APPLICATIONS:

MACHINE TRANSLATION I

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

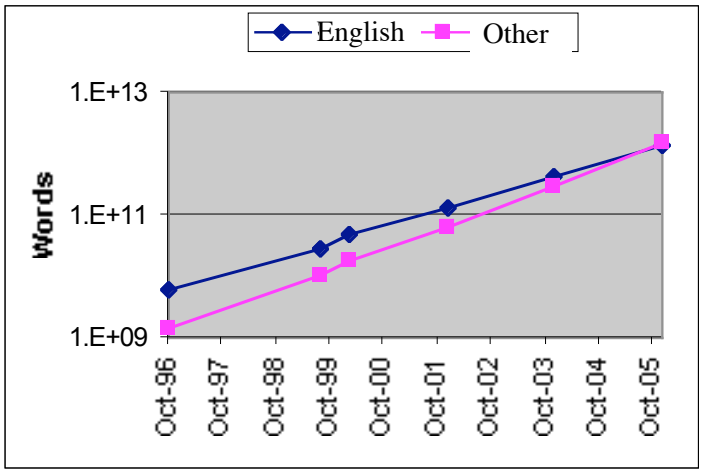
1. History

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

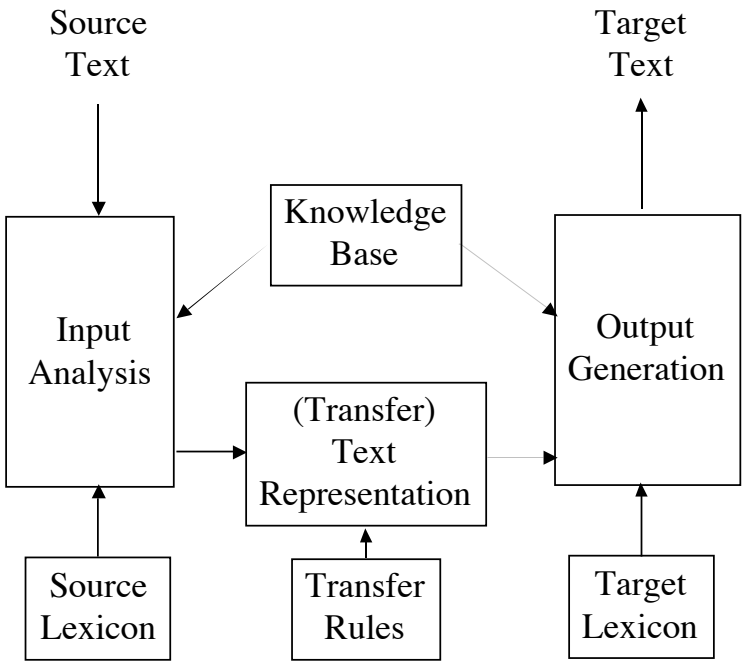
Automation Tradeoffs

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

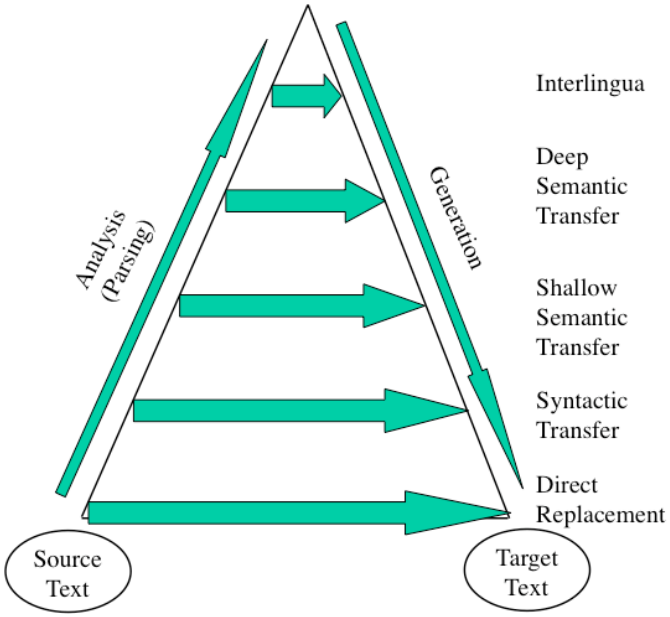
Human-assisted
- More expensive (>10c a page...)
- Slow
- High(er) quality
- Editing tools, dictionaries, etc., required
- Most common commercial configuration

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.

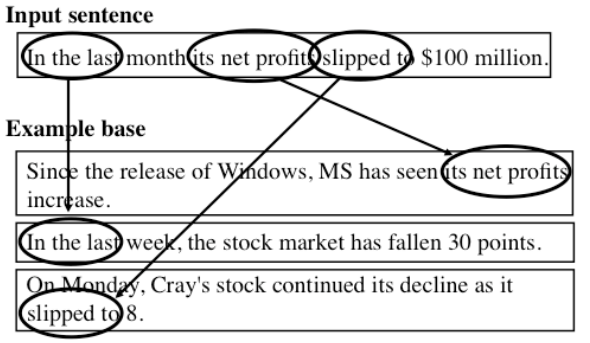
If you are smart, you notice that some words are replaced by no words, and some by more than one word (the 'word fertility'), and also that some words move around relative to others ('distortion'). The earliest simple MT systems of the 1950s rapidly went beyond Direct Replacement, but in the early 1990s 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



Observations

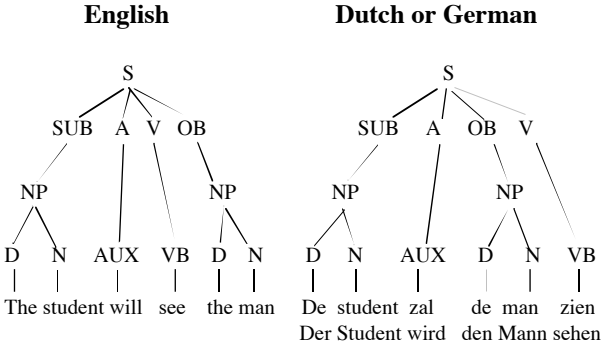
- Normalizing gives better coverage
- BUT: Greedy algorithm does not always find best covering
- Open research problem

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).

The future: Web-based MT.

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.
- 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.
- 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.

5. 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

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).