SMT Errors Requiring Grammatical Knowledge for Prevention

Yukiko Sasaki Alam
Department of Digital Media
Hosei University
3-7-2 Kajino-cho, Koganei,
Tokyo, Japan
sasaki@hosei.ac.jp

Abstract

This paper introduces three types of Statistical Machine Translation (SMT) output errors that would require grammatical knowledge for prevention. The first type is due to words that are negative in meaning but not in form. Problems arise when the negative forms are obligatory in target languages. The second type of errors is derived from the rigidity of pattern phrases or correlatives which do not allow for intervening elements. The third type is caused by ellipses in input sentences which must be reinstated for output sentences when so required by rules of omission in target languages or the difference in Head-Complement order between source and target languages.

1 Introduction

Machine translation (MT) output errors are varied, and have been discussed and classified by many researchers. Flanagan (1994) classifies and ranks errors into three levels according to improvability and intelligibility. Elliot et al. (2004) identify fluency- and adequacy-related errors for automatic MT evaluation. Vilar et al. (2006) make a comprehensive classification of SMT output errors. Farrús et al. (2010) present a linguistic-based evaluation of a variety of SMT output errors. Popović and Burchardt (2011) attempt to provide methods for an automatic error analysis of MT output errors that overcome weaknesses of automatic evaluation metrics.

This paper introduces three particular types of errors made at the online English-to-Japanese translation on the Google Language Tools, a cutting-edge SMT system, and demonstrates that grammatical knowledge is required for preventing such errors.

2 Negative in Meaning but Not in Form

There are several English determiners and adverbs that are negative in meaning but not in form, which are termed *negation-implying* words in this paper. Negation-implying determiners are *little* and *few*, whereas negation-implying adverbs include *little*, *seldom*, *rarely*, *scarcely*, *hardly* and *barely*. This discrepancy between meaning and form causes problems with translation into such languages as Japanese, the grammars of which require the explicit forms of negation.

2.1 Negation-implying determiners

The word *few* used as a determiner[1] as in *Few men* in (1a) below means "not many", emphasizing how small a number of people is.:

(1a) *Few people turned up for work.*

In Japanese, such an emphasis on scarcity is usually expressed both by a negation-implying determiner or adverb and the negative form of the predicate that follows it. (1a) could be roughly translated as the following:

(1b) *hotondo-no hito-ga* *shokuba-ni*
almost-POSS[2] people-SUBJ[3] workplace-LOC[4]
araware-nakatta.
show-up NOT-DID

Notice that the predicate of the corresponding Japanese sentence is in the negative form.

Translation of an English sentence with such a negation-implying word into Japanese is problematic, because it requires additional tasks in-

[1] Determiners in English are located at the beginning of noun phrases. They include articles, demonstrative, quantifiers and possessives.
[2] *POSS* indicates the possessive marker.
[3] *SUBJ* denotes the subject marker.
[4] *LOC* stands for the locus marker indicating place, goal or point of time.

cluding the recognizing of the relevant predicate and the changing of it into the negative form, let alone the identification of the part of speech of the word in question. Failure in carrying out such tasks leads to mistranslation. Take the following for instance:

(2a) *Few decisions will have as lasting an impact on your life as your choice of profession.*

The SMT system translates (2a) as a Japanese sentence that means:

(2b) A few decisions turn out to have a lasting impact on your life as the choice of profession.[5]

At this translation, the SMT system probably mistook *few* indicating "not many" for *a few* denoting "several". The resulting Japanese sentence does not convey the same importance on the rarity of such decisions as implied in the input sentence. The output sentence also fails to express the core meaning of the input sentence.

It is necessary to identify the part of speech of *few* in use, because it does not have a negative connotation when used in other parts of speech. It can be used as a pronoun, as illustrated below:

(3) *A lucky few will be able to enjoy the new low monthly payment.*

In addition, it can be used as an adjective, as below:

(4) *Read the first few pages*.

A tripartite distinction should be made in a noun phrase (NP) with *few*: (i) determiner in $_{NP}$[*few* … noun],[6] (ii) pronoun in $_{NP}$[determiner ... *few*], and (iii) adjective in $_{NP}$[determiner … *few* … noun].

A similar word *little* is more problematic, because it can belong to the adverb category as well as the same parts of speech for *few*. Failure to recognize the negation-implying determiners is costly, because the output sentence could indicate an opposite meaning to that of the input sentence. Consider the following:

(5a) *This has received little attention since 1983*.

The SMT system in question translates (5a) as a Japanese sentence that means:

(5b) This has received almost attention since 1983.[7]

The meaning of the output is contrary to that of the input. This suggests that the SMT system requires grammatical knowledge for identifying the parts of speech for *few* and *little*, and, when used as a determiner, it needs to recognize the relevant predicate and transform it into the negative form of the equivalent of the target language.

2.2 Negation-implying adverbs

Adverbs such as *little*, *seldom*, *rarely*, *scarcely*, *hardly* and *barely* are negative in meaning but not in form. They seem to be treated by the SMT system better than the counterpart determiners, but a closer examination reveals that the treatment is indeed erratic.

As long as an input sentence with a negation-implying adverb is short and in the present tense, and the predicate is not a *be* verb, the SMT system in question generates an output with the negative form of the predicate, thus conveying the same emphasis on the rarity of the event. For instance, the sentence *It seldom works out that way* is translated as a sentence with the intended meaning,[8] only with a minor error of the wrong negative inflection of the predicate.

However, the translation of a short sentence fails, when the system is unable to identify the parts of speech of the adverb and the surrounding words, as illustrated below:

(6a) *The world will little note, nor long remember what we say here, but it can never forget what they did here*.[9]

The above sentence is translated as below:

(6b) The world will a short note, and also remember what we who are long say here, but it can never forget what they did here.[10]

The SMT system fails to identify the parts of speech of *little* and *note* in the initial clause, resulting in an ungrammatical output clause. It also fails to recognize that *long* is an adverb, and places the corresponding Japanese adjective immediately before *we*. That means that *long* modifies *we*. The negative conjunction *nor* is totally skipped, thus giving rise to the corresponding clause with the opposite meaning to the input clause

The SMT's treatment of the tense is puzzling for short sentences with a negation-implying adverb. It sometimes does not make a distinction between the present and past tenses. For instance, *He rarely eats red meat* and *He rarely ate red meat* are translated as the same Japanese sen-

[5] The output is "いくつかの決定は、職業の選択としてあなたの人生に影響を持続として持つことになります。".

[6] $_{NP}$*[…]* indicates that the phrase between brackets is a noun phrase.

[7] The output is "これは 1983 年以来、ほとんど注目されています。".

[8] The output is "それはほとんどそのようにうまくいくありません。".

[9] This is from the Gettysburg address by Abraham Lincoln.

[10] The output is "世界は少しノート、また長い私たちがここで言うことを覚えているだろうが、それは彼らがここで何をしたか忘れることはできません。".

tence in the present tense,[11] indicating that he rarely eats red meat. The treatment of the tense is correct sporadically, though.

Near perfect translations were produced when the predicates of sentences begin with *could hardly* or *can hardly*, and the verbs are not a *be* verb, as in:

(7) *She could hardly wait for her child coming back*.[12]

(8) *This fact can hardly be used to explain present patterns of crime*.[13]

The treatment of the tense is also correct in the two examples above.

With a *be* verb, however, even simpler input sentences in (9a) and (10a) than in (7) and (8) are translated as wrong output sentences, shown in (9b) and (10b):

(9a) *This was hardly surprising*.

(10a) *He was barely aware of the feeling*.

The above two sentences are respectively translated as Japanese sentences meaning:

(9b) This was something almost surprising.[14]

(10b) He was aware of almost every feeling.[15]

The resulting outputs convey opposite meanings to the English input sentences.[16]

We have seen that the SMT system in question is erratic in the treatment of negation-implying determiners and adverbs. For a comprehensive treatment, it must identify the part of speech of such a word, and, if it is used as a determiner or an adverb, find the predicate and change it into the negative form of the corresponding predicate.

3 Intervening Elements

It has been observed that when an English sentence consists of a pattern phrase containing an intervening element such as an adverb or a noun phrase (NP) of time, the SMT system in question generates incomprehensible Japanese outputs.

3.1 Adverbials causing discontinuous constituents

Adverbs and adverbial phrases (called hereafter *adverbials*) can be located at several positions of sentences. This mobility of adverbials often creates problems with translation by pattern matching. Take the following for instance:

(11a) *The government (may be paying incorrect subsidies to more than 1 million Americans for their health plans in the new federal insurance marketplace and)*[17] *has been unable so far to fix the errors.*

(11a) is roughly translated as below:

(11b) The government (can pay incorrect subsidies to more than 1 million Americans for its own health plans in the new federal insurance marketplace, and) has not so far been able to do it, in order to fix the errors.[18]

The SMT system in question fails to recognize the pattern phrase *unable to fix the errors*, because the phrase contains an adverbial *so far* between *unable* and *to fix the errors*.[19] Another example follows:

(12a) *Investments in the health systems of low-income countries have long been geared toward treating infectious disease, he said*.[20]

The sentence in (12a) is translated as a Japanese sentence roughly meaning:

(12b) Investments in the health systems of low-income countries are geared toward treating long infectious disease, he said.

The part of speech of *long* in the input sentence is an adverb meaning "for a long time", but in the output sentence, it is an adjective, resulting in a phrase denoting "a long disease". This error is probably due to the presence of *long* in the middle of the pattern phrase *have been geared*. With *long* taken off, (12a) is translated correctly by the SMT system in question.

To deal with adverbs and other unexpected interruptions such as interjections and rephrasing, a controlled skip parser was proposed (Yamada, 1996). The parser uses statistical information (N-

[11] The output is "彼はめったに赤肉を食べません".

[12] The output is "彼女はほとんど彼女の子供が戻ってくるのを待つことができませんでした。".

[13] The output is "この事実はほとんど犯罪の現在のパターンを説明するために使用することができません。".

[14] The output is "これはほとんど驚くべきことでした。".

[15] The output is "彼は感情のほとんどを知っていました。".

[16] It should be noted that when *barely* modifies a number as in *The industry employed barely 150,000 men*, it is translated as a word denoting "almost", and the output Japanese sentence is a little awkward, but intelligible. The output is "業界はほとんど 15 万人を採用します".

[17] The parentheses in the sentence indicate that the elements inside them are not targets for our discussion.

[18] The output is "政府は新たな連邦政府の保険市場で自分の健康の計画のために 100 万人以上のアメリカ人に誤った補助金を支払うことができ、エラーを修正するために、これまでできませんでした。".

[19] The present focus is on translation errors caused by the presence of an adverbial located inside a pattern phrase, and thus the other errors found in the Japanese output are not discussed here.

[20] The output is "低所得国の保健システムへの投資が長い感染性疾患の治療に向けられている、と彼は言いました。".

grams) to determine what to skip in newspaper articles. As far as the skipping of adverbs is concerned, it does not affect the meaning of the sentence when they are such adverbs as *so far* and *long* respectively in (11a) and (12a), but it would be detrimental if they are such adverbs as *little* and *hardly* in (6a) and (9a). A careful research would be conducted on the skipping of adverbials. The skipping of a constituent of an input sentence at machine translation (MT) should be a last resort.

A *to*-infinitive can be used either as an adverbial modifier as well as a nominal one. The following example shows that an adverbial use breaks a pattern.

(13a) *Exercise does more to bolster thinking than thinking does*.

(13a) is translated as a sentence roughly with the following meaning:

(13b) Exercise carries out than, in order to bolster thinking, than thinking.[21]

The Japanese word equivalent to *than* is followed by the object marker, indicating that *than* is interpreted as the object of the predicate, resulting in an intelligible output.

The removal of *to bolster thinking* from (13a) gives rise to a grammatical and comprehensible sentence. This demonstrates that the SMT system is able to handle the pattern phrase x (Subject) + verb + *more than* + y + *does*, but not one with an intervening element.

3.2 Elements intervening relative clauses and the heads

We have seen that the SMT system in question goes awry when an adverbial cuts into a pattern phrase. A similar error occurs when an element intervenes between usually adjacent elements such as a relative clause and the head noun phrase (NP). Take the following for instance:

(14) *"I want to marry Mii, but I can't do that," Marini said in a video posted online that attracted the attention of gaming blogs and online forums this week.*[22]

The SMT system is unable to recognize the head NP and the relative clause *a video … that attracted …* because of the intervening phrase *posted online* between. The removal of the intervening phrase produces a fairly grammatical and intelligible output.[23]

In addition, the following relatively short and simple sentence is translated incorrectly for the same reason as above:

(15) *There were nine men that year who had run faster than 10.2 seconds*.[24]

Again, the SMT system is unable to recognize the relative clause and the head NP because of the presence of *last year* between them. It misinterprets *last year* as the head NP of the relative clause. Furthermore, *last year* is regarded as the subject NP of the predicate *were*. As a result, the output is an unintelligible sentence. The removal of *last year* from (15) produces a grammatical output,[25] recognizing the head NP and the relative clause.[26]

These examples indicate that the system in question needs grammatical knowledge to identify discontinuous constituencies and relations with unexpected intervening elements such as adverbials and NPs of time.

4 Omission and Recoverability

In languages, constituents of sentences which are predictable from context are often omitted. This omission, however, causes problems for machine translation (MT). An MT system must (i) detect if the sentence contains an ellipsis, and if it does and if the reinstatement is required by the target grammar, it must (ii) recover the word or phrase. Such reinstatement would require grammatical knowledge.

4.1 Verbal omissions

In (16a) below, the main predicate *changed* of the final clause is omitted because it is a repetition of the predicate of the immediately preceding clause. As the SMT system fails to recognize the omission, it mistook the auxiliary verb *have* for a regular verb denoting possession:

(16a) *Values have shifted, the population has changed, and the cities have too*.

[21] The output is "運動は思考よりも思考を強化するために、よりを行います。".

[22] The output is "「私はミイと結婚したいが、私はそれを行うことはできません、「マリーニは、ビデオが今週のゲームブログやオンラインフォーラムの注目を集め、オンラインを掲載で述べています。".

[23] The output is "「私はミイと結婚したいが、私はそれを行うことはできません、「マリーニは、今週のゲームブログやオンラインフォーラムの注目を集めたビデオの中で述べています。".

[24] The output is "9 男性はより速く 10.2 秒を実行していたその年がありました。".

[25] The output is "速い 10.2 秒よりも実行していた 9 人の男性がありました。".

[26] However, the verb *run* is translated as a Japanese word meaning "operate" instead of "move at a speed faster than a walk", probably because the frequency of the chosen meaning is higher on the SMT system in question.

The output roughly means:

(16b) The numbers,[27] the population has been changed, has shifted, and[28] the cities possess too much.[29]

The sentence in (16a) is a flat one, conjoining three simple clauses. The SMT system fails to recognize the initial clause *Values have shifted* as a clause. It also fails to understand that the main verb of the final clause has been omitted. The word *too* is treated as the object NP of the verb *have*. The system does not have knowledge that *too* is an adverb, and that it cannot be the object of a verb. Nor does it have grammatical means of distinguishing *have* between uses of the auxiliary and the main verbs.

The following sentence in (17a) contains a relative clause in which both *did* and *did not* share the main verb phrase *watch television*. The system fails to identify the sharing, resulting in an unintelligible output roughly meaning (17b).

(17a) *Conversely the difference between those who did and did not watch television widened.*[30]

(17b) Conversely one conducted television, and difference from those who did not watch it widened.

It is difficult to imagine how such an output is produced. An MT system should be provided with grammatical knowledge for recognizing that the first *did* is not a verb meaning "conduct" or "perform", but is an auxiliary verb.

The omission of the repeated verb phrase poses problems for translation between head-initial languages in Verb-Object (VO) order and head-final languages in Object-Verb (OV) order. In VO languages like English, the object NP shared by two verbs can appear at the final position together with the second verb, creating the configuration of "Verb1 (=Head) and Verb2 (=Head) + Object (=Complement)". On the other hand, in OV languages like Japanese, the object NP should appear at the initial position of the verb phrase, generating the configuration of "Object (=Complement) + Verb1 (=Head) and Verb2 (=Head)". In (17a), the shared complement is the verb phrase *watch television*. (18a) is a partial phrase cut off from (17a) for the sake of discussion:

(18a) *those who did and did not watch television*

The configuration showing the omission in (18a) is "AUX1 [31] (*did*=Head) and AUX2 (*did not*=Head) + Verb Phrase (*watch television*=Complement)". The Japanese configuration that corresponds to (18a) is "Verb Phrase (television watch=Complement) + AUX1 (PAST=Head) and AUX2 (NOT PAST=Head)". The grammatical output of (18a) is roughly as below:

(18b) television-OBJ watch-PAST people and watch-NOT-PAST people

Notice that the verb phrase *watch television* is moved to the initial position. For a proper treatment of this type of omission, the system needs to identify the missing element (the verb phrase in this example), and move it to the appropriate place, according to the target language grammar.

4.2 Other problematic omissions

The omission of repeated verb phrases is fairly common and could be easy to handle, but repeated other categories can be omitted, too. Special measures should be taken for them when source and target languages differ much in grammars of omission. For instance, with recovered missing constituents, (18a) would look like the following:

(18a') *those who watched television and those who did not watch television*

In (18a), *watched television* in (18a') is replaced with *did*, and the repeated phrase *those who* is omitted. A Japanese grammar of omission, however, does not permit such an omission, and therefore *those who* must be reinstated in translation into Japanese. It would be difficult to solve this type of problem by statistical means. It would need several rules of recognition and generation to bridge the gap in this regard between the two grammars.

Another example is an omission that takes place in an idiomatic phrase. English has the phrase *as* + adjective /adverb + *as* + x, which expresses a comparison in relation to the same degree. When the final part of the phrase is equivalent or similar in semantics to *as before*, it can be omitted, as follows:

(19) *When she was running this whole show, and had her own money, she didn't need me as much from that standpoint*.[32]

---

[27] *Value* has at least two meanings: principles or standards of behavior and a numerical amount or number. The meaning of numbers is chosen in the output, probably because of the higher frequency stored at the SMT in question.

[28] Japanese uses different conjunctions for conjoining NPs and clauses. The conjunction used in the output is one for conjoining NPs, generating an awkward translation.

[29] The output is "値は、人口が変更された、シフトしている、と都市があまりにも持っています。".

[30] The output is "逆にテレビを行なったし、見なかった人との差が拡大しました。".

[31] *AUX* stands for auxiliary.

[32] The output is "彼女はこの番組全体を実行し、自分のお金を持っていたとき、彼女はその観点から、私は同じくらい必要はありませんでした。".

(20) *The journey from London to Bath took forty hours in 1720, but only half as long in 1770.*[33]

The output of (19) is unintelligible, because the object *me* of the verb *need* is translated as the subject. The phrase *as much* is regarded as an idiomatic phrase denoting the same, as in *I am sure she would do as much for me*. As a result, the clause *she didn't need me as much from that standpoint* is translated to mean "she from that standpoint I didn't need about the same". Probably the quality of translation would have been better by skipping the phrase *as much*.

The SMT system in question also fails to produce a comprehensible output of (20). The relevant portion in (20) to the topic of omission is *only half as long in 1770*. The adverb *only* is translated as a word meaning "unique", while the phrase *half as long* as a word meaning "half" followed by a stem denoting "long". The system cannot recognize the part of speech of *long*. An accumulation of these errors gives rise to an intelligible output. The removal of *as long*, however, generates a little ungrammatical but understandable Japanese sentence.[34] While looking for a general solution, it might be a realistic alternative to skip problematic words or phrases. But even for this approach, research would be required for identifying what to skip.

5 Conclusion

This paper has shown three types of English-to-Japanese SMT output errors, and demonstrated that the solution of these errors needs grammatical knowledge. The first type is caused by difference in negative implications of words in source and target languages. The second type of output errors is derived from the rigidity of pattern phrases. Adverbs and adverbial phrases can appear at one of several positions of sentences, resulting in discontinuous constituents. The inclusion of adverbials in pattern phrases would cause the proliferation of pattern phrases and probably be not a feasible solution. Skipping them could be an alternative, but the investigation of which one to skip and not to skip would require much research. The third type of errors concern the omission of the constituents of sentences. It is difficult for an SMT system to find a missing constituent, but the target grammar sometimes requires the reinstatement of the omitted element. The solution of these problems would need syntactic parsing and grammatical knowledge.

[33] The output is "バースへのロンドンからの旅は 1770 年に唯一の半分の長 1720 年 40 時間かかりましたが、。".

[34] The output is "バースへのロンドンからの旅は、1720 年 40 時間かかりましたが、1770 年に半分だけ。".

Reference

Bunt, H. and A Van Horck. 1996. *Discontinuous Constituency*. Mouton De Gruyter.

Condon, Sherri, Dan Parvaz, John Aberdeen, Christy Doran, Andrew Freeman, and Marwan Awad. Evaluation of Machine Translation Errors in English and Iraqi Arabic. *The MITRE Corporation*. 7 pages. 2010

Elliot, Debbie, Anthony Hartley, and Eric Atwell. A Fluency Error Categorization Scheme to Guide Automated Machine Translation Evaluation. *AMTA-2004*. 64-73, 2004.

Farrús, Mireia, Marta R. Costa-jussa, Jose B. Marino and Jose A. R. Fonollosa. Linguistic-based Evaluation Criteria to Identify Statistical Machine Translation Erros. *Annual Conference of the European Association for Machine Translation*. 167-173, 2010.

Farrús, Mireia, Marta R. Costa-jussa, Jose B. Marino, Marc Posh, Adolfo Hernandez, Carlos Henriquez, and Jose A. R. Fonollosa. Overcoming statistical machine translation limitations: error analysis and proposed solutions for the Catalan-Spanish language pair. *Language Resources and Evaluation* (Springer). Vol. 45 Issue 2. 181-,. 2011.

Flanagan, Mary A. Error classification for MT evaluation. *-1994*. 65-72, 1994.

Popović, Maja and Aljoscha Burchardt. From Human to Automatic Error Classification for Machine Translation Output. *Proceedings of the 15th Conference of the European Association for Machine Translation*. 265-272, 2011.

Stymne, Sara and Lars Ahrenberg. On the practice of error analysis for machine translation evaluation. *LREC*. 1785-1790. 2012.

Talbot, David, Hideto Kazawa, Hiroshi Ichikawa, Jason Katz-Brown, Mranz J. Och. A Lightweight Evaluation Framework for Machine Translation Reordering. *Proceedings of the 6th Workshop on Statistical Machine Translation*. 12-21. 2011.

Vilar, David, Jia Xu, Luis Fernaudo D'Haro, and Hermann Ney. Error Analysis of Statistical Machine Translation Output. *Proceedings of the LREC*. 697-702. 2006.

Yamada, Ken. 1996. A controlled Skip Parser. *Proceedings of the 2nd AMTA Conference*.