



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`

Problems with NL Evaluations

- ◆ current evaluation methods:
 - time consuming and labour intensive
 - requires translation/linguistic experts
 - cognitively demanding tasks
 - too often influenced by evaluator's intelligence
 - inconsistent (unreliable) results
 - does not extend across systems or domains
- ◆ current accuracy metrics:
 - no single correct answer
 - misleading/uninformative for the untrained consumer or average user

Want direct feedback on user acceptance of translation quality and areas of focus for researchers or developers.

Heuristic Evaluation (Nielsen 1993)

- ◆ evaluator = potential user
- ◆ inspection method that walks through system by carrying out meaningful task in real-use scenario
- ◆ gives evaluator good sense of system capabilities
- ◆ evaluation centred around a set of principles
- ◆ resulting comments are a list of identified system problems
- ◆ benefits:
 - easy experimental set-up
 - low demand of evaluator experience or time
 - requires few evaluators (~5 to identify 75% of all system problems)

Adapting for NL Evaluation

- ◆ evaluation of system *output* quality, not interface
- ◆ in addition to commenting on principles, also want quantitative *score*
- ◆ for MT domain, we need:
 - principles
 - meaningful task
 - test materials
- ◆ second test-bed for generality in TS domain

MT Principles

- ◆ **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.
- ◆ **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.
- ◆ **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.
- ◆ **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.

MT Principles cont.

- ◆ **Coherence:** Each sentence is meaningful on its own. The role of a sentence with respect to the entire text can be identified.
- ◆ **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.
- ◆ **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".
- ◆ **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.

MT Task

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

MT Test Materials

- ◆ 4 themes:
 - comic descriptions – humour, irony, satire
 - fairy tale – narrative, figurative, dialogue
 - medicinal instructions – technical, special terms
 - movie review – colloquial, dialogue, slang

	1A	1B	2A	2B	3A	3B	4A	4B	Total
w	146	180	326	342	87	90	245	242	11,606
s	8	9	19	15	13	9	17	13	721

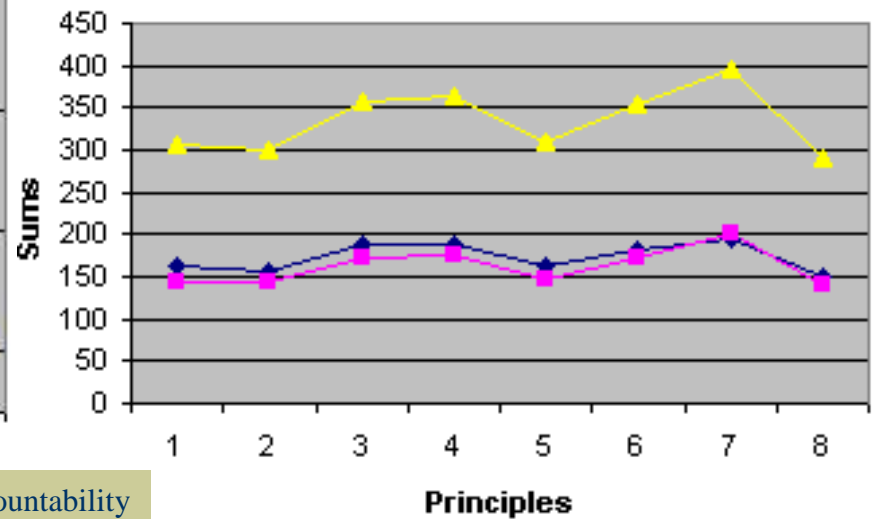
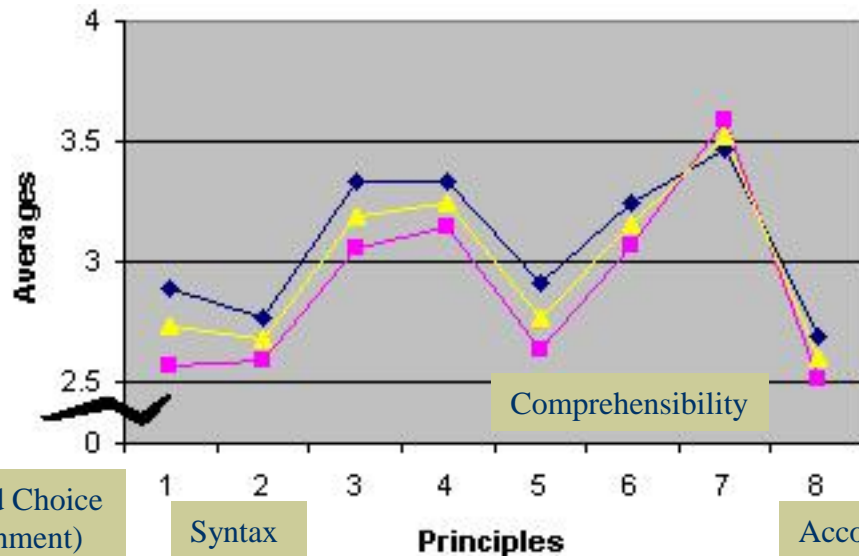
- ◆ 2 samples per theme (A,B)
- ◆ 4 groups x 7 participants

Experiments

- ◆ 2 systems: BabelFish, Pratique
- ◆ System 1:
 - group 1 evaluated samples 1A,2A,3A,4A
 - group 2 evaluated samples 1B,2B,3B,4B
- ◆ System 2:
 - group 3 evaluated samples 1A,2A,3A,4A
 - group 4 evaluated samples 1B,2B,3B,4B
- ◆ complete Q/A task for 4 French samples
- ◆ for each sample, rate each principle out of 5
- ◆ comments, with access to English and French texts

Results

- ◆ time ranged from 30 min – 2 hours
- ◆ System 1 more acceptable than System 2 ($p < 0.05$)
- ◆ themes contributed to this significance ($p < 0.001$)



Word Choice
(alignment)

Syntax
(rules)

Accountability

◆ System 1 ◆ System 2 ◆ Both

◆ System 1 ◆ System 2 ◆ Both

Results cont.

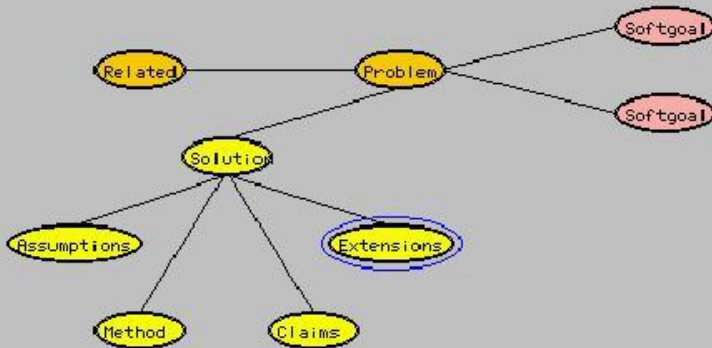
- ◆ most problematic: Principles 1,2,5,8
- ◆ “Big problems in conjugating the verbs”
- ◆ “Important words are translated so wrong that the point is completely missed”
- ◆ “no agreement”
- ◆ consequently, consistency and accountability suffered “because of word translations”
- ◆ “syntax problems have to be overcome first to ensure easy comprehensibility”
- ◆ several evaluators found some principles *redundant*
- ◆ more linguistically inclined evaluators found 8 principles *overwhelming*

Second Test-Bed: Text Summarization

- ◆ Treat document modeling system as basic summarizer

Title: Future Directions For Human-Computer Interaction

Graphical View



Extensions:

It is dangerous, but necessary, to dream about the future.
Dangerous because misguided dreams mislead designers, necessary because without vision navigation is difficult.
Without dreams we risk stagnation, and lose the chance to make a better world.
There is excitement in the user interface research and development community. ideas are emerging daily and the reduction to practice is rapid.
Beyond the technology, I strongly encourage vigorous discussion of the high-level goals and the dangers.
Discussing the future is an important part of the process of creating it.

TS Principles

- ◆ **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.
- ◆ **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?
- ◆ **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.
- ◆ **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.

TS Principles cont.

- ◆ **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.
- ◆ **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.
- ◆ **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.
- ◆ **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).

TS Task

- ◆ What is the problem addressed by this work? Does it describe why the problem is significant?
- ◆ 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?
- ◆ What are the contributions of this work? Are the benefits and limitations clear? Are the results positive or negative?

TS Test Materials

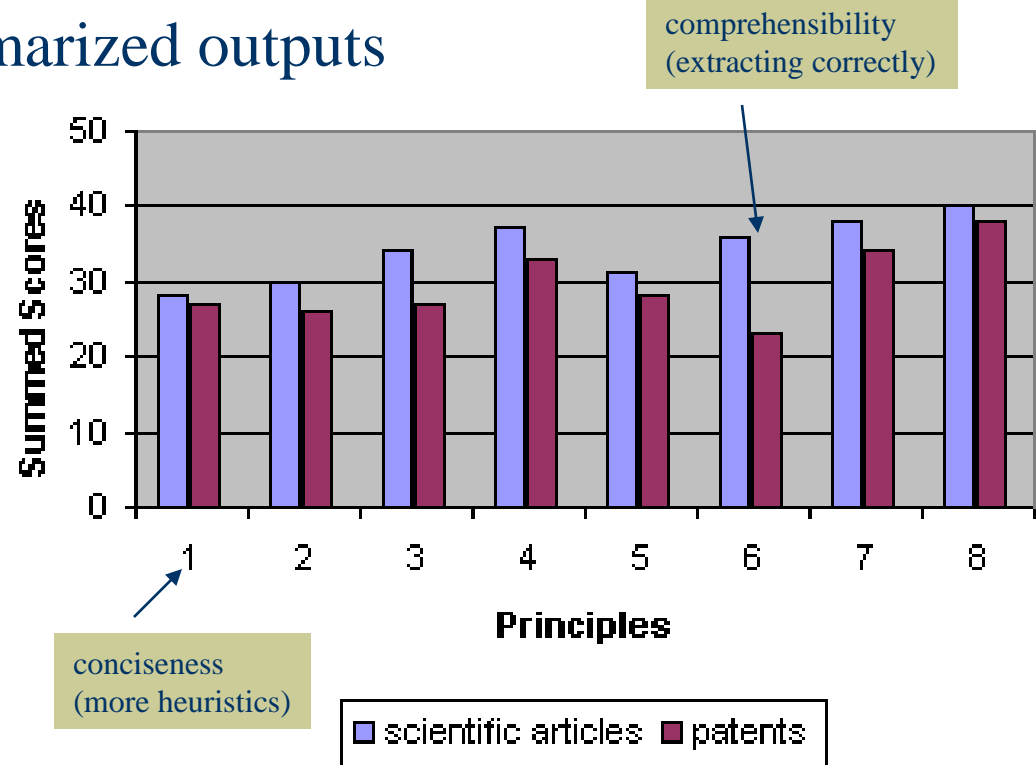
- ◆ 2 themes:
 - patents = abstract + summary
 - scientific articles = abstract + introduction + conclusion

	P1	S1	P2	S2	P3	S3	Total
w	828	257	1041	688	512	709	15,055
s	27	16	27	34	23	22	553

- ◆ 3 samples per theme (1,2,3)
- ◆ 7 participants

Experiments and Results

- ◆ group 1 (3 participants) evaluated P1,S1
- ◆ group 2 (4 participants) evaluated P2,P3,S2,S3
- ◆ complete task for summarized outputs
- ◆ for each output, rate each principle out of 5
- ◆ comments, access to original and summarized texts
- ◆ time spent between 1 hour – 1.5 hour



Summary

- ◆ need for measuring usability and getting at development problems of NL systems
- ◆ adapted heuristic evaluation
 - 1. comparative MT evaluations
 - 2. extended framework to TS
- ◆ MT experiment:
 - 2 systems, 4 themes, 2 samples per theme, 28 human evaluators
- ◆ TS experiment:
 - 2 themes, 3 samples per theme, 7 human evaluators
- ◆ recall difficulties often cited in literature
- ◆ heuristic evaluation method is remarkably encouraging

Advantages

- ◆ effectively assesses user acceptance of translation quality
- ◆ compares user preferences of multiple systems
- ◆ not time consuming for the experimenter
- ◆ requires about 1 to 2 hours of a non-expert evaluator's time
- ◆ not cognitively overwhelming for evaluators
- ◆ quantitative data analysis can be automated
- ◆ qualitative analysis gives insight to sys. problems for developers
- ◆ generates summary survey results for consumers
- ◆ easy evolution of principles and task according to application
- ◆ works well for NL systems that do not have a gold standard

Future Work

- ◆ Establish translation principles
 - through usage and standardization
- ◆ Component selection
 - combining the “best” components of different systems
- ◆ User profiling
 - collect data where users rank the evaluation principles to reflect which criteria are more important to them
 - this also gives an indication of errors that are more *forgiving*
 - create user groups based on criteria (vs. demographics)



Acknowledgements

Special thanks to all the 35 (un)paid participants!

Appendix A: Output Characteristics

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

References

- ◆ N. Bohan, E. Breidt, & M. Volk. Evaluating Translation Quality as Input to Product Development. Proceedings of 2nd International Conference on Language Resources and Evaluation, 2000.
- ◆ J.B. Carroll. An experiment in evaluating the quality of translations. Mechanical Translation, 9(3-4), 1966.
- ◆ B. Dorr, P.W. Jordan, & J.W. Benoit. A Survey of Current Research in Machine Translation. Advances in Computers, M. Zelkowitz (ed), 49, Academic Press, London, 1999.
- ◆ EAGLES: Interim Report. Obtainable from Center for Language Technology, Njalsgade 80, DK 2300 Copenhagen, 1994.
- ◆ T. Hirao, Y. Sasaki, & H. Isozaki. An Extrinsic Evaluation for Question-Biased Text Summarization on QA Tasks. NAACL Workshop on Automatic Summarization, 2001.
- ◆ E. Hovy & D. Marcu. Automated Text Summarization: Tutorial Notes. COLING-ACL'98, Montreal, Canada, 1998.
- ◆ E. Hovy. Toward Finely Differentiated Evaluation Metrics for Machine Translation. EAGLES Workshop on Standards and Evaluation, 1999.
- ◆ B. Hui & E. Yu. Extracting Conceptual Relationships from Specialized Documents. The 21st International Conference on Conceptual Modeling (ER 2002), 2002.
- ◆ H. Jing, R. Barzilay, K. McKeown, & M. Elhadad. Summarization Evaluation Methods: Experiments and Analysis. AAAI Intelligent Text Summarization Workshop, 1998.
- ◆ M. King & K. Falkedal. Using test suites in evaluation of machine translation systems. Proceedings of the 19th Conference of COLING, 1990.
- ◆ M. King. Evaluating translation. C. Hauenschild & S. Heizmann (eds.), Machine Translation and Translation Theory, Walter de Gruyter & Co.: Berlin, 1997.
- ◆ J.C. Loehlin. Latent Variable Models. Erlbaum Associates, Hillsdale NJ, 1992.

References cont.

- ◆ K. Miller. *The Lexical Choice of Prepositions in Machine Translation*. Georgetown University, Maryland, USA, 2000.
- ◆ J. Nielsen. *Usability Engineering*. Academic Press, Inc., 1993.
- ◆ E.H. Nyberg, T. Mitamura, & J.G. Carbonnell. *Evaluation Metrics for Knowledge-Based Machine Translation*. Proceedings of the 15th International Conference on Computational Linguistics (COLING'94), 1994.
- ◆ F. Reeder & E. Hovy. *Workshop on Machine Translation Evaluation (AMTA-00)*. 2000.
- ◆ K. Sparck-Jones & J.R. Galliers. *Evaluating Natural Language Processing Systems: An Analysis and Review*. New York: Springer, 1996.
- ◆ K. Sparck-Jones. *Towards Better NLP System Evaluation*. Proceedings of the Human Language Technology Workshop, 1996.
- ◆ S. Teufel. *Task-Based Evaluation of Summary Quality: Describing Relationships Between Scientific Papers*. NAACL Workshop on Automatic Summarization, 2001.
- ◆ J.S. White, T. O'Connell, & F.E. O'Mara. *The ARPA MT evaluation methodologies: Evolution, lessons and further approaches*. Technology partnerships for crossing the language barrier: Proceedings of the first conference of the Association for Machine Translation in the Americas, 1994.

Other References

- ◆ W. Black (Rapporteur). Evaluation Methods for Summarization. Summarizing Text for Intelligent Communication, 1993.
- ◆ M. King. Les belles infideles: fidelity as a criterion of good translation. B.H. Pattee & P. Sgall (eds.), Discourse and meaning. Papers in honour of Eva Hajiova, John Benjamins, 1995.
- ◆ D. Marcu. From discourse structures to text summaries. Proceedings of the ACL/EACL workshop on Intelligent Scalable Text Summarization, 1997.
- ◆ MUC-7. Proceedings of the 7th Message Understanding Conference. Morgan Kaufmann, 1998.
- ◆ D. Radev, H. Jing, & M. Budzikowska. Centroid-Based Summarization of Multiple Documents: Sentence Extraction, Utility-Based Evaluation, and User Studies. ANLP/NAACL Workshop, 2000.
- ◆ M. Walker, C. Kamm, & D. Litman. Towards Developing General Models of Usability with PARADISE. Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems, 2000.
- ◆ Y. Wilks. Systran: It obviously works but how much can it be improved? J. Newton (ed.), Computers in Translation: A Practical Appraisal, Routledge, London, 1992.