EVALUATION OF ENDOGENOUS SYSTEMS

H. JOHN CAULFIELD
Distinguished Research Scientist
Fisk University
1000 17th Avenue, N.
Nashville, TN 37208
hjc@dubois.fisk.edu

FLORENCE M. REEDER^
George Mason University /
The MITRE Corporation
1820 Dolley Madison Blvd.
McLean, VA 22102
freeder@mitre.org

ABSTRACT
In system development we are faced with the necessity of evaluation. Evaluation measures our success relative to other: a) theories of development or domains; b) implementations of similar theoretical principles; or c) increments of a given system. For successful software engineering evaluation, progress is measured against a model, a task frequently accomplished through requirements analysis. This model is lacking in natural language processing (NLP) systems (not to be confused with neurolinguistic programming). NLP systems defy evaluation in part because they model an endogenous process – where the whole process is irreducible. Therefore, while specific feature-based evaluations appear reasonable, they fail to capture an overall measure of success. In this paper, we look at the part/whole aspects of evaluation in more detail with regard to one language system type – machine translation.

INTRODUCTION

In system development, of any type of system, we are faced with the necessary evil of evaluation. Formal evaluation measures the success we are having relative to: a) other theories of development or domain modeling; b) other implementations of the same theoretical principles; c) other systems for the purpose of purchase or funding; or d) previous increments of a given system. For successful evaluation in software engineering, one needs a model against which progress can be measured, a task frequently accomplished through requirements analysis. For many domains, this is a straight-forward process: a correct answer exists, such as, "Pushing F10 results in program termination."

Many types of NLP systems, however, have defied rigorous evaluation because of this lack of defining criteria or requirements. Part of the difficulty lies in the fact that NLP systems are modeling an endogenous process – where the whole of the process is irreducible, context-dependent and lacking a unique right answer. Therefore, while specific feature-based evaluations appear reasonable, they fail to capture an overall measure of success. In this paper, we look at the part/whole aspects of evaluation in more detail with regard to one language system type – machine translation (MT).

This paper starts with a description of the MT process as traditionally developed. It then describes the resulting evaluation strategies that follow from these views of MT. The features of MT which categorize it as an endogenous are presented as a precursor to showing a new view of the original MT vision. Finally, the questions of evaluation are explored within this new view.

MACHINE TRANSLATION

Machine Translation (MT) systems attempt to replicate the very human process of translating between human languages. The idea for attempting this comes from Weaver (1947) when he describes translation as a kind of noisy channel decryption process: a French document is really an English document which has been encrypted in the strange symbols of French. Unfortunately, language learners and language translators know that the process is much more than the simple substitution of symbols.

Traditional MT systems have been developed following the Chomsky ideal (1957) of processing language through successive steps, each providing another level of information which isolates words from meaning. These levels are pragmatic (mechanics of language in use), discourse (rules of interaction and coherence), semantic (attribution of meaning to words), syntactic (assignment of structural interaction of words), morphological (the features of words), and phonemic (realization of words in sound). This understanding of language led to the development of MT systems based on these layers of processing, often characterized as a pyramid where complexity increases through ascending layers of processing.

An MT system performs a level of analysis of increasing detail as the type of system advances in the pyramid. That is, a lexical transfer system will perform only morphological processing. A syntactic transfer system will do morphological analysis and syntactic analysis on the source text before generating text in the target language. The interlingua level represents a "pure" translation process, much like human translators perform, where analysis yields a language-independent representation from which generation can be performed. Systems in this paradigm typically have modules and information sources as shown in Figure 1.[1]

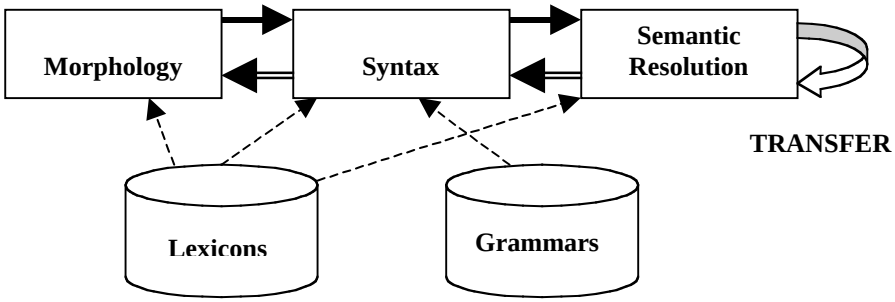


Figure 1: A Typical MT Processing Flow

MACHINE TRANSLATION EVALUATION (MTE)

In software engineering practice, system evaluation reflects how well the system performs the requirements for which it was designed. Derived in advance of system design, these requirements cover system aspects such as

[1] Note that most current systems fall far short of discourse and pragmatic processing.

coding language, mean-time-between-failure, number of records processed in X time, the procedure for updating records, etc. However, MT systems are frequently developed in the absence of requirements, and reflect a notion that good MT, like pornography, "can't be defined but we'll know it when we see it." The many MTE metrics reflect these broad and ill-defined criteria.

While basic software engineering metrics (disk usage, execution time, throughput) apply to MTE, MT systems have a wide range of additional parameters for evaluation, each having different importance to different users who have a large set of possible uses. Other NLP tasks like speech transcription have a right answer, or "gold standard". Not so for MT as there is no single right answer for a translation even when humans do it. What constitutes an acceptable level of translation varies. For users in legal domains, even minor failures are unacceptable. For other domains, such as intelligence analysis, a 40% solution is good enough.

Accordingly, MTE has had a long and often painful history. From the early days (ALPAC, 1966), excessive optimism or pessimism and misconceptions – about language, MT usage and system requirements – have been predominant. The various notions of MTE result in too broad a scope for reasonable effort: everything from interface, to scalability, to faithfulness of translation, to mean-time-between-failures are fair game for MT system evaluation. Claims of 90% accuracy abound without an indication of what the system is accurate about.

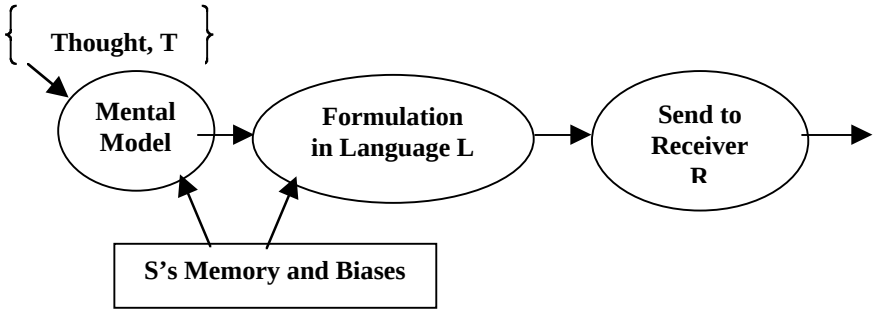
The set of MTE approaches reflects the varying uses and requirements for MT, in addition to the multiple stakeholders in evaluation – developers, funders, users (White, 1998). Evaluation strategies have included: a) evaluating MT systems as software (e.g., EAGLES, 1996); b) black-box evaluations on fidelity and intelligibility of output only (e.g., Van Slype, 1979); c) measuring accuracy of input/output pairs; d) glass-box evaluations of components; e) measuring speed, cost,quality in process and others. These have relied, primarily, on human judgments as to the "correctness" of a translation on two basic metrics: fidelity (amount of the message conveyed) and intelligibility (the fluency of the message as translated). These efforts require a large number of evaluators, looking at, and holistically rating, sentences on a 5, 7 or 9 point scale (i.e., DARPA, 1994). The difficulty is that aside from human factors issues, the results give little information to either users or developers of MT.

In reaction to this, recent efforts at evaluating system abilities to translate have focused on pieces of the language puzzle. Specific grammatical phenomena are measured such as the ability of a system to translate prepositions (Miller, 2000) or named-entities (Reeder et al., 2001). These approaches seek to measure the specific contribution of MT to the preservation of very specific pieces of information. These strategies have great appeal – they are objective, replicable, informative and indicate the ability to meet a process "bottom line." However, recent exercises in MTE for named entities (Reeder et al, 2001) show a definite gap between the specific piece indicators and more holistic scoring. How do we explain this?

COMMUNICATING EFFECTIVELY USING MT

The answer lies in re-thinking the problem in terms of endogenous systems. An information theoretic view of NLP is not new, but its utility encourages us to

revisit MTE, looking for better evaluation strategies. To this end, we define effective communication and describe ways to test if it has been accomplished.



As a start, we examine communication between two people using one mode, voice, in one language in conversation. Then, we will be ready to discuss MT.

Figure 2: Basic Model of Utterance Generation

We assume something (a scene, situation, desire, etc.) to be communicated – in NLP terms, a discourse purpose. In some cases, the target something (**T**) will be external to the sender, **S**. In others, it exists only as a mental model within **S**. If external, it is sensed by **S**, which is to say that **S** forms a mental model of it. **S**'s mental model of what others might agree is the same thing may differ from everyone else's because he brings different memories, assumptions and biases, to the model-forming task than others. **S** represents the mental model with words organized into one or more coherent sentences. Chomsky (1957) described this process (Figure 2) as transforming deep-structure into surface-structure. The resulting message is sent as agreed to the receiver, **R**.

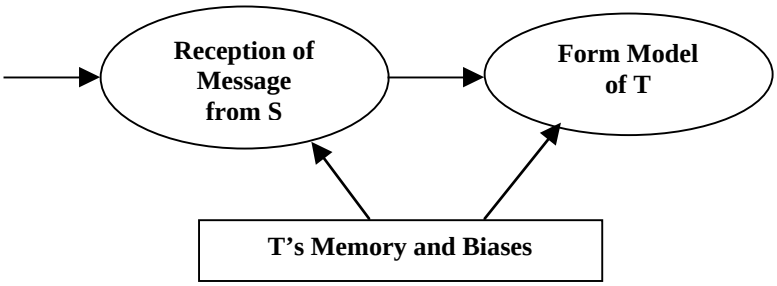


Figure 3: R's reception and processing of the message

The receiver forms his own mental image of something that may or may not be **T**, bringing his own memory and biases to the task. This is the analysis phase or conversion from surface-structure to deep-structure, Figure 3. And, of course, the words are frequently ambiguous, often inadequate to convey the full model content or a literal mismatch from the intended message.[2] Now, we have

[2] If a man and woman are sitting in a room with an open window, her utterance, "Gee, there's a draft" is considered a command, not a statement of fact.

a problem. Both **S** and **R** have mental models of something (presumably **T**), but how can we know that they are the same or even consistent? Translation between languages adds another layer of complexity. Instead of the message **T** being consistent, it is transformed, sometimes at the word level, between two languages, each with its own biases, strengths and weaknesses (Figure 4). The sought-after congruence of mental models (identity is hopeless in most cases) is what characterizes what we call "effective communication" and represents what evaluation of MT systems must measure.

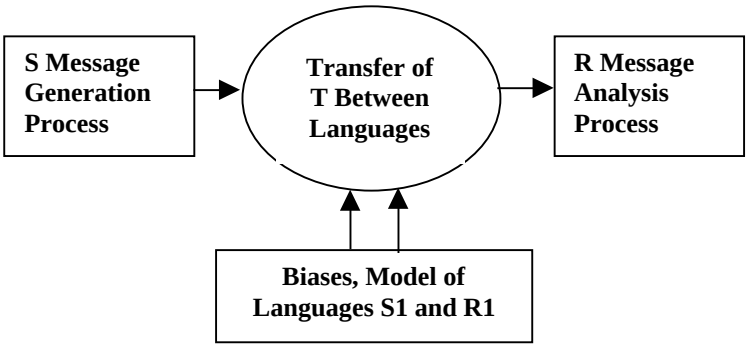


Figure 4: Translation Process with Model

EVALUATING EFFECTIVE COMMUNICATION TOOLS

The problem is clear. **S**'s model and **R**'s model of **T** cannot be compared, because they exist in two different minds and because of their endogenous nature. If **T** is a command, and if **R**'s action seems an appropriate response to that command, then we may have had effective communication. If the correct response does not occur, we cannot infer ineffective communication. A two-year old may completely get the message, "Don't eat that" even if he eats the bug. Likewise, **R** need not have fully "understood" the message to have correct response, due to the redundant nature of language. From this we can see that we need a surer, more general test. Having established different variables, we go back to our desire to effectively measure MT success. We show two methods deriving from communications theory and show how they relate to current MTE.

Method 1: Closing the Loop. We ask **R** to send his understanding of **T** to **S** in words different from the words used by **S**. **S** will form a model of **T** on the basis of **R**'s message. Now **S** has two models of **T** in his own mind and can judge whether they are sufficiently congruent. A corresponding MTE strategy is the "round-trip translation" test.

Method 2: Redundant Messages. We ask **S** to send two messages about **T** using different words. If **R** (who now has two models of **T**) judges them to be congruent, **S** can feel assured that he has communicated the message effectively. This corresponds to human judgments and we are back to the beginning in MTE.

As stated earlier, recent MTE focuses on pieces of messages and usually applies Method 2. Name translation evaluation uses method 2 specifically for one part of the message – the peoples, places and things represented in the

message. A system translation is compared with two human translations. Unfortunately, no strong correlation can be drawn between this metric and the "whole" rating given in previous evaluations with the same data. That is, systems may or may not have good scores on the named entity task which do not correlate with their quality scores on a sentence by sentence basis (White, et al, 2001). This implies a part-whole disjunction which must be accounted for.

AN APPROXIMATION & CONCLUSION

We are left, then, with the goals of MTE – reliably, objectively determining the effectiveness of an MT system to support the communicative task. Where we differ, however, is in providing a model for communication which explicitly represents the many factors of communication. This is the point where smart engineering and endogenous views allow us to approximate a solution. We have human assigned scores based on Method 2 (a 1→5 rank of message reception). This holistic scoring does reliably capture system quality information.

Given that we have a body of data for which we have human ratings, our slightly different look at MTE makes the rating, not the human translations, the "gold standard". The problem then becomes one of identifying and measuring the criteria which contribute to the ratings. It is a classical machine learning problem of categorizing items based on objective and reasonably measured criteria. In this way, MTE becomes a more replicable, automatable task while at the same time capturing the human intuition of quality output. While we will not find all of the criteria or a complete correspondence until we develop endogenous evaluation abilities, we can at least approximate current measures.

REFERENCES

Chomsky, N., 1957, *Syntactic Structures*, Mouton Publishers, The Hague.

EAGLES. 1996. The EAGLES MT Evaluation Working Group. 1996. EAGLES Evaluation of Natural Language Processing Systems Final Report. EAGLES Document EAG-EWG-PR.2, ISBN 87-90708-00-8. Center for Sprogteknologi, Copenhagen.

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

Pierce, J., Chair. 1966 *Language and Machines: Computers in Translation and Linguistics*. Report by the Automatic Language Processing Advisory Committee (ALPAC). Publication 1416. National Academy of Sciences National Research Council.

Reeder, F., Miller, K., Doyon, J., & White, J., 2001, The Naming of Things and the Confusion of Tongues: An MT Metric. *Proceedings of the Machine Translation Evaluation Workshop, MT-Summit 2001*. Campostela de Santiago, Spain.

Van Slype, G. 1979. *Critical Methods for Evaluating the Quality of Machine Translation*. EC Directorate: General Scientific and Technical Information and Information Management. Report BR-19142. Bureau Marcel van Dijk.

Weaver, W., 1947, Letter to Norbert Wiener, March 4, 1947, (Rockefeller Foundation Archives).

White, J., et al. 1992. ARPA Workshops on Machine Translation. Series of 4 workshops on comparative evaluation. PRC Inc. McLean, VA.

White, J. 1998. Evaluation of Machine Translation. A Tutorial. *AMTA-98*.

---

^ The views expressed in this paper are those of the author and do not reflect the policy of the MITRE Corporation.