Don 't use big words with me:

an evaluation of English-Thai statistical-based machine translation

Sanoooch S.NATHALANG, Peerachet PORKEAW, Thepchai SUPNITHI

Human Language Technology

National Electronics and Computer Technology Centre (NECTEC)

Abstract: The main purpose of this study is to evaluate **S**tatistical-based **M**achine translation for **E**ngish-**T**hai (**SMET)** by a human evaluator. We look in particular for potential areas of difficulty that may cause problems to SMET. The data is a 200,000-English-Thai-aligned-sentence (around 1.3 million words) corpus. We consider the English sentences as the source, and the Thai sentences as the target. Our investigation showed that simple words in English that can find their equivalences in Thai do not pose major problems in translation. However, problematic cases occur where an English word corresponds to more than one word in Thai. Explanations to these problems can be drawn from a number of approaches, ranging from language typology, morphology and syntax, to lexicography. The results of the investigation lead us to conclude that linguistic differences between the source language and the target language play a significant role in developing and improving SMET.

1. Introduction

In the world where information is a key to success, gaining access to this information has become a goal of many individuals and industries. Through translation, especially with recent advancements in machine translation, language barrier has been breached and information can

be accessed more easily now. So far, a number of machine translation systems have been developed to accommodate translation to and from many languages. There are also a number of ways to evaluate the quality of these systems too. Much research on the evaluation of MT systems has been done by automatic methods such as BLEU. These methods can provide a certain degree of objectivity in terms of how "close" the target sentence is from the source sentence.Nevertheless, they fall short of offering a clear picture of what makes the MT system 'good' or 'bad'.

Statistical-based machine translation for English-Thai (SMET for short) was developed by Human Language Technology Laboratory (HLT) at National Electronics and Computer Technology Center (NECTEC) as part of a research project called Speech to Speech (S2S) translation, which aimed at developing an electronic translator system between English and Thai to support communication at the basic level in the tourism domain. SMET is made up of two types of corpora. One is the English-Thai parallel corpus from which some data were extracted for current analysis. It consists of 200,000 English-Thai aligned sentences (around 1.3 million words) originally taken from bilingual sources that are mainly educational in nature such as dictionaries and phrase books. Examples of the sources of the sample sentences/ phrases are *English by Example: A Dictionary of English Collocations with Thai Translations* (Wattanapichet, 2004), *English Usage through Sentences* (Wanda, 2007), and Dictionary in Action English-Thai by Example Usage (Khamying, 2007). The other is the corpus of the Thai language. The translation model is trained from a bilingual (English-Thai) sentence- aligned text corpus, while the language model is trained from a monolingual text corpus. The architecture of SMET is shown in Figure 1 below.

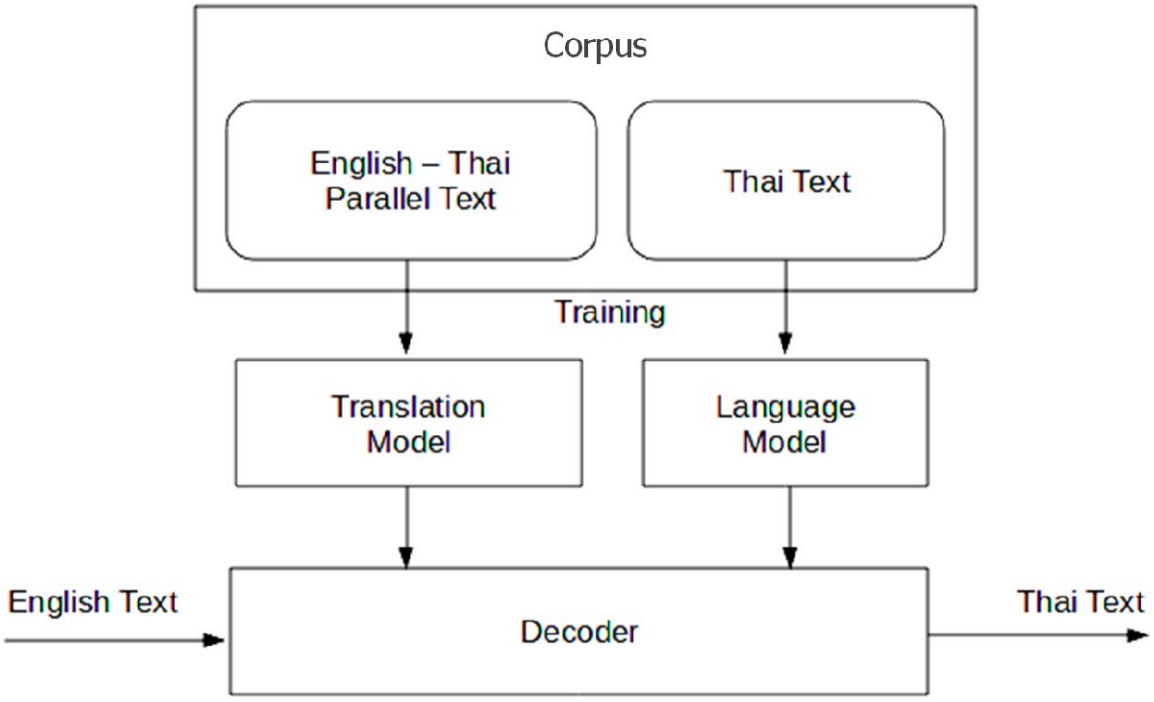

Figure 1 Architecture of SMET

As previously mentioned, SMET has adopted a statistical-based approach to machine translation. The translation is generated on the basis of the frequencies of the word translations in a target language corpus which yield probability estimates, irrespective of the linguistic context. Figure 2, for example, shows 3 word translations from the Thai corpus as equivalents of 'they' in English, i.e. พวกเขา, พวกหล่อน and เขาเหล่านั้น. The system then selected เขาเหล่านั้น based on its highest probability score. The same process is repeated with all the other words in the sentence. Therefore, SMET translates 'They go to school by car every day' as 'เขาเหล่านั้น ไปโรงเรียน ด้วย รถยนต์ ทุกวัน'.

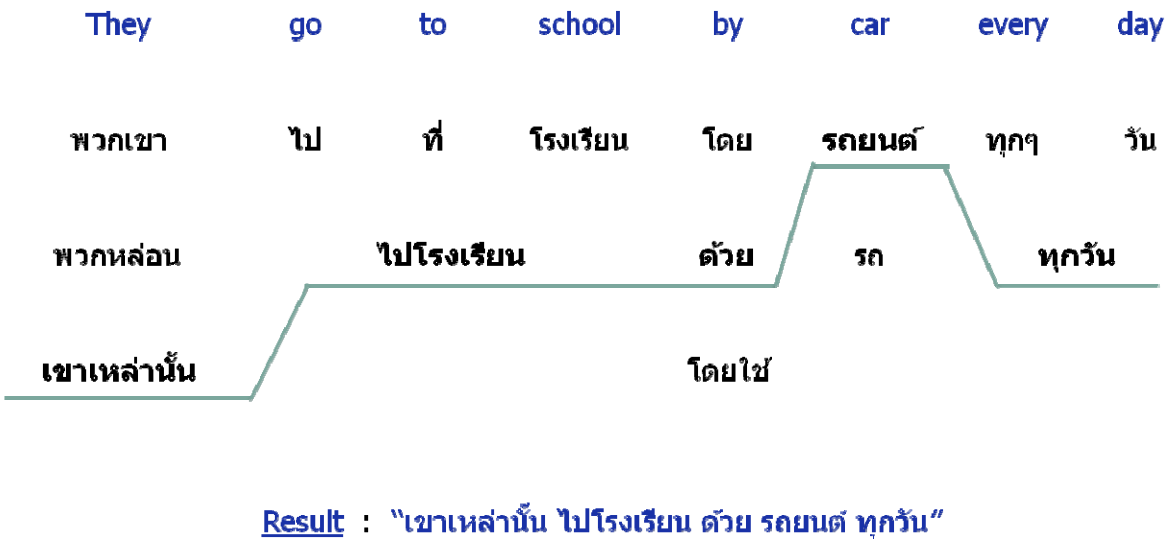

Figure 2 Probabilistic and frequency-based model of SMET

In the past, an attempt was made to evaluate the translations generated by SMET (Porkeaw, Supnithi, Wutiwiwatchai, 2007), using the BLEU metric. The result yielded was a low BLEU score of 0.13. This immediately calls for an in-depth analysis of the difficulties that challenge the system and with the hope to finally establish a standard test set for the English-Thai machine translation system.

The main purpose of this study is to investigate problems in translation by SMET which have resulted in the low BLEU score of 0.13. In so doing, we will try to identify areas of difficulty that SMET has yet to overcome. The list of problems faced by SMET is by no means comprehensive since we limit our investigation to an evaluation of SMET at the word level. In what follows, a theoretical framework in translation studies adopted in current research will be introduced. A brief overview of Thai will later be provided in order to help the reader establish some typological differences between English and Thai. Finally, some translation problems will be presented.

2. Theoretical framework

Much discussion in translation studies to date is centered around the notion of equivalence. One oft-cited definition of translation was offered by Catford (1965:20) who views the translation process as involving "[T]he replacement of textual material in one language by equivalent material in another language". McArthur (1992) states that translation is called for when two people of different languages or dialects communicate and transmission (of the message) is likely to be broken. The translation process, according to McArthur (ibid: 1052) then involves "convert[ing] the message in Language A (*the source language*) into a more or less equivalent message in Language B (*the target language*). Catford's definition is problematic, as many have acknowledged, simply because it implies that one could find the matching 'material' of one language in another. McArthur's use of the term 'message' is deemed more appropriate for the purpose of the present investigation since, as one shall see later, the source language, i.e. English, cannot always find its equivalent 'material' in the target language, i.e. Thai due to differences in linguistic characteristics between the two languages. Here, we define the term 'message' loosely as referring not only to equivalence[1] in vocabulary and syntax, which can be observed directly, but also to equivalence in the meaning which requires a contextual and/or cultural interpretation. As such, we follow Larson (1998:3) in assuming that translation

> … consists of studying the lexicon, grammatical structure, communication situation, and cultural context of the source language text, analyzing it in order to determine its meaning, and then reconstructing this same meaning using the lexicon and grammatical structure which are appropriate in the receptor language and its cultural context.

The above discussion on translation and equivalence is mostly relevant to the task of translation where a human translator is involved. Human translators can vary in their translation outputs, depending on their linguistic knowledge, both in the source and the target language, as well as their extra-linguistic knowledge to help interpret the message implied in the source

language. They are capable of not only dealing with overtly observable aspects of language (e.g. morphological and grammatical properties of words in the sentence), but also with the hidden levels of meaning and thought (e.g. the speaker's intention) that are not readily translatable without context.

All things being equal, human translators possess the two categories of knowledge which enable them to understand the language. The same however is not true of machine translation. Some kind of programming is required to train the computer to understand both knowledge categories. It has been widely accepted nowadays that linguistic properties such as word morphology, syntax, semantics or even pragmatics are easier to codify than extra-linguistic knowledge (Ping, 2008). Within the linguistic knowledge category itself, morphology and syntax seem easier to codify than semantics and pragmatics as evidenced by a greater number of research studies on the former topic than the latter.

In our study, we will focus on how well our SMET does in generating translation from English to Thai. Given the challenges for machine translation as Ping (ibid) points out, we will limit our evaluation to some morpho-syntactic features of a word. By morpho-syntactic features, we appeal to Baker (1992)'s 4 levels of equivalence, namely word, grammar, text and pragmatics, the first two of which are directly relevant to the present investigation. Textual and pragmatic levels are dealt with at a discourse level, both of which are somewhat irrelevant to the present analysis as we are dealing with individual clauses in disconnected discourse. Since word and grammar are sometimes not mutually exclusive, features germane to morphology and syntax will be referred to globally as morpho-syntactic features.

According to Baker, potential problems in translation in relation to these morpho-syntactic features boil down to 2 main issues: the one-to-one relationship between words and meaning, and the lexical meaning itself. Regarding the first issue, it now is a widely held

premise that words are not the smallest unit of meaning in every language. In some languages, meanings are carried by structures and linguistic devices much more complex than a single word. In English, for instance, morphemes are the smallest units of meaning which cannot be further analysed as shown in [1]. When these morphemes are to be translated into languages with no overt morpho-syntactic realisation, such as Thai [2], they may cause some difficulty for machine translation. For ease of reference, [2] is a rough equivalent of [1].

[1] English

a) inconceivable → *in-* + conceive + *-able*

b) girls → girl + *-s*

c) hated → hate + *-ed*

[2] Thai

a) **นึกไม่ถึง** → **นึก** **ไม่** **ถึง**[2]
think *not* *arrive*

b) เด็กผู้หญิงหลายคน → เด็ก ผู้ หญิง หลาย คน
child *CFY-WD*[3] *woman* *many* *human*

c) เกลียด → เกลียด
hate (no indication of tense or aspect)

A few observations must be made clear regarding [1] and [2]. English is an inflectional language. The formation of a word [1a] and the realisation of grammatical features on words such as number [1b] or tense [1c] are done by means of affixation, i.e. adding prefixes and suffixes to a word stem. In Thai, however, as an isolating language, words are not typically formed by means of affixation but by compounding words together [2a-b] and there is no overt marking of tense or aspect on the verb [2c]. According to Baker (1992), the diversity of grammatical features across languages must not be overlooked since "[D]ifferences in the

grammatical structure of the source and the target languages often result in some change in the information content in the message during the process of translation" (p. 86). Some major grammatical categories that could pose potential problems include number, tense and aspects, voice, person and gender.

Another problem concerning the mismatch between words and meanings is that one-to-one correspondence cannot always be established between orthographic words and meaning both within and across languages. This can be further elaborated by two linguistic possibilities. That is, one word in English may be represented by several words in Thai and one word in Thai may be represented by several English words.

[3] English → Thai

English	Thai	
findings	a)	ข้อมูล ที่ ค้น พบ *information that search find* (**source**: English-Thai parallel corpus)
	b)	ผล วิจัย *result research* (**source**: Google translate)
	c)	ผล ของ การ สืบ ค้น *result of/thing NOM[4] investigate search* (**source**: LEXiTRON)

It can be seen from [3] that one word in English corresponds to at least three noun phrases in Thai. Two issues are worth considering here. First, we assume that the equivalent NPs given are determined by both linguistic and extra-linguistic knowledge of each translator. As such, there is no 'right' or 'wrong' answer, just 'acceptable'. With respect to machine translation, usually one alternative is given. The question is which among the 3 alternatives in [3a-c] would

be the most acceptable. Secondly, English and Thai are typologically different, specifically in relation to [3], in terms of head directionality. Failure to take into account linguistic typological differences is likely to result in both inaccurate and unintelligible translations.

As for the second issue on the lexical meaning, Baker listed 4 main areas that could give rise to problems in translation, but we will only consider two since they are more relevant to the present study, namely, the propositional meaning vs. the expressive meaning, and the presupposed meaning. The propositional-expressive meaning distinction may be simply interpreted as the fact-feeling distinction. By that, it means a word normally carries propositional meanings, or facts, that distinguish them from the others. The propositional meaning of 'head' as 'the part of the body above the neck that contains the eyes, nose, mouth and ears and the brain' makes it distinct from 'bottom' which is 'the part of your body that you sit on'. The expressive meaning of 'head' in 'use your head' is a certain degree of irritation, as opposed to 'think' which is rather neutral. With respect to the presupposed meaning, two kinds of restrictions are involved: selectional restrictions and collocational restrictions. The word 'pretty' prototypically selects 'girl' or 'woman' while 'handsome' selects 'boy' or 'man'. A violation of such selectional restrictions may imply specific meanings as often inferred from 'pretty boy'. As for collocational restrictions, Baker showed a range of verbs used with 'teeth' in different languages, e.g. 'brush teeth (English), 'polish teeth' (German), 'wash teeth' (Polish) and 'clean teeth' in Russian. In translating words that can be subject to such restrictions as described, it is important to know when to treat them as individual words and when to treat them as unanalysed chunks.

In conclusion, we have looked at some related theories that can help us in the evaluation of translation outputs generated by SMET which is a statistical-based machine translation for English and Thai. An operative definition of translation has been offered and some areas where translation could have problems have been pointed out. The discussion has led us to believe that

some morpho-syntactical differences between English and Thai words are likely to cause problems in translation for SMET. As words in English are ‘big’, i.e. containing more information that they appear, due to the presence of derivational affixes, inflectional markers, grammatical markers, and various lexical meanings as stated, translating these English words and other associated morpho-syntactic properties into a target language like Thai which is ‘simpler’ given its low morpheme-per-word ratio is by no means an easy task.

3. The analysis

To evaluate the quality of the machine translation system in question means we have to find out how much ‘equivalence’ between the source and the target languages can be established. To find equivalence between the source and target language cannot be limited to form mapping given the mismatch between orthographic form and meaning as previously shown. To find equivalence therefore means to retain the same ‘message’ from the source language when translating it into the target language. The higher the equivalence, the better the translation results.

Two criteria used to establish such equivalence are intelligibility and accuracy. Intelligibility refers to the degree to which the translated sentence is intelligible without reference to the source sentence. We divide it up into 3 scales, 1 indicating unintelligibility, 2 more or less intelligible and 3 highly intelligible. We do not use more than 3 scales, as many research studies have done, because it is easier to evaluate since consistency in the judgement can be maintained. It must be noted also that intelligibility depends to a great extent on native speakers’ intuition. When we ‘feel’ that the translated sentence ‘sound’ intelligible, we mark the sentence as intelligible or 3. On the contrary, if the sentence sounds unintelligible, we assign the sentence the value of 1.

By ‘accuracy’, we compare the result of the translation with the source sentence. If the ‘message’ from the source sentence can be retained in the output, then the translation is deemed accurate, and vice versa. Like intelligibility, we grade the accuracy of the translation outputs on the basis of a 3-scale criterion: 1 not accurate, 2 moderately accurate and 3 highly accurate. One difficulty in grading accuracy is when the translation is considered unintelligible due to the unnaturalness of the language. In a case like that, word-for-word translation is sometimes more or less accurate but the way words are put in a sentence is not grammatically correct. Therefore, to make the evaluation consistent, when a sentence is marked 1 or unintelligible, the accuracy is also 1, inaccurate.

In the evaluation of SMET, two native speakers of Thai worked together to agree on the intelligibility and accuracy of the translation outputs. One is an engineer with no background in linguistics. The other is a linguist by profession and has been teaching English for at least 10 years. Both evaluators spent 6 years in England doing a degree in Electrical Engineering and Linguistics respectively. As a result, they are sufficiently qualified to judge the translation outputs for their accuracy.

We randomly judged 150 sentences in the corpus for intelligibility and accuracy. One could argue that this is too small a number to investigate given a corpus of 200,000 words. However, our purpose is not to evaluate the translation generated by SMET quantitatively but qualitatively. Let us recall that the corpus has been evaluated previously by BLEU, a commonly used automatic evaluation method of machine translation, which yielded a low score of 0.13. It is not therefore as significant to see the degree of bad translation generated by SMET as to see what causes problems for SMET. The latter is what we are more interested to see, and it is the purpose of the present study.

4. Results of the study

4.1 Evaluation of SMET on intelligibility and accuracy

As discussed in Section 3 above, both intelligibility and accuracy are measured from a 3-scale ranking, 1 being unintelligible and inaccurate, while 3 being intelligible and accurate. The results of the evaluation by both criteria are displayed in Table 1 below.

Table 1 Summary of the evaluation of SMET outputs by intelligibility and accuracy

Intelligibility	**Number (150)**	**%**	**Accuracy**	**Number (150)**	**%**
1 (unintelligible)	85	56.6	1 (inaccurate)	102	68
2 (moderately intelligible)	20	13.3	2 (moderately accurate)	16	10.6
3 (intelligible)	45	30	3 (accurate)	32	21.3

Table 1 shows the percentage of the evaluation of the SMET outputs considered separately in relation to the intelligibility and accuracy criteria. The evaluation revealed relative high percentages of translation that is unintelligible (56.6%) and inaccurate (68%), while those that are intelligible and accurate are 30% and 21.3 % respectively. It is too soon, however, to conclude that SMET generates outputs that are unintelligible or inaccurate. A closer look at the results suggests that 'intelligibility' should be considered in conjunction with 'accuracy' in order to obtain a more meaningful interpretation of the results. This is because there is always a possibility that a highly intelligible sentence is in fact inaccurate. The results of such analysis are displayed in Figure 3.

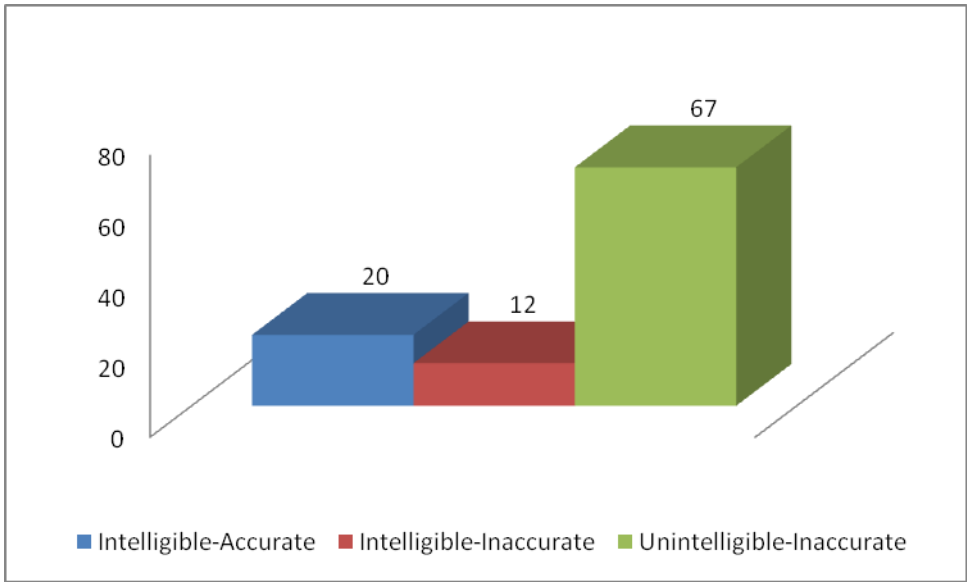

Figure 3 Evaluation of SMET outputs by intelligibility and accuracy (percentage)

From Figure 3, it appears that as much as 67% of the translation generated by SMET is both unintelligible and inaccurate while 12% of the outputs is intelligible but inaccurate. Only 20% of the outputs is both intelligible and accurate. Based on these results, an in-depth analysis of what constitutes problems in translation for SMET is in order. Figure 4 shows the results of the analysis.

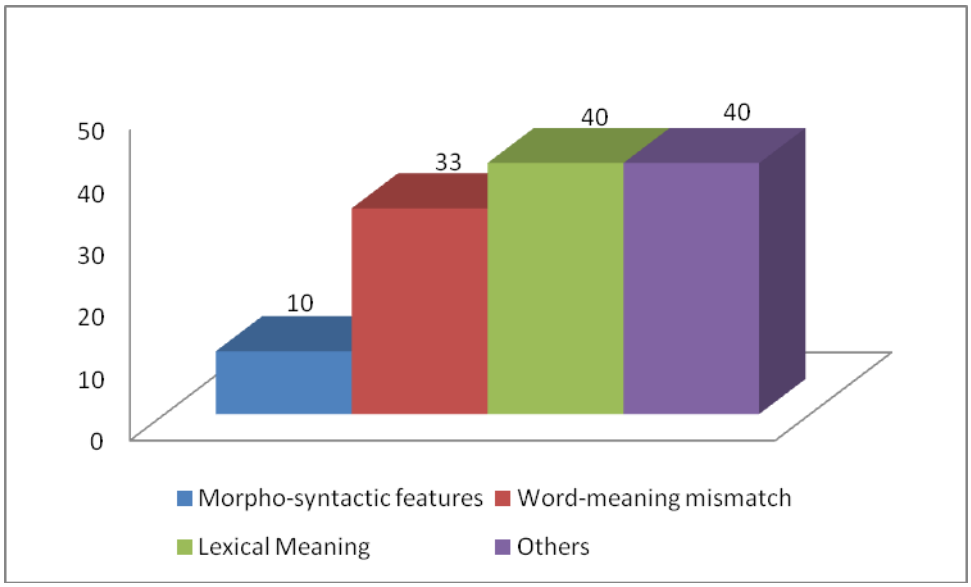

Figure 4 Percentage of problems identified in SMET translation outputs by categories

From Figure 4, lexical meanings pose the greatest potential problem in translation (40%), followed by the mismatch between words and meaning (33%) and morpho-syntactic features of a word (10%). Other problems include mistranslation, differences in sentence structures and word order between English and Thai, and the presence of words that cannot be traced back to any words in the source language. The last problem falls outside the scope of this present study.

4.2 Areas of potential problems in translation for SMET

In this section, we will give some examples of words that pose some difficulty in translation for SMET. It is not our intention to list all the problems we have encountered but to illustrate each category more clearly.

a) Morpho-syntactic features of words

Morpho-syntactic features posed the least problem for SMET (10%). This is not surprising if one considers that some features affect meaning more than others. Without question, meaning is carried more by affixes than grammatical features such as tense markers or number morphemes. Based on the corpus analysis, affixes present a greater challenge than grammatical properties on words such as tense and number.

- **Affixation**

Prefixes and suffixes in English are sometimes not translated into Thai in the parallel corpus. The analysis shows that, on a number of occasions, root words get translated while their derivational words are not. Examples are *worse-worsen, monk-monkhood, wild-wildly-wilderness, tied-untied, touched-untouched, tasted-tasteless*.

- **Syntactical properties of words**

The meaning implied in tense and number, for example, cannot always be preserved in translation. For example, '*she works well with other*' is translated as 'เธอทำงานได้ดีกับผู้อื่น', where 'ทำงาน' or '*work*' can be interpreted as past, or present. In some cases, the context can help clarify whether the 'past' or 'present' interpretation is needed. Still, evidence from the corpus indicates that there are cases where tense is ambiguous and yet that does not interfere with the comprehensibility of the translation output on the whole. Given that the corpus under study contains individual sentences in disconnected discourse, the interpretation of tense as referring to the past event or the present one becomes less rigid.

b) Word-meaning mismatch

The mismatch between words and meanings is the biggest problem for SMET, based on the results presented above. This is partly because translation is directly relevant to the conveyance of meaning from one language to another while preservation of the original meaning should be attempted. As language is the result of human communication, people's perception of the world, including for example their experiences, beliefs, religion, ideology and cultural orientation, is reflected in language. By this virtue, ideas that are familiar to one community may be foreign to another. It is therefore not surprising to see that some words in English do not exist in Thai and vice versa. An entity or concept represented by one word in English then may be represented by more than one word in Thai. As the corpus has revealed, there are quite a number of words in English that cannot be directly translated into Thai as the one word in English corresponds to more than one word

translation in Thai. A case in point is the word '*access'*. There are 12 occurrences in the corpus, 10 of which results in unintelligible and hence inaccurate translation. English vocabulary from the corpus belonging to this category is somewhat huge, hence a high percentage of unintelligible and inaccurate translation outputs.

c) Lexical meaning

There are quite a number of examples where words are translated inaccurately in the corpus such as '*The boy must have worms'*. Polysemous words like '*worm*' pose potential problems since in translation, we have to know which meaning is appropriate in that context. In this particular case, '*worm'* is translated as 'หนอน' which typically co-occurs with trees or soils.

With reference to collocational restrictions, most errors in SMET are produced because the translation is done on a word-for-word basis, instead of taking into account word collocation or the context in which a particular collocation occurs. For example, the English verb '*deliver'* can be used with a range of nouns such as *goods, a report, a speech, a baby,* and *a* blow. In Thai, however, different verbs are used with each of the nouns mentioned. Based on the analysis of the corpus, the word 'ส่ง' is constantly used in the translation where '*deliver'* or '*delivery'* is found. As expected, the only context where 'ส่ง' is accurately translated is when it is in collocation with the word signifying some kind or goods or services.

5. Conclusions, suggestions and future work

The main purpose of this study is to evaluate SMET which is statistical-based machine translation for English and Thai. The results of the investigation seem to suggest that typological differences between English and Thai with reference to the morpho-syntactic

features may be accountable for some difficulties SMET has faced. In addition, the results of the study confirms that there are still quite a number of issues SMET has to deal with, which is consistent with the evaluation using BLEU, which concluded that the quality of the translation is low. The evaluation has shed some light on where SMET needs to improve. First of all, problems in translation by SMET are largely due to the fact that in the corpus English and Thai words are aligned with one another obviously without any consideration of linguistic restrictions specifically imposed on the two languages. To improve the quality of translation, SMET must be able to distinguish between literal meanings of each individual word and specific meanings of words in collocation. Moreover, syntactic rules of Thai must be built into the system of SMET to accommodate for structural differences between English and Thai. In other words, a combination of statistical-based and rule-based machine translation may enable SMET to translate more effectively. As this study does not extend beyond the word level, further studies could be carried out to evaluate the problems found at the higher levels such as grammar, text and discourse.

References

Baker, M. (1992). *In other words: Coursebook in translation***.** London: Routledge

Catford, J. C. (1965). *A Linguistic Theory of Translation*. Oxford**:** Oxford University Press.

Cuellar, S.B. (2002) "Equivalence Revisited: A Key Concept in Modern Translation Theory". *Forma y function* 15: 60-88. Retrieved 1 October 2010 from http://redalyc.uaemex.mx/redalyc/pdf/219/21901504.pdf

Iwasaki, S. and Ingkaphirom, P. (2005). *A reference grammar of Thai.* Cambridge: Cambridge

University Press.

Khamying, S. (2007). *Dictionary in Action: English-Thai by Example Usage*. Bangkok: VJ Printing.

Kenny, D. (2008) "Equivalence" In M. Baker and G. Saldanha (eds.) *Routledge encyclopedia of translation studies* (2nd ed.). London: Routledge, 96-99

Larson, Mildred L. (1998). *Meaning-based translation: A guide to cross-language equivalence.* Lanham, MD: University Press of America and Summer Institute of Linguistics.

Ping, K. (2008). machine translation In M. Baker and G. Saldanha (eds.) *Routledge encyclopedia of translation studies* (2nd ed.). London: Routledge, 162-168

Porkeaw, P., Supnithi, T. and Wutiwiwatchai, C. (2008). Statistical machine translation for Thai-English Electronic Translator. *NECTEC Technical Journal, NECTEC-ACE2008 Special Edition.*

McArthur, T. (ed.) (1992). *The Oxford Companion to the English Language (Oxford Companion to English Literature).* Oxford: Oxford University Press.

Smyth, D. (2002). *Thai: an essential grammar*. London: Routledge.

Wanda. (2007). *Quick learn: English usage through sentences*. Bangkok: Chulalongkorn University Press.

Wattanapichet, W. (2004). *English by Example*. Bangkok: Thaiways Publications.

Notes

[1] For more controversies and issues on the term ‘equivalence’ (see e.g. Cuellar (2002) Kenny (2008) among others)

[2] Thai writing does not show word boundaries. In other words, words in Thai are written continuously without separation. Spaces may serve as punctuation markers, where commas and/or full stops are usually required. (Smyth, 2002)

[3] CFY-WD stands for a classifying word. Iwasaki and Ingkaphirom (2005) call it a classifying prefix.

[4] NOM stands for a nominaliser, or what Iwasaki and Ingkaphirom (2005) call word noun-forming prefix