ERROR CLASSIFICATION FOR MT EVALUATION

Mary A. Flanagan
CompuServe
1000 Massachusetts Ave.
Cambridge, MA 02138
email: mflanagan@csi.compuserve.com

Abstract

A classification system for errors in machine translation (MT) output is presented. Translation errors are assigned to categories to provide a systematic basis for comparing the translations produced by competing MT systems. The classification system is designed for use by potential MT users, rather than MT developers. The error categories can be ranked according to criteria which are important to the user, such as improvability and intelligibility. The results of CompuServe's use of the error classification methodology are presented as a case study.

1. Introduction

As machine translation gains increasing acceptance as a translation productivity tool, corporations, translation bureaus and translators are confronting the need to evaluate the quality of machine-generated translations. Translation quality is only one consideration in the decision to purchase MT software, but for most MT consumers it is both the most important and the most difficult to assess. Factors such as cost, speed, size of dictionaries and skill levels of staff, though important to the operational success of MT, do not insure that the end product, the translation, will meet the user's standards for translation quality. Thus, most potential users of MT find it necessary to measure the quality of output before choosing an MT system.

Machine translation quality can be difficult to quantify for a number of reasons:

- A text can have several different translations, all of which are correct.

- Defining the boundaries of errors in MT output is often difficult. Errors sometimes involve only single words, but more often involve phrases, discontinuous expressions, word order or relationships across sentence boundaries. Therefore, simply counting the number of wrong words in the translation is not meaningful.

- One error can lead to another. For example, if the part of speech of a word is identified incorrectly by the MT software, the entire analysis of the sentence may be affected, creating a chain of errors.

- The cause of errors in MT output is not always apparent. The evaluator usually does not have access to a trace of the software's tests and actions. Thus it can be difficult to identify what went wrong in the translation of a sentence.

This paper presents a system for classifying errors in MT output as a means of evaluating output quality. Classification of errors provides a basis for comparing translations produced by different MT systems and formalizes the process of error counting. Error classification can provide a descriptive framework that reveals relationships between errors. For example, if subject and verb do not agree in person or number, the error can be classified as one of agreement, rather than an incorrect noun inflection, or verb inflection or both. Error categorization can also help the evaluator to map the extent of the effect in chains of errors, allowing comparison among MT systems.

2. Error categories

Although some error categories may apply to many languages, a unique category set should be developed for each language pair to reflect the error types that actually occur. The categories used in CompuServe's evaluation were identified by observing the most frequent error types in machine translations of the Hewlett-Packard[1] test suite. For English-to-French MT the following categories were used:

[1] The Hewlett-Packard test suite is a comprehensive set of English sentence patterns developed by Dan Flickinger, Ivan Sag, John Nerbonne and Tom Wasow at Hewlett-Packard Labs. Flickinger, Dan, John Nerbonne, Ivan Sag and Tom Wasow. *Toward Evaluation of NLP Systems.* Hewlett Packard Laboratories, Palo Alto, CA. 1987.

Category	Description
Spelling	Misspelled word
Not Found Word	Word not in dictionary
Accent	Incorrect accent
Capitalization	Incorrect upper or lower case
Elision	Illegal elision or elision not made
Verb inflection	Incorrectly formed verb, or wrong tense
Noun inflection	Incorrectly formed noun
Other inflection	Incorrectly formed adjective or adverb
Rearrangement	Sentence elements ordered incorrectly
Category	Category error (e.g. noun vs. verb)
Pronoun	Wrong, absent or unneeded pronoun
Article	Absent or unneeded article
Preposition	Incorrect, absent or unneeded preposition
Negative	Negative particles not properly placed or absent
Conjunction	Failure to reconstruct parallel constituents after conjunction, or failure to identify boundaries of conjoined units
Agreement	Incorrect agreement between subject-verb, noun-adjective, past participle agreement with preceding direct object, etc.
Clause boundary	Failure to identify clause boundary, or clause boundary unnecessarily added
Word Selection	Word selection error (single word)
Expression	Incorrect translation of multi-word expression

For English to German evaluation, the category set was revised. The elision category was removed, since elision does not occur in German. The following categories were added:

Rel. Pronoun	The relative pronoun is absent or incorrect
Case	The case ending is incorrect
Punctuation	Punctuation is incorrect, absent or unneeded

A case error was assigned when an incorrect case ending was generated by the MT system. The relative pronoun error was assigned when the relative pronoun was ordered incorrectly, or when an implicit relative pronoun in English was not made explicit in German, but should have been. The punctuation error was assigned when the system failed to insert a comma between a noun phrase and a subordinate clause, as is required in German or when an unneeded comma was inserted.

In addition, the meaning and impact of error categories can differ with the language pair. Capitalization errors in English-to-French were due to dictionary errors and had little impact on the output, while in English-to-German, an uncapitalized noun could indicate a category error, and would undermine the translation quality substantially.

3. Category Assignment

It can be difficult to determine the best classification for an error in MT output. Because this method of evaluation is intended for MT users, rather than developers, the focus was placed on how errors are realized in the output, rather than the error's cause. Most users evaluating an MT system do not have access to the software's sequence of operations, and therefore cannot determine with certainty how an error was introduced into the translated sentence. Thus, in our evaluation, error classes were assigned according to their realization in the output text. For example:

(1E) We interviewed Abrams.
(1F) Nous avons interviewé l'Abrams.

the error of inserting the article 'le' before the proper noun 'Abrams' was classified as an Article error even though the likely cause for the error was that the word 'Abrams' was not found in the MT system's dictionary, and was treated as a common noun, requiring an article in French.

Many machine translated sentences contained multiple, linked errors, which occasionally pose questions for category assignment. Since the causes of errors could not be determined with certainty, our strategy was to count each of the errors individually. Thus the sentence:

(2E) Abrams bet Browne five dollars that Chiang hired Devito.
(2F) Pari d'Abrams Browne cinq dollars que Chiang engagé Devito.

was assigned two category errors, for analysis of 'bet' as a noun and 'hired' as an adjective, and a rearrangement error, even though the rearrangement error was probably a result of the misanalysis of 'bet'.

4. Error Ranking

The error categories can be ranked to facilitate comparison among systems. Ranking error categories has two primary benefits:

A) Error rankings can be used to resolve uncertain category assignments. The user can consistently resolve errors upward or downward in the error hierarchy. If errors are resolved upward, the best case analysis is made. If errors are resolved downward, the worst case analysis is made.

B) Error rankings can help the user assess how well a system meets the user's priorities. For example, errors can be ranked according to improvability and intelligibility. Improvability is an important consideration when evaluating MT. Since purchasing an MT system is a considerable investment in both time and money, a change can be costly. Customized dictionaries built for one MT system are rarely transferable to others, and thus changing MT systems usually means that the often substantial effort of customization is lost. If most of the errors made by the software have low improvability, the system may have reached its limit for output quality, to be surpassed by other systems in a few years. The potential user should assess the system's errors to find a product which can be improved over time.

Errors may also be ranked according to their effect on the intelligibility of the translation. Assessment of translation intelligibility is of particular importance if the MT output is to be used without postediting, as in CompuServe's application.

5. Case Study— CompuServe's MT Evaluation for English-to-French systems

Error classification was used as the second phase of a two-part evaluation of English-to-French and English-to-German machine translation software at CompuServe. The classification scheme was applied to the translated output of the Hewlett-Packard test suite, a comprehensive set of English sentence patterns. The test suite sentences were numbered, and ill-formed sentences were deleted. The edited test suite consisted of 910 sentences. No words were added to the MT system's dictionaries in preparation for the test.

MT evaluation can be costly and time consuming. The error classification portion of CompuServe's evaluation required approximately 160 person/hours. In addition there were training and access costs for each system evaluated. The entire two-part evaluation involved ten English-to-French systems and required approximately 6 months to complete.

Three MT systems were evaluated. The three systems were chosen in the first phase of testing. This phase involved measurements of the intelligibility, accuracy and style of translations of CompuServe texts. The results of the first phase of evaluation are available in "Evaluating MT for Message Translation", Mary A. Flanagan, Proceedings of the 34th Annual Conference of the American Translators Association, 1993.

In CompuServe's evaluation, categories of errors were grouped into classes to assess improvability and intelligibility. For the improvability measure, Class 1 errors were most improvable, and Class 3 errors were least improvable. For the intelligibility measure, Class 1 errors had the least impact on intelligibility, and Class 3 errors had the greatest effect.

Measure: improvability

Class 1
spelling
not-found-word
accent
capitalization
expression

Class 2
verb inflection
noun inflection
other inflection
pronoun
preposition
article
agreement
elision

Class 3
category
rearrangement
conjunction
clause boundary
negative
word selection

Measure: intelligibility

Class 1
elision
accent
capitalization
spelling
article

Class 2
not-found-word
verb inflection
noun inflection
other inflection
pronoun
preposition
agreement
negative clause
boundary

Class 3
expression
category
rearrangement
word selection
conjunction

Of particular importance for improvability are category, rearrangement, conjunction and clause errors, since these reflect the strength of the system's sentence analysis. Word selection, or the ability to identify the correct sense of a word based on context and the corresponding translation, may also have low improvability in systems which use little semantic disambiguation. Improvability is least concerned with not-found-word, spelling, accent, capitalization, and expression, since these errors can be readily resolved by dictionary additions.

Intelligibility was affected most by expression, word selection, category, conjunction and rearrangement errors, and was least affected by elision, accent and capitalization, spelling and article errors.

The results of the error classification exercise for English-to-French systems are presented at the end of this paper.

6. Conclusions

Evaluation of MT quality is necessarily a subjective process because it involves human judgments. Classification of errors can allow these judgments to be made in a more consistent and systematic manner. The error classification framework presented offers two benefits; it is flexible, allowing for the deletion or addition of categories, and the ranking categories according to the user's priorities; and it is simple in design. Error classification was a central part of CompuServe's evaluation of MT systems, and may be adaptable to other organizations considering deployment of an MT system.

References

1. Flickinger, Dan, John Nerbonne, Ivan Sag and Tom Wasow. *Toward Evaluation of NLP Systems.* Hewlett Packard Laboratories, Palo Alto, CA. 1987.
2. Hutchins, W. John & Harold L. Somers. *An Introduction to Machine Translation.* London: Academic Press. 1992.
3. Quirk, Randolph and Sidney Greenbaum. *A Concise Grammar of Contemporary English.* New York: Harcourt Brace Jovanovich. 1973.
4. Rolling, Loll. "EC Evaluation Activities". Presentation at the MT Evaluation Workshop. San Diego, November 2-3,1992.
5. Van Slype, Georges. *Critical study of methods for evaluating the quality of machine translation: Final report.* Prepared by Bureau Marcel van Dijk for the Commission of the European Communities Directorate General Scientific and Technical Information and Information Management. Brussels. 1979.

Appendix. Error classification Results, English-to-French systems

	Spell	NFW	Ace	Caps	Bis	Vinfl	Ninfl	Oinfl	Rear	Cat	Pron	Art	Prep	Neg	Con
Sysl	0	4	2	59	7	57	7	4	166	98	124	220	116	33	71
Sys2	7	6	1	2	3	38	1	1	118	85	35	7	76	21	42
Sys3	18	5	0	0	5	44	0	1	123	91	45	386	79	30	41

Key:

Spell	Misspelled word
NFW	Word not in dictionary
Ace	Incorrect accent
Cap	Incorrect upper or lower case
Elis	Illegal elision or elision not made
Vinfl	Incorrectly formed verb, or wrong tense
Ninfl	Incorrectly formed noun
Oinfl	Incorrectly formed adjective or adverb
Rearr	Sentence elements ordered incorrectly
Cat	Category error (e.g. noun vs. verb)
Pron	Wrong, absent or unneeded pronoun
Art	Absent or unneeded article
Prep	Incorrect, absent or unneeded preposition
Neg	Negative particles not properly placed or absent
Conj	Failure to reconstruct parallel constituents after conjunction, or failure to identify boundaries of conjoined units
Agr	Incorrect agreement between subject-verb, noun-adjective, past participle agreement with preceding direct object, etc.
Clse	Failure to identify clause boundary, or clause boundary unnecessarily added
WS	Word selection error (single word)
Expr	Incorrect translation of multi-word expression