On-going Cooperative Research towards Developing Economy-Oriented Chinese-French SMT Systems with a New SMT Framework

Yidong CHEN[1], Lingxiao WANG[2], Christian BOITET[2], Xiaodong SHI[1]
(1) School of Information Science and Technology, Xiamen University, Xiamen, Fujian, China
(2) GETALP, Laboratoire d'Informatique Grenoble (LIG), Université Joseph Fourier, Grenoble, France
ydchen@xmu.edu.cn

Résumé. Nous présentons un projet collaboratif en cours mené par l'université de Grenoble et l'université de Xiamen, et visant à créer des instances d'un nouveau type de système de traduction automatique statistique utilisant des ressources lexico-sémantiques et discursives. Le but concret est de développer des systèmes de TAS chinois-français pour des sites boursiers et économiques. Comme très peu de corpus et de dictionnaires bilingues chinois-français sont disponibles sur Internet, l'anglais est utilisé comme "pivot" pour construire les équivalents chinois-français par transitivité. Outre la description générale de ce projet, nous décrivons les progrès sur deux axes de recherche liés à ce projet. Pour cela, nous utilisons une méthode, proposée par XMU, d'induction de probabilité fondée sur la similarité thématique, qui produit des tables de traduction C-F à partir de tables de traduction C-E et E-F. Pour disposer de bons corpus parallèles C-F, nous utilisons un système Web de post-édition collaborative qui peut déclencher l'amélioration incrémentale du système de TA en utilisant des métriques d'évaluation de TA et en extrayant la "meilleure partie" de la mémoire de traductions courante.

Abstract. We present an on-going collaborative project pursued by Grenoble University and Xiamen University and aiming at creating instances of a new kind of SMT system using semantics and discourse-related resources. The concrete goal is to develop Chinese-French systems specialized to stock option and economic websites. Since very few Chinese-French bilingual corpora and dictionaries are freely available on Internet, English is used as a "pivot" for constructing the Chinese-French translation equivalents by transitivity. For this, we use a method, proposed by XMU, of probability induction based on topic similarity, which produces C-F translation tables from C-E and E-F translation tables. For getting good C-F parallel corpora, we use a web-based collaborative post-editing system that can trigger the incremental improvement of the MT system by using MT evaluation metrics and extracting the "best part" of the current translation memory.

Mots-clés : traduction automatique statistique (SMT), chinois-français, domaine économique
Keywords: SMT, Chinese-French, Economic Domain

1 Introduction

The desire and need for cross-cultural communication between China and Europe, both officially and non-governmentally, are on the increase. Especially, France is an important strategic partner of China and has deep relationships with China in many fields such as economy, science and technology, culture and education, etc. But the language barrier between Chinese and French is impassable in most situations.

Concerning the economy, the cooperation between China and France grew rapidly in the past decades. France is now China's fourth largest trade partner in the EU, behind Germany, the Netherlands and the UK. According to data from the National Bureau of Statistics of China, bilateral trade between China and France increased from $13.4 billion in 2003 to $51 billion in 2012. Two-way direct investments are also on the rise. With the continuous improvement of China's investment environment, more and more French investors intend to invest in China. However, most portals of China, such as the website of Shanghai Stock Exchange and the website of Shenzhen Stock Exchange, only provide a Chinese version and an English version, but no French version. Moreover, Chinese investors are also more and more on the look for opportunities in France and other francophone areas, notably in Africa.

Although French-Chinese bilingual applications are urgently needed in many situations, not much work has been done yet on French-Chinese MT, and French-Chinese is still an under-resourced language pair as there are no large good

quality freely available bilingual lexico-semantic resources, and no sufficiently good MT systems. Most French-Chinese MT systems, in particular GT (GoogleTranslate[1]) use English text as a pivot, which introduces more errors, often degrading translation quality to uselessness, as illustrated below. Results may be useful for guessing the topic of a text, but are otherwise bad or misleading, while investors need precise translations to decide whether to invest and where. Here are two examples, both selected from the announcements of the Shanghai Stock Exchange. We show the result of GT: output of GT-zh-en and then of en-fr (calling GT on zh-fr gives the same outputs) and en-PE-fr (post-edited en).

Ex1 (Source): 恢复交易后，如该证券交易中再次出现异常情况，本所可实施第二次停牌，停牌时间持续至今日收盘前五分钟。
(GT-zh-en) After the resumption of trading in the securities trading as abnormal situation occurs again, this can be implemented by the second suspension, suspension lasted until today the closing five minutes.
(GT-zh-en-PE) If the trading of this security appears to be abnormal again after resumption of its trading, we may perform a second suspension upon it and this suspension will last until five minutes before today's closing.
(GT-zh-en-fr = GT-zh-fr): Après la reprise de la négociation dans le négoce de titres comme situation anormale se produit de nouveau, ce qui peut être mis en œuvre par la deuxième suspension, la suspension **a duré** jusqu'à aujourd'hui les cinq dernières minutes.
(GT-zh-en-PE-fr): Si la négociation de ce titre semble être anormal à nouveau après la reprise de ses opérations, nous pouvons effectuer une deuxième suspension sur elle et cette suspension durera jusqu'à cinq minutes avant la clôture d'aujourd'hui.

Ex2 (Source): 四、凡 2013 年 12 月 31 日前在本所上市的公司债券，以及 2014 年在本所上市时未披露 2013 年年度报告的公司债券，应于 2014 年 4 月 30 日前完成本次年度报告的披露工作。
(GT-zh-en) Fourth, where the corporate bond December 31, 2013 are listed in this, as well as the 2014 listed in the 2013 annual report of undisclosed corporate bonds should be April 30, 2014 disclosed in the annual report is completed work.
(GT-zh-en-PE) Fourth, the bonds which were listed before December 31, 2013 or listed in 2014, but have not yet disclosed their annual reports of 2013, should complete the disclosure of their annual reports before April 30, 2014.
(GT-zh-en-fr = GT-zh-fr): Quatrièmement, lorsque **le lien** de l'entreprise 31 Décembre , 2013 **figurent dans cet , ainsi que de 2014** figurant dans le rapport annuel 2013 d'obligations de sociétés non divulgués devrait être de 30 Avril , 2014 communiquées dans le rapport annuel **est terminé travail**.
(GT-zh-en-PE-fr): Quatrièmement, les **liens** qui ont été répertoriés avant le 31 Décembre 2013 répertoriés en 2014, mais n'ont pas encore divulgué leurs rapports annuels de 2013, devrait compléter la divulgation de leurs rapports annuels avant le 30 Avril 2014.

Examples above show that GT uses English as a pivot for the zh-fr direction. However, it is not the case for the fr-zh direction: outputs are different. Google Translate is claimed to be one of the "state-of-the-art" MT system. The truth is however that "state-of-the-art" for general MT systems means "very bad to useless": not knowing Chinese, one is at a loss to guess the exact meaning. Also, even if an output looks good (due to the use of a good target language model), its reliability may be very low, especially when negation appears or disappears at random (ex. 2 in French), or, as in these examples, when essential words are badly translated ("If" → "après" [after] and not "si", "bond" → "lien" [link] and not "obligation", "last" → "dernière" instead of "durer"), grammar is wrong leading to ununderstandability (ex. 2 in French, at several points), dependencies have been modified ("its resumption of" should be "resumption of its"), and bad handling of time in English leads to incorrect tenses in French (ex. 1 and 2). Actually, Example 2 is total gibberish in French. No sense can be extracted from it. Example 1 has actually a *negative* adequacy[2] in French, as it gives counterfactual and misleading information (the suspension "has lasted until…" instead of "will last until…").

This illustrates the need of building direct MT systems between French and Chinese, specialized to the domains (stock option markets, economy) and to the corresponding sublanguages. That will certainly considerably improve MT quality. But, even if one augments the BLEU by 25% or more[3], the problems above remain. In situations where exactness of meaning is crucial and post-editing would require professionals working round the clock (because flash reports have a life expectancy of only a few hours). To improve quality, semantic and discourse information should be used.

Section 2 will now give an overall description of this project. Then two aspects where progress has been made will be described in Section 3 in more detail.

[1] http://translate.google.com/

[2] Adequacy is usually measured on a scale from 0 to 5, but should rather be similar to a kappa coefficient, ranging from -1 to +1.

[3] In an experiment conducted at the EU, a Moses system was built on 50,000 sentences of the corpus of the Council. Compared to the unspecialized system built on 20 M sentences (400 times more!), it showed an *increase by 25%* of the BLEU measure.

2 Overview of the Project

As mentioned in Section 1, the overall goal of this project is to construct economy-oriented Chinese-French computational linguistic resources, e.g. bilingual semantic resources, parallel corpora, and discourse structure annotation banks, etc., then propose a new SMT model making use of semantic and discourse information, and finally build a web-based collaborative post-editing platform. We distinguish three levels: data development, fundamental research, and tool building. FIGURE 1 shows the overall organization of this project, in which three points should be noted.

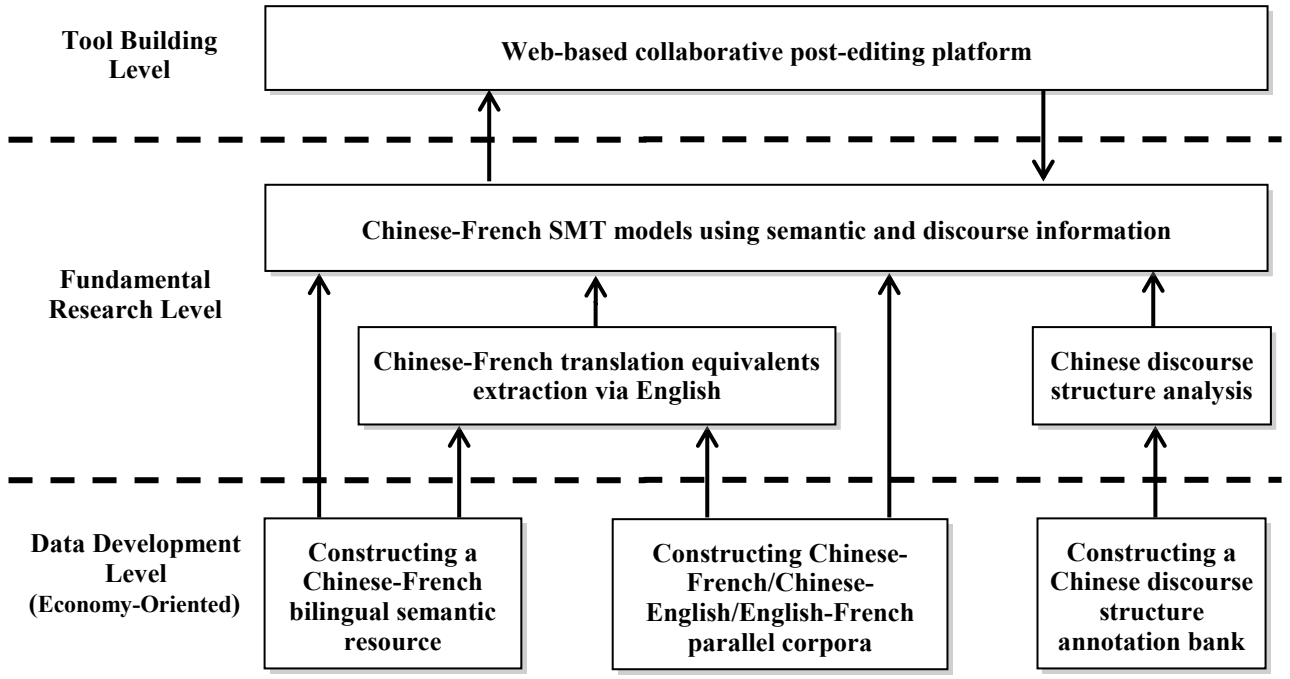

FIGURE 1 : Overall picture of the project organization

Firstly, a pivot-based method is used for extracting the Chinese-French translation equivalents. We choose to use English as a pivot, because it is much easier to gather large-scale Chinese-English and English-French parallel texts than to directly collect Chinese-French ones. Moreover, pivot-based methods (Wu and Wang, 2007; Paul et al., 2009; Wu and Wang, 2009) have been proven to be effective when building SMT systems for under-resourced language pairs.

Secondly, the translation model uses discourse-level information. Actually, recent research has shown that discourse-level information is quite important for machine translation systems, especially those concerning Chinese (Gong et al., 2011, 2012; Tu et al., 2013). As Chinese does not have grammatical markers of tense, the time at which an action takes place usually has to be inferred from the context. Consider again example 1 in Section 1. The source sentence contains three clauses, namely "恢复交易后，如该证券交易中再次出现异常情况 (If the trading of this security appears to be abnormal again after resumption of its trading,)", "本所可实施第二次停牌 (we may perform a second suspension upon it)" and "停牌时间持续至今日收盘前五分钟 (and this suspension will last until five minutes before today's closing.)". Because of the omission of the object in the second clause, we don't know what the second suspension will be executed for, and the unclearness of the tense in the third clause also contributed to the inadequacy of the translation. Suppose now that we have the discourse graph shown in FIGURE 2 below. Both the omitted information and the tense information could then be inferred. Therefore, we decide to conduct research on Chinese-French MT making use of discourse-level information.

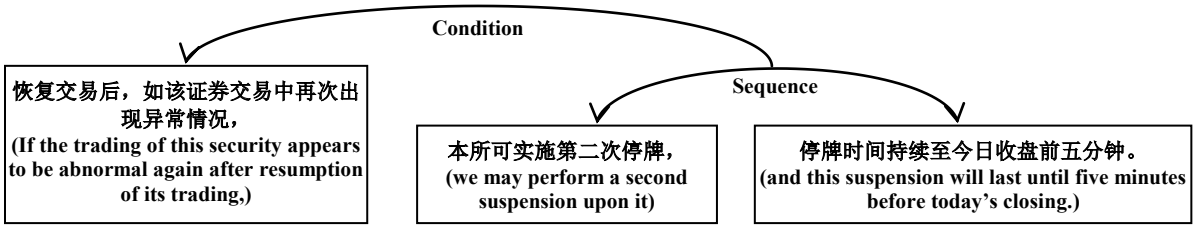

FIGURE 2 : The discourse graph of the source sentence of Example 1

Thirdly, semantic knowledge is used when extracting the translation equivalents. We would do so because we found that the previous pivot-based approaches, which do not take semantic knowledge into account, may produce incorrect translation equivalents due to the lack of sufficient context. For example, suppose we have a phrase pair "银行 ↔ bank" in the Chinese-English bilingual phrase table, and phrase pairs "bank ↔ la banque" and "bank ↔ la rive" in the English-French bilingual phrase table. Then, using the transfer method (Paul et al., 2009; Wu and Wang, 2009), we get two Chinese-French phrase pairs, i.e. "银行 ↔ la banque" and "银行 ↔ la rive". However, the latter is obviously wrong. We expect that, by incorporating semantic information, this problem could be overcome.

We distinguish seven subtasks in this project (see FIGURE 1), briefly described below.

1) Constructing Chinese-French bilingual lexico-semantic resources

This subtask aims at creating bilingual lexico-semantic data by enriching HowNet (Dong and Dong, 2006), a Chinese-English conceptual database, with French data. In this subtask, economy-related bilingual terms will be integrated into the semantic resource.

2) Constructing Chinese-French/Chinese-English/English-French parallel corpora

In order to carry out SMT research, parallel corpora are needed. Since very few Chinese and French bilingual data are freely available on Internet, it is not easy to construct large-scale Chinese-French parallel corpora, especially economy-oriented ones. Therefore, it is reasonable to build the resources of Chinese-French SMT systems using English as a pivot. To this end, we are building Chinese-English and English-French bilingual corpora related to the domain of economy. Then, to support learning structure transformation knowledge between Chinese and French, a medium-scale Chinese-French parallel corpus will also be created.

3) Constructing Chinese discourse structure annotation bank

In order to incorporate discourse-level information in our Chinese-French SMT system, we first need a Chinese discourse structure analysis module trained on Chinese discourse structure annotation banks. However, the Chinese discourse structure annotation banks, currently under construction to support discourse-based Chinese-related MT research, are not yet available now (Li et al., 2012). Hence, this subtask aims at constructing a Chinese discourse structure annotation bank based on the Chinese part of the economy-oriented parallel corpora constructed in Subtask 2. The course of constructing the annotation bank could be summarized as two steps. Firstly, sentences are partitioned into sentence groups according to their topic similarities. Secondly, the relationships among the sentences in the same groups, as well as the time information of the sentences or the time relationships between sentence pairs are humanly tagged. The annotation result for example 1 in Section 1 may look like as follows:

(Discourse (Sentences (S1, 0-21), (S2, 22-32), (S3, 33-48)),
(Sentence-relationships (Condition S1, S2), (Sequence S2, S3)),
(Time-information Time(S1)=present, Time(S2)>Time(S1), Time(S3)=Time(S2)))

4) Chinese-French translation equivalents construction via English

Given the constructed Chinese-English and English-French parallel corpora, the objective of the subtask is to construct Chinese-French translation equivalents. In the course of equivalents extraction, semantic information based on the lexico-semantic resources constructed in Subtask 1 will be considered.

5) Chinese discourse structure analysis

The objective of this subtask is to research and propose a Chinese discourse structure analysis model based on the Chinese discourse structure annotation bank constructed in Subtask 3.

6) Chinese-French SMT models using semantic and discourse information

The goal of this subtask is to propose a new Chinese-French SMT model using semantic and discourse information. In the course of translation, the tense and semantic consistency across sentences will be considered, based on the discourse graphs produced by the analysing module in Subtask 5 and the time model trained with the annotation bank in Subtask 3.

7) Web-based collaborative post-editing platform

The objective of this subtask is to build a collaborative web-based post-editing platform. This platform is built based on iMAG/SECTra (Wang and Boitet, 2013), and incorporates the economy-oriented SMT system developed in Subtask 6.

3 Some preliminary progresses

This project started in September 2013. Since then, we progressed on two fronts: data collection and pivot-base construction of bilingual equivalents.

1) Data collection

From websites related to stock exchange, about 1000 bilingual web pages have been crawled and handled so far (Table 1 shows the statistics of this dataset).

Direction	# of pages	Size
zh-en	761	39.4M
en-fr	250	13.5M

TABLE 1: Statistics of the data collected so far

At the same time, with the participation of 2 Chinese students, we started the process of creating a Chinese-French parallel corpus via post-editing, using an iMAG collaborative gateway, as in (Wang and Boitet, 2013).

2) Research on the pivot-based construction of bilingual equivalents

In previous pivot-based methods, such as (Wu and Wang, 2007), the translation probability induction may become inaccurate if the semantic tendencies of the source-pivot (SP) corpus and the pivot-target (PT) corpus are different. To overcome this problem, an effective method is to use context information to measure the semantic similarity of the rule pairs. To solve this problem, we have proposed and validated a pivot probability induction method based on topic similarity information. It consists of two steps.

Firstly, we use the topic model (Blei, 2003) to discover the latent topic structure for the pivot language document and each synchronous pivot phrase rule is then attached with the topic distribution according to the document it comes from. Formula 1 is used to assign topic distribution to a given rule.

$$P(z_i \mid r) = \frac{\sum_{D \in \sum_*} count(I, D) P(z_i \mid d_p)}{\sum_{z \in Z} \sum_{D \in \sum_*} count(I, D) P(z \mid d_p)} \quad (1)$$

where D is a bilingual document in the given bilingual corpus $\sum_*$. $P(z_i \mid d_p)$ represents the probability of the ith topic of the pivot language document d_p in D. In addition, the function $count(I, D)$ denotes the frequency of the rule instance I in D.

Secondly, the probability induction is executed by computing the topic similarity for the rule pairs instead of using simple phrase translation probability multiplication between the SP and PT translation models. Formulas 2-3 are used:

$$\phi(\bar{t} \mid \bar{s}) = \frac{\sum_{\bar{p}} sim(P(Z \mid r_{\bar{s}\bar{p}}), P(Z \mid r_{\bar{p}\bar{t}}))}{\sum_{\bar{t}} \sum_{\bar{p}} sim(P(Z \mid r_{\bar{s}\bar{p}}), P(Z \mid r_{\bar{p}\bar{t}}))} \quad (2)$$

$$\phi(\bar{s} \mid \bar{t}) = \frac{\sum_{\bar{p}} sim(P(Z \mid r_{\bar{t}\bar{p}}), P(Z \mid r_{\bar{p}\bar{s}}))}{\sum_{\bar{s}} \sum_{\bar{p}} sim(P(Z \mid r_{\bar{t}\bar{p}}), P(Z \mid r_{\bar{p}\bar{s}}))} \quad (3)$$

where $\bar{s}$, $\bar{p}$, $\bar{t}$ and Z are the source phrase, pivot phrase, target phrase and topic distribution, respectively. $sim(x, y)$ is the similarity function, and is defined as the cosine (pseudo-distance) of vectors x and y (Formula 4).

$$sim(x, y) = \cos(x, y) = \frac{\langle x, y \rangle}{\|x\| \|y\|} \quad (4)$$

The experimental results showed that our method greatly outperforms the baseline. A paper related to this work was published in the *Journal of Computational Information Systems* in 2013 (Huang et al., 2013).

4 Conclusion

This paper gives an introduction of a joint project between GETALP and the NLP lab of XMU on building economy-oriented Chinese-French SMT systems, as well as its preliminary progress. We demonstrate the need of building direct Chinese-French MT systems specialized field to that field and its sublanguages, and also explain why it seems to be necessary to incorporate semantic and discourse-based information to obtain a quality (of raw MT) sufficient for the needs of users, in a situation when usually there is no time for post-editing. This project is still in its initial stages and many efforts are required in the future. However, specialized versions not yet using the new semantic- and discourse-related features should be demonstrable at the time of the conference. We hope and believe that the future results of this project will be useful in overcoming the language barrier of the communications between China and France in the economy-related field.

Acknowledgements

This work was supported by the State Scholarship Fund of China (Grant No. 201208350055), the National Natural Science Foundation of China (Grant No. 61005052) and the Natural Science Foundation of Fujian Province of China (Grant No. 2011J01369). It was also partly supported by the French ANRT agency and the Lingua et Machina firm.

References

BLEI D. M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning 3*, 993-1022.

DONG, Z.,AND DONG, Q. (2006). HowNet and the Computation of Meaning. *World Scientific Publishing Co.Pte.Ltd*.

GONG, Z., ZHANG, M., ZHOU, G. (2011). Cache-based Document-level Statistical Machine Translation. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, Edinburgh, UK, pp. 909-919.

GONG, Z., ZHANG, M., TAN, C., ZHOU, G. (2012). N-gram-based Tense Models for Statistical Machine Translation. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju, Korea, pp. 276-285.

HUANG, Y., SHI, X., SU, J., CHEN Y., HUANG, G. (2013). Pivot Probability Induction for Statistical Machine Translation with Topic Similarity. *Journal of Computational Information Systems*, 20(9), 8351-8359.

LI, Y., ZHU, K., ZHOU, G. (2012). Summary of research on English discourse parsing. *Application Research of Computers*, 29(6).

PAUL, M., YAMAMOTO, H., SUMITA, E., AND NAKAMURA S. (2009). On the Importance of Pivot Language Selection for Statistical Machine Translation. *Proceedings of NAACL-HLT'09*, 2009, pp. 221-224.

TU, M., ZHOU Y., AND ZONG C. (2013). A Novel Translation Framework Based on Rhetorical Structure Theory. *The 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, Sofia, Bulgaria.

WANG, L., AND BOITET, C. (2013), Online production of HQ parallel corpora and permanent task-based evaluation of multiple MT systems: both can be obtained through iMAGs with no added cost. *Proceedings of MT Summit XIV, The 2nd Workshop on Post-Editing Technologies and Practice.* Nice, France 2 - 6 September 2013.

WU, H. AND WANG H. (2007). Pivot Language Approach for Phrase-Based Statistical Machine Translation. *Proceedings of ACL'07*, 2007, pp. 856-863.

WU, H. AND WANG, H. (2009). Revisiting Pivot Language Approach for Machine Translation. *Proceeding of ACL'09*, 2009, pp. 154-162.