Chapter 13

Evaluation

(Following section is taken from Chapter 13 "Evaluation")
of the book: "Survey of the state of the art in human language technology")

13.2 Task-Oriented Text Analysis Evaluation

Beth Sundheim

Naval Command, Control and Ocean Surveillance Center RDT&E Division (NCCOSC/NRaD), San Diego, California, USA

The type of text analysis evaluation to be discussed in this section uses complete, naturally-occurring texts as test data and examines text analysis technology from the outside; that is, it examines technology in the context of an application system and treats the system as a black box. This type of evaluation is in contrast with ones that probe the internal workings of a system, such as ones that use constructed test suites of sentences to determine the coverage of a system's grammar. Two types of task-oriented text processing system evaluations have been designed and carried out on a large scale over the last several years:

1. Text retrieval has been evaluated in the context of:
   - a document routing task, where the system is tuned to match a statement of a user's persistent information need against previously unseen documents;
   - an *ad hoc* retrieval task, where the system is expected to match a user's one-time query against a more or less static (previously seen) text database.
2. Text understanding has been evaluated in the context of an information extraction task, where the system is tailored to look for certain kinds of facts in texts and to represent the output of its analysis as a set of simulated database records that capture the facts and their interrelationships. More recently, evaluations have been designed that are less domain-dependent and more focused on particular aspects of text understanding.

The forums for reporting the results of these evaluations have been the series of Text REtrieval Conferences (TREC) (Harman, 1993; Harman, 1994) and Message Understanding Conferences (MUC), particularly the more recent ones (DARPA, 1991b; DARPA, 1992b; ARPA, 1993b). The TRECs and MUCs are currently sponsored by the U.S. Advanced Research Projects Agency (ARPA) and have enjoyed the participation of non-U.S. as well as U.S. organizations.

The methodology associated with evaluating system performance on information extraction tasks has developed only in recent years, primarily through the MUC evaluations, and is just starting to mature with respect to the selection and exact formulation of metrics and the definition of readily evaluable tasks. In contrast, text retrieval evaluation methodology is now quite mature, having enjoyed over thirty years of development especially in the U.K. and U.S., and has been further developed via the TRECs, which have made substantial contributions to the text retrieval corpus development methodology and to the definition of evaluation metrics. With a fairly stable task definition and set of metrics, the TRECs have been able to measure performance improvements from one evaluation to the next with more precision than has so far been possible with the MUCs.

There are many similarities between TREC and MUC, including the following:

- Inclusion of both ARPA-sponsored research systems and other systems in the evaluations. Participation has included sites from North America, Europe, Asia and Australia.
- Use of large, naturally-occurring text corpora.
- Objective of end-to-end (*black box*) performance assessment.
- Evaluation metrics that are notionally similar, though different in formulation, reflecting the differences in the nature of the tasks.

The most enduring metrics of performance that have been applied to text retrieval and information extraction are termed *recall* and *precision*. These may be viewed as judging effectiveness from the application user's perspective, since they measure the extent to which the system produced all the appropriate output (recall) and only the appropriate output (precision). In the case of text retrieval, a correct output is a relevant document; in information extraction, a correct output is a relevant fact.

```
Recall = #relevant-returned/#relevant
Precision = #relevant-returned/#returned
```

In the above formulas, *relevant* refers to relevant documents in retrieval and to relevant facts in extraction; *returned* refers to retrieved documents in text retrieval and to extracted facts in information extraction. As will be explained below, text retrieval and information extraction represent fundamentally different tasks; therefore, the implementation of the recall and precision formulas also differs. In particular, the formulation of the precision metric for information extraction includes a term in the denominator for the number of *spurious* facts extracted, as well as the number of correct and incorrect facts extracted.

Typically, text retrieval systems are capable of producing ranked results, with the documents that the system judges more likely relevant ranked at the top of the list. Evaluation of the ranked output results in a recall-precision curve, with points plotted that represent precision at various recall percentages. Such a curve is likely to show very high precision at 10% recall, perhaps 50% precision at 50% recall (for a challenging retrieval task), and a long tail-out toward 100% recall.

A simple information extraction task design might involve a fixed number of data elements (attributes) and a fixed set of alternative values for each attribute. If the system was expected always to produce a fixed number

of simulated database records (sets of attributes), and a fixed number of facts per attribute from a fixed set of possible facts, it would be performing a kind of classification task, which is similar to the document routing task. In the document routing task performed by text retrieval systems, the routing queries represent categories, and the task is to determine which, if any, category is matched by a given text. However, an information extraction task typically places no upper bound on the number of facts that can be extracted from a text—the number of facts could conceivably even exceed the number of words in the text. In addition, a given fact to be extracted is not necessarily drawn from a predetermined list of possibilities (categories) but may instead be a text string, such as the name of a victim of a kidnapping event.

Thus, since texts offer differing amounts of relevant information to be extracted and the *right answers* often do not come from a closed set, it is probably impossible for an information extraction system to achieve 100% recall except on the most trivial tasks, and its false alarms are likely to include large amounts of spurious data (as well as simply erroneous data) if it is programmed to behave aggressively, in an effort to enable it to miss as little relevant information as possible. Current information extraction systems are not typically based on statistical algorithms, although there are exceptions. Therefore, evaluation typically does not produce a recall-precision curve for a system, but rather a single measure of performance.

One of the major contributions of both the TREC and the MUC evaluations has been the use of test corpora that are large enough to yield statistically valid performance figures and to support corpus-based system development experiments. The TREC-1 collection contained two-hundred times the number of documents found in a prior standard test collection (Harman, 1993). The MUCs have gradually brought about a similar revolution in the area of information extraction, which started in 1987 with a combined training and test corpus numbering just a few hundred, very short texts, and now uses several thousand longer texts; the number of test articles has increased from tens to hundreds.

To judge the correctness of the retrieval and extraction system outputs, the outputs must be compared with *ground truth*. Ground truth is determined by humans. In text retrieval, where the system may be evaluated using corpora consisting of tens of thousands of documents, it would be almost literally impossible to judge the relevance of all documents with respect to all queries used in the evaluation. Instead, one effective method in a multisystem evaluation on a corpus of that size is to pool the highest-ranked documents returned by each system and to judge the relevance of just those documents. For TREC-3, the 200 highest-ranked documents were pooled. It has been shown that different systems produce significantly different sets of top-ranking documents, and the pooling method can be fairly certain to result in a reasonably complete list of relevant documents (perhaps over 80% complete, on average across queries).

Information extraction systems have been evaluated using relatively small corpora (perhaps 100-300 documents) and just one or two extraction tasks. Ground truth is created by manually generating the appropriate database records for each document in the test set. Ground truth is not perfect truth in either retrieval or extraction, due not only to human factors but also to incomplete evaluation task explanations provided by the evaluators and to the inherent vagueness and ambiguity of text.

Widely varying system architectures, processing techniques, and tools have been tried, tested, and refined in the context of the MUC and TREC evaluations, accelerating progress in the robust processing of naturally-occurring text. There have been exciting innovations in technologies, including hybrid statistical/symbolic techniques and refined pattern-matching techniques. The infrastructure provided by the conferences and evaluations—shared corpora, evaluation metrics, etc.—encourage the interchange of ideas and software resources and help participants understand which techniques work.

The need to isolate system strengths and weaknesses is one of the motivations underlying recent TREC and MUC efforts. These efforts have resulted in a greater range of evaluation options for participants. For example, the range of MUC evaluations has broadened from a single, complex, domain-dependent information extraction task

to include also a simple, domain-independent task, and other tasks have been developed to test component-level technologies, such as identification of coreference relations and recognition of special lexical patterns such as person and company names. Various corporate and government organizations in Europe and the U.S. have sponsored similar component-technology, multisite evaluation efforts. These have focused especially on grammars and morphological processors as, for example, did the 1993 Morpholympics evaluation, coordinated by the Gesellschaft für Linguistische Datenverarbeitung (Hausser, 1994).

13.3 Chapter References

AMTA (1992). *MT evaluation: basis for future directions*, Washington, D.C. Association for Machine Translation in the Americas.

Arnold, D. et al. (1994). *Machine translation: an introductory guide*. NCC/Blackwell, Manchester, Oxford.

Arnold, E. et al. (1993). Special issue on evaluation of MT systems. *Machine Translation*, 8(1-2):1–126.

ARPA (1993a). *Proceedings of the 1993 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

ARPA (1993b). *Proceedings of the Fifth Message Understanding Conference*, Baltimore, Maryland. Morgan Kaufmann.

ARPA (1994). *Proceedings of the 1994 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Baird, H. S. (1992). Document image defect models. In Baird, H. S., Bunke, H., and Yamamoto, K., editors, *Structured Document Analysis*, pages 1–16. Springer-Verlag.

Balkan, L. et al. (1994). Test suites for natural language processing. *Translating and the Computer*, 16:51–58. papers presented at a conference.

Bimbot, F. et al. (1994). Assessment methodology for speaker identification and verification systems: an overview. Technical Report SAM-A Project 6819, Task 2500, SAM-A, Martigny, Switzerland.

Black, E. (1993). Parsing english by computer: The state of the art. In *Proceedings of the 1993 International Symposium on Spoken Dialogue*, Waseda University, Tokyo.

Black, E., Abney, S., Flickenger, D., Gdaniec, C., Grishman, R., Harrison, P., Hindle, D., Ingria, R., Jelinek, F., Klavans, J., Liberman, M., Marcus, M., Roukos, S., Santorini, B., and Strzalkowski, T. (1991). A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, Pacific Grove, California. Defense Advanced Research Projects Agency, Morgan Kaufmann.

Black, E., Garside, R., and Leech, G., editors (1993). *Statistically-Driven Computer Grammars of English: The IBM/Lancaster Approach*. Rodopi, Amsterdam, Atlanta.

Brill, E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy.

CCW (1991). Research achievements on Chinese character and voice recognition. *China Computer World*, 349. Written in Chinese.

Chollet, G., Bimbot, F., and Paoloni, A., editors (1994). *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, Switzerland. ESCA.

Chollet, G., Capman, F., and Daoud, J. F. A. (1991). On the evaluation of recognizers—statistical validity of the tests. Technical Report SAM-ENST-02, SAM.

Cohen, P. R. and Oviatt, S. L. (1994). The role of voice in human-machine communication. In Roe, D. B. and Wilpon, J., editors, *Voice Communication Between Humans and Machines*, pages 34–75. National Academy of Sciences Press, Washington, DC.

Cole, R. A., Hirschman, L., Atlas, L., Beckman, M., Bierman, A., Bush, M., Cohen, J., Garcia, O., Hanson, B., Hermansky, H., Levinson, S., McKeown, K., Morgan, N., Novick, D., Ostendorf, M., Oviatt, S., Price, P., Silverman, H., Spitz, J., Waibel, A., Weinstein, C., Zahorian, S., and Zue, V. (1995). The challenge of spoken language systems: Research directions for the nineties. *IEEE Transactions on Speech and Audio Processing*, 3(1):1–21.

Council, N. R. (1966). Appendices 9–15. In *Languages and Machines: Computers in Translation and Linguistics*. National Academy of Sciences, Washington, DC.

DARPA (1989). *Proceedings of the Second DARPA Speech and Natural Language Workshop*, Cape Cod, Massachusetts. Defense Advanced Research Projects Agency.

DARPA (1990). *Proceedings of the Third DARPA Speech and Natural Language Workshop*, Hidden Valley, Pennsylvania. Defense Advanced Research Projects Agency, Morgan Kaufmann.

DARPA (1991a). *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, Pacific Grove, California. Defense Advanced Research Projects Agency, Morgan Kaufmann.

DARPA (1991b). *Proceedings of the Third Message Understanding Conference*, San Diego, California. Morgan Kaufmann.

DARPA (1992a). *Proceedings of the Fifth DARPA Speech and Natural Language Workshop*. Defense Advanced Research Projects Agency, Morgan Kaufmann.

DARPA (1992b). *Proceedings of the Fourth Message Understanding Conference*, McLean, Virginia. Morgan Kaufmann.

Eagles (1995). Report of the spoken language systems working group 5. Technical report, EAGLES, EAGLES Secretariat, Istituto di Linguistica Computazionale, Via della Faggiola 32, Pisa, Italy 56126, Fax: +39 50 589055, E-mail: ceditor@tnos.ilc.pi.cnr.it. In press.

Eurospeech (1993). *Eurospeech '93, Proceedings of the Third European Conference on Speech Communication and Technology*, Berlin. European Speech Communication Association.

Falkedal, K., editor (1994). *Proceedings of the of the Evaluators' Forum, 1991*, Les Rasses, Vaud, Switzerland. ISSCO, Geneva.

Fourcin, A. et al. (1992). ESPRIT project 2589 (SAM) multi-lingual speech input/output assessment, methodology and standardization. Technical Report SAM-UCL-G004, SAM.

Galliers, J. R. and Sparck Jones, K. (1993). Evaluating natural language processing systems. Technical Report 291, University of Cambridge Computer Laboratory. To appear in *Springer Lecture Notes in Artificial Intelligence*.

Harman, D., editor (1993). *National Institute of Standards and Technology Special Publication No. 500-207 on the The First Text REtrieval Conference (TREC-1)*, Washington, DC. National Institute of Standards and Technology, U.S. Department of Commerce, U.S. Government Printing Office.

Harman, D. (1993). Overview of the first Text REtrieval Conference (TREC-1). In Harman, D., editor, *National Institute of Standards and Technology Special Publication No. 500-207 on the The First Text REtrieval Conference (TREC-1)*, pages 1–20, Washington, DC. National Institute of Standards and Technology, U.S. Department of Commerce, U.S. Government Printing Office.

Harman, D., editor (1994). *National Institute of Standards and Technology Special Publication No. 500-215 on the The Second Text REtrieval Conference (TREC-2)*, Washington, DC. National Institute of Standards and Technology, U.S. Department of Commerce, U.S. Government Printing Office.

Harrison, P., Abney, S., Black, E., Flickenger, D., Gdaniec, C., Grishman, R., Hindle, D., Ingria, R., Marcus, M., Santorini, B., and Strzalkowski, T. (1991). Evaluating syntax performance of parser/grammars of English. In *Proceedings of the Workshop On Evaluating Natural Language Processing Systems*. Association For Computational Linguistics.

Hausser, R. (1994). The coordinator's final report on the first Morpholympics. *LDV-Forum*, 11(1):54–64.

Höge, M., Hohmann, A., and Mayer, R. (1992). Evaluations of TWB: Operationalization and test results. Final Report of the ESPRIT I Project 2315 Translators' Workbench (TWB).

Höge, M., Hohmann, A., van der Horst, K., Evans, S., and Caeyers, H. (1993). User participation in the TWB II project: The first test cycle. Report of the Esprit II Project 6005 Translators' Workbench II (TWB II).

House, A. S., Williams, C. E., Hecker, M. H. L., and Kryter, K. D. (1965). Articulation testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America*, 37:158–166.

Houtgast, T. and Steeneken, H. J. M. (1984). A multi-lingual evaluation of the Rasti-method for estimating speech intelligibility in auditoria. *Acustica*, 54:185–199.

Hutchins, W. J. and Somers, H. L. (1992). An introduction to machine translation. In *An introduction to Machine Translation*. Academic Press, London.

ICDAR (1993). *Proceedings of the Second International Conference on Document Analysis and Recognition*, Tsukuba Science City, Japan. AIPR-IEEE, IAPR.

Ishii, K. (1983). Generation of distored charaters and its applications. *System, Computer, Controls*, 14(6):1270–1277.

ISO (1991). Information technology—software product evaluation, quality characteristics and guidelines for their use. Technical Report 9126, International Organization for Standardization.

ITU (1993). ITU-TTS draft recommendation p.8s: Subjective performance assessment of the quality of speech voice output devices. Technical Report COM 12-6-E, International Telecommunication Union.

Jekosch, U. (1993). Speech quality assessment and evaluation. In *Eurospeech '93, Proceedings of the Third European Conference on Speech Communication and Technology*, volume 2, pages 1387–1394, Berlin. European Speech Communication Association. Keynote address.

Jones, K. and Mariani, J., editors (1992). *Proceedings of the 1992 Workshop of the International Committee on Speech Databases and I/O Systems Assessment*. COCOSDA.

Kanai, J., Rice, S. V., Nartker, T. A., and Nagy, G. (1993). Performance metrics for document understanding systems. In *Proceedings of the Second International Conference on Document Analysis and Recognition*, pages 424–427, Tsukuba Science City, Japan. AIPR-IEEE, IAPR.

Kanungo, T., Haralick, R. M., and Phillips, I. (1993). Global and local document degradation models. In *Proceedings of the Second International Conference on Document Analysis and Recognition*, pages 730–736, Tsukuba Science City, Japan. AIPR-IEEE, IAPR.

Karis, D. and Dobroth, K. M. (1991). Automating services with speech recognition over the public switched telephone network: Human factors considerations. *IEEE Journal of Selected Areas in Communications*, 9(4):574–585.

King, M. and Falkedal, K. (1990). Using test suites in evaluation of MT systems. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 211–216, Pittsburgh, Pennsylvania. Association for Computational Linguistics.

Klaus, H., Klix, H., Sotscheck, J., and Fellbaumn, K. (1993). An evaluation system for ascertaining the quality of synthetic speech based on subjective category rating tests. In *Eurospeech '93, Proceedings of the Third European Conference on Speech Communication and Technology*, volume 3, pages 1679–1682, Berlin. European Speech Communication Association.

Kryter, K. D. (1962). Methods for the calculation and use of the articulation index. *J. of the Acoustical Society of America*, 34:1689–1697.

Lehrberger, J. and Bourbeau, L. (1988). *Machine translation: linguistic characteristics of MT systems and general methodology of evaluation*. John Benjamins, Amsterdam, Philadelphia.

Li, Y., Lopresti, D., and Tomkins, A. (1994). Validation of document image defect models for optical character recognition. In *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 137–150, University of Nevada, Las Vegas.

Logan, J. S., Greene, B. G., and Pisoni, D. B. (1989). Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America*, 86(2):566–581.

Matsui, T., Noumi, T., Yamashita, I., Watanabe, T., and Yoshimuro, M. (1993). State of the art of handwritten numeral recognition in Japan—the results of the first IPTP character recognition competition. In *Proceedings of the Second International Conference on Document Analysis and Recognition*, pages 391–396, Tsukuba Science City, Japan. AIPR-IEEE, IAPR.

Moore, R. C. (1994). Semantic evaluation for spoken-language systems. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Nagy, G. (1994). Validation of simulated OCR data sets. In *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 127–135, University of Nevada, Las Vegas.

Nerbonne, J., Netter, K., Diagne, A. K., Klein, J., and Dickmann, L. (1993). A diagnostic tool for German syntax. *Machine Translation*, 8:85–107.

Nomura, H. and Isahara, H. (1992). JEIDA's criteria on machine translation evaluation. In *Proceedings of the International Symposium on Natural Language Understanding and AI*, Kyushu Institute of Technology, Iizuka, Japan. part of the International Symposia on Information Sciences.

Oviatt, S. L. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language*, 9:19–35.

Oviatt, S. L., Cohen, P. R., and Wang, M. Q. (1994). Toward interface design for human language technology: Modality and structure as determinants of linguistic complexity. *Speech Communication*, 15(3–4):283–300.

Pallett, D., Fiscus, J., Fisher, W., Garofolo, J., Lund, B., and Prysbocki, M. (1994). 1993 benchmark tests for the ARPA spoken language program. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*, pages 49–74, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Pols, L. C. W. (1991). Quality assessment of text-to-speech synthesis-by-rule. In Furui, S. and Sondhi, M. M., editors, *Advances in speech signal processing*, chapter 13, pages 387–416. Marcel Dekker, New York.

Pols, L. C. W. (1994a). Speech technology systems: Performance and evaluation. In Asher, R. E., editor, *The Encyclopedia of Language and Linguistics*, volume 8, pages 4289–4296. Pergamon Press, Oxford.

Pols, L. C. W. (1994b). Voice quality of synthetic speech: Representation and evaluation. In *Proceedings of the 1994 International Conference on Spoken Language Processing*, volume 3, pages 1443–1446, Yokohama, Japan.

Pols, L. C. W. and Jekosch, U. (1994). A structured way of looking at the performance of text-to-speech systems. In *Proceedings, ESCA/IEEE Synthesis Workshop*, pages 203–206, New Paltz, New York.

Pols, L. C. W. and SAM-partners (1992). Multi-lingual synthesis evaluation methods. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, volume 1, pages 181–184, Banff, Alberta, Canada. University of Alberta.

Rhyne, J. R. and Wolf, C. G. (1993). Recognition-based user interfaces. In Hartson, H. R. and Hix, D., editors, *Advances in Human-Computer Interaction*, volume 4, chapter 7, pages 191–250. Ablex Publishing Corp, Norwood, New Jersey.

Rice, S. V. (1993). The OCR experimental environment, version 3. Technical Report ISRI TR-93-04, University of Nevada, Las Vegas, Nevada.

Rice, S. V., Kanai, J., and Nartker, T. A. (1993). An evaluation of OCR accuracy. Technical Report ISRI TR-93-01, University of Nevada, Las Vegas, Nevada.

Rice, S. V., Kanai, J., and Nartker, T. A. (1994). The third annual test of OCR accuracy. Technical Report ISRI TR-94-03, University of Nevada, Las Vegas, Nevada.

Rinsche, A. (1993). Evaluationsverfahren für maschinelle übersetzunngssysteme: zur methodik und experimentellen praxis. Technical report, Kommission der Europaeischen Gemeinschaften, Bericht EUR 14766 DE.

Sinaiko, H. W. and Klare, G. R. (1972). Further experiments in language translation: readability of computer translations. *ITL*, 15:1–29.

Sinaiko, H. W. and Klare, G. R. (1973). Further experiments in language translation: a second evaluation of the readability of computer translations. *ITL*, 19:29–52.

Sorin, C. (1994). Towards high-quality multilingual text-to-speech. In *Proceedings of the CRIM/FORWISS Workshop on Progress and Prospects of Speech Research and Technology*, pages 53–62, Münich.

Sparck Jones, K. (1994). Towards better NLP system evaluation. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Spiegel, M. F. (1993). Using the ORATOR synthesizer for a public reverse-directory service: Design, lessons, and recommendations. In *Eurospeech '93, Proceedings of the Third European Conference on Speech Communication and Technology*, volume 3, pages 1897–1900, Berlin. European Speech Communication Association.

Spitz, J. (1991). Collection and analysis of data from real users: Implications for speech recognition/understanding systems. In *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, Pacific Grove, California. Defense Advanced Research Projects Agency, Morgan Kaufmann.

Steeneken, H. J. M. (1992). Quality evaluation of speech processing systems. In Ince, N., editor, *Digital Speech Coding: Speech coding, Synthesis and Recognition*, chapter 5, pages 127–160. Kluwer Norwell, USA.

Steeneken, H. J. M. and Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *J. Acoustical Society of America*, 67:318–326.

Steeneken, H. J. M., Verhave, J., and Houtgast, T. (1993). Objective assessment of speech communication systems; introduction of a software based procedure. In *Eurospeech '93, Proceedings of the Third European Conference on Speech Communication and Technology*, volume 1, pages 203–206, Berlin. European Speech Communication Association.

Thompson, H., editor (1992). *The Strategic Role of Evaluation in Natural Language Processing and Speech Technology*. Human Communication Research Centre, University of Edinburgh.

Van Slype, G. (1982). Conception d'une méthodologie générale d'évaluation de la traduction automatique. *Multilingua*, 1(4):221–237.

White, J. S. et al. (1994). The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In *Technology partnerships for crossing the language barrier: Proceedings of the 1st Conference of the Association for Machine Translation in the Americas*, pages 193–205, Washington, DC. Association for Machine Translation in the Americas.

Wilkinson, R. A., Geist, J., Janet, S., Grother, P. J., Burges, C. J. C., Creecy, R., Hammond, B., Hull, J. J., Larsen, N. J., Vogl, T. P., and Wilson, C. L. (1992). The first census optical character recognition systems conference. Technical Report NISTIR-4912, National Institute of Standards and Technology, U.S. Department of Commerce.

Zoltan-Ford, E. (1991). How to get people to say and type what computers can understand. *International Journal of Man-Machine Studies*, 34:527–547.