Post-MT Term Swapper: Supplementing a Statistical Machine Translation System with a User Dictionary

Masaki Itagaki, Takako Aikawa
Microsoft
One Microsoft Way, Redmond, WA 98052, U.S.A.
(mitagaki@microsoft.com, takakoa@microsoft.com)

Abstract

A statistical machine translation (SMT) system requires homogeneous training data in order to get domain-sensitive (or context-sensitive) terminology translations. If the data consists of various domains, it is difficult for an SMT system to learn context-sensitive terminology mappings probabilistically. Yet, terminology translation accuracy is an important issue for MT users. This paper explores an approach to tackle this terminology translation problem for an SMT system. We propose a way to identify terminology translations from MT output and automatically swap them with user-defined translations. Our approach is simple and can be applied to any type of MT system. We call our prototype "Term Swapper." Term Swapper allows MT users to draw on their own dictionaries without affecting any parts of the MT output except for the terminology translation(s) in question. Using an SMT system developed at Microsoft Research, called MSR-MT (Quirk, et al., (2005); Menezes & Quirk (2005)), we conducted initial experiments to investigate the coverage rate of Term Swapper and its impact on the overall quality of MT output. The results from our experiments show high coverage and positive impact on the overall MT quality.

1. Introduction

Terminology translations need to be context-sensitive. For instance, the Japanese translation corresponding to the term "memory resources" in (1a) is appropriate in the IT domain, whereas the one in (1b) would be appropriate in the general domain but would be awkward in the IT domain.

(1) There is a shortage of memory resources.
a. メモリ リソースが不足しています。
"(computer) memory resources" =>OK in the IT domain

b. 記憶資源が不足してます。
"(brain) memory resource" => OK in the general domain

Selecting context-sensitive terminology translations would be challenging for an SMT system, especially when its training data is not homogeneous in terms of domain. Even when the system has homogeneous training data, this is difficult. For instance, at Microsoft, we have been localizing technical documents over the last couple of years, using an SMT system called MSR-MT (Quirk, et al, 2005). The MSR-MT training data comes solely from the IT domain, yet we find terminology translations challenging as many product groups use different terminology translations for the same English terms (Itagaki, et al, 2007).

Such domain differences in training data for SMT systems may have led to different term translations by different systems. For instance, using several Web-based translation systems, we translated the simple terms, "spot light" and "traffic light," into Japanese. As shown in Table 1, all the systems return different translations, which are all correct in isolation.

	spot light	traffic light
System A	スポット ライト	トラフィックライト
System B	光スポット	交通信号灯
System C	点照明	交通信号

Table 1: Various terminology translations in Japanese across different systems

As we extend the use of MT across different domains or across different systems, we need a way to make terminology translations context-sensitive. One solution is to allow users to apply their own dictionary to an MT system, so that they can use specific terminology translations. Term Swapper provides a way to achieve this.

2. Overview of Term Swapper

2.1 Terminology Translation Variations based on Contexts

One of the difficulties in applying a user-defined dictionary to an SMT system is to identify the locations of specific terminology translations in its output. For instance, using MSR-MT, we tested sentences that contain the term "Family Safety" (a security feature name in Windows Vista). Table 2 lists the Japanese translations from MSR-MT.

Source	Target
Family Safety exists.	家族の安全が存在します。 ***'The safety of family exists.'***
This is Family Safety.	これはファミリ安全です。 ***'This is Family Safely.'***
Click on the Family Safety button.	ファミリ Safety をクリックします。 ***'Click Family Safety.'***

Table 2: Various terminology translations for 'Family Safety", depending on a given context"

The underlined parts are the translations for the term, "Family Safety." As shown, depending on a given context, the translation varies. This is because an SMT system is designed to return terminology translations probabilistically. Predicting the location of a particular terminology translation in a given MT output is thus extremely difficult. Yet, unless we can identify the location of the terminology translation in question, it is impossible to swap it with the user-defined terminology translation**.**

The simplest way to identify the MT-generated translation should be to retranslate the term itself. See the example in (2), in which "Family Safety" is in question.

(2)

a. See the chapter entitled **Family Safety**.
家族の安全の章を参照してください。

Then you could re-translate the term itself to get a corresponding translation.

b. **Family Safety**
家族の安全

If this translation matches the one in the first translation sentence, it is trivial to swap it with a user dictionary entry.

c. See the chapter entitled **Family Safety**.
[User Translation] の章を参照してください。

However, as we discussed earlier, term translations in an SMT system change dynamically depending on the context. To examine this issue, we translated 500 sample sentences from game software with MSR-MT. Then we translated all noun terms (both compounds and singletons) from the sentences in isolation in order to see if the term translations are the same as the ones in the original sentence translations. We found that only 65.9% of the noun terms were the same.

This indicates that if you want to obtain a term translation by retranslation, you need to pass the term as well as the sentence form to the system rather than just the term itself.

2.2. Method

We tried to find some simple sentence patterns that could induce term translations as in the original translation sentences. We manually analyzed 454 randomly selected MT-generated sentences from one of the Web-based applications as a test data set. By analyzing sentence structures and grammatical features, we came up with 15 sentence patterns that may induce various term translations that could not be obtained by translating a term itself (Appendix I). Table 3 shows some samples.

Then we identified a way to automatically strip off the parts that are not relevant to the terminology translation(s). In the above samples, we strip off translations for the parts "exists", "This is" and "is a word." This way we can isolate the translation for "X" in Table 3.

	Pattern Notations	Patterns *[Examples]*	*Descriptions [Examples]*
A	SUBJ+V	X exists.	A term as the subject of an intransitive verb.
		Family Safety exists.	*家族の安全が存在します。*
B	SUBJ_THIS_IS	This is X.	A term as the predicate of a copula
		This is Family Safety.	*これはファミリ安全です。*
C	SUBJ+BE	X is a word.	A term as the subject of a copula.
		Family Safety is a word.	*ファミリ* Safety *は、単語です。*

Table 3: Sample templates for Term Swapper

Once we identify an MT-generated term translation, we replace it with a translation that a user desires. The examples in (3) below demonstrate swapping translations for "Family Safety." Let us suppose that a user wants to translate Family Safety as 家族のための安全設定 (lit., 'the Setting of Family Safety') and she wants to translate (3a) with this term translation. Based on the list of the terminology translation candidates, we can identify the location of the terminology translation for "Family Safety" in the raw MT output (i.e., the underlined part in (3b)) and can swap it with the user defined translation, resulting in (3c).

(3)

a. [English Source] Your parent has not finished setting up Family Safety for your account.

b. [MT Output] 親はアカウントに、家族の安全をセットアップすることを終わらなかった。

c. [Output with Term Swapper] 親はアカウントに、家族のための安全設定をセットアップすることを終わらなかった。

3. Term Swapper Templates

3.1. The Nature of Swapper Templates

The linguistic patterns in Term Swapper templates (see Appendix I) need to be simple and predictable as the simplicity of the templates is what allows us to predict possible terminology translation candidates for a given MT system. This is also very important in ensuring good system performance. The more complicated a template gets, the longer it takes to strip off unrelated text in an output translation. To see this more concretely, let us examine Template C in Table 3.

(4) This is Family Safety.
~~これは~~ファミリ安全~~です。~~

The parts struck out are irrelevant to the term translation in question. Let us walk though how Term Swapper can predict such parts, using example (4) above. Example (4) uses Template C in Table 3 (i.e., "This is X."). Given this template, we can assume that potential translation strings from any MT system would be one of the following provided in (5).

(5) これ{は/が/は、/が、}TERM{です/だ/である}。

The variants of the case marker in (5) (i.e., {は/が/は、/が、}) or those of the predicate 'be" (i.e., です/だ/である) are necessary as these parts (as well as terminology translations themselves) vary, depending on a given MT system or even depending on a given term. Table 4 and Table 5 illustrate this point.

	This is Family Safety.
System A	~~これは~~ファミリ安全~~です。~~
System B	~~これは~~家族の安全~~である。~~
System C	~~これは、~~ご家族の安全~~です。~~

Table4: Translation variations across different systems

This is TERM.	The output from MSR-MT
This is Microsoft.	~~これは~~ Microsoft ~~です。~~
This is NULL.	~~これが~~NULL ~~です。~~
This is a Manager Error ID."	~~これは、~~マネージャ エラー ID~~です。~~

Table5: Translation variations depending on Terminology

Table 4 lists the translations of the sentence, "This is Family Safety", from three different systems. Table 5, on the other hand, lists the translations of the same template with three different terms from MSR-MT. The underlined parts are the terminology translations in question and the parts struck out are those that Term Swapper needs to strip off. As shown, the translations of function words or predicate parts vary, depending on a given term (see Table 5) as well as on a given system (see Table 4). In order to achieve the goal of stripping off the parts irrelevant to terminology translations, it is critical for our templates to be as simple as possible.

As explained earlier, we used the training set of 454 sentences to identify simple template codes that could achieve this goal for three language pairs; namely, English -> Japanese, Chinese (Simplified), and Korean. The same templates were used across the three target languages, but different ways of stripping off case markers, predicates, etc. were implemented by each language.

4. Coverage Experiment

4.1. Design and Results

We conducted an investigation to determine the coverage rate of Term Swapper, using MSR-MT. To this end, we first created a dummy user dictionary that consists of 634 English lexical items (all were nouns), each of which was translated as "[DUMMY]" to easily identify it as swapped text. We then took 500 sentences from a game product ("Age of Empires") as the test data. Since the game content is not in our training data, a large part of the terms are unknown to the SMT system. Each of the test sentences contains at least one of the lexical items in our user dictionary. Using MSR-MT, we translated these 500 sentences into the three languages (i.e., Japanese, Chinese, and Korean) with Term Swapper and counted how many sentences contained [DUMMY] string(s) in their MT outputs. Table 6 provides the percentage of the occurrence of the sentences with [DUMMY] for the three languages.

EJ	EC	EK
90.6%	92%	86%

Table 6: The current coverage rates of EJ, EC, and EK systems.

4.2. Error Analysis

As shown in the previous section, there are terms that were not swapped successfully. To find out why Term Swapper failed to swap the terminology translations in question, we looked at some examples, using MSR-MT. Here, we would like to provide a brief analysis of such cases.

The English terms used for the above experiment are all nouns but some of them involve compound nouns. When such compound nouns are translated non-adjacently in the original raw MT output, Term Swapper is destined to fail. Examine (6), for instance.

(6)

a. [Input] Choose what shader model to use.

b. [Dictionary] shader model : シェーダのモデル

c. [MT Output] どのシェーダを使用してモデルを 選択します。

The term in the user dictionary is "shader model". While the raw MT output in (6c) is awkward, the point we want to make is the fact that the translation of 'shader' (i.e., シェーダ) and that of 'model' (i.e., モデル) do not occur in adjacent positions. In such cases, there is no way for Term Swapper to identify the location of the entire terminology translation of a compound noun such as "shader model." Note that this is not a problem intrinsic to Term Swapper. Rather it is a problem intrinsic to the analysis of the MT system. If the translations of the two terms occur adjacently as in (7c), Term Swapper can successfully fix the terminology translation as in (7d).

(7)

a. [Input] Choose a shader model.

b. [Dictionary] shader model : シェーダのモデル

c. [MT Output] シェーダ モデルを選択します。

d. [Fixed Output] シェーダのモデルを選択します。

Admittedly, we still have a lot of room to improve our template codes. However, cases like (6), where compound term translations are split in the original MT output (most likely, due to the misanalysis by a given MT system), would lead to too much complication that would drastically impact system performance.

5. Impact of Term Swapper on MT Quality

5.1. Experiment Design

Term Swapper should give us a big boost in MT quality as it can swap (wrong/unwanted) terminology translations for those that users want to use. To test this objectively, we decided to measure the differences in quality between raw MT outputs (MT outputs without Term Swapper) and fixed MT outputs (those with Term Swapper), using two tools to measure MT quality: (i) Bleu and (ii) Edit-Distance (character-based). We used the same test data we used for the coverage experiment. This time, however, we created a real user dictionary; that is, we assigned proper target translations to all of the 634 English terms. For this experiment, we tested only the English->Japanese language pair.[1]

5.2. Results

Table 8 presents the differences in Bleu and Edit-Distance quality measures when using each system with Term Swapper and without it, respectively. These results indicate the magnitude of the impact of using Term Swapper on MT quality.[2]

Measures	Without Term Swapper	With Term Swapper
Bleu	12.43	22.51
Edit-Distance	0.630429	0.521621

Table 7: Impact of Term Swapper using Bleu and Edit-Distance using MSR-MT

We provided some of the examples from our test data in Appendix II.

5.3. Applying Term Swapper to Other MT Systems

Term Swapper's templates need to be adjusted to some extent for each MT system since the template translations may be slightly different for any given system . However, we simplified the template sentences and structures to try to account for this issue.

We then tested two other translation systems to investigate the applicability of the templates that were designed based on MSR-MT (Table 8).

System A		
Measures	Without Term Swapper	With Term Swapper
Bleu	6.39	13.80
Edit-Distance	0.655887	0.595938

System B		
Measures	Without Term Swapper	With Term Swapper
Bleu	5.93	18.26
Edit-Distance	0.659491	0.562336

Table 8: Impact of Term Swapper using Bleu and Edit-Distance using 3rd-Party MT Systems

While translation quality improved for both of the tested MT systems, the coverage rate was as low as 76.8%. This was due to a high failure rate of irrelevant text cleanup, which indicates that processing templates need to be fine tuned for each MT system. Nevertheless, quite a large number of terms were successfully identified, which contributed to an improvement of Bleu and Edit-Distance scores. [3]

6. Conclusion and Future Work

In this paper, we presented a simple, yet effective way to integrate a user-defined dictionary into an MT system as a post-process approach. The proposed method is pragmatic but admittedly, there is much room for future work. First, the templates of our current prototype are manually created and their coverage rates are not perfect. We would like to explore a way to build such templates automatically and leverage them, so that we can achieve better coverage rates across different languages. Second, Term Swapper needs to be tested using languages with inflections or agreements (e.g., Romance languages such as Spanish, French, Italian and German). Term Swapper only swaps terms with user-defined translations. The current Term Swapper prototype obviously raises inflection or agreement issues for such languages. We would like to investigate ways to handle such languages in the future. Last but not least, in this paper, we focused only on terminology translations of nouns. We would like to explore a way to expand the usage of Term Swapper to other part-of-speech lexical items.

References

Itagaki, M., Aikawa, T., He, X. (2007). Automatic Validation of Terminology Translation Consistency with Statistical Method, *Machine Translation Summit XI*, Copenhagen, Denmark

Menezes, A. and Quirk, C. (2005). Microsoft Research Treelet Translation System: IWSLT Evaluation. In *Proceedings of IWSLT 2005*, Pittsburgh, PA, USA, October 2005.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001).

[1] We did not create such a dictionary for Chinese and Korean, which is why we limited ourselves only to the English -> Japanese pair.

[2] The low Bleu scores for the baseline systems seem to be due to the fact that the test data we used are all from the game domain and they contain lots of game-domain specific terminology.

[3] It is important to note that our interest is not to compare different systems in terms of their quality.

BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of *the 40th Annual Meeting of ACL*, Philadelphia, PA, pp. 311- 318.

Quirk, C, Menezes, A. and Cherry, C. (2005). Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of the 43rd ACL*, Ann Arbor, Michigan, pp. 271–279.

Appendix I: Templates in Term Swapper

Templates	**Patterns [X = a term]**	**Descriptions**
AS IS	X	A term itself
SUBJ+V	X exists.	A term as the subject of an intransitive verb.
SUBJ+BE	X is a word.	A term as the subject of a copula.
SUBJ+A_UNK_1	A X damns it.	A term as the subject of an unknown verb (with indefinite article)
SUBJ+A_UNK_2	An X damns it.	A term as the subject of an unknown verb (with indefinite article)
SUBJ+A_UNK_3	Thg X damns it.	A term as the subject of an unknown verb (with definite article)
OBJ_THIS_IS_1	This is X.	A term as the predicate of a copula (with a period)
OBJ_THIS_IS_2	This is X	A term as the predicate of a copula (without a period)
OBJ_V	Select X.	A term as an object of a transitive verb.
OBJ_HAVEITS	I have its X.	A term as an object with a possessive pronoun.
PREP_WITH	with X	A term following a common preposition, “with”
PREP_ONEOF	one of X	A term following a preposition, “of”
PREP_RARE	onto X	A term following a infrequent preposition, “onto”.
PARENTHESIS	(X)	A term in parenthesises.
ENGLISH	*A source term*	A term as in the source text.

Appendix II: Example of Swapped Terms

English	A natural **formation** of rock and ice.
Raw MT	岩と氷の自然な**フォーメーション**をします。
User Dict.	formation =構造物
Fixed MT	岩と氷の自然な**構造物**をします。
Human Translation	岩や氷でできた自然の構造物です。

English	Provides **Experience** Points and other **benefits**.
Raw MT	**エクスペリエンス** ポイントとその他の**利点**を提供します。
User Dict.	Experience=経験値, benefits=利益
Fixed MT	**経験値**ポイントとその他の**利益**を提供します。
Human Translation	経験値ポイントやその他の利益をもたらします。

English	**Maurice** of **Orange**, also known as **Maurice** of **Nassau**, was born in 1567.
Raw MT	Maurice の Orange、別名 Maurice の Nassau、1567 で誕生しました。
User Dict.	Maurice = マウリッツ, Orange=オラニエ、Nassau=ナッサウ
Fixed MT	**マウリッツ** の **オラニエ**、別名 **マウリッツ** の **ナッサウ**、1567 で誕生しました。
Human Translation	ナッサウ伯マウリッツ (オラニエ公マウリッツ) は、1567 年に生まれた。