Toward Determining the Comprehensibility of Machine Translations

Tucker Maney, Linda Sibert, and Dennis Perzanowski
Naval Research Laboratory
4555 Overlook Avenue, SW
Washington, DC
{tucker.maney|linda.sibert|dennis.perzanowski}@nrl.navy.mil

Kalyan Gupta and Astrid Schmidt-Nielsen
Knexus Research Corporation
163 Waterfront Street, Suite 440
National Harbor, MD
{kalyan.gupta.ctr|astrid.schmidtnielsen.ctr}@nrl.navy.mil

Abstract

Economic globalization and the needs of the intelligence community have brought machine translation into the forefront. There are not enough skilled human translators to meet the growing demand for high quality translations or "good enough" translations that suffice only to enable understanding. Much research has been done in creating translation systems to aid human translators and to evaluate the output of these systems. Metrics for the latter have primarily focused on improving the overall quality of entire test sets but not on gauging the understanding of individual sentences or paragraphs. Therefore, we have focused on developing a theory of translation effectiveness by isolating a set of translation variables and measuring their effects on the comprehension of translations. In the following study, we focus on investigating how certain linguistic permutations, omissions, and insertions affect the understanding of translated texts.

1. Introduction

There are numerous methods for measuring translation quality and ongoing research to improve relevant and informative metrics (see http://www.itl.nist.gov/iad/mig/tests/metricsmatr) (Przybocki et al., 2008). Many of these automated metrics, including BLEU and NIST, were created to be used only for aggregate counts over an entire test-set. The effectiveness of these methods on translations of short segments remains unclear (Kulesza and Shieber, 2004). Moreover, most of these tools are useful for comparing different systems, but do not attempt to identify the most dominant cause of errors. All errors are not equal and as such should be evaluated depending on their consequences (Schiaffino and Zearo, 2005).

Recently, researchers have begun looking at the frequencies of errors in translations of specific language pairs. Vilar et al. (2006) presented a typology for annotating errors and used it to classify errors between Spanish and English and from Chinese into English. Popovic and Ney (2011) used methods for computing Word Error Rate (WER) and Position-independent word Error Rate (PER) to outline a procedure for automatic error analysis and classification. They evaluated their methodology by looking at translations into English from Arabic, Chinese and German and two-way English-Spanish data (Popovic and Ney, 2007). Condon et al. (2010) used the US National Institute of Standards and Technology's NIST post-editing tool to annotate errors in English-Arabic translations

These methods have all focused on finding frequencies of individual error categories, not on determining their effect on comprehension. In machine translation environments where post-editing is used to produce the same linguistic quality as would be achieved by standard human translation, such a focus is justified. A greater reduction in the time needed to correct a translation would be achieved by eliminating errors that frequently occur.

However, there are situations in which any translation is an acceptable alternative to no translation, and the direct (not post-edited) content is given to the user. Friends chatting via in-

NAACL-HLT 2012 Workshop on Predicting and Improving Text Readability for target reader populations (PITR 2012)., pages 1–7,
Montréal, Canada, June 7, 2012. ©2012 Association for Computational Linguistics

stant messaging tools or reading foreign-language e-mail mainly want to understand roughly what is being said. When a Marine is out patrolling and needs to interact with the local inhabitants to get information, it is “far better to have a machine [translation] than to not have anything” (Gallafent, 2011). For such purposes, automated translation can provide a “gist” of the meaning of the original message as long as it is comprehensible. In such situations, errors that affect comprehension trump those that occur frequently and should receive a greater focus in efforts to improve output quality.

Recently, companies have begun customizing translation engines for use in specific environments. IBM and Lionbridge’s GeoFluent (http://en-us.lionbridge.com/GeoFluent/GeoFluent.htm) uses customization to improve translation output for online chatting and other situations where post-editing is not feasible. TranSys (http://www.multicorpora.com/en/products/product-options-and-add-ons/multitrans-prism-transys/) from Mutlicorpora and Systran also uses customization to deliver translations ready for immediate distribution or for human post-editing. Knowing the major factors for creating understandable text can play a role in perfecting such systems.

Research has not settled on a single methodology for classifying translation errors. Two of the five categories proposed by Vilar et al. (2006), missing words and word order, are the focus of this project. Missing word errors fall into two categories, those essential to the meaning of the sentence and those only necessary for grammatical correctness. Only the first of these is addressed here. Likewise, there is a distinction between word- or phrase-based reordering. The results of the experiment presented in this paper are concerned only with the latter.

The present research seeks to determine the impact of specific error types on comprehension. We contend that research efforts should focus on those errors resulting in misinterpretation, not just on those that occur most often. This project therefore focuses on the use of linguistic parameters, including omissions and changes in word order, to determine the effect on comprehensibility of machine translations at the sentence and paragraph level.

2. Methodology

The first step in this research was determining the linguistic parameters to be investigated. Nine sentence types exhibiting the following characteristics were selected:

- Deleted verb
- Deleted adjective
- Deleted noun
- Deleted pronoun
- Modified prepositions *in, on, at* to an alternate one (e.g. *in* → *at*)
- Modified word order to SOV (Subject, Object, Verb)
- Modified word order to VOS
- Modified word order to VSO
- Retained SVO word order (control).

The one additional parameter, modifying a preposition, was added to the original list because it is a frequent error of translations into English (Takahaski, 1969).

The next step was to identify a means to test comprehension. Sachs (1967) contends that a sentence has been understood if it is represented in one’s memory in a form that preserves its meaning, but not necessarily its surface structure. Royer’s (Royer et al., 1987) Sentence Verification Technique (SVT) is a technique for measuring the comprehension of text paragraphs by determining if such a representation has been created. It has been used for three decades and been shown to be a reliable and valid technique for measuring comprehension in a wide variety of applications (Pichette et al., 2009).

In composing SVT tests, several paragraphs, each containing approximately 12 sentences, are chosen. For each of the sentences appearing in the original text, four test sentences are created. One is an exact copy of the original sentence and another, a paraphrase of that sentence. A “meaning change” test sentence is one in which a few words are changed in order to alter the meaning of the sentence. The fourth test sentence is a “distractor” which is consistent with the text of the original, but is not related in meaning to any sentence in the original passage (Royer et al., 1979).

We used a similar measure, a variation of the Meaning Identification Technique (MIT) (Marchant et al., 1988), a simpler version of the test that was developed out of the SVT and corrected for some of its shortfalls. Here, there are only two test sentence types presented, either a paraphrase of the original sentence or a "meaning change" sentence. In the description of the MIT technique for sentence creation, a paraphrase is created for each sentence in the original text and altering this paraphrase produces the "meaning change" sentence. In this experiment, the original sentence, not the paraphrase, was used to produce a sentence using many of the same words but with altered meaning.

In the test, readers are asked to read a passage, in our case a passage in which the linguistic parameters have been manipulated in a controlled fashion (see Section 3 (2)). Then with the text no longer visible, they are presented with a series of syntactically correct sentences shown one at a time in random order and asked to label them as being "old" or "new", relative to the passage they have just read (see Section 3 (3)). A sentence should be marked "old" if it has the same meaning as a sentence in the original paragraph and "new" otherwise. "New" sentences contain information that was absent from or contradictory to that in the original passage.

3. Experiment

The first requirement of the study was developing paragraphs to be used for the experiment. Eleven passages found on the WEB, many of which were GLOSS (http://gloss.dliflc.edu/search.aspx) online language lessons, were edited to consist of exactly nine sentences. These paragraphs, containing what will be referred to as the original sentences, served as the basis for building the passages to be read by the participants and for creating the sentences to be used in the test.

The next step was to apply the linguistic parameters under study to create the paragraphs to be read initially by the reader. One of the linguistic parameters listed above was randomly chosen and applied to alter a sentence within each paragraph, so that each paragraph contained exactly one of each of the parameter changes. However, pronouns and prepositions were not present in all sentences. When one of these was the parameter to be changed in a given sentence but was not present, adjustments had to be made in the original pairing of sentences with the other linguistic parameters. The changes were done as randomly as possible but in such a way that each paragraph still contained one of each type of parameter modification.

In sentences in which the change was an omission, the word to delete was chosen randomly from all those in the sentence having the same part of speech (POS). For sentences in which the preposition needed to be modified, the choice was randomly chosen from the two remaining alternatives as listed above in Section 2.

In creating the test sentences, the original sentences were again used. For each sentence within each paragraph, a committee of four, two of which were linguists, decided upon both a paraphrase and a meaning change sentence. Then, within each paragraph, the paraphrase of four randomly chosen sentences and the meaning change alternative for four others, also randomly picked, were selected. The ninth sentence randomly fell in either the paraphrase or meaning change category.

After reading the altered paragraph, the participant saw four or five sentences that were paraphrases of the original sentences and four or five sentences that were "meaning change" sentences, all in random order. The following is (1) an example of part of an original paragraph and (2) the same section linguistically altered. In (2), the alterations are specified in brackets after each sentence. Participants in the study did not, of course, see these identifiers. In (3), the sample comprehension questions posed after individuals read the linguistically altered passages are presented. In (3), the answers are provided in brackets after each sentence. Again, participants did not see the latter.

(1) World powers regard space explorations as the best strategy to enhance their status on the globe. Space projects with cutting-edge technologies not only serve as the best strategy to enhance their status on the globe. Korea must have strong policies to catch up with the space powers. The nation needs an overarching organization that manages all its space projects, similar to the National Aeronautics and Space Administration (NASA) and the European Space Agency (ESA). In addition, a national consensus must be formed if a massive budget is to be allocated with a long-term vision. Only under these

circumstances can the nation’s brightest minds unleash their talent in the field.

(2) World powers regard space explorations as the best strategy to enhance status on the globe. [PRO] Space projects with cutting-edge technologies not only as the driver of growth in future industries and technological development, but play a pivotal role in military strategies. [VERB] Korea strong policies space powers the to catch up with have must. [SOV] Needs an overarching organization that manages all its space projects, similar to the National Aeronautics and Space Administration (NASA) and the European Space Agency (ESA) the nation. [VOS] In addition, a national consensus must be formed if a massive budget is to be allocated with a vision. [ADJ] Can unleash, only under these circumstances, the nation’s brightest minds their talent in the field. [VSO]

(3) World powers regard space explorations as a viable, but expensive strategy to enhance their status among other countries. [NEW] Though space projects can be important for military purposes, the long-term costs can hamper a country’s development in other areas. [NEW] To perform on a par with the predominate players in space exploration, Korea must develop robust policies. [OLD] Managing all of the nation’s space projects will require a central organization, similar to the United States’ National Aeronautics and Space Administration (NASA). [OLD] Securing the necessary budget and allocating these funds in accordance with a long-term vision will require national consensus. [OLD] The nation’s brightest minds will be expected to work in the aerospace field. [NEW]

20 people volunteered as participants, consisting of 11 males and 9 females. All were over 25 years of age. All had at least some college, with 15 of the 20 holding advanced degrees. Only two did not list English as their native language. Of these, one originally spoke Polish, the other Farsi/Persian. Both had learned English by the age of 15 and considered themselves competent English speakers.

Participants were tested individually. Each participant was seated at a computer workstation equipped with a computer monitor, a keyboard and mouse. The display consisted of a series of screens displaying the passage, followed by the test sentences and response options.

At the start, participants completed two training passages. The paragraph read in the first had no linguistic alterations, while the second was representative of what the participants would see when doing the actual experiment. For both passages, after selecting a response option for a test sentence, the correct answer and reason for it was shown. There was an optional third training passage that no one elected to use.

During the experiment, participants were asked to read a passage. After finishing, with the text no longer in view, they were asked to rate a series of sentences as to whether they contained “old” or “new” information, relative to the information presented in the passage. Every participant viewed the same passages, but the order in which they were shown was randomized. Likewise, the sentences to be rated for a given passage were shown in varied order. Participants’ keyboard interactions were time-stamped and their choices digitally recorded using software specifically designed for this experiment.

After completing the test session, participants were asked to complete a short online questionnaire. This was used to obtain background information, such as age, educational level, and their reactions during the experiment.

4. Software

The interface for the experiment and final questionnaire were developed using QuestSys, a web-based survey system that is part of the custom web application framework, Cobbler, licensed by Knexus Research Corporation. Cobbler is written in Python and uses the web framework CherryPy and the database engine SQLite, both from the public domain.

5. Results

During the test, participants choose either “old” or “new” after reading each sentence. The number they correctly identified out of the total viewed for that condition in all paragraphs was determined. This score, the proportion correct (pc) for each condition, is as follows:

SVO	0.788 (control)
PREP	0.854
PRO	0.800
SOV	0.790
NOUN	0.769
VOS	0.769
VSO	0.757
ADJ	0.689
VERB	0.688

The average performance for SVT is about 75% correct. In a valid test, one at the appropriate level for the population being tested, overall group averages should not fall below 65% or above 85% (Royer et al., 1987). The results of this experiment were consistent with these expectations.

Because pc does not take into account a person's bias for answering yes or no, it is considered to be a poor measure of one's ability to recognize a stimulus. This is because the response chosen in a discrimination task is known to be a product of the evidence for the presence of the stimulus and the bias of the participant to choose one response over the other. Signal Detection Theory (SDT) is frequently used to factor out bias when evaluating the results of tasks in which a person distinguishes between two different responses to a stimulus (Macmillan and Creelman, 1991). It has been applied in areas such as lie detection (truth/lie), inspection (acceptable /unacceptable), information retrieval (relevant /irrelevant) and memory experiments (old/new) (Stanislaw and Todorov, 1999). In the latter, participants are shown a list of words and subsequently asked to indicate whether or not they remember seeing a particular word. This experiment was similar: users were asked, not about remembering a "word", but to determine if they had read a sentence having the same meaning.

The unbiased proportion correct, $p(c)_{max}$, a metric provided by SDT was used to generate unbiased figures from the biased ones. For yes-no situations, such as this experiment,
$p(c)_{max} = \Phi\ (d'/2)$, where $d' = z(H) - z(F)$, H being the hit rate and F, the false alarm rate.

Larger d' values indicate that a participant sees a clearer difference between the "old" and "new" data. The d' values near zero demonstrate chance performance. Perfect performance results in an infinite d' value. To avoid getting infinite results, any 0 or 1 values obtained for an individual user were converted to 1/(2N) and 1-1/(2N) (Macmillan and Creelman, 1991). Negative values, which usually indicate response confusion, were eliminated.

The results of Single Factor Anova of $p(c)_{max}$ are shown below (Table 1). Since the F value exceeds the F-crit, the null hypothesis that all treatments were essentially equal must be rejected at the 0.05 level of significance.

Dunnett's t statistic (Winer et al., 1991) (Table 2) was used to determine if there was a significant difference between any of the eight sentence variations and the control (SVO). The results are given below.

The critical value for a one-tailed 0.05 test: $t_{0.95}$ (9,167) ≈ 2.40. The results in Table 2 indicate that, in this experiment, adjective (ADJ) and verb deletions (VERB) had a significant effect on the understanding of short paragraphs. Other deletions and changes in word order were not shown to significantly alter comprehension.

6. Discussion

Though translation errors vary by language pair and direction, this research focused on two areas that cause problems in translations into English: word deletion and alterations in word order. It looked at how these errors affect the comprehension of sentences contained in short paragraphs.

In the research cited above (Vilar et al. (2006), Condon et al. (2010), and Popovic and Ney (2007; 2011)), wrong lexical choice caused the most errors, followed by missing words. For the GALE corpora for Chinese and Arabic translations into English, Popovic and Ney (2011) categorized missing words by POS classes. The POS that predominated varied by language but verbs were consistently at the top, adjectives near the bottom. Our study showed that both significantly affect the comprehension of a paragraph. Deleted nouns, prepositions and pronouns did contribute to the overall error rate, but none proved important to the reader in interpreting the text. Word order modifications were not a major cause of errors in the research above, nor did they appear to cause problems in our experiment. These results lead us to argue that in situations where there may be no or limited post-editing, reducing errors in verb translation should be a

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
SVO	19	15.75532	0.829227	0.01104
PREP	20	17.12685	0.856343	0.017096
PRO	20	16.17873	0.808936	0.013273
SOV	20	16.24132	0.812066	0.0135
NOUN	20	16.04449	0.802225	0.010088
VOS	20	15.9539	0.797695	0.011276
VSO	19	15.13767	0.796719	0.020403
ADJ	19	13.78976	0.725777	0.010103
VERB	19	13.88158	0.730609	0.015428

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.27809	8	0.034761	2.563014	0.011608	1.994219813
Within Groups	2.264963	167	0.013563			
Total	2.543053	175				

Table 1. Anova Single Factor of $p(c)_{max}$

PREP	PRO	SOV	NOUN	VOS	VSO	ADJ	VERB
0.736215	-0.55093	-0.46596	-0.73316	-0.85615	-0.86029	-2.7377	-2.60981

Table 2. Dunnett's t statistic

major focus in machine translation research. Though missing adjectives also significantly affected comprehension, a commitment of resources to solve an infrequently occurring problem may be unwarranted. It must be noted, however, that the data used in reporting error frequencies was limited to Chinese and Arabic. Further research is still required to determine the applicability of these findings for translating from other languages into English.

7. Conclusion

In this experiment, the paragraph appears to have provided enough context for the reader to correctly surmise most missing words and to understand an altered word order. The deletion of an adjective or verb, however, caused a significant decline in comprehensibility. In research by others dealing with error frequencies, verbs were frequently missing in English translation output, adjectives rarely.

This suggests that translation of verbs should receive more attention as research in machine translation continues, particularly in systems designed to produce "good enough" translations.

This was a small test and the part of speech chosen for elimination was not necessarily the most salient. It is unknown if a longer test, involving more passages, or passages in which the missing word was always significant, would have amplified these results.

This study used the Sentence Verification Technique in a novel way. Though constructing the test requires some expertise, it provides a way to test the comprehensibly of translation output without the use of experienced translators or reference translations produced by such translators.

Acknowledgements

The authors would like to thank Rachael Richardson for her research contributions. The US Naval Research Laboratory supported this research through funding from the US Office of Naval Research.

References

Condon, Sherri, Dan Parvaz, John Aberdeen, Christy Doran, Andrew Freeman, and Marwan Awad. (2010). English and Iraqi Arabic. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10),*Valletta, Malta, May 19-21.

Gallafent, Alex. (2011). Machine Translation for the Military. In *The World,* April 26, 2011.

Gamon, Michael, Anthony Aue, and Martine Smets. (2005). Sentence-level-MT evaluation without reference translations: Beyond language modeling. In *EAMT 2005 Conference Proceedings*, pp. 103-111, Budapest.

Kulesza, Alex and Stuart Shieber. (2004). A learning approach to improving sentence-level MT evaluation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore, MD, October 4–6.

Lavie, Alon, Kenji Sagae, and Shyamsundar Jayaraman. (2004). The Significance of Recall in Automatic Metrics for MT Evaluation. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004),* pp. 134–143. Washington, DC.

Macmillan, Neil and C. Douglas Creelman. (1991). Detection theory: A User's guide. Cambridge University Press, pp. 10 &125.

Marchant, Horace, James Royer and Barbara Greene. (1988). Superior reliability and validity for a new form of the Sentence Verification Technique for measuring comprehension. In *Educational and Psychological Measurement*, 48, pp. 827-834.

Pichette, François, Linda De Serres, and Marc Lafontaine. (2009). Measuring L2 reading comprehension ability using SVT tests. *Round Table Panels and Poster Presentation for the Language and Reading Comprehension for Immigrant Children (LARCIC),* May, 2009.

Popovic, Maja and Hermann Ney. (2007) Word Error Rates: Decomposition over POS Classes and Applications for Error Analysis. In *Proceeding of the Second Workshop on Statistical Machine Translation*, pp. 48-55, Prague.

Popovic, Maja and Hermann Ney. (2011) Towards Automatic Error Analysis of Machine Translation Output. *Computational Linguistics*, 37 (4): 657-688.

Przybocki, Mark, Kay Peterson, and Sébastien Bronsart. (2008). *Official results of the NIST 2008 "Metrics for MAchine TRanslation" Challenge (MetricsMATR08)*, http://nist.gov/speech/tests/metricsmatr/2008/results/

Royer, James, Barbara Greene, and Gale Sinatra. (1987). The Sentence Verification Technique: A practical procedure teachers can use to develop their own reading and listening comprehension tests. *Journal of Reading*, 30: 414-423.

Royer, James, Nicholas Hastings, and Colin Hook. (1979). A sentence verification technique for measuring reading comprehension. *Journal of Reading Behavior*, 11:355-363.

Sachs, Jacqueline. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics*, 1967(2): 437-442.

Schiaffino, Riccardo and Franco Zearo. (2005). Translation Quality Measurement in Practice. 46th ATA Conference, Seattle Technologies.

Stanislaw, Harold and Natasha Todorov. (1999). Calculation of Signal Detection Theory Measures, *Behavior Research Methods, Instruments, & Computers*, 31(1): 137-149.

Takahaski, George. (1969). Perceptions of space and function of certain English prepositions. *Language Learning*, 19: 217-234.

Vilar, David, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. (2006). Error Analysis of Statistical Machine Translation Output. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, pp. 697–702, Genoa, Italy

Winer, B., Donald Brown, and Kenneth Michels. (1991). Statistical Principles in Experimental Design. 3rd Edition. New York: McGraw–Hill, Inc. pp. 169-171.