APPLICATIONS 1:

MACHINE TRANSLATION I: MACHINE TRANSLATION THEORY AND HISTORY

Theme

The first of two lectures on Machine Translation: the oldest application of NLP. Background and historical developments. 3 application niches. The MT Triangle: increasing depth and difficulty. Examples of each level, incl. EBMT, transfer, and KBMT.

Summary of Contents

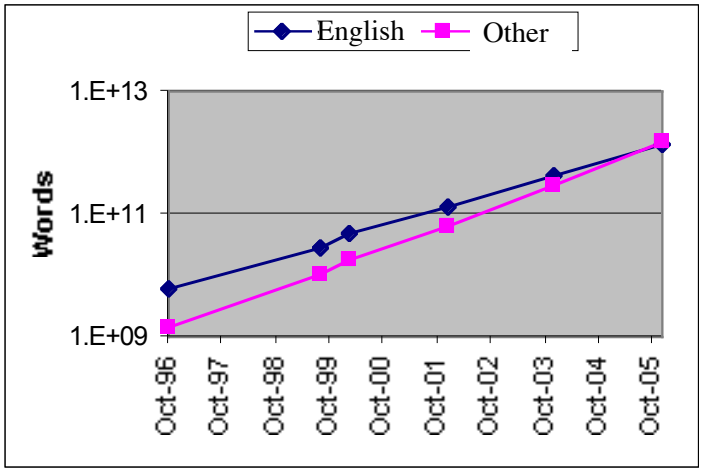
1. History

History: MT from the late 1940's and the Cold War, to 2000 and the Web. The ALPAC report. Some examples: early SYSTRAN, METEO, Eurotra, late SYSTRAN, CANDIDE.

The general trend: the larger and more established the lexicon, the better the system.

Current: growth of languages on the internet. Study of languages (Grefenstette, Xerox Europe 2000), augmented by Oard (Maryland) for projection and Hovy (ISI) for Asian languages.

	Sample (thousands of words)			Exponential Growth Assumption		
Language	**Oct-96**	**Aug-99**	**Feb-00**	**Dec-01**	**Dec-03**	**Dec-05**
English	**6,082.09**	**28,222.10**	**48,064.10**	**128,043.57**	**419,269.14**	**1,375,098.05**
German	228.94	1,994.23	3,333.13	13,435.07	65,161.79	316,727.36
Japanese	228.94	1,994.23	3,333.13	9,375.41	40,070.32	171,600.89
French	223.32	1,529.80	2,732.22	9,375.41	40,070.32	171,600.89
Spanish	104.32	1,125.65	1,894.97	8,786.78	48,968.42	273,542.30
Chinese	123.56	817.27	1,338.35	8,786.78	48,968.42	273,542.30
Korean	123.56	817.27	1,338.35	4,507.93	18,206.81	73,675.11
Italian	123.56	817.27	1,338.35	4,507.93	18,206.81	73,675.11
Portuguese	106.17	589.39	1,161.90	3,455.98	13,438.26	52,350.71
Norwegian	106.50	669.33	947.49	3,109.04	11,474.59	42,425.27
Finnish	20.65	107.26	166.60	480.19	1,628.87	5,534.62
Non-English	**1,389.49**	**10,461.70**	**17,584.48**	**65,820.52**	**306,194.61**	**1,454,674.58**
Non-English%	18.60%	27.04%	26.79%	33.95%	42.21%	51.41%



2. Usage

Three basic patterns of usage; these determine what user wants and likes.

- Assimilation: the user gathers information from out there, at large. Need wide coverage (many domains), low (browsing-level) quality, and speed. Often hooked up to IR engine and classification engine. Usually used for triage, with the selected material passed on to humans for professional translation. Typical user: information gathering office of company or Government.
- Dissemination: the user produces information to be sent to others. Need narrow coverage (just the domain of specialization) but high quality; speed is less important. Typical user: large manufacturer (Xerox, Caterpillar).
- Interaction: the user interacts with someone else via the web, bboards, or email. General domain, browsing quality needed but not too important; speed and dialogue support are important (slang, bad grammar, funny face icons, etc.). Typical user: shoppers and chatters on the web.

The interaction of editor and MT system:

in → fully automated (no editing) → out	cheap but low quality
in → pre-editing → MT → out	only with limited domains
in → MT → post-editing → out	usual method; costs about 10c / page
in → MT with in-editing → out	experimental; need bilingual assistants

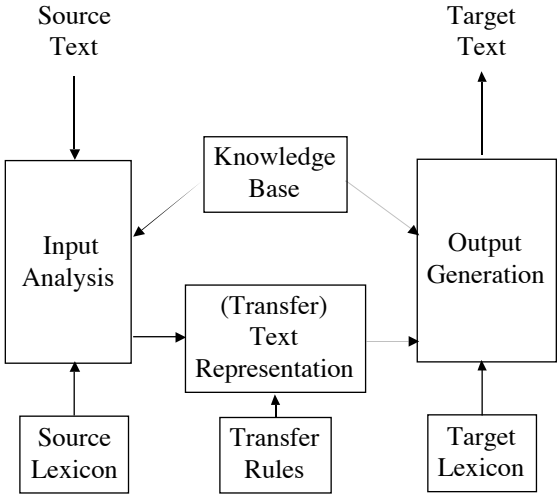
Automation Tradeoffs

Fully automated
- Cheap
- Fast; background/batch mode
- Low quality, unless small domain
- Best example: METEO (weather report translation)

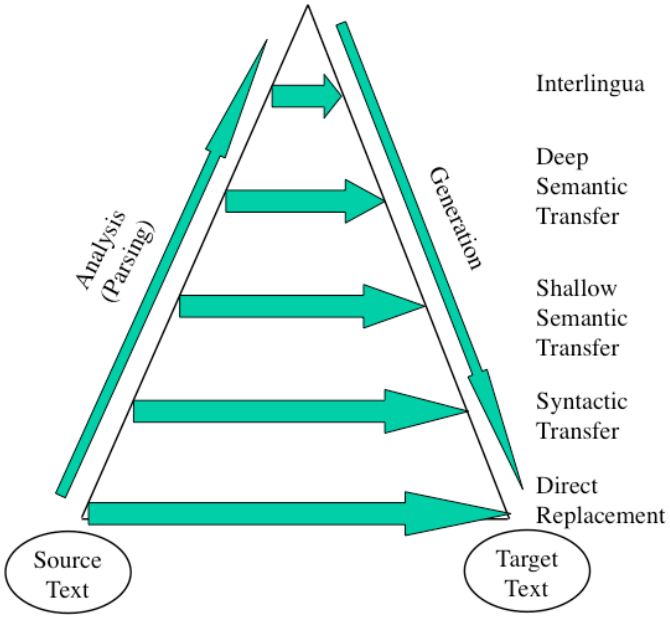
Human-assisted
- More expensive (>10c a page...)
- Slow
- High(er) quality
- Editing tools, dictionaries, etc., required
- Most common commercial configuration

3. Theory

The basic MT architecture.



The MT Triangle (Vauqois). Increasing levels of complexity move the internal processing gradually away from words and closer to concepts. This is accompanied by more and more processing, both in parsing/analysis and in generation.



The Vauqois (MT) Triangle.

Lowest level: direct replacement systems. Simply replace each word, or multi-word phrase, with its equivalent in the target language. For this you need a large bilingual lexicon (or its generalization, some kind of phrasal correspondence table). At its simplest, there is no local word reordering, no change of word morphology (conjugation of verbs, declension of nouns and adjectives, etc.). The main two problems are:

- you get essentially the source language syntax using target language words;
- there is no word-sense disambiguation, so for words with more than one possible meaning (sense), the system has to guess which word to replace it with.

This gives very low-quality output; unreadable in the case of distant languages. Many handheld 'translators' you buy today work this way.

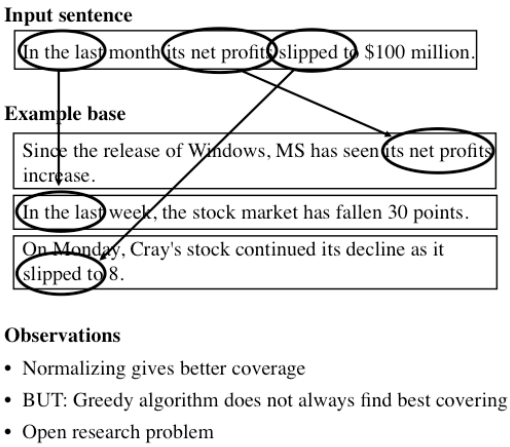
If you are smart, you notice that some words are replaced by no words, and some by more than one word (the 'word fertility'), and also that some words move around relative to others ('distortion'). The earliest simple MT systems of the 1950s rapidly went beyond Direct Replacement, but in the early 1990s IBM's 1986–94 CANDIDE statistical MT system was a later example of this method. (We discuss CANDIDE and the systems that built upon it in the next lecture.)

Also if you are smart, you notice that the very large 'correspondence tables' (bilingual lexicons) are highly redundant for inflected languages (why should you have separate entries for "go", "went", "gone", "going", "goes", etc.? If you replace each source word by its root form and add a pseudo-word that carries the other information (tense, number, politeness, etc.), then you can greatly reduce the size of the translation table. This leads to the next level.

Next-lowest level: perform a small amount of processing for morphological and other term standardization (thereby reducing bilingual lexicon size). To carry the remaining information (tense, number, etc.) you introduce pseudo-words. This requires a small generation module at the output end to reassemble the necessary features into the root word and produce the correct form. Now you have left the surface level and are starting to work with abstractions, and have begun to move toward grammar.

Example-Based MT (EBMT) is a special case of this method. Here you comb over a parallel corpus (source and target language texts aligned sentence-by-sentence) to find snippets of text that correspond to one another. How large is a snippet?—from two words to a sentence, or even a paragraph in cases of technical manuals. In the latter case, this is called 'Translation Memory' in the MT industry. The main problem for short snippets is to combine the target snippets grammatically.

EBMT EXAMPLE 2

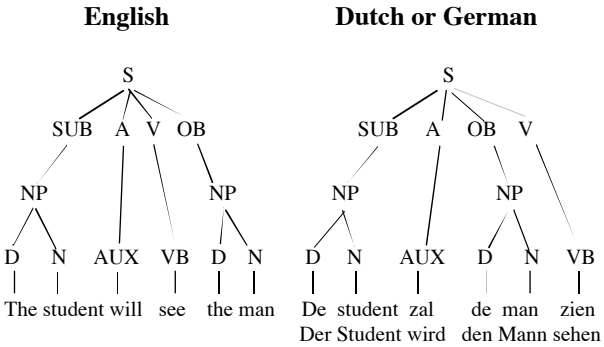


Selecting and combining the snippets into coherent sentences is a nontrivial problem: you may find two long snippets that don't fit well together vs. four short ones that do, but are not guaranteed to make a good sentence altogether. Typically, people use a dynamic programming algorithm to find the best solution within certain bounds.

Middle level I: syntactic transfer systems. In order to get the target sentence grammar right, you produce the source sentence syntax tree and then map it into the target form. So you need a parser and a set of transfer (mapping) rules, in addition to the bilingual lexicon. Getting a good parser with wide coverage, as you know, is difficult! Still, this is probably the most popular MT method, augmented with some semantic transfer as well.

SYNTACTIC TRANSFER EXAMPLE

English: The student will see the man
Dutch: *De student zal zien de man
De student zal de man zien
German: *Der Student wird sehen den Mann
Der Student wird den Mann sehen



Only need one (syntactic) transfer rule:

$$\{A\ V\ O\}_{English} \rightarrow \{A\ O\ V\}_{Dutch,German}$$

Middle level II: semantic transfer (shallow semantic) systems. Still there are many phenomena that do not map across languages via syntax alone. In Japanese you say "my head hurts" for "I have a headache"; in German and French you say "I have hunger" for "I am hungry"; in Spanish you say "I cross the river swimmingly" for "I swim across the river". To get this right you have to understand something of the meaning of what is being said, and of the idiomatic ways of expressing that meaning in the target language. So you have to invent some kind of meaning representation that reflects some of the meanings of what you want to say in the languages you are handling (shallow semantics), and have to build a semantic analyzer to follow (or replace?) the parser. You also have to extend the lexicon to include semantic features, such as *animacy*, manner ("swimmingly"), etc. You need rules of demotion (a syntactico-semantic constituent is demoted from higher to lower in the syntax tree (*verb*: "swim" to *manner*: "swimmingly") and promotion (the opposite). If you thought syntactic parsing was hard, then just try this! We talk about representations at the next level. Most commercial systems use a combination of syntactic and semantic transfer, and an internal representation that combines features of both. Examples: SYSTRAN, Logos (both commercial); Eurotra (research).

Top level: Interlingua systems. And finally, you say: let's just go for the ultimate: the true, language-neutral, non-syntactic, meaning. Now you have a real problem! What must go into the representation? How do you guarantee that it carries all the pertinent aspects of meaning, and what do you do when you discover it doesn't? (For example, in Hebrew and Arabic you have not only singular and plural, but also a

special form of the noun for dual, for paired things like eyes and arms. If you didn't know this, would your interlingua record the fact that people have exactly two eyes and arms, so that when the English input is "he swung his arms" you would record the fact that there are two, and so get the correct output form in Arabic?) Eventually you are led to invent a set of symbols for meanings and to taxonomize them so that you get feature inheritance: you build an 'ontology' (theoretical model) of the world, or of your domains. In this picture, a big question is the content of the lexicon: how much information is truly world knowledge and how much is language-specific? (The color example, and the drinking tea in England, China, and Texas example.) No large-scale Interlingua systems have ever been built, despite many attempts. Still, some very domain-specific ones are used (CMU's KANT system for Caterpillar) and have been researched (CMU's KBMT, CICC in Asia, Pangloss, etc.).

Definition of an Interlingua:

An Interlingua is a system of symbols and notation to represent the meaning(s) of (linguistic) communications with the following features:

- language-independent
- formally well-defined
- expressive to arbitrary level of semantics
- non-redundant

4. Practice

The tradeoffs in constructing MT systems: small perfect toys vs. large low-quality robust engines. METEO vs. SYSTRAN.

The n^2 argument against transfer systems and for interlinguas.

A newer idea is multi-engine MT: combine several MT systems in one, send the input through them all, and combine their outputs (by selection, say) into a single one (or a set of alternatives, ranked). The problem is to determine the snippet size (granularity) of the result, to compare the alternative snippets, and to assemble them into (a) grammatical sentence(s).

Today: Web-based MT.

5. What Makes MT Hard?

Process-related Problems

1. Lexical level

Goal: locate each source word correctly in lexicon

Problems:

- morphology: spelling errors, bad segmentation (missing or extra delimiters) "run together" or "run to get her"
- ambiguity: idioms, proper name delimitation

Tasks:

- de-inflection (morphology analysis)
- correct lexicon access

2. Syntactic level

Goal: build correct parse tree

Problems:

- extra- and non-grammaticality
- syntactic ambiguity (symbol and attachment: *time flies like an arrow*)

Tasks:
- part of speech analysis
- find appropriate grammar rule
- assemble parse tree

3. Semantic level

Goal: build correct semantic interpretation

Problems:
- ambiguity (symbol and attachment) (the man with the child vs. the man with the glasses)
- context: underspecified input
- metonymy (Washington announced that...) & metaphor

Tasks:
- find correct symbol in symbol lexicon
- find correct combination/mapping rules and constraints assemble semantic structure correctly

4. Discourse level

Goal: build complete interpretation of text as a whole

Problems:
- nature of discourse structure
- Speech Acts and Implicatures
- pragmatics: interlocutor goals, interpersonal effects, etc.

Tasks:
- resolve anaphora (pronouns, etc.)
- assemble discourse structure

Resource-related Problems

Core problem: the size of language!

Lexical coverage
- number of words and phrases (commercial systems: 250,000 to 1 million)
- number of proper names (easily 100,000+)

Syntactic coverage
- number of grammar rules (commercial systems: 500–1,000)

Semantic coverage
- number of internal representation symbols—more, for greater delicacy of representation and better quality: 200,000+
- number of inference rules: no-one knows

6. Commercial MT

- Alis Technologies Inc. provides web-based translation for Netscape.
- ASTRANSAC focuses on translation to and from Japanese.

- Bowne Global Solutions.
- Google Translation engines. Created by Franz Och (formerly ISI) and his team. Statistical pattern-based replacement approach.
- IBM's WebSphere translation server is available.
- Language Engineering Corporation, the makers of LogoVista products, was established in 1985.
- LanguageWeaver, created by our own Kevin Knight and Daniel Marcu. About 100 people. Statistical pattern- (and now tree)-based approach. .
- Lingvistica, b.v. develops computer translation software and dictionaries.
- Microsoft does a lot of research on MT.
- Sakhr Software focuses on translation to and from Arabic.
- SDL International is a corporate member of the AMTA and a leading developer of machine translation. Their products include the Enterprise Translation Server, Transcend, EasyTranslator, and they are the company behind FreeTranslation.com.
- SYSTRAN Software is one of the oldest MT companies in the world, and the creator of AltaVista's Babelfish.
- Translation Experts Limited, manufactures natural language translation software.
- WorldLingo Inc.: Translation, Localization, Globalization.

A very comprehensive summary is available at http://www.hutchinsweb.me.uk/Compendium.htm .

There are increasingly many apps and products that provide limited MT on handheld devices. These systems either perform simple phrasal lookup into phrasal dictionaries for given domains (like travel, lost&found, buying&selling, etc.), or they access a server via the internet.

7. Speech translation

'Translating telephone'. Still experimental; first commercial product (workstation) announced in 2002 by NEC Japan. They are working on a handheld version. 50,000 words, J↔E, travel domain.

Problem: compound errors of speech recognition and MT. Use only in very limited domains.
- C-STAR consortium—CMU (USA), U of Karlsruhe (Germany), ATR (Japan)
- Verbmobil—various places in Germany
- NEC system, NEC, Tokyo
- SL-Trans—ATR, Kyoto
- JANUS—CMU, Pittsburgh

8. Evaluation

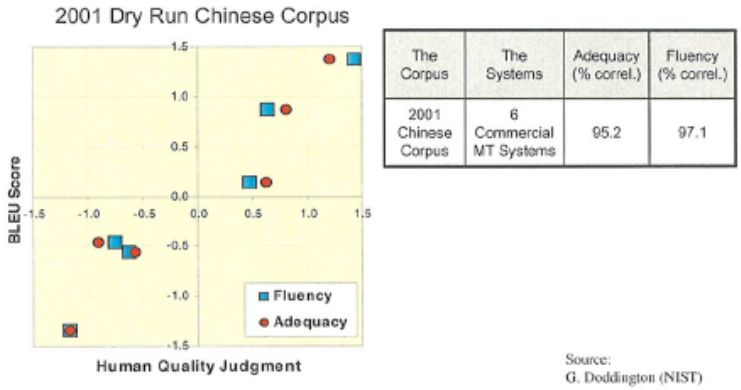
"MT Evaluation is older than MT itself" — a famous quote from Yorick Wilks. A great deal of work has been done on this topic. In the last 15 years, research MT has used the BLEU metric, which we will discuss in the next lecture. But real-world MT uses many other metrics besides, many of them focusing on practical issues such as cost and editor-friendliness.

In the early 1990s, DARPA funded a very large-scale evaluation using three metrics:
- **Fluency**: How good is the language? (Regardless of whether the facts are correct, is it readable?)

- **Adequacy**: How true is the translation? (Regardless of the readability, is the information conveyed?)
- **Comprehensibility**: Regardless of either of the above, can the reader understand the point of the text?

Three language pairs (Spanish-English, Japanese->English and French->English) were tested, mainly on newspaper-genre text.

Various studies were performed to try to find a single, easy, comprehensive metric that says it all. Ultimately, the consensus was that Fluency and Adequacy are sufficient, and even that they correlate quite well: if the output is fluent it is probably pretty correct as well, and vice versa.



The Corpus	The Systems	Adequacy (% correl.)	Fluency (% correl.)
2001 Chinese Corpus	6 Commercial MT Systems	95.2	97.1

Fluency vs. Adequacy from the DARPA MT Evaluations, 2001.

In order to make sense of all the various evaluation proposals, Hovy, King, and Popescu-Belis conducted a large-scale survey and integration effort around 2000, which resulted in the FEMTI framework, available at http://www.issco.unige.ch/en/research/projects/isle/femti/

FEMTI's backbone is made of two classifications or taxonomies. The first one is intended to be used in determining what the user's needs are. We suggest that if you are using FEMTI in order to design your own evaluation, you start by working through this section. You may well find that as you work through, the information solicited by later steps causes you to modify your responses at earlier steps. In that sense, the first taxonomy is to be thought of as forming a feedback loop which you may want to go through several times. The steps given here are meant to conform to those proposed in ISO/IEC 14598 for designing an evaluation.

The second taxonomy concerns quality characteristics of MT systems that are potentially of interest. In the light of your selections from the first taxonomy, it should be possible to pick out those characteristics that are pertinent to the needs you have made explicit. Following the pointers from these quality characteristics will lead you to possible metrics that can be applied to measure a particular system's performance with respect to that characteristic. The system characteristics are structured at the top level in conformity with the ISO/IEC 9126 standard, which concerns quality characteristics of software products. There is however one major divergence from the ISO/IEC 9126 standard in the inclusion of costs as a top level quality characteristic, as well as in the separation of internal (2.1) and external (2.2) quality characteristics. ISO/IEC 9126 considers cost to be a management factor, which is taken into account only after the evaluation proper has been done rather than an inherent factor influencing the quality of the system. In the specific case of MT systems, where cost may be a critical factor in determining whether it is even worthwhile designing and carrying out a full scale evaluation, we have thought it sensible to promote costs to the status of a quality characteristic.

In the future, we would like to be able to provide automatic pointers from specific user needs to specific system characteristics as well as from there to specific metrics. Indeed, some such pointers are already included. However to do this fully requires substantial user validation of our classification scheme, as well as much reflection on the relationship between needs and system features. You can contribute to this reflection by providing us with feedback on your experience of using FEMTI in its current state, by following the **Comments** link, at the top level or for each individual taxon.

COMPACT FEMTI CLASSIFICATION

1 Evaluation requirements

- 1.1 The purpose of evaluation
- 1.2 The object of evaluation
- 1.3 Characteristics of the translation task
    - 1.3.1 Assimilation
    - 1.3.2 Dissemination
    - 1.3.3 Communication
- 1.4 User characteristics
    - 1.3.1 Machine translation user
    - 1.3.2 Translation consumer
    - 1.3.3 Organizational user
- 1.5 Input characteristics (author and text)

2 System characteristics to be evaluated

- 2.1 System internal characteristics
    - 2.1.1 MT system-specific characteristics
    - 2.1.2 Translation process models
    - 2.1.3 Linguistic resources and utilities
    - 2.1.4 Characteristics of process flow
- 2.2 System external characteristics
    - 2.2.1 Functionality
        - 2.2.1.1 Suitability
        - 2.2.1.2 Accuracy
        - 2.2.1.3 Wellformedness
        - 2.2.1.4 Interoperability
        - 2.2.1.5 Compliance
        - 2.2.1.6 Security
    - 2.2.2 Reliability
    - 2.2.3 Usability
    - 2.2.4 Efficiency
    - 2.2.5 Maintainability
    - 2.2.6 Portability
    - 2.2.7 Cost

Each bullet above expands to a great deal of detail. For example, the first bullet becomes the following.

1.1 Purpose of evaluation

Definition

The purpose of evaluation is to provide a basis for making decisions.

Relevant qualities - from part 2

No qualities are listed at present for this taxon.

Stakeholders

If there is development, the development organization is always a stakeholder in every evaluation purpose.

References

Reeder and White 2002
White 2000
Arnold et al. 1994
Spark-Jones and Galliers 1996

Notes

- **1.1.1 Feasibility evaluation**
  Definition
  According to White (2000) a feasibility study is an evaluation of the possibility that a particular approach has any potential for success after further research and implementation. Feasibility evaluations provide results of interest to researchers and to sponsors of research. The characteristics that a feasibility evaluation typically tests for are functionality attributes such as the coverage of sub-problems particular to a specific language pair and the possibility of extending to more general phenomena (changeability).
  Relevant qualities - from part 2
  Coverage (2.2.1.1.2.1/504)
  Accuracy (2.2.1.2/177)
  Stakeholders

  Researcher, research sponsor.
- **1.1.2 Requirements elicitation**
  Definition
  Requirements discovery is often an iterative process in which developers create prototypes in order to elicit reactions from potential stakeholders. In so-called "rapid prototyping" approaches to requirements discovery, developers create prototypes designed to demonstrate specific aspects of functional capabilities that might ultimately be implemented. Scenario-based observational studies are often used to assess the utility of the functions demonstrated by the prototype.
  Relevant qualities - from part 2
  Characteristics of the intended mode of use (2.1.4/160)
  Utility (2.1.1.1.5/176)
  Usability (2.2.3/603)
  Stakeholders

  Developers. Project managers. End-users.

  References
  Connell and Shaffer, 1995.
- **1.1.3 Internal evaluation**
  Definition
  According to White 2000, internal evaluation occurs on continual or periodic bases in the course of research and development. Internal evaluations test whether, for example, the components of an experimental prototype or pre- release system work as they are intended.
  This type of evaluation mainly concerns functionality and needs to show coverage of the fundamental contrastive phenomena of the language pair, just like feasibility evaluation. However, at this point in a system's life cycle, it must also be shown that the system is actually improving as a result of development (changeability), and that improvement in one area does not make something else worse (stability). (In terms of EAGLES 1996, this is a progress evaluation).
  Relevant qualities - from part 2
  Translation process models (2.1.1/402)
  Coverage (2.2.1.1.2.1/504)
  Readability (2.2.1.1.1.1/172)
  Terminology (2.2.1.2.3/175)
  Accuracy (2.2.1.2/177)

Well-formedness (2.2.1.3/186)
Stakeholders

Developers. Sponsors. Development managers.

References
White 2000
Reeder dissertation (forthcoming)
King 1990
Notes

Diagnostic evaluation is arguably a subset of the internal evaluations; however, segregating the diagnostic needs at the same level as internal does improve the distinction between iterative progress tests and tests to find a specific problem.

- **1.1.4 Diagnostic evaluation**
Definition

EAGLES 1996 distinguishes a type of evaluation whose purpose was to discover why a system did not give the results it was expected to give. Typically performed by a researcher developing a prototype system, such an evaluation is almost exclusively concerned with functionality characteristics and will also often make use of internal metrics based on the intermediate results the system produces. Diagnostic evaluation typically uses glass-box evaluation principles.

Relevant qualities - from part 2
Translation process models (2.1.1/402)
Coverage (2.2.1.1.2.1/504)
Readability (2.2.1.1.1.1/172)
Terminology (2.2.1.2.3/175)
Accuracy (2.2.1.2/177)
Stakeholders

Developers. Development managers.

Notes

Internal evaluation may also be glass box, and it is also possible to use black-box evaluations to do diagnostics.
The property "glass/black box" does not spawn children, but is rather a property which distinguishes certain methods under more than one taxon.

- **1.1.5 Declarative evaluation**
Definition
According to White 2000, the purpose of declarative evaluation is to measure the ability of an MT system to handle texts representative of an actual end-user. It purports to measure the actual performance of a system external to the particulars of the feasibility of the approach or of the development process.
As with feasibility and internal evaluation, we look at coverage of linguistic phenomena and handling of samples of real text. However, these generally do not use constrained test patterns, and they are not directly used to determine the extensibility of the system, but how good it is right now. Declarative evaluations generally test for the functionality attributes of intelligibility, (how fluent or understandable it appears to be) and fidelity (the accurateness and completeness of the information conveyed).
Relevant qualities - from part 2
Translation process models (2.1.1/402)
Linguistic resources and utilities (2.1.2/403)
Suitability (2.2.1.1/168)
Accuracy (2.2.1.2/177)
Well-formedness (2.2.1.3/186)
Stakeholders

Researcher, research sponsor, developer, investor, marketing, purchasing

- **1.1.6 Operational evaluation**
Definition

According to White 2000, operational evaluations generally address the question of whether an MT system will actually serve its purpose in the context of its operational use. The primary factors include the cost-benefit of bringing the system into the overall process (costs).

Relevant qualities - from part 2

Linguistic resources and utilities (2.1.2/403)
Interoperability (2.2.1.4/192)
Reliability (2.2.2/600)
Maintainability (2.2.5/620)
Portability (2.2.6/622)
Cost (2.2.7/624)

Stakeholders

Researcher, research sponsor, developer, investor, marketing, purchasing

Notes

A variety of issues are considered here, including such things as software and hardware compatibility with the incumbent office automation system (interoperabililty). However, the more fundamental question to ask for operational use is whether the MT system enhances the effectiveness of the down stream task, or whether the end-to-end process is better off without it.

As an example, consider cross lingual information retrieval. Evaluation of MT embedded into a cross lingual information processing environment takes into account the measures that are germane to the downstream task. So if we want to know whether an MT system helps information extraction we compare the recall and precision (metrics germane to extraction) of the MT plus extraction configuration to an expert translation plus extraction process, or to an extraction without any translation at all. Note that we do not measure functionality characteristics of the MT system itself, such as fidelity and intelligibility, but rather the effect of the MT (good or bad) on the downstream task in term of that task's metrics. To a large extent then, operational evaluation lies outside the bounds of this classification, which is concerned only with the classification and evaluation of MT systems.

- **1.1.7 Usability evaluation**

Definition

According to White (2000), the purpose of usability evaluation is to measure the ability of a system to be useful to people who are actually going to use it. ISO 9126 talks of "quality in use" characteristics, which are the combination of other characteristics which will enable a user to achieve specified goals with effectiveness, productivity, safety and with satisfaction in a specified context use.

Relevant qualities - from part 2

Usability (2.2.3/603)
Efficiency (2.2.4/606)

Stakeholders

developer, investor, marketing, purchasing, operations mgt., end user

Notes

Usability evaluation is a domain in its own right, which involves kinds of testing such as scenario and laboratory testing which are common to many kinds of software product. It is often undertaken by the manufacturers of products before the product is launched on the market. It falls outside the scope of the current classification. However, much information about usability evaluation can be found by consulting the European Usability Support Centres home page.

Optional further reading

www.amtaweb.org; click on Links and Job Offers

Overview:

Hutchins, J.H. and H. Somers: *Machine Translation*. Academic Press, 1992.

Hovy, E.H. Overview article in MITECS (*MIT Encyclopedia of the Cognitive Sciences*). 1998.

Hovy, E.H. Review in *BYTE magazine*, January 1993.

Knight, K. 1997. Automating Knowledge Acquisition for Machine Translation. *AI Magazine* 18(4), (81–95).

Transfer:

Nagao, M., 1987, Role of Structural Transformation in a Machine Translation System. In *Machine Translation: Theoretical and Methodological Issues*, S. Nirenburg, ed. Cambridge: Cambridge University Press, pp. 262-277.

EBMT:

Nirenburg, S., S. Beale and C. Domashnev. 1994. A Full-Text Experiment in Example-Based Machine Translation. *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, England.

Somers, H. 2000. A Review of EBMT. *Machine Translation* 15(4).

Interlinguas:

Dorr, B.J. 1994. Machine Translation Divergences: A Formal Description and Proposed Solution. *Computational Linguistics* 20(4) (597-634).

Nirenburg, S., J.C. Carbonell, M. Tomita, and K. Goodman. 1992. *Machine Translation: A Knowledge-Based Approach*. San Mateo: Morgan Kaufmann.

Multi-Engine MT:

Frederking, R., S. Nirenburg, D. Farwell, S. Helmreich, E.H. Hovy, K. Knight, S. Beale, C. Domanshnev, D. Attardo, D. Grannes, R. Brown. 1994. Integrating Translations from Multiple Sources within the Pangloss Mark III Machine Translation System. *Proceedings of the First AMTA Conference*, Columbia, MD (73-80).