Measuring User Acceptability of Machine Translations to Diagnose System Errors: An Experience Report

Bowen Hui
Department of Computer Science
University of Toronto
Canada
bowen@cs.utoronto.ca

Abstract

Conventional ways of measuring machine translation quality compares the accuracy of system output without clearly specifying what "accuracy" entails. Many current evaluation methods suffer from requiring too much time commitment from expert human evaluators. Moreover, these methods do not give direct feedback on user acceptability of the system, and do not hint on areas of focus for researchers or developers. In this work, we explore an output inspection method that measures user acceptance and pokes at system errors so that developers and researchers can walk away knowing what was acceptable and what to improve on. The evaluation framework for machine translation is described and experimental results for two systems are presented. The results of the experiments are very encouraging. We provide a discussion on identifying important translation quality factors for users, a pilot study of running this evaluation in the text summarization domain, and ideas on how to use the gathered data to create user profiles.

1 Introduction

Many researchers have criticized and proposed evaluation techniques for natural language (NL) systems (Sparck-Jones, 1996; Dorr et al., 1999), and especially for areas such as machine translation (MT) where there is no single *correct* answer for a given text. Thus, conventional ways of measuring precision and recall become misleading and uninformative for the untrained consumer or average user of MT. In fact, "so-called evaluations of MT technology ... [give] claims of upwards of 90% accuracy for systems, without a clear specification of what 'accuracy' entails" (Miller, 2000). Moreover, most existing evaluations suffer from two major deficiencies: they do not measure user acceptance of translation quality and do not "provide the slightest hint about the ease with which the system can be extended or modified" (King and Falkedal, 1990). Possible audiences who are interested in system evaluation outcomes include users, developers, and managers; however, current methods tend to concentrate on getting results for developers. Most existing MT evaluations focus on gathering fine-grained results that either require too much time or are cognitively overwhelming and labour-intensive for non-expert translators to complete. For example, some methods ask evaluators to identify and correct errors of many translation passages, some ask evaluators to rank translation quality based on finely differentiated criteria, and some assess system performance indirectly via evaluators' intelligence (Carroll, 1966; White et al., 1994; Bohan et al., 2000; Hovy, 1999). From her critical account of MT evaluations (King, 1997), King proposes that researchers focus on developing methods that allow users and developers understand the quality of an MT system and allow developers to relate the evaluation results to fixing system errors. We believe these are exceedingly important factors in research technologies, and proposes to approach MT evaluation that attempts to address these issues.

In this work, we explore an inspection method, called the *heuristic evaluation*, that measures user acceptance by implicitly asking users to diagnose system misbehaviour. The evaluation method is presented to the user as a free-trial of a system and survey – a concept that non-experts are already familiar with. Also, the evaluation groups the results in terms of system functionality, so that the results can be used quickly by developers to fix the problems. This method is an attempt to directly address the need of assessing user acceptance of NL systems and to provide useful development directions for researchers at the same time. Adapting this framework for MT is described in Sections 2 and 3 and experimental results for two systems are presented in Section 4. The results suggest that heuristic evaluation has clear advantages over existing NL evaluations and is worth investigating further. A discussion on identifying translation quality factors that are important to users is provided in Section 5. We believe that this framework can be generalized to other NL domains, and demonstrate this with a pilot study for a text summarization (TS) system in Section 6.

An interesting way of using the evaluation data is to create user profiles. By collecting data in our evaluation scheme, we have, in essence, elicited a database of user preferences. Also, the results are grouped in terms of system functionality. Therefore, by comparing user acceptance among multiple systems, what we get is what people like most about each system. Section 7 outlines ways of exploiting both kinds of data via cluster analysis.

2 Properties of the Heuristic Evaluation Method

Human-computer interaction (HCI) has long advocated for user evaluations of computer systems. Among the different methods that have developed in this field, one that is particularly notable is called the *heuristic evaluation* (Nielsen, 1993). In brief, heuristic evaluation requires several evaluators to "walkthrough" the system under real-use scenarios while performing a meaningful task so that they can get a good sense of the system design and responses to various user actions. Thereafter, the evaluator comments on the system based on a set of design principles that have been generally accepted as standard criteria in interface research. The outcome of the evaluation is a list of interface problems identified by potential users. This method has the benefit of easy experimental set-up, low requirements on user's experience or time, and few number of users as evaluators. In fact, Nielsen empirically shows that 5 evaluators typically identifies 75% of the system problems (see Figure 1, taken from (Nielsen, 1993)). This method is especially effective in carrying out quick evaluations and getting iterative feedback for developers.

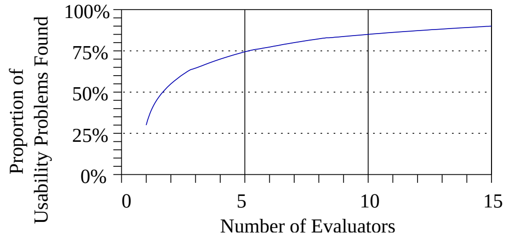

Figure 1: Usability Problems Found By Heuristic Evaluation As A Function Of Evaluators

The obstacle with directly applying an HCI evaluation method to NL is that it has been designed for assessing user interfaces, not system output. The trick therefore is to tailor the heuristic evaluation so that it allows users to evaluate the quality of output rather than the interface. In the following section, we will explain the steps we took to tailor the method towards evaluating MT systems.

3 Evaluating MT Systems

As mentioned in Section 2, the framework is already defined but the details need to be refined. The three steps involved in realizing the method are determining the principles, the task, and the test material.

3.1 Identifying Principles

In order to facilitate the inspection of system problems, the design of principles should group important characteristics that are relevant to particular system modules together. Different researchers have proposed using various metrics (EAGLES, 1994; Bohan et al., 2000; Nyberg et al., 1994). We have concentrated ours on characteristics that concern the quality of translation output. Thus, most of our metrics came from the "output characteristics" subset of the overall effort to standardize evaluation criteria by AMTA (Reeder and Hovy, 2000). (The full list is listed in Appendix A.) In an attempt to simplify terminology for non-expert evaluators, we reduced to using 8 principles:

1. **Word Choice:** Individual words are translated correctly in its context. Special terminology is translated with the same level of difficulty. Words are meaningful and consistent in the provided context.
2. **Syntax:** Translated sentences are grammatical. The structure of sentences may differ from the original if changing the structure can effectively deliver the style of the original text.
3. **Style:** Each paragraph maintained a similar style (e.g., tone, mood, level of formality) than that of its original text. Readers should be able to read the translated sample only and get the same reaction towards the message that the author was trying to deliver.
4. **Coherence:** Each sentence is meaningful on its own. The role of a sentence with respect to the entire text can be identified.
5. **Comprehensibility:** All information should be grammatical and coherent within each sentence as well as each paragraph. Idioms and dialogues preserve their meaning and mood in the translations. Words, phrases, or idioms that could not be translated or that were not translated correctly do not create distortion to the overall meaning of text. Overall, the text is clear and readable.
6. **Consistency:** Information should be expressed clearly in words, phrases, and concepts consistent with those in the original document. Readers should not have to wonder whether different pronouns, words, situations, or actions mean the same thing. The amount of information in

the original text is reproduced in the translations.

7. **Fit For Audience:** The information and the style of presentation fit the intended audience. The same group of audience (e.g., children, politicians) intended in the original language is also the audience of the translated language. Cultural or linguistic differences are therefore also "translated".

8. **Accountability:** The kinds and frequency of errors (punctuation, words, syntax, style) are tolerable. Readers are generally satisfied with the translation and are likely to recommend the system to other users.

In this way, a principle receiving an unfavourable score and comments indicates specific modules for further development. We intend for these principles to be refined through iterations of applying this method in MT. For example, special-purpose MT systems may include a principle that addresses the quality of the translation of domain-specific terminology. In Section 5, we reflect on our evaluators' experience using this method, and suggest that there is a smaller set of translation criteria that are relevant to users.

3.2 Defining the Task

To help evaluators examine translations, they were asked to answer the following questions:

1. What is the genre exhibited in the writing (e.g., story, advertisement, instructions, diary entry, job posting, etc.)?
2. What is the purpose of this writing (intended by the author)?
3. Suggest some intended audience for this writing (e.g., children, students, athletes, computer users, photographers, etc.).
4. List the entities (people or objects) involved or discussed by the author.
5. What would be a coherent sentence that follows the excerpt, based on what you have read?

In essence, well-known tasks, such as the Shannon Game and the Classification Game (Hovy and Marcu, 1998; Teufel, 2001; Hirao et al., 2001), can be used as well, so long as the task allows evaluators to go through enough of the output to comment on each principle afterwards. This convenience holds for designing tasks for other NL systems.

Experimenters may decide to ask the evaluators to provide answers to the questions in the task, if they would like to measure the accuracy of the answers and the time taken for them to complete the task. However, in a heuristic evaluation, the focus is on the principles – which is the novel aspect that current NL evaluations lack. Therefore, we focus our results and discussion on the use of these principles only.

3.3 Selecting Test Material

In total, we selected passages that exhibit a wide range of styles and language usage from four genres:

1. comic descriptions – humour, irony, satire
2. fairy tale – narrative, figurative, dialogue
3. medicinal instructions – technical, special terms
4. movie review – colloquial, dialogue, slang

For each genre, we took 2 samples, labeled A and B. Table 1 shows the number of words (w) and sentences (s) for each sample.

	1A	**1B**	**2A**	**2B**	**3A**	**3B**	**4A**	**4B**
w	146	180	326	342	87	90	245	242
s	8	9	19	15	13	9	17	13

Table 1: Number of Words/Sentences Per Sample

4 Experiments and Results

The systems we evaluated were Babel Fish[1] (hereafter referred to as System 1) and Pratique[2] (hereafter, System 2) focusing on translations from English to French. A total of 28[3] participants were divided into 4 groups evenly – group 1 evaluated samples 1A,2A,3A,4A for System 1; group 2 evaluated 1B,2B,3B,4B for System 1; group 3 evaluated 1A,2A,3A,4A for System 2; and group 4 evaluated 1B,2B,3B,4B for System 2. Thus, each system had a total of 11,606 words or 721 sentences as evaluation material.

First, participants were given a description of the principles and question-answering task (cf. Sections 3.1 and 3.2). Then, they were asked to complete the task for 4 French samples. For each sample, they were asked to rate each principle on a scale of 5 and provide comments while having access to both the English and French texts. Participants spent between 30 minutes to 2 hours to complete the evaluation.

Figure 2 shows the total acceptability score for the two systems with respect to each principle. The score is calculated by $\sum_{i=1}^{n} s_i$, where $n = 14$ is the number of evaluators per system and s_i is the

[1] Available at `http://babelfish.altavista.com/`.

[2] Available at `http://chaines.free.fr/traduction/`.

[3] Although Nielsen demonstrated that 5 evaluators is enough for identifying 75% of system problems (cf. Figure 1), we decided to use a large number of evaluators because we want to be able to achieve statistical convergence on our results.

score given by an individual evaluator. The maximum score is 560 (4 groups x 7 participants per group x 4 samples each x 5 points). Figure 3 shows the average acceptability score under the same perspective. These results indicate that principles 1, 2, 5, and 8 are especially low. Indeed, many evaluators made lists of corrections to words and phrases and commented: "Big problems in conjugating the verbs", "Important words are translated so wrong that the point is completely missed", "no agreement". Consequently, consistency and accountability suffered (e.g., accompanying comments such as "because of word translations", "syntax problems have to be overcome first to ensure easy comprehensibility").

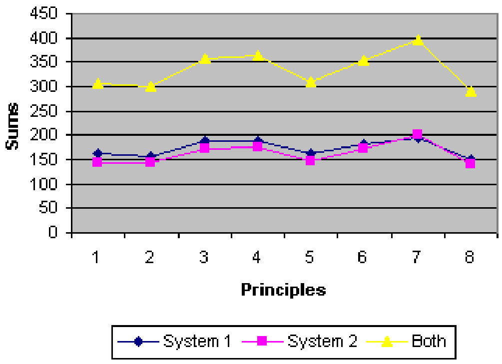

Figure 2: MT Evaluation Results

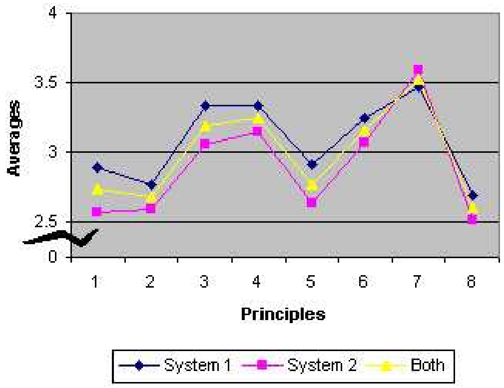

Figure 3: Average System Scores: System 1 (blue), System 2 (red), Both (yellow).

With the exception of Principle 7, Fit for Audience, System 1 scored higher than System 2. In fact, calculating the statistical significance of the average of all the scores showed that System 1 is more acceptable than System 2 ($p < 0.05$) and that the four genres contributed to this significance ($p < 0.001$). We wanted to also compare the two systems when genre is conditioned. We found that System 1 is significantly better than System 2 ($p < 0.01$) in translating the fairy tale, medicinal instructions, and movie review genres.

5 Discussion

Finding fluent speakers of French and English to volunteer 1 to 2 hours of time for this experiment was not difficult. In most cases, we imagine that if English were one of the languages under investigation, it would not be very hard to find bilingual speakers. Due to the nature of these experiments, evaluators need not be translation experts nor domain experts of the test materials. Although the quantitative results obtained from these experiments are objective, King (King, 1997) suggests that using a large sample population may be amenable. In this experiment, we used 28 evaluators and showed statistical convergence on their agreements with the acceptability of a system.

Conducting this experiment proved that the procedure and analysis was not demanding on the experimenter at all. Once ready, the procedure can be repeated by different evaluators without changing the set-up. Depending on the scale of the tests, the experimenter may wish to add more test materials, which is also easy to modify. However, one evaluator found the experiment very overwhelming in terms of the terminology used. When asked for further feedback on the experimental procedure, about $\frac{1}{4}$ of the evaluators voiced concern of having redundant principles or encountering difficulty in attributing an error to a particular principle. When assigning scores and determining which principles explained a mistake in the translation, evaluators found that they only had an intuition as to why a translation is bad. Thus, some principles that they considered to be similar were treated the same and received similar scores. This problem has been documented before although we did not expected it to surface with as few as 8 metrics. To address this issue, common factor analysis was used as a first step to minimizing the number of factors that play a role in explaining the findings. Usually, the number of factors is determined by a combination of several criteria (Loehlin, 1992). Based on the Kaiser criterion, scree test, and interpretability, factor analysis suggests that the acceptability data can be explained by 4 factors underlying 8 principles. Further work to identify the relevant factors and their inter-correlations needs to be done.

6 Second Test-bed: Summarization

To test the generalizability of the evaluation framework, we defined the approach for TS systems. A pilot experiment was conducted using a system that automatically extracts information from specialized text documents and presents the information in point-form organized by a graphical hierarchy of concepts (Hui and Yu, 2002). The main objective of this system is to allow users to learn and find information in documents quickly and easily. Next, we turn to the design of principles, task, and test materials, following the same format as in Section 3.

6.1 Principles

After amalgamating criteria for the "worthiness" of text summaries (Sparck-Jones and Galliers, 1996; Sparck-Jones, 1996; Hovy and Marcu, 1998), the following set of principles were chosen.

1. **Conciseness:** Components should not contain information that is irrelevant or redundant. Every extra unit of information competes with the relevant units of information and diminishes their relative visibility. All information should appear in a natural and logical order.
2. **Retention:** Information retained in the system output should be representative of the key concepts and main points made in the original document. Are the major objectives of the paper captured in the summary? What about the major steps in the proposed solution and the results?
3. **Coherence:** All information should be coherent within each component as well as the overall summary. Sentences need not be perfectly grammatical, but each point should make sense in its context.
4. **Consistency:** Each component should be expressed clearly in words, phrases, and concepts consistent with those in the original document. Users should not have to wonder whether different words, situations, or actions mean the same thing.
5. **Informativeness:** Information should be presented in a useful and easily accessible way. Some interface issues may be influential here as well. Irrelevant information should be omitted and words should not clutter the display of the information.
6. **Comprehensibility:** Each point of information should be easy to understand. Users should not have to look up related information in another part of the system in order to understand a particular component.
7. **Fit For Audience:** The information and the style of presentation fits for the intended audience. Audience may vary in their experience with domain knowledge. Access to different kinds of information should be easy and clear. The ability to show, modify, and hide information should be made obvious to the users.
8. **Fit For Purpose:** The information and the style of presentation fits for the intended task (e.g., question-answering) or purpose (fast learning, easy to read).

6.2 Task

In this experiment, we designed a question-answering task that is modeled after the job of a conference referee. This way, evaluators acted as reviewers using only the system output. The questions used follow:

1. What is the problem addressed by this work? Does it describe why the problem is significant?
2. Does the work present the approach taken to solve the problem targeted? Is the design or implementation of a system described in terms of key ideas of the approach?
3. What are the contributions of this work? Are the benefits and limitations clear? Are the results positive or negative?

6.3 Test Materials

Excerpts were extracted as input to the system. In particular, abstract and summary sections from 3 patents (P1,P2,P3) and abstract, introduction, and conclusion sections from 3 scientific articles (S1,S2,S3) were used. The number of words (w) and sentences (s) for each excerpt are shown in Table 2.

	P1	**S1**	**P2**	**S2**	**P3**	**S3**
w	828	257	1041	688	512	709
s	27	16	27	34	23	22

Table 2: Number of Words/Sentences Per Sample

Excerpts P1 and S1, totaling to 3255 words, or 129 sentences, were evaluated by 3 participants and the others, totaling 11,800 words, or 424 sentences, were evaluated by 4 participants. Thus, a total of 7 participants evaluated the system output which yields a total of 15,055 words, or 553 sentences.

6.4 Results

Seven participants were presented with the description of the principles and the question-answering task (cf. Sections 6.1 and 6.2). Then, they were asked to complete the task for each of the summarized samples. For each sample, they were asked to rate each principle on a scale of 5 and provide comments while having access to the original document and the graphical summary. Each participant spent about 1 to 1.5 hour for the entire session. Figure 4 shows the sum of all the scores (9 documents x 5 points = 45 maximum) for each principle in terms of scientific articles and patents.

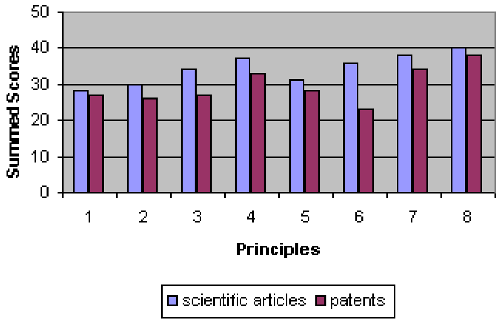

Figure 4: TS Evaluation Results

Figure 4 shows that scientific articles were generally more accepted than patents. Principle 6, comprehensibility, has the lowest score which indicates that the inaccuracy of extracting sentences into the right concepts dampened the acceptance of this criteria. The low score on Principle 1 (conciseness) indicates that more heuristics should be incorporated so that the resulting text is more condensed.

Due to the small number of observations in this experiment, statistical significance was not found between the two document types. However, when the same data is duplicated 4 times to increase data size (but preserving the pattern of the results), scientific articles were statistically significantly more acceptable than patents ($p < 0.001$). Furthermore, evidence from the Kaiser criterion, scree test, and interpretability suggest that the acceptability data can be explained by 5 factors underlying 8 principles. Further work is necessary to identify these factors and how they interplay.

7 Conclusions and Extensions

We advocated for the need of measuring usability of NL system output to assess user satisfaction with the translation quality. We adapted a heuristic evaluation method for comparative MT evaluation and extended the framework to evaluate text summarizers. Through experimenting with 28 human evaluators for MT and 7 for TS, our experience point out many features that this method effectively assesses user acceptance of a system; compares user preferences of multiple systems; is not time consuming for the experimenter; requires about 1 to 2 hours of an evaluator's time; is not cognitively overwhelming for evaluators; that the quantitative data analysis can be automated; the qualitative analysis gives insight to system problems for developers; a summary of results can be used to generate survey results for consumers; and changing the principles and task of the framework according to application works well for NL systems that do not have a gold standard. Comparing with the difficulties faced by existing evaluations (Sparck-Jones and Galliers, 1996; King, 1997; Reeder and Hovy, 2000; Hovy and Marcu, 1998; Jing et al., 1998) as mentioned in Section 1, the effectiveness of this method is remarkably encouraging.

Establishing translation principles. Currently, as an extension to the discussion in Section 5, we are working on using methods such as confirmatory factor analysis and principle components analysis to identify the core set of evaluation principles that are important to human users.

User profiling. We are very interested in using the collected data in these experiments as preferences of translational criteria elicited from users. Discussion on user demographics (e.g. native vs. non-native French speakers) was not provided because we wanted to focus on how the data may be explained by individual translational preferences, not by personal background as is usually done. Although not designed in our experiments, we are currently collecting data where users rank the evaluation principles to reflect which criteria are more important to them. (Doing so also gives an indication of errors that are more forgivable.) To indicate an emphasis on certain principles, weights that designate "importance" can be assigned to them. In particular, for k principles, weights w_1, w_2, ..., w_k must satisfy $0 < w_j < 1$ and $\sum_{j=1}^{k} w_j = 1$. Then, to determine a single, overall score for a system, we take:

$$\sum_{j=1}^{k} \frac{\sum_{i=1}^{n} w_j s_i}{n}$$

Therefore, each user's criteria preference can be represented by a weight vector. Cluster analysis can create meaningful user groups based on user preferences over language task rather than demographics and software can be customized for users who share language preferences for an MT application.

Component selection. A syntactic module may score well in one MT system while a stylistic module may do better in another. Ideally, one would want to pick and choose the individual components that perform well in different systems. Since the evaluation results are grouped in terms of system functionality, by comparing user acceptance across multiple systems allows us to identify which module from which system works well. From there, we can select the "best" (most widely accepted) components from different MT systems and combine them into one abstract machine. Obviously there will be incompatibility issues, however, this analysis will provide interesting insights on analyzing current MT methodology.

Appendix A: Output Characteristics

Quality of translation: (a) quality of text as a whole – acceptability to the end user, clarity, coherence, comprehensibility, consistency, fidelity, informativeness, readability, style, terminology, utility of output; (b) quality of each individual sentence – morphology, syntax (sentence and phrase structure). Errors: (a) diction errors; (b) punctuation errors; (c) syntax errors; (d) stylistic errors.

References

N. Bohan, E. Breidt, and M. Volk. 2000. Evaluating translation quality as input to product development. In *Proceedings of 2nd International Conference on Language Resources and Evaluation*, Athens, Greece.

J.B. Carroll. 1966. An experiment in evaluating the quality of translations. *Mechanical Translation*, 9(3-4):55–66.

B. Dorr, P.W. Jordan, and J.W. Benoit. 1999. A Survey of Current Research in Machine Translation. *Advances in Computers, M. Zelkowitz (ed)*, 49:1–68.

EAGLES, 1994. *Interim Report.* Obtainable from Center for Language Technology, Njalsgade 80, DK 2300 Copenhagen.

T. Hirao, Y. Sasaki, and H. Isozaki. 2001. An Extrinsic Evaluation for Question-Biased Text Summarization on QA Tasks. In *NAACL Workshop on Automatic Summarization*, pages 61–68.

E. Hovy and D. Marcu, 1998. *Automated Text Summarization: Tutorial Notes.* COLING-ACL'98, Montréal, Canada.

E. Hovy. 1999. Toward Finely Differentiated Evaluation Metrics for Machine Translation. In *EAGLES Workshop on Standards and Evaluation*, Pisa, Italy.

B. Hui and E. Yu. 2002. Extracting Conceptual Relationships from Specialized Documents. In *21st International Conference on Conceptual Modeling (ER 2002)*, To appear. Tampere, Finland.

H. Jing, R. Barzilay, K. McKeown, and M. Elhadad. 1998. Summarization Evaluation Methods: Experiments and Analysis. In *AAAI Intelligent Text Summarization Workshop*, pages 60–68.

M. King and K. Falkedal. 1990. Using test suites in evaluation of machine translation systems. In *Proceedings of the 19th Conference of COLING.*

M. King. 1997. Evaluating translation. In *C. Hauenschild & S. Heizmann (eds.), Machine Translation and Translation Theory.* Walter de Gruyter & Co.: Berlin.

J.C. Loehlin. 1992. *Latent Variable Models.* Erlbaum Associates, Hillsdale NJ.

K. Miller. 2000. *The Lexical Choice of Prepositions in Machine Translation.* Ph.D. thesis, Georgetown University, Maryland, USA.

J. Nielsen. 1993. *Usability Engineering.* Academic Press, Inc.

E.H. Nyberg, T. Mitamura, and J.G. Carbonnell. 1994. Evaluation Metrics for Knowledge-Based Machine Translation. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, pages 95–99, Kyoto, Japan.

F. Reeder and E. Hovy, 2000. *Workshop on Machine Translation Evaluation.* AMTA-00, October.

K. Sparck-Jones and J.R. Galliers. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review.* New York: Springer.

K. Sparck-Jones. 1996. Towards Better NLP System Evaluation. In *Proceedings of the Human Language Technology Workshop*, pages 102–107. ARPA.

S. Teufel. 2001. Task-Based Evaluation of Summary Quality: Describing Relationships Between Scientific Papers. In *NAACL Workshop on Automatic Summarization.*

J.S. White, T. O'Connell, and F.E. O'Mara. 1994. The ARPA MT evaluation methodologies: Evolution, lessons and further approaches. In *Technology partnerships for corssing the language barrier: Proceedings of the first conference of the Association for Machine Translation in the Americas*, pages 193–205, Columbia, USA.