International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:11, No:2, 2017

Knowledge Required for Avoiding Lexical Errors at Machine Translation

Yukiko Sasaki Alam

***Abstract*—**This research aims at finding out the causes that led to wrong lexical selections in machine translation (MT) rather than categorizing lexical errors, which has been a main practice in error analysis. By manually examining and analyzing lexical errors outputted by a MT system, it suggests what knowledge would help the system reduce lexical errors.

***Keywords*—**Error analysis, causes of errors, machine translation, outputs evaluation.

I. Introduction

THIS paper explores the causes of output errors in a statistical machine translation (SMT) system from English to Japanese, in particular, lexical errors, and aims to arrive at recommendations to improve the quality of MT in terms of better lexical selection.

A main reason for choosing output errors from a SMT system is due to an anticipation of the great potential of statistics applicable to MT. Furthermore, the reason for choosing *Google Language Tools* as a system for investigation is its enormous amount of lexical resources.

The methods of this study are examining each pair of input and output sentences, detecting errors, determining the causes, and rating output sentences. Although syntactic errors gravely affected the readability or intelligibility of output sentences, they have not been discussed in this study.

II. Related Study

Research for error analysis of MT outputs has a long history [4]-[7], [14], [17], [19]. In the past, however, more emphasis has been placed on the classification of error categories rather than on their causes. This study investigated the causes of the selection of wrong outputs.

Errors involving word selection have been chosen for a detailed study, since they are a most serious and predominant error category. Reference [7], for instance, regards it as a Class 3 type, a most serious error category in terms of both "improvability" and "intelligibility".

III. Data

The data used for investigation comprises seven articles on the stock market which were collected from online news and financial magazines in early 2016. The total number of sentences is 226, and the total number of words 4,440. The average length of sentences is 19.6 words[1], with the shortest of four words and the longest of 48 words. The intelligibility of each sentence was manually measured according to the criteria illustrated in Table I, in which the degree of intelligibility is broken down into five levels from a totally incomprehensible sentence (0 point) to a completely understandable sentence (4 points).

TABLE I
Criteria for Evaluating the Intelligibility of MT Outputs

Five Degrees of Intelligibility	Score
Correctly conveys the intended meaning	4
Possible to guess the intended meaning in spite of several unintelligible portions	3
Half of the intended meaning is intelligible	2
Most words and phrases are unintelligible in spite of several intelligible ones	1
The output sentence is totally incomprehensible	0

Table II illustrates the intelligibility-based distribution of measured output sentences. The results show that 35.1% of the total number of sentences was rated unintelligible, 18.2%, sporadically intelligible, about a quarter (24%), half understandable, 13.8%, approximately intelligible, and 8.9%, completely intelligible.

Causes of this low intelligibility could be divided largely into syntactic and lexical errors. It is hard to speculate how much syntactic errors or how much lexical errors contribute to the low intelligibility. Syntactic errors irrevocably change the grammatical relations of sentence components, while lexical errors give rise to sentences full of confusing words. Short sentences had less syntactic errors, but suffered from wrong lexical selections.

The results of evaluation in Table II indicate that the SMT system was not able to produce a satisfactory level of translation outputs, with 35.1% of the outputs totally incomprehensible. In fact, as 18.2% was rated only sporadically comprehensible, more than half (53.3%) of the evaluated sentences were unintelligible. As the sentences were collected from relatively serious online articles related to the stock market, they were fairly long, with an average length of about 20 words.

Table III illustrates the relation of sentence length (incremented by five words) and intelligibility, and the distribution of sentences in each length group. The number of sentences with 10 words or less is 35 out of 226, and the intelligibility scored 1.97 points, which means that most of

Yukiko Sasaki Alam, Ph.D. is a full professor at the Department of Digital Media, Faculty of Information and Sciences, Hosei University, Tokyo, 184-8584, Japan (phone: 81-42-387-4549; fax: 81-42-387-4560; e-mail: sasaki@hosei.ac.jp).

[1] A multiword such as *take care of* is counted as three-word long.

them were half understandable. The number of sentences with 11 to 15 words is 50. About a quarter of all the sentences were in about this range of length; the intelligibility was 1.58 points, indicating a little less than 50% of intelligibility. The number of sentences with 16 to 20 words is 48; almost a quarter of all the sentences, and the score of intelligibility was 1.48, indicating less than 50% of intelligibility. The number of sentences with 20 words or less was about 60% of all the sentences, and the number of sentences with 25 words or less accounted for a little over 70% of all the sentences. As expected, the shorter the sentence, the higher the intelligibility. Still, even the short sentences of 10 words or less just achieved about 50% of intelligibility. Causes of failure to translate short sentences were no doubt various, including syntactic, morphological and lexical errors, but lexical errors contributed much to the low intelligibility of short sentences. Lexical errors are a category of error that was discussed much in previous literature on error analysis [4]-[7], [14], and [19].

TABLE II
DISTRIBUTION OF OUTPUTS BY INTELLIGIBILITY CRITERIA

Total no. of sentences	Average sentence length (words)	Average Intelligibility	Intelligibility				
			0 point (not at all)	1 point (sporadically)	2 points (half)	3 points (approximately)	4 points (completely)
			No. of sentences (percentage of the number)				
226	19.6	35.18%	79 (35.1%)	41 (18.2%)	55 (24.0%)	31 (13.8%)	20 (8.9%)

TABLE III
INTELLIGIBILITY DIFFERENCES IN TERMS OF LENGTH OF SENTENCES

Sentence length (no. of words)	<= 10	11-15	16 - 20	21 - 25	26 - 30	31 - 35	36 - 40	41 - 45	46 -50
No. of sentences	35	50	48	34	28	19	8	3	1
Intelligibility (max = 4)	1.97	1.58	1.48	1.29	1.18	1.11	0.75	0.33	1.00

While syntactic error categories such as word order, conjunction, apposition, clause and phrase boundary, discontinuous construction, modification relation, grammatical relation, and sentence pattern were detected in the course of the current investigation, this paper has focused on lexical errors, because they were as pervasive and significant (rated one of the most serious error categories in [7]) as syntactic errors, and it was easier to take a look into them than syntactic errors.

IV. RESULTS

The selection of incorrect output words was due to a lack of various kinds of knowledge. Table IV lists required knowledge, the absence of which primarily led to the choice of wrong output words. This list is not an exhaustive one, but includes a majority of the findings in this research. It should be noted, however, that some errors overlap each other to a certain degree because of the interactive nature inherent in knowledge involving language. In the following, a detailed discussion is presented on some of the causes listed in Table IV.

A. Semantic Coherence

By far the most prevalent type of lexical error was those caused by the failure to observe the semantic coherence of the relevant constituents. This type of error accounted for about 24% of the total lexical errors detected in this study. Such an error is found in the noun *appetite* in *Honeywell's appetite for expansion*. The SMT system chose the Japanese term denoting 'desire for food', a wrong equivalent because *expansion* does not stand for food. Another example is the verb *see* in *see negative pressure*. It was translated into a Japanese term denoting 'refer to'. The verb *see* has several meanings, and the meaning of experiencing or the like should have been chosen. The noun phrase *brightest minds* in *the brightest minds in their respective product categories* was translated into a phrase that means 'lively spirit'. The verb *beat* in *beat the price* was translated into a word for 'hit (a physical object)', even though *price* does not refer to a physical object. The verb *cut* in *cut $1.3 billion in expenses* was translated into a word indicating 'make an incision'. The verb *charge* in *charging customers half the cost of their bills with other carriers* was translated into a term standing for 'store electrical energy'.

Errors involving sense incompatibility are divided into two groups: incompatibility between: (a) head and complement (e.g. verb and the argument), and (b) head and modifier (e.g. a noun and the adjective that modifies it). The relation (a) is found in a verb phrase or a noun phrase with the deverbal head noun, while (b) in a noun phrase or in an extended verb phrase that contains an adverbial modifier.

The first task of the system, whether a SMT or not, is to recognize the boundaries for a noun phrase or a verb phrase. It is hard to recognize it without a syntactic parser. Some might say that a list of translation pairs serves the purpose, as is the case with a current SMT. But how long should the list be? A less frequent phrase, such as *see negative pressure*, is less likely to get paired with a corresponding translation.

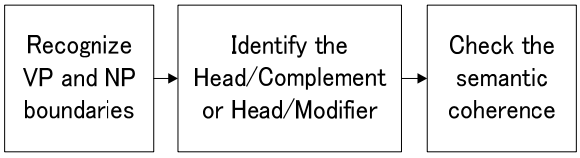

Fig. 1 Process for checking the semantic compatibility of the relevant constituents

TABLE IV
KNOWLEDGE REQUIRED FOR AVOIDING LEXICAL ERRORS (LES)

	Required Knowledge	No. of LEs	% of the LEs	Examples of Lexical Errors
1	semantic coherence	66	23.9	*cash-generating power* (not 'electrical power') *more hefty dividend hikes* (not 'long walking')
2	verb patterns	32	11.6	*leave* X *undervalued* (not 'abandon')
3	domain terminology	22	8.0	*19 percent rally* (not 'mass meeting') *short more of his shares premarket* (not adjective)
4	MWEs - nouns	11	4.0	*nest egg* (not 'eggs lying in the nest')
5	MWEs - verb + object or subject + verb	12	4.3	*hit a record* (not 'disc') *fortunes turn* (not 'destiny gets switched on')
6	MWEs - phrasal verbs	14	5.1	*sell off* (not 'sell while a person is off')
7	MWEs - quantities	12	4.3	*a bit of help* (not 'help's information unit')
8	MWEs - idiomatic phrases/adverbials	13	4.7	X *instead of* Y (not 'Y instead of X') *as low as $56.30* (generating an incorrect output) *go all the way back to X* (not 'return to X and go every way')
9	clause/phrase boundaries	11	4.0	*What it saw was* … (not 'tool for cutting') *as the strong dollar devalued foreign results* (*as* not to mean 'in the capacity of')
10	grammar – missing verbal suffixes	18	6.5	X *that* Y *expects* Z *to earn in the current fiscal year* (not 'X that Y expectation to earn …': *expects* was translated as a noun without the verbal suffix)
11	grammar – failures to identify the grammatical roles of V-*ing*	9	3.3	[main clause], *topping the $734.6 million analysts had predicted* (not 'the topping that 734. 6 million analysts predicted')
12	grammar - modality	5	1.8	*It offers a service that many people simply couldn't live without* (not 'it offers a service that many people were not able to live without')
13	grammar and other knowledge	28	10.1	X's *prospects* (not 'likelihood'), *as it has so far this year* (not 'possess'), *its 2015 high* (not adjective)
14	missing equivalents	9	3.3	*record price*, *pullback*, *low end*, *flywheel*, *tact, lure*
15	rules for quotes	11	4.0	Translation deteriorated much without the rules.
16	noun patterns	3	1.1	*part of the portfolio's confidence in the company* (not 'part of confidence in the portfolio within the company')
	TOTAL	276	100	

The second task is to find, for instance, the head and the complement(s) of the verb phrase or the head and the modifier of a noun phrase. The third task is to check the semantic coherence of head/complement or head/modifier. The required semantic feature of the complement does not have to be complex. For instance, in cut $1.3 billion in expenses, the semantic feature of the object of the verb cut required for denoting reducing can be a quantity. Would SMT systems be able to implement such checking tasks in their engines? For instance, the translation pair such as cut X in Y → REDUCE THE AMOUNT OF X in Y is not adequate, because cut X in Y → DIVIDE X in Y is possible. A SMT will be capable of making a correct lexical selection among a relatively short list of possible selections, given that a SMT system is able to find X, that it has a list of such translation pairs, and that it is able to identify the similarity of the two nouns.

B. Verb Patterns

Many verbs have different meanings when used in different verb patterns or subcategorizations [8]. By identifying the verb pattern used in the sentence, it is often possible to identify the meaning in use. The SMT system's failure to recognize verb patterns caused wrong lexical selections. Such errors made up 11.6% of all the lexical errors. For instance, verbs such as *leave*, *keep*, *hold, make*, and *rate* were used in the verb pattern of "subject + verb + object + complement," but the system failed to recognize that pattern. The verb *left* in *The collapse of oil and natural gas prices has left much of the energy industry awash in red ink* was wrongly recognized as a transitive verb denoting leaving a place, and the sentence was translated into one indicating 'The collapse of oil and natural gas prices is awash in red ink, and has left the energy industry'.

Another example is its failure to identify an idiomatic verb pattern of *keep*. In *helped keep the stock from a deeper decline*, *keep* in the pattern of '*keep* + NP1 + *from* + NP2' stands for 'protect', but it was incorrectly translated into a term denoting 'retain possession of', which is a meaning used as a transitive verb. In addition, due to a lack of knowledge about the verb pattern of '*characterize* NP1 *as* NP2', the system translated *as a small hiccup* in *characterizing this news as a small hiccup* into an incorrect phrase denoting 'in the same manner as a small hiccup'. Some verbs occur in verb patterns that would not be listed in the lexical entries in a regular dictionary. The following verb pattern of *hit* was frequent in the articles on the stock market: *hit a record* + a quantity of money, as in *per-share earnings hit a record $7.58*. The sentence was translated into one indicating 'per-share earnings stroke $7.58 with a phonograph record'.

The system should be provided with information on domain-specific idiomatic verb patterns that appear frequently in texts on the domain. Furthermore, hopefully, it should be able to handle some unregistered verb patterns as well, because, for instance, most verbs of creation can take double objects, even though the double object verb patterns are not registered in the dictionary entries. For instance, in *have carved themselves wide economic moats*, the system translated it into a phrase connoting 'wide economic moats (are) engraved', and displaced the equivalent for *themselves* so far away from the output verb phrase that it became difficult to understand the relation of *themselves* with the verb phrase.

To identify verb patterns, the system needs a syntactic parser as well as the lexicon containing information on verb patterns for the verbs together with the meanings associated with the patterns. Otherwise, it needs to store a large number of possible combinations of words, and the number will increase exponentially.

C. Domain Terminology

Identification of domain terminology is important in natural language processing including MT. Much research has been devoted to this effect ([1], [13], and [20], to name just a few). All the texts examined in this study are related to the stock market, and errors involving the domain terminology accounted for 8% of all the lexical errors.

The SMT system in question was not consistent with the handling of the technical terms, the reason for which is hard to figure out. The word *stock* was sometimes incorrectly translated into a term for 'goods in a warehouse', while sometimes into a proper equivalent. The word *unwind* in *investors who borrowed to buy shares had to unwind trades* was incorrectly translated into a term for 'untie' or 'unfasten', resulting in a confusing output phrase. The phrase *unwind a trade* is a domain-specific one, indicating 'reverse a securities transaction'. Another example is the verb *short*, which is a domain-specific term for 'sell'. The verb in *short more of his shares premarket* was translated into a Japanese adjective denoting 'short in length'. The wrong selection of the meaning for the adjective use could have been prevented due to the syntactic anomaly if the SMT system had been able to understand *more of his shares* as a complete noun phrase or due to the semantic incoherence between modifier and modified if the system had been equipped with a module checking for coherence.

The word *compile* in *as compiled by FactSet* was translated into a *Hiranaga* character-transcribed *konpairu*, a Japanese technical term for 'compile' used in computer science. The system could have been able to avoid the use of the technical term in other domains, if it had knowledge about domain-specific meanings.

D. Multiword Expressions (MWEs) – Compound Nouns

Ever since the publication of [16], more and more research has been devoted to MWE problems that confront natural language processing [2], [3], [9]-[12], [15], [18]. MWEs refer to compound words which should not be translated compositionally. Errors relating to nominal MWEs occupied 4% of the total lexical errors.

Several MWEs were translated into wrong words because they were translated compositionally. For instance, *nest egg* was translated into a phrase denoting 'eggs lying in the nest'. In *characterizing this news as a small hiccup*, *a small hiccup* in this context stands for 'a temporary setback', but was translated into a phrase referring to an involuntary spasm of organs of the human body.

Knowledge on nominal MWEs are helpful, not only in the selection of correct translation words, but also in syntactic parsing. In the long noun phrase *its operational efficiency, track record of execution and rational capital allocation decisions*, the system failed to recognize *track record* as a MWE, and treated *track* as a verb, resulting in an incorrect output sentence that means that its operational efficiency tracks records of execution and rational capital allocation decisions. The incorrect output also reveals that the system lacks a grammar stating that in the present tense with the third person singular subject, *s* should be added to the end of the verb. With this knowledge, *track* would not have been dealt with as the predicate verb, because the incorrectly identified subject *its operational efficiency* is third person singular.

E. MWEs – Bare-Bones Verb Phrases

Not only nominal MWEs, but some verb phrases in idiomatic use should not be translated compositionally. An example often cited in this regard is *kick the bucket*, which means 'die' when used idiomatically. Errors of this type accounted for 4.3% of all the lexical errors. For instance, *ate my words* in *I used to be skeptical about things like the "death cross" but I ate my words a few months ago* was translated into a phrase denoting 'digested my words'. The verb phrase *ran more numbers* in *I went back and ran more numbers on the "death cross" going all the way back to the 1920s* is another of this kind. The verb phrase *run the numbers* means 'make numerical calculations'. This particular example is more difficult to recognize, because the comparative adjective *more* replaces *the*, modifying *numbers*. There must be a rule for dealing with possible variations of idiomatic verb phrases, because a variation like this example occurs often, and because verbs conjugate into several forms.

Attention should be paid not only to a combination of a verb and the object, but to a combination of a verb and the subject when the verb is intransitive. In the sentence *"Trees don't grow to the sky," the old Wall Street line goes*, the system translated *the old Wall Street line goes* into a phrase that means that the long, narrow line of the old Wall Street goes in. When the subject of *go* refers to a speech, proverb, song and the like, the verb means 'say' or 'state'. In this example, it is also difficult to figure out the correct meaning of the noun *line*. A possible approach is to recognize (a) that the head noun of the subject noun phrase of *the old Wall Street line goes* is *line*, (b) that the predicate verb is *goes*, and thus, (c) that a combination of the subject head noun *line* and the predicate verb *go* denotes 'the proverb says'. The following is another example in which a combination of the meanings of the subject and the verb matters. In *the company's fortunes have turned*, the turning of one's fortunes indicates the change of one's fortunes. The system, however, failed to recognize the meaning generated by the combination, and outputted a sentence indicating 'the destiny of the company has been switched on', the meaning of which is not intelligible.

F. MWEs – Phrasal Verbs

A phrasal verb is a verb followed by a preposition, an adverb or both, and the combination creates a meaning different from the meaning of the verb used alone. Failure to recognize phrasal verbs caused serious problems. Such failures occupied 5.1% of

the total lexical errors found in this research. Some examples follow.

The phrasal verb *sell off* in *brands representing about $30 million in sales that it wants to sell off* was not identified as a phrasal verb, and translated into a Japanese phrase that roughly means 'brands that present $30 million as sales that we want to sell while we are off'. The adverb *off* was treated independently from the verb *sell*. In *Exxon stepped up with a 5.8% increase*, *stepped up* was not translated as a phrasal verb, but only the pronunciation of the phrase was transcribed in *Katakana* characters, which are customary characters to transcribe loan words. A similar treatment was observed in *rack up* in *The world's largest retailer has continued to rack up greater sales*. Only the difference from the treatment of *stepped up* was that *up* was transcribed before *rack*, resulting in a word *appu rakku* in *Katakana* characters. In *The death cross is saying to stay away from U.S. stock*, the system failed to recognize the phrasal verb *stay away from*, and the sentence was translated into a sentence roughly indicating 'the death cross is saying to be away, and makes a stay from U.S. stock', the meaning of which is difficult to guess.

Much more difficult to handle was a phrasal verb the components of which are discontinuous. The phrasal verb *take up* in *That would take the index up near 17,935* means raising, but only *take* was translated into a term for 'acquire', resulting in an unintelligible sentence. For the identification of discontinuous phrasal verbs, the system must identify the object noun phrase between the verb and the preposition, and thus it requires a syntactic parser.

G. MWEs – Expressions of Quantity

As the texts under investigation are on stock market, they tend to contain many expressions of quantity. The SMT system in question was weak and not consistent with handling them. Such errors accounted for 4.3% of all the lexical errors. In particular, it failed to treat a phrase consisting of a quantity word followed by the preposition *of* followed by a noun phrase (NP): a NP in a sequence of a quantity word + *of* + NP. Some examples are *years of strong growth*, *a bit of help*, *57% of analysts*, and *66% of Apple's total sales*. In all such phrases, the NP part following *of* was first translated, *of* was next, and the quantity word preceding *of* was last.

The system seemed to translate such a quantity expression in the same manner as a regular noun phrase containing *of* such as *the destruction of the city*. Although *a bit of help* means a little help, it was translated into a phrase denoting 'help's information unit'. The phrase *a fraction of the volatility and risk* was translated into a phrase standing for 'rate of the volatility and risk'. When dealing with noun phrases in the construction of NP1 + *of* + NP2, it is necessary to check if NP1 is a quantity word or not. In addition, the system often failed to translate simple numerical expressions such as $8 billion, $8.1 billion, and $2 billion, although it was successful in other cases. This inconsistency is puzzling.

H.MWEs – Other Idiomatic Phrases and Adverbials

Other idiomatic phrases and adverbial multiword expressions such as *as much as*, *instead of* and *kind of* were not properly translated. Such errors accounted for 4.7% of all the lexical errors. The phrase X *instead of* Y was translated into a phrase denoting the opposite meaning, namely 'Y instead of X'. In X *is really kind of concerning*, the system translated it into a sentence indicating 'X is very kindly concerning'. It failed to identify the idiomatic adverbial *kind of* denoting 'to some extent'.

It is relatively easy to fix problems of such phrases when the components are contiguous, but it is problematic when they are discontinuous. *Thanks to* in *thanks in part to falling demand from China and emerging economies* allows an intervening element *in part* in the middle, and hence, the input sentence was broken down into incoherent components, resulting in an incomprehensible output sentence. A solution to such a discontinuous phrase is the use of a syntactic parser. Otherwise, since a possible intervening element like this example is not numerous, several phrases with a possible intervening element should be prepared to handle such cases.

I. Clause and Phrase Boundaries

Although the system's failure to recognize clause or phrase boundaries caused more syntactic errors than lexical ones, they also affected word selection. Such errors accounted for 4.0% of all the lexical errors. For instance, in *What Honeywell saw was the chance to wring from United the kind of efficiencies Honeywell has achieved in its own businesses*, *What Honeywell saw* is a clause, and *saw* is the predicate verb. The system, however, failed to recognize that, and the verb was translated into a noun standing for 'tool for cutting wood'. In *57% of the analysts who cover the stock rate it as a "buy"*, because the system was unable to identify the clause boundaries, [*who cover the stock*], it outputted an expression for 'stock rate'.

The selection of an appropriate meaning of *as* from among many requires knowledge on the clause boundaries or phrase boundaries as well as on the semantic relation between the preceding and following clauses or phrases. In *The dividend provided a firewall, of sorts, as the Wal-Mart shares crumbled to as low as $56.30 in December*, it is difficult to identify the clausal conjunction *as* without recognizing the semantic relation between the two (preceding and following) clauses involved. The system translated it into a term standing for 'in the capacity of', which was a wrong translation word.

Following is another instance requiring knowledge on the relationship of the two clauses involved in order to choose the correct meaning of a clausal conjunction. In *3M shares are expected to stay flat, based on analysts' price targets as compiled by FactSet, while IBM is actually overvalued by 11%, going by the median analyst price target*, *while* was translated into a word with a temporal sense standing for 'during the time', although it originally meant 'whereas' indicating a contrast of the two situations denoted by the preceding and following clauses. Because the selection of the proper meaning involves semantics of the clauses involved, this treatment is a difficult one. A solution to this is to pay attention to the verb types and

tenses of the predicates such as whether they are a state type or an action type. When the verb type is a state, and the tense is present, the meaning of *while* is likely to denote 'whereas'. Semantic information in this regard is in need, but to formalize such information requires a lot of linguistic investigation to create workable rules.

J. Grammar – Missing Verbal Suffixes

A significant number of verbal suffixes were missing in output sentences. Verbal suffixes typically indicate tense, voice (passive or active voice), and modality. Their absence indicates the system's weakness with grammar, in particular concerning syntax and morphology. Errors of this type occupied 6.5% of all the lexical errors. In *If the 19% rally implied by that target arises*, the past participle *implied* was translated into a noun denoting implication, and as a result, the relation of modifier and modified was not recognizable in the output sentence, causing an incomprehensible *if*-clause. In *just 23% of the $9.07 per share that Wall Street expects Apple to earn in the current fiscal year*, the predicate verb *expects* of the relative clause was translated into a noun denoting expectation, and the relation of the antecedent and the relative clause became unclear, thus generating a confusing output sentence. Japanese grammar sometimes allows a sentence to end with a deverbal noun when the verb is the predicate of the matrix clause, but that style is not allowed for the predicate verb of a dependent clause such as a relative clause and a *when*-clause. The SMT system in question, however, seems to fail to distinguish between a matrix clause and a dependent clause.

K. Grammar – Failure to Identify the Grammatical Role of the V-ing form of the Verb

A verb suffixed with *ing* (termed V-*ing* here) has many functions. It can be a present participle or a gerund in form. It can be used as a noun (in case of a gerund), a part of the progressive form of a predicate, and an adjective (both in case of a present participle). It can also appear at the beginning of a dependent clause. The usages are various, and it is difficult to identify the function in use without resorting to grammatical (in particular, syntactic and morphological) knowledge. Errors involving V-*ing* accounted for 3.3% of all the lexical errors. In fact, they are closely related to syntactic errors, and therefore only those errors involving lexical selection were discussed here. In …, *topping the $734.6 million analysts had predicted*, the clause beginning with *topping* was translated into a phrase indicating 'a topping that $734.6 million analysts had predicted'. That is, *topping* was treated wrongly as a noun. In *one encouraging sign*, *encouraging* was treated as a verb, and *sign* as the object, resulting in a Japanese phrase denoting 'to encourage signs'. A similar example is *falling demand* in *thanks in part to falling demand from China and emerging economies*. The phrase *falling demand* was translated into the predicate verb phrase of the matrix sentence, denoting '(the subject of the matrix sentence) drops demand'. In the verb phrase *signal growing confidence*, *growing* was treated as a verb, resulting in a phrase indicating 'to grow and inform people of confidence'. In *unmatched depth and breadth in growing global health care markets*, *growing* was mistaken for a noun, outputting a phrase denoting 'unmatched depth and breadth in the growth of global health care markets'. These errors are closely related to the system's consistent weakness with syntactic parsing.

L. Grammar and Other Knowledge

Errors due to a lack of grammatical knowledge other than the areas discussed above accounted for 10.1% of all the lexical errors. Errors in this group are varied, and only some types are taken up here. The plural noun *prospects* in *Wal-Mart's prospects* was treated as a singular noun, and translated into a word standing for 'likelihood of some future event occurring'. But the noun in plural form means 'chances or opportunities for success or wealth'. English has words like *prospects* that have different meanings in singular and plural forms. Another example is *fortune*. In *the company's fortunes have turned*, *fortunes*, a plural noun, was incorrectly translated into an equivalent denoting destiny. *Fortunes* stand for 'success or failure', not 'destiny'. The system does not seem to prepare for nouns of this type.

In the following example, the present perfect aspect was not identified, and the auxiliary verb *have* was translated into a regular verb for 'possess'. In *If the dollar continues to weaken against many foreign currencies, as it has so far this year*, *has* was translated into a word standing for possession. The auxiliary verb *have* in *CVS shares have more than tripled* was also translated into a term for 'possess', resulting in an unintelligible output.

The preposition *with* has many functions, among which are to indicate a partner as in *He hung out often with high school friends* as well as an accompaniment as in *cameras with night vision*. The preposition *with* in each example must be translated into two different Japanese words because the complement of one use of *with* refers to a human while that of the other to an inanimate object. In *Exxon stepped up with a 5.8% increase*, *with* was translated into the word used only for a human accompaniment, resulting in an awkward sentence.

V. Conclusion and Future Work

This research has identified a variety of causes of lexical errors. One of the most common causes was lack of knowledge about semantic coherence between verb and object, verb and subject, or modifier and modified. This is one of the most difficult problems to solve. Just to increase the number of corresponding collocations is not a realistic approach, because that number exponentially adds up to enhance coverage, and it will soon be unmanageable. A proper and useful approach is to provide words with minimal semantic features. After all, in a similar vein, bags of words have been used for practical purposes in natural language processing to detect semantic coherence in a broad sense.

Some areas of knowledge are relatively easy to implement: domain terminology, verb patterns, noun patterns, multiword expressions, general vocabulary, and domain-specific verb patterns. Texts on the stock market are frequented by such verbs of movement as slip, remain, leave, fall, hit a record, drop,

and the verb patterns of such verbs should be well taken care of. Knowledge about verbal suffixes, and phrase and clause boundaries involves syntactic knowledge, and thus the improvement in syntactic rules helps reduce lexical errors.

The current research has detected many syntax-related errors such as word order, conjunction, apposition, clause and phrase boundaries, discontinuous construction, modification, grammatical relations and sentence patterns. It has been observed that weaknesses in syntactic parsing led to lexical errors, and that the lack of lexical knowledge caused syntactic errors. An in-depth study of the causes of syntactic errors remains to be seen.

REFERENCES

[1] Anick, P, Verhagen, M., and Pustejovsky, J. 2014. Identification of Technology Terms in Patents. LREC 2014. 2008-2014.

[2] Baldwin, T., Bannard, C., Tanaka, T. and Widdows, D. 2003. An Empirical Model of Multiword Expression Decomposaibility. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions Analysis, Acquisition and Treatment*. 89-96.

[3] Church, K. 2013. How Many Multiword Expressions Do People Know? *ACM Transactions on Speech and Language Processing*. 10(2), Article 4: 1-13.

[4] Elliot, D, Hartley, A., and Atwell, E. 2004. A Fluency Error Categorization Scheme to Guide Automated Machine Translation Evaluation. *AMTA 2004*. Pages 64-73.

[5] Farrús, M., Costa-jussa, M., Marino, J., and Jose Fonollosa, J. 2010. Linguistic-based Evaluation Criteria to Identify Statistical Machine Translation Errors. *EAMT 2010*. Pages 167-173.

[6] Farrús, M., Costa-jussa, M., Marino, J., Posh, M., Hernandez, A., Henriquez, C., Jose A., and Fonollosa, J. 2011. 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. 181-208.

[7] Flanagan, M. 1994. Error classification for MT evaluation. *AMTA 1994*. 65-72.

[8] Hunston, S. and Francis, G. 2000. *Pattern Grammar A corpus-driven approach to the lexical grammar of English*. Benjamins Publishing Co.

[9] Hurskainen, A. 2008 Multiword Expressions and Machine Translation. Technical Reports in Language Technology Report No 1, 2008. http://www.njas.helsinki.fi/salam.

[10] Kim, S. and Baldwin, T. 2013. Word Sense and Semantic Relations in Noun Compounds. *ACM Transactions on Speech and Language Processing*. 10(3), Article 9: 1-17.

[11] Kordoni, V. and Simova, I. 2014. Multiword Expressions in Machine Translation. LREC 2014. 1208-1211.

[12] Lau, J., Baldwin, T., and Hewman, D. 2013. On Collocations and Topic Models. *ACM Transactions on Speech and Language Processing*. 10(3), Article 10: 1-14.

[13] Nadeau, D. and Sekine, S. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*. 30(1):3-26.

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

[15] Ramisch, C., Villavicencio, A., and Kordoni, V. 2013. Introduction to the Special Issue on Multiword Expressions: From Theory to Practice and Use. *ACM Transactions on Speech and Language Processing*. 10(2), Article 3: 1-10.

[16] Sag, I., Baldwin, T., Bond, F., Copestake, A, and Flickinger, D. 2002. Multiword Expressions: A Pain in the Neck for NLP, In *Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico.

[17] Stymne, S. and Ahrenberg, L. 2012. On the practice of error analysis for machine translation evaluation. *LREC*. 1785-1790.

[18] Shutova, E., Kaplan, J., Teufel, S., and Korhonen, A. 2013. A Computational Model of Logical Metonymy. *ACM Transactions on Speech and Language Processing*. 10(3), Article 11:1-28.

[19] Vilar, D., Xu, J., D'Haro, L., and Ney, H. 2006. Error Analysis of Statistical Machine Translation Output. *Proceedings of the LREC*. 697-702.

[20] Wong W, Liu W, and Bennamoun M. 2007. Determining Termhood for Learning Domain Ontologies in a Probabilistic Framework. In *6th Australasian Conference on Data Mining (Isbn 978-1-920682-51-4*).