Are numbers good enough for you?
A linguistically meaningful MT evaluation method

Takako Aikawa
Microsoft Research,
Machine Translation team
One Microsoft Way
Redmond, WA 98052, USA
takakoa@microsoft.com

Spencer Rarrick
University of Washington,
Department of Linguistics
PO Box 354340
Seattle, WA, 98195, USA
rarricks@u.washington.edu

Abstract

This poster explores a way to qualitatively identify linguistic properties of a particular MT system, so that we can identify its strengths and weaknesses more readily. The paper provides preliminary results for two English-to-Japanese SMT systems. We demonstrate that comparison of n-gram frequencies between human translations and machine-translated outputs can lead us to linguistically meaningful information about a given MT system. We argue that our method has the potential to guide the research and development process in a way that numeric scores alone cannot and that it can shed new lights on how we assess MT quality.

1 Introduction

Over the last 10 years, a number of different kinds of metrics for quantifying the quality of machine translation (MT) systems have been proposed in the literature (BLEU [Papineni et al., 2002], NIST [Doddington, 2002], METEOR [Banerjee and Lavie, 2005], Word Error Rate, etc.). The dominant approach of such metrics involves computing the distance between reference(s) and MT output(s). BLEU, for instance, is one of the first metrics to be adopted by the MT community as a 'standard' metric, and we (Microsoft Research, Machine Translation team) have been using it to measure the improvement of our MT system. We also use BLEU to compare our own statistical machine translation (SMT) systems to our competitors, so that we can gauge the overall quality difference(s) with respect to other MT systems.

The primary advantage of statistical/automatic measurements is that they are free and fast. Furhtermore, no human intervention is necessary and hence, such metrics are scalable. The validity of such automatic metrics has also been verified in the literature. For instance, Coughlin (2003) argues that BLEU indeed correlates with human evaluations. The problem, however, is that the scores from such automatic metrics do not reveal any specific characteristics of a particular MT system. For example, let us suppose that we used BLEU to compare two MT systems (SMT1 and SMT2), yielding respective scores of 0.256 and 0.261. Based on these score, one is able to make relatively broad statements such as "SMT2 is slightly better than SMT1 but the difference is not statistically significant." While useful for general comparisons between MT systems, such evaluation techniques are less meaningful in addressing questions such as "how are two MT systems different?" or "what are the strengths and weaknesses of a particular MT system?"

The method we propose in this paper explores a way to retrieve 'linguistic characteristics' of a particular MT system, so that we can identify its strengths and weaknesses more readily. This paper provides preliminary results for two English-to-Japanese SMT systems. We argue that our method has the potential to provide the MT community with a new angle for examining MT quality and a new tool for guiding research efforts.

2 Experiments

2.1 Overview

In this experiment, we evaluated two English-to-Japanese web-based translation services (SMT1 and SMT2).[1] The intuition behind our method is as follows: human translations (HT) consist of well-formed n-grams. Comparing n-grams from MT and those from HT then should let us identify differences in translation patterns between the two. We assume that many discrepancies in n-grams counts between HT and a given MT system's output are indicative of linguistic inaccuracies or biases of that MT system. Using our method in conjunction with other automatic metrics (e.g., BLEU), we can not only quantify the quality of MT systems in numbers but can also analyze differences in linguistic behavior among MT systems.

2.2 Data and Methodology

We sampled 100K English-Japanese parallel sentence pairs from a large corpus of data[2] and translated the English side of each of the sampled sentences into Japanese using the two SMT systems mentioned above. Then, for English and Japanese we counted the frequency of various order n-grams across the three versions of the data set. Taking trigrams as an example, we calculate the frequency of each trigram in each of the three Japanese corpora: the human translated corpus and the two corpora generated by translating the English corpus into Japanese using SMT1 and SMT2. We ignore per-sentence counts and look only at aggregate counts over an entire corpus.

Because all three Japanese corpora are translated from the same set of English source sentences, we expect counts for n-grams to be roughly the same in each corpus, except for cases where a word or phrase is somewhat consistently translated differently by two translators (machine or human). As the human translations are in most cases perfect or near-perfect, we can therefore attribute discrepancies in these n-gram counts to peculiarities of the two SMT systems.

N-grams that appear much more frequently in the human translations than in the output of one of the SMT systems can indicate areas where the SMT system has difficulty producing the correct grammatical wording. N-grams that appear frequently in the MT output and infrequently in human translations, on the other hand, may be ungrammatical, unnatural, or otherwise awkward, as these are unlikely to be written by a human. In some cases an SMT system consistently chooses one particular wording when translating a specific phrase, while a human translator would vary wording depending on context. This would result in high counts for relevant n-grams in the SMT output for the system's preferred wording, and relatively high counts in the human translations for the alternate wordings.

While we count fully lexicalized unigrams and bigrams, we encounter sparsity issues with higher order n-gram counts. To solve this, we transform each sentence before counting 3-grams and 4-grams. We leave the hundred most frequent words for that language lexicalized, and replace any less frequent words with an <UNK> (for "unknown"). We also replace punctuation tokens with a generic <PUNC> token.[3] We hypothesized that interesting patterns for these higher order n-grams would generally relate to function words, and our list of frequent words is intended as a proxy for function words. However, because the corpus used to generate the Japanese list had a heavy technical domain bias, it includes a few words that are common in the tech domain, but relatively less common elsewhere (e.g. '削除'[*sakujo*, 'to delete'].) Because the majority of function words come from parts of speech which form closed classes, the remaining words that fall under <UNK> tend to come from open class categories, such as nouns, verbs, adjectives and adverbs.

[1] SMT1 is Microsoft Translator, and SMT2 is another third-party web service. For technical details on Microsoft Translator, see Menezes and Quirk (2005) and Quirk, et al. (2005).

[2] We hoped to create a balanced corpus, so that the results would not be skewed due to the data issues. However, much of our English-Japanese parallel data is technical, and so admittedly, there is a bias toward technical domain in our sample.

[3] Different MT systems tend to translate particular punctuation characters into Japanese differently from humans (e.g., half-width vs. full-width periods). Replacing all punctuation with a generic token allows us to consider how word token interact with punctuation, without differences in n-gram counts being dominated by inconsistent use of such character.

SMT1	1. できる よう <UNK>	‘in order to <UNK>’
	2. いる こと も	‘be nominalizer also’
	3. れる の は	‘passive nominalizer Top’
	4. こと も ある	‘nominalizer also exist’
	5. する の を	‘do nominalizer Acc’
	6. ある の は	‘exist nominalizer Top’
	7. <UNK> が つい	‘Nom unintentionally’
SMT2	1. いる こと も	‘be nominalizer also’
	2. こと も できる	‘nominalizer also can’
	3. ない 場合 も	‘Neg case also’
	4. ない <UNK> も	‘Neg <UNK> also’
	5. サービス に より	‘service according-to’
	6. の に <PUNC>	‘in order to <PUNC>’

Table 1: Trigrams that appeared frequently in HT but never in the respective MT

SMT1	1. <s> <s> ない	‘<s> <s> Neg’
	2. て ください する	‘V-Gerund please do’
	3. を でき ます	‘Acc can-polite’
	4.で を 使用	‘with Acc use’
	5. が ある ない	‘Nom exist Neg’
	6. または が	‘or Nom’
	7. に は が	‘in Top Nom’
	8. する いる と	‘do be if/when’
SMT2	1. <s> <s> ない	‘<s> <s> Neg’
	2. ください <UNK> の	‘please <UNK> Gen’
	3.です <UNK> の	‘be <UNK> Gen’
	4.ます <UNK> する	‘V-Polite <UNK> do’
	5.ます を <UNK>	‘V-Polite Acc <UNK>’
	6.て ください に	‘V-Gerund please to’
	7. ます こと を	‘V-Polite nominalizer-Acc’

Table 2: Trigrams that appeared frequently in MT but never in HT.

3 Results and Analyses

We made a number of interesting observations based on the data we extracted using the methods described in Section 2.

3.1 Trigrams

Table 1 and Table 2 above provide some trigram sample data that have huge discrepancies in terms of frequency between HT and MT. In these tables, a sequence of one or more ‘<s>’ tokens indicates the beginning of sentence boundary.

The items in Table 1[4] were seen frequently in HT but never in the output of the respective SMT systems. The items in Table 2, on the other hand, were seen frequently in the output of one of the SMT systems but never in the HT corpus. Thus, one could say that the n-grams in Table 1 are characteristic of human translated text, whereas those in Table2 are indicative of the weaknesses of the two SMT systems.

Let us briefly examine the items in these two tables here. In Table1, SMT1’s output contains the items that involve the Japanese nominalizer ‘の/こと’ (*no/koto* ‘-ing/the fact that’), whereas SMT2 contains those that involve the Japanese postposition ‘も’ (*mo* ‘also’). This implies that SMT1 and SMT2 do not use these constructions/expressions in the contexts where humans are most likely to use them. Notice that this does not necessarily mean that SMT1 and SMT2 are simply omitting these nominaizers and the postposition respectively; it may be that the two systems are using other constructions to realize these structures and meanings.

On the other hand, the items seen in Table2 are expected to be ill-formed, as they are n-grams that have not been observed anywhere in the human-translated sentences. Therefore, they are indicative of the linguistic mistakes that these SMT systems tend to make.

Looking at the items for SMT1, we can make a couple of inferences. First, it seems that SMT1 fails to use the correct morphology for negation when translating English phrases such as ‘*do/does not have*’ or ‘*do/does not exist*’. In general, the morpheme ‘ない’ (*nai*, ‘not’), which is underlined in Table 2, combines with an inflected verb stem to indicate that the verb is negated. The combination of the infinitive form of a verb such as ‘ある’ (*aru*, ‘to exist/to have’) and this negation morpheme ‘ない’ (*nai*) (e.g. SMT1-#5) is thus totally ill-formed. Furthermore, ‘ある’ is irregular in that its negated form is simply the bare negation morpheme ‘ない’. It would appear from our data that SMT produces a form that is incorrect on both accounts.

To validate our hypothesis, we looked at the original source English sentences and the corresponding outputs from SMT1. Examples below support our hypothesis here.

(1) If you don't have a saved game, Mahjong Titans starts a new game.
保存したゲームが あるない 場合 、 マージャン
saved game-Nom exist-Neg if/when Mahjong
タイタンは、新しいゲームを 開始します 。
Titans-Top new game-Acc start

[4] Abbreviations include Top: ‘Topic’; Acc: ‘Accusative’; Nom: ‘Nominative’; Gen: ‘Genitive’; Neg: ‘Negation’

(2) If you do not have a backup, perform the steps in resolution 1.
バックアップがあるない場合は、　手順の
backup-Nom exist-Neg if/when-Top steps-Gen
解像度　で　1が　適用されます。
resolution in 1-Nom perform-passive

Second, the usage of the postpositions, case markers, or conjuncts seems to be mishandled in some contexts; multiple occurrences of postpositions such'で を' (*de wo* *'*with Acc') (SMT1-#4) or 'に は が' (*ni wa ga 'in* Top Nom') (SMT1-#7) or the co-occurrence of 'または が' (*matawa ga* 'or Nom') (SMT1-#6) are ill-formed, and hence, the weakness of SMT1. We found examples like (3)-(4) to support this hypothesis.

(3) Include all forms of information.
すべての形態　情報　に は が　含まれます。
all-Gen form information in-Top-Nom include-passive

(4) Packets may be reordered or duplicated before they arrive.
パケットの 順序が　変更　または が　到着する
packets-Gen order-Nom change or Nom arrive
前に　複製されます。
before duplicate-passive

Looking at the items for SMT2 in Table2, we can infer one prominent characteristic: it appears that SMT2 is not handling relative clause (including reduced relative clause) or adjectival modifier constructions properly. Japanese does not allow the occurrence of 'ます' (*masu*, politeness suffix for a verb) or that of 'です' (*desu*, polite form of the copula) in prenominal relative clauses. If we assume that <UNK> is most likely to be a noun in the cases shown in Table 2, cases like SMT2-#3 or SMT2-#4 constitute violations of this rule.

At this point, we would like to note a couple of patterns observed in both of these SMT systems. One pattern involves the occurrence of the Japanese negation morpheme 'ない' (*nai*) (the first item in both SMT's) at the beginning of a sentence.[5] Intuitively, this is most likely due to the discrepancy in nominal negation constructions between English and Japanese: in English, one can directly negate a nominal by inserting the word *'no'* immediately before, whereas in Japanese, negation must happen morphologically on the predicate. We confirmed our intuition by examining a number of the MT sentences containing this n-gram. This is illustrated in the following example from our data, where (5) is the English source sentence, (6) is the correct human translation and (7) is the output of SMT2:

(5) No folder will be created.

(6) フォルダー は 作成されません。
folder-Top create-passive-Neg

(7) ない フォルダが　作成されます。
Neg folder-Nom create-passive

The English negation occurs at the beginning of the sentence whereas the Japanese one should occur at the end of the sentence. Our results indicate that such a discrepancy between two languages might still be very challenging for SMT systems.

Another interesting case observed in both SMT systems involves 'ください' (*kudasai* 'please'). This expression is used together with a verb gerund form and it occurs at the end of a sentence. So patterns like SMT1-#2, SMT2-#2, or SMT2-#6 in the above tables are ill-formed. We did not have any intuition on what type(s) of English constructions would trigger such ill-formed translations. While looking at the data, however, we observed that when the input English sentence contains the verb "see" or "refer" or some sort of modal expression (e.g., "can", "should", etc.), the usage of 'ください' often seems to be mishandled. The following examples illustrate this point.

(8) You can at least see how to use the two methods.
少なくとも 2 つの　方法　を使用する方法を
at least two-Gen methods-Acc use way-Acc
参照してくださいする ことができます。(SMT1)
see please do is possible

(9) You should see that your device is now running.
デバイスが　現在　実行されていることを
device-Nom now running the fact that-Acc
参照してくださいする必要があります。(SMT1)
see please do is necessary

(10) But visibility is poor and there's nothing much to see.
しかし、可視性が　乏しく、何も参照して
but visibility-Nom poor anything see
くださいに 多くありません。(SMT2)
please to much there is-Neg

[5] We also have seen this sequence in other n-gram tables (e.g., <s> ない for their bigram tables).

3.2 Unigrams

The results for unigrams reveal some other interesting characteristics of the two SMT systems. First, a large portion of the unigrams that are found frequently in SMT2 but rarely in HT involve so-called Katakana long vowel words. Table 3 below lists some such examples.

Katakana words found in SMT2	Alternatives
コンバータ ‘convertor’	コンバーター
アダプタ ‘adaptor’	アダプター
エミュレータ’emulator’	エミュレーター
プリンタ ‘printer’	プリンター
スキャナ ‘scanner’	スキャナー

Table 3: Katakana short vowels found in SMT2

Katakana is one of the Japanese writing (alphabet) systems and is typically used to transcribe foreign loanwords. In some cases, there is variation in the choice of katakana characters used to represent a particular foreign sound from a loan word. For instance, in the above examples, the character ‘ー’ indicates a long vowel (e.g., タ is /ta/ whereas ター is /taa/). All of the words in the first column use short vowels, and though they are understandable, the fact that they are hardly found in HT indicates that SMT2 might be overgeneralizing the use of short vowels. SMT1, on the other hand, appears to use short vowels much less frequently.

The unigram results for SMT1, on the other hand, show another type of characteristics associated with SMT1; namely, SMT1 seems to overgenerate personal pronouns. Table 4 provides such pronouns.

私たち/私達 ‘we’	私 ‘I’
あなた ‘you’	彼女 ‘she’
彼ら ‘they’	彼 ‘he’

Table 4: Pronouns found in SMT1

Japanese hardly uses overt pronouns unless there is some specific reason to use them. Again, the fact that such personal pronouns are hardly found in HT indicates that SMT1 seems to be generating too many unnecessary pronouns.

4 Concluding Remarks

This poster paper explores a different way to assess the quality of an MT system and identify its weaknesses. We have demonstrated that comparison of n-gram frequencies between HT and MT output can provide us with linguistically meaningful information about a given MT system. While our evaluation technique is not completely automatic, we believe that this kind of qualitative output has the potential to guide the research and development process in a way that numeric scores alone cannot.

In addition to their usefulness in qualitative MT evaluation, the MT-only and HT-only n-gram lists used in our method have a number of other potential applications. First, they could be used in an MT-hypothesis re-ranker to penalize or forbid hypotheses that contain n-grams known to appear primarily in MT output. By doing so, one may be able to reduce or eliminate certain commonly seen MT errors. Second, these n-gram lists may be useful in automatic MT evaluation. As many of the items in these lists contain disfluencies, high frequency of these items should be indicative of serious grammatical errors. A numeric score could be generated by simply counting the number of known MT-only n-grams appearing in a corpus translated by a certain MT system. This could also serve useful in identifying problematic sentences for post-editing when the MT output is intended for dissemination.

Last but not least, these n-gram lists can be used for the purpose of MT detection. In recent years, the prevalence of machine-translated content on the web has increased dramatically. One may wish to include web-scraped parallel data in a training corpus for an MT system or other application, but inclusion of content that has been output by an MT system is likely to introduce a lot of noise. We may be able to identify such problematic documents by looking for a high occurrence of these n-grams that are found only in MT output.

Although we have investigated only one language pair in this paper, we are confident that our method can be applicable to other language pairs. Further, it might be interesting to use our method to compare different types of MT systems (e.g., statistical vs. rule-based MT systems) as most automatic metrics currently in use do poorly at this. We hope that this paper will encourage the MT community to reexamine the way that they assess the quality of MT systems, so that they will pay closer attention to qualitative differences and not focus solely on optimization for quantitative metrics.

References

Satanjeev Banerjee and Alon Lavie. 2005. M*ETEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.* In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization.*

Deborah Coughlin. 2003. *Correlating Automated and Human Assessments of Machine Translation Quality*. In *Proceedings of MT Summit IX, New Orleans, USA.*

George Doddington. 2002. *Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics.* In *Proceedings of the Second Human Language Technologies Conference(HLT).*

Arul Menezes and Chris Quirk. 2005. *Microsoft Research Treelet Translation System: IWSLT Evaluation.* In *Proceedings of the International Workshop on Spoken Language Translation.*

Kishore A. Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. *BLEU: A Method for Automatic Evaluation of Machine Translation.* In *Proceedings of ACL.*

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. *Dependency Tree Translation: Syntactically Informed Phrasal SMT.* In *Proceedings of the 43rd Annual Meeting of the ACL, Ann Arbor, MI.*