美台关系问题向美国政府提出强烈抗议。
Increase the beautiful set to relate to the question to hand in to the American
government strong protest.

刘华秋说, 美国政府不顾中方的多次交涉和坚决反对,
The Lou 华秋 talk, American government regardless inside square of many negotiat
ion and resolute oppose,

公然宣布将采取一系列提高美台关系的措施。
Openly to declare to will adopt a series increase beautiful set relation of meas
ure.

这是美方蓄意制造“两个中国”、“一中一台”的政治行动,
This is boughten deliberately manufacturin; " the political action of two inside
country ", " an inside a set"s ,

不仅严重违反了中美间三个 联合公报确定的原则, 而且粗暴干涉了
Not only went against seriously Central America a three principle that joint comm
unique make sure, but also coarse interfered with

中国内政, 践踏了中国的主权, 中国政府和人民对此表示不满和愤慨。

Chinese domestic affairs, trampled the Chinese sovereignty, Chinese government and people rightness this show dissatisfied and excited and bitter.

References

- [BaCr97] Ballesteros, L & Croft, W.B. (1997). Phrasal Translation and Query Expansion Techniques for Cross- Language Information Retrieval. In: N.J. Belkin, D. Narasimhalu & P. Willett (Eds.), Proc. of 20th Ann. Intl. ACM SIGIR Conf. on R&D in IR. pp.84-91, ACM Press: NY.
- [DaOg97] Davis, M.W. & Ogden, W.C (1997). QUILT: Implementing a Large-Scale Cross-Language Text. . In: N.J. Belkin, D. Narasimhalu & P. Willett (Eds.), Proc. of 20th Ann. Intl. ACM SIGIR Conf. on R&D in IR. pp.92-98, ACM Press: NY.
- [Gref98] Cross language Information Retrieval. G. Grefenstette (ed.) Kluwer, 1998.
- [Kwok97a] Kwok, K.L. (1997) Evaluation of an English-Chinese Cross-Lingual Retrieval Experiment. AAAI-97 Symposium on Cross Language Text & Speech Retrieval. TR: SS-97-05, pp.133-137.
- [Kwok97b] Kwok, K.L. (1997). Comparing Multiple Representations for Chinese Information Retrieval. In: C. Cardie & R. Weischedel (Eds.), Proc. of 2nd Conf. on Empirical Methods in NLP. pp.141-148, ACL: NJ.
- [Kwok97c] Kwok, K.L. (1997). Lexicon Effects on Chinese Information Retrieval. In: N.J. Belkin, D. Narasimhalu & P. Willett (Eds.), Proc. of 20th Ann. Intl. ACM SIGIR Conf. on R&D in IR. pp.34-41, ACM Press: NY.
- [OaDr96] Oard, D.W & Dorr, B.J (1996) A Survey of Multilingual Text Retrieval. CS-TR-3615, Univ. of Maryland, Institute for Advanced Computer Studies.
- [ThDo97] Thompson, P. & Dozier, C. (1997) Name searching and information retrieval. In: C. Cardie & R. Weischedel (Eds.), Proc. of 2nd Conf. on Empirical Methods in NLP. pp.134-140, ACL: NJ.

	1	2	3	4	5	6	7	8	9	10	11	12
	basis	corup bas	" 0"	"1"	0 e15	1 e15	0 e25	1 e25	0 e35	1 e35	0e25dict	1e25dict
Rel_ret:	4852	4713	2980	3333	3731	3714	3850	3909	3963	3927	4056	4128
Interpolated Recall - Prec. Av.:												
at 0.1	0.792	0.75	0.465	0.481	0.484	0.54	0.498	0.549	0.541	0.56	0.541	0.581
at 0.3	0.662	0.636	0.374	0.382	0.386	0.42	0.405	0.419	0.426	0.417	0.451	0.434
at 0.5	0.568	0.541	0.316	0.308	0.322	0.342	0.338	0.352	0.353	0.352	0.379	0.368
at 0.7	0.438	0.422	0.242	0.241	0.258	0.255	0.267	0.267	0.274	0.265	0.302	0.286
at 0.9	0.279	0.271	0.143	0.141	0.156	0.14	0.16	0.145	0.161	0.145	0.178	0.152
Av. Prec (non-interpolated):												
	0.535	0.515	0.298	0.301	0.31	0.331	0.325	0.341	0.343	0.341	0.36	0.357
Prec. @												
10 docs	0.733	0.698	0.406	0.426	0.439	0.483	0.452	0.524	0.493	0.53	0.494	0.543
20 docs	0.691	0.657	0.386	0.404	0.397	0.462	0.421	0.478	0.448	0.476	0.462	0.497
30 docs	0.646	0.619	0.356	0.388	0.382	0.436	0.399	0.444	0.42	0.436	0.436	0.464
100 docs	0.464	0.45	0.248	0.26	0.27	0.298	0.29	0.306	0.312	0.312	0.316	0.321
R-Precision:	0.519	0.499	0.304	0.313	0.314	0.346	0.333	0.357	0.351	0.355	0.363	0.37

"0" unique translation; "1" multiple meaning translation, e25 25 expansion terms

Table 1: CLIR Retrieval Results

Query	Words	Query	Words	Query	Words	Query	Words
#2	reunification (4)	#12	4th world conference	#25	ecoprotection		liaohe
#3	daya		(2)	#27	robotic		songhua
	qinshan	#14	yunnan*	#28	psdn	#43	lama (5)*
#6	wto (3)*		hiv	#30	between+		indepence+
#7	prc (2)	#15	un (4)		1983-1993	#44	resettlement (4)
	spratly (3)*		multination*	#31	castro	#46	sino (2)
	dongsha	#16	un (7)	#32	traffikers (3)+		vietnamese (2)*
	xisha	#17	apcc (3)*		cali		nongovernmental
	asean*		wto (3)*		medina		campuchea
#8	richter		signatory		traffiers+	#47	pinatubo
#9	cocaine	#19	benefited	#33	hijackings		minatubo
	marijuana*	#20	mia's	#34	arrid+		subic
	trafficking		vietnamese*		acerage+		clark
#10	xinjiang*	#21	prc (3)	#35	mandela*	#51	formaulated+
	uigur*		reunification (2)	#38	sunveillance+	#53	sino (3)
	trading*		dingkang (2)	#39	assasination+		f-16 (3)
#11	un (2)	#23	un	#41	kowloon (3)		
	bosnia (2)*	#24	bosnian (2)*	#42	yangtze*		
	nato*		bosnia (2)*		huaihe		
	bosniaun		herzogovenia (2)		haihe		

Table 2: Un-translated or Misspelled Words