Treatment of Complex Sentences, Modality and Verbal Structures in Linguistics-Based MT

Alexis Kauffmann(1,2)　Daisuke Kawahara(1)　Sadao Kurohashi(1)
(1) Kyoto University
(2) University of Geneva
alexis.kauffmann@unige.ch
{dk,kuro}@i.kyoto-u.ac.jp

Abstract

In this paper, we present a series of enhancements to the English-Japanese version of a multilingual Linguistics Based Machine Translation system, for the translation of complex sentences, modality and complex verbal structures. The system is using a classical transfer-based architecture and dedicated lexical databases. Relying on linguistic data acquired on large corpora or compiled from the web, corrections have been done in constituent reordering, lexical selection and verb conjugation. Even if the system is not, so far, as efficient as state-of-the-art English-Japanese MT systems, the results show a clear progress and underline the interest of using syntactic information in MT.

1 Introduction

Linguistics based MT (LBMT,[2]) is based mainly on linguistic knowledge, and usually relies on hand made programming, whereas Corpus based MT (CBMT, which includes Statistical based MT[3] and Example-Based MT[5]) relies on machine learning of aligned bilingual corpora. Recently, methods based on statistics and machine learning have recently shown impressive results[10], and many of them try to incorporate syntactic data to their model[11].

We try here to combine both methods, but instead of using a CBMT system, we use a classical LBMT system and improve it with linguistic data acquired from large corpora or from the web by statistical techniques.

First programmed to translate simple sentences, the English-Japanese version of the Its2 MT system[9] needed to be improved, and to be able to translate complex sentences or sentences expressing modality or containing other complex verbal structures involving English infinitives or gerunds. This work also gives the occasion to test the adaptability of this multilingual system to Japanese, as it was first designed for Western languages.

In this article, we first give a brief overview of the system architecture, in section 2. Then, section 3 describes the treatment of complex sentences. Sections 4 and 5 focus on the translation of modality and verbal structures. Section 6 presents the experiments and their results.

2 Description of the MT system

Its2 is a LBMT system. It relies on an implementation of theoretical translation rules, based on grammatical and linguistic knowledge, and on the use of dedicated lexical databases[9].

Its2 is a multilingual system. It was first made to translate between Western languages such as French, English, German, Italian and Spanish. Since 2009 it has been adapted for English-to-Japanese and French-to-Japanese translation too.

It is using a classical transfer-based architecture. The first step in the translation process is the syntactic parsing of the source sentence, achieved by the syntactic parser Fips[8]. It is followed by a bilingual transfer phase and by the generation of the target sentence. The system is built in an object-oriented architecture, implemented in Component Pascal language. The core multilingual modules interact with specific monolingual modules for parsing and generation, and bilingual ones for lexical transfer.

There are three kinds of lexical database tables used by the system: the monolingual lexeme[1] tables, monolingual word tables and bilingual correspondence tables. Lexeme tables store morphologic, syntactic and semantic information about lexemes, word tables store conjugated forms of each lexemes, and bilingual correspondence tables store and rank the translation between lexemes or collocations. The databases also contain collocations or multi-word expressions.

3 Complex sentences

Translation of complex sentences, from English to Japanese, is often subject to clause reordering, which varies depending on the type of sentence structure. Thus,

[1] Lexemes are the canonical forms of words, the ones that are expected to be found in dictionnaries[7]. The opposite of a canonical form is a conjugated form.

Japanese Conj. Type	Rule
だから Type	Cut the sentence between clauses.
けど Type	Comma after the conjunction.
と Type	Delete conjunction. Add a coma. Put the left clause verb at "て" form.
そして Type	Put the left clause verb at "て" form.

Figure 1: Clause Coordination Translation Rules

we have studied the different possible cases, such as juxtaposition and different types of coordination of subordination. Those cases can also involve a modification of verb conjugation in the target sentence. We wrote theoretical transfer rules based on this linguistic study and implemented most of them to the system.

This led us to notice the need for an accurate classification of Japanese conjunctions, conjunctive adverbs and conjunctive particles. Then, in a corpus of 90000 sentences, we checked the number of occurrences and percentage of sentence head and sentence final occurrences of every Japanese conjunctive word. This resulting data was used to annotate Its2 Japanese lexical database, and enabled the system to produce correct clause reordering or verb conjugation in the target Japanese sentence. For example, for the input sentence(1), instead of incorrect translation(2), we could generate correct translation(3). Figure 1 shows the translation rules for clause coordination, depending on the conjunction type.

(1) I ate okonomiyaki and I drank beer.

(2) お好み焼きを 食べた と ビールを
okonomiyaki wo tabeta to biiru wo
okonomoyaki ate and beer
飲んだ。
nomda
drank
(I) ate okonomiyaki and (I) drank beer.
(incorrect grammar)

(3) お好み焼きを 食べて、 ビールを
okonomiyaki wo tabete, biiru wo
okonomoyaki ate beer
飲んだ。
nomda
drank
(I) ate okonomiyaki and (I) drank beer.
(correct grammar)

Moreover, as some English conjunctions have several possible translations in Japanese, we sometimes had to implement dedicated lexical selection procedures, which choose the correct one, depending on the source sentence syntactic context.

4 Modality

English modals and other verbs related to modality expression have various possible translations in Japanese[4]. Japanese expression of modality usually requires the use of some specific governing verbs or multi word expressions, and sometimes specific moods for the verb of the object clause.

(4) 私 は 働かなければ
watashi ha hatarakanakereba
I would to work
ならない。
naranai
not happen
I have to work

Studying the results of a statistical analysis of English-Japanese modality translation on 1 300 000 sentence long travel domain aligned corpora and manually collecting more evidence in Wall Street Journal article aligned corpora, we found out what should be the most frequent translations[2] for every English modal or semi-modal, depending on the syntactic context.

We wrote the corresponding theoretical translation rules and tried to implement them to the system. We have been able to implement over 65 percent of those lexicalised transfer rules, but failed to implement the other ones because of time constraints. Still, the resulting modality translations on implemented cases seemed to be correct.

5 Other Verbal Structures

Apart from modals, many other English verbs take verbal or sentential objects. Verbal and sentential object are a part of verbs' arguments, which are defined in verbal subcategorisation frames. We show here how we have set verb subcategorisation data in the lexical databases, and then focus on the generation of Japanese verbal and sentential objects.

5.1 Verb Subcategorisation

Considering subcategorisation[3] or case frames can be useful for a better translation of verbs and their arguments. In the lexical databases we used, every verb lexeme corresponds to a unique subcategorisation[9]. Thus, if a same verb can take, for example, 5 several subcategorisation frames, 5 lexemes will be defined in the monolingual lexicon. So far, about 11600 verb lexemes have been recorded in our English lexical database.

[2]We found out that all English modals have a most frequent Japanese possible translation in a determined syntactic context, except ``could`` which remains highly ambiguous, especially in an affirmative sentence, when there is a lack of semantic and pragmatic information.

[3]The verb subcategorisation describes the syntactic behaviour of a verb: if it is transitive or intransitive, if it can take a sentence as a complement, if it should have an indirect object...

	Verb	Arg 1 (Subject)	Arg 2	Arg 3
English Verb Subcat	go	NP		
	go	NP	PP(to)	
	go	NP	PP(from)	PP (to)
	write	NP	NP	
Japanese Verb Case Frame	行く	が		
	行く	が	に	
	行く	が	から	まで
	書く	が	を	

Figure 2: Verb Subcategorisation in English and Japanese

In the bilingual lexicon, correspondences between English and Japanese verbs should always associate lexemes taking the same arguments in their subcategorisation. These arguments may have different syntactic descriptions in the two languages. For example, an indirect object in English may become a direct object in Japanese.

Our English lexicon already had a detailed description of many subcategorisations for its verbs. Our Japanese one didn't have such detailed data about subcategorisation. So, we extracted data from the Case Frame file, that stores Japanese verb case frames and that was compiled from web-extracted data[1]. Table 2 shows some English verb subcategorisations as they are recorded in LATL monolingual database, Japanese equivalent case frame, as in Case Frame file.

Applying a series of SQL queries, we added the extracted subcategorisation data to our Japanese monolingual lexicon and added about 5000 subcategoriation frames to the 2500 that it already contained.

By another set of SQL queries, we attempted to create right bilingual correspondences between English and Japanese verb subcategorisations. When two different Japanese subcategorisation were possible for a same English one, we gave a higher ranking to the one with the higher count score in the Case Frame file. We added then about 8000 bilingual correspondences in the bilingual database. Finally, a Japanese speaker checked and corrected by hand 2400 correspondences that took both an English prepositional object and a Japanese object with a postpositional particle, in order to avoid semantical errors.

5.2 Verbal and Sentential Objects

Many English verbs take verbal or sentential objects. Depending on verb subcategorisation, verbal objects can either be infinitives or gerunds, and sentential objects can also have other tense conjugations.

In Japanese, several conjugations are possible for verbal or sentential objects, depending on the verb and the context. The data extracted from Case Frame file showed that verbs with an argument with conjunctive particle と("to") usually take a sentential object,

(5) 彼は 帰った と
kare ha kaetta to
he came back home that
思います 。
omoimasu
think
(I) think that he came back home.

and that transitive verbs can take a nominalized sentential object.

(6) 梅酒を 飲んだのを
umeshu wo nomda no wo
plum liquor drank(the action of)
覚えっている 。
oboetteiru
remember
I remember drinkinking plum liquor.

Other data showed that Japanese semi-auxiliaries and light verbs take verbal objects. Their conjugation can either be base form or gerundive in -て("Te" form).

(7) 私 は やって みます 。
watashi ha yatte mimasu
I doing try
I'll try to do it.

We extracted conjugation parameters from those files and added them, completing subcategorisation frames, to our Japanese monolingual database. We implemented generation procedures, to give the correct conjugation to the output, depending on the parameters.

The overall resulting translation output was satisfying in many cases. However, some further work would be necessary to reach a perfect output. For example, a use of language models would enable to choose between several candidates when the tense of the target sentence verb cannot be clearly determined by syntactical rules.

6 Experiments and Results

We tested all the improvements of the system on basic sentence sets and the improvements appeared clearly. Then, we compared BLEU scores on a sample of 500 sentences of a scientific paper abstract corpus. The results showed a progression of +0.3% with the treatment of complex sentences and +0.6% with the treatment of complex sentences, modality and verbal objects, reaching a final BLEU score of 2.52% . BLEU scores are usually lower for rule-based or linguistic based MT than for SMT or methods based on machine learning. However, this is not a sufficient reason to explain the low scores we obtained. The major reason is that the translation quality of our system, remains so far insufficient, and not as

good as state-of-the-art ones[11]. A linguistic analysis of the generated output showed both a clear improvement on the treatment of the mentioned syntactic structures, and a difficulty to generate fluent output for complex input sentences.

7 Conclusion

In this paper, we have shown how the use of grammar and syntax combined with data acquired by statistical methods from large corpora or from the web can lead to an improvement of the quality of translation in English to Japanese MT. However, even if this improvement is very clear on some specific types of syntactic structures and on basic sentences, we have not been able to obtain a good quality of translation for a wide range of complex sentences.

This can be explained by several factors, such as errors occurring at the syntactic parsing level, insufficiencies in the English-Japanese bilingual lexicon, syntactic rules that have not been implemented yet in the system, and a lack of bilingual data for the translation of collocations and multi-word expressions.

In a future work, we should try to solve these problems. We may also be interested in expanding our model to some statistical corrections, referring to language models to improve the lexical selection, or using machine learning or by automatic post-editing [6]. Another possible expansion would be to enable the user to choose a Japanese politeness level for the target sentence, and to specify a verb conjugation and lexical selection that depends on this politeness level.

Acknowledgements

We would like to thank Nobuhito Tamaki, Toshiaki Nakazawa and Michiaki Mochizuki for their help. The research described in this paper has been supported by a grant from the Swiss National Science Foundation.

References

[1] Daisuke Kawahara and Sadao Kurohashi. Case frame compilation from the web using high-performance computing. In *Proceedings of The 5th International Conference on Language Resources and Evaluation (LREC-06)*, 2006.

[2] Satoshi Kinoshita, John Phillips, and Jun-Ichi Tsujii. Interaction between structural changes in machine translation. In *Proceedings COLING-92, Nantes, France*, 1992.

[3] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), Prague*, 2007.

[4] Masaki Murata, Kiyotaka Uchimoto, Qing Ma, Toshiyuki Kanamaru, and Hitoshi Isahara. Error analysis of translation of tense, aspect, and modality in machine translation system. In *Proceedings of FIT 2005, pp.77-80*, 2005.

[5] Toshiaki Nakazawa and Sadao Kurohashi. Fully syntactic ebmt system of kyoto team in ntcir-8. In *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-8), pp.403-410, Tokyo, Japan*, 2010.

[6] Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn. Rule-based translation with statistical phrase-based post-editing. In *Proceedings of the Second Workshop on Statistical Machine Translation, pp. 203-206, Prague*, 2007.

[7] Lonneke VanDerPlas. *Automatic lexico-semantic acquisition for question answering*. PhD thesis, University of Groningen, 2008.

[8] Eric Wehrli. Fips, a "deep" linguistic multilingual parser. In *Proceedings of ACL 2007, Prague, Czech Republic*, 2007.

[9] Eric Wehrli, Luka Nerima, and Yves Scherrer. Deep linguistic multilingual translation and bilingual dictionaries. In *Proceedings of Fourth Workshop on Statistical Machine Translation, pp. 90-94*, 2009.

[10] Deyi Xiong, Min Zhang, and Haizhou Li. Learning translation boundaries for phrase-based decoding. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pp. 136-144, Los Angeles, California*, 2010.

[11] Xiaoqiang Luo Young-Suk Lee, Bing Zhao. Constituent reordering and syntax models for English-to-Japanese statistical machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pp. 626-634, Beijing*, 2010.