Introduction to Computational Linguistics

Frank Richter

fr@sfs.uni-tuebingen.de.

Seminar für Sprachwissenschaft

Eberhard-Karls-Universität Tübingen

Germany

How to Choose the Best MT Strategy

- If low quality translation is acceptable and if source and target language have similar syntax, then a direct translation system may be acceptable.
- If the system will only translate between two languages and good-quality translation is necessary, a transfer system is all that is needed.
- If the system will have to translate among several languages, an interlingua approach may be preferable, especially if the languages are from the same language family and have similar patterns of word meanings.

The Impossibility of FAHQMT

The Impossibility of Fully Automatic, High Quality Machine Translation (FAHQMT):

Little John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy.

(Bar-Hillel 1959)

Machine Translation (1)

- full machine translation (MT)

Machine Translation (1)

- full machine translation (MT)
- human-aided machine translation (HAMT)

Machine Translation (1)

- full machine translation (MT)
- human-aided machine translation (HAMT)
- machine-aided human translation (MAHT)

Full Machine Translation

- machine is responsible for the entire translation process.

Full Machine Translation

- machine is responsible for the entire translation process.
- minimal pre-processing by humans, if any.

Full Machine Translation

- machine is responsible for the entire translation process.
- minimal pre-processing by humans, if any.
- no human intervention during the translation process.

Full Machine Translation

- machine is responsible for the entire translation process.
- minimal pre-processing by humans, if any.
- no human intervention during the translation process.
- post-processing by humans may be required.

Human-aided Machine Translation (HAMT)

- machine is responsible for translation production

Human-aided Machine Translation (HAMT)

- machine is responsible for translation production
- translation process may be aided by human monitor; e.g. for:

Human-aided Machine Translation (HAMT)

- machine is responsible for translation production
- translation process may be aided by human monitor; e.g. for:
  - part-of-speech disambiguation

Human-aided Machine Translation (HAMT)

- machine is responsible for translation production
- translation process may be aided by human monitor; e.g. for:
  - part-of-speech disambiguation
  - resolving for phrase attachment

Human-aided Machine Translation (HAMT)

- machine is responsible for translation production
- translation process may be aided by human monitor; e.g. for:
  - part-of-speech disambiguation
  - resolving for phrase attachment
  - choosing appropriate word for the target language from a set of candidate translations

Machine-aided Human Translation (MAHT)

- human is responsible for translation production

Machine-aided Human Translation (MAHT)

- human is responsible for translation production
- human translation is aided by on-line tools; e.g. by

Machine-aided Human Translation (MAHT)

- human is responsible for translation production
- human translation is aided by on-line tools; e.g. by
  - a corpus of sample translations

Machine-aided Human Translation (MAHT)

- human is responsible for translation production
- human translation is aided by on-line tools; e.g. by
  - a corpus of sample translations
  - electronic dictionaries for source and target language

Machine-aided Human Translation (MAHT)

- human is responsible for translation production
- human translation is aided by on-line tools; e.g. by
  - a corpus of sample translations
  - electronic dictionaries for source and target language
  - a terminology database

Machine-aided Human Translation (MAHT)

- human is responsible for translation production
- human translation is aided by on-line tools; e.g. by
  - a corpus of sample translations
  - electronic dictionaries for source and target language
  - a terminology database
  - word processing support for text formatting

The History of Machine Translation (1)

1629 René Descartes proposes a universal language, with equivalent ideas in different tongues sharing one symbol.

1933 Russian Petr Smirnov-Troyanskii patents a device for transforming word-root sequences into their other-language equivalents.

1949 Warren Weaver, director of the Rockefeller Foundation's natural sciences division, drafts a memorandum for peer review outlining the prospects of machine translation (MT).

The History of Machine Translation (2)

1952 Yehoshua Bar-Hillel, MIT's first full-time MT researcher, organizes the maiden MT conference.

1954 First public demo of computer translation at Georgetown University: 49 Russian sentences are translated into English using a 250-word vocabulary and 6 grammar rules.

1960 Bar-Hillel publishes his report arguing that fully automatic and accurate translation systems are, in principle, impossible.

The History of Machine Translation (3)

1964 The National Academy of Sciences creates the Automatic Language Processing Advisory Committee (Alpac) to study MT's feasibility.

1966 Alpac publishes a report on MT concluding that years of research haven't produced useful results. The outcome is a halt in federal funding for machine translation R&D.

The History of Machine Translation (4)

1968 Peter Toma, a former Georgetown University linguist, starts one of the first MT companies, Language Automated Translation System and Electronic Communications (Latsec).

1969 In Middletown, New York, Charles Byrne and Bernard Scott found Logos to develop MT systems.

Machine Translation Systems

North America and Canada

- SYSTRAN
  - Originated from GAT (Georgetown Machine Translation project)
  - Founded in 1968 by Peter Toma, a principal member of the GAT project
  - Versions for English, German, Russian, French, Spanish, Dutch and Portugese
  - Purchased by Major Corporations and Government Agencies for further development, including General Motors, Xerox, Siemens, European Commission

Machine Translation Systems

- TAUM-METEO
  - TAUM: Traduction Automatique de l'Universtite de Montreal
  - Fully-automatic MT system METEO
  - Fully integrated into the Canadian Meteorological Center's (CMC) nation-wide weather communications network by 1977
  - Translates appr. 8.5 million words/year with 90-95% accuracy. Mistakes mainly due to misspelled input or unknown words

Machine Translation Systems: Europe

- EUROTRA
- Long-term MT research and development program funded by the European Commission (1982-92)
- EUROTRA 1 - Research and development programme (EEC) for a machine translation system of advanced design, 1982-1990
- EUROTRA 2 - Specific programme (EEC) concerning the preparation of the development of an operational EUROTRA system, 1990-1992

MT Systems: EUROTRA 1

- EUROTRA 1 - Research and development programme (EEC) for a machine translation system of advanced design, 1982-1990
  - Main Goal: To create a machine translation system of advanced design capable of dealing with all (nine) official languages at the time (Danish, Dutch, English, French, German, Greek, Italian, Spanish and Portuguese) of the Community by producing an operational system prototype in a limited field and for limited categories of text, which would provide the basis for subsequent development on an industrial scale.

MT Systems: EUROTRA 2

- EUROTRA 2 - Specific programme (EEC) concerning the preparation of the development of an operational EUROTRA system, 1990-1992
  - Main Goal: To create, starting from the EUROTRA prototype, the appropriate conditions for a large-scale industrial development, including the development of methods and tools for the re-usability of lexical resources in computer applications as well as the creation of standards for lexical and terminological data.

Machine Translation Systems: GETA

- GETA (Group d' Etudes pour la Transduction Automatique) at the University of Grenoble, France
- MT research group with longest history in Europe, if not world-wide,
- headed by Bernard Vauquois and later by Christian Boitet
- Systems developed:
  - 1967-1971 development of CETA (Russian/French):
  - ARIANE -78

Machine Translation Systems: CETA

- CETA (Russian/French):
  - first large-scale second-generation system (first-generation systems aimed at direct translation) with finite- state morphology, augmented context-free syntactic analysis with assignment of dependency relations, procedural semantic analysis tranforming tree structures into an interlingua (pivot language), lexical transfer, syntactic generation and morphological generation.

MT Systems: ARIANE-78

- ARIANE-78
  - emphasis on flexibility and modularity
  - powerful tree-transducers written in transfer-rule formalism ROBRA
  - conception of static and dynamic grammars
  - Different levels and types of representation (dependency, phrase structure, logical) incorporated on single labelled tree structures and thus considerable flexibility in multilevel transfer representations.

MT Systems: Verbmobil

- Verbmobil
  - A speaker-independent and bidirectional speech-to-speech translation system for spontaneous dialogs in mobile situations.
  - Recognizes spoken input, analyses and translates it, and finally utters the translation.
  - The multilingual system handles dialogs in three business-oriented domains (appointment scheduling, travel planning, remote PC maintenance) with context-sensitive translation between three languages (German, English, and Japanese).

MT Systems: Verbmobil

- Verbmobil
  - Travel planning scenario with a vocabulary of 10 000 words was used for the end-to-end evaluation of the final Verbmobil system
  - integrates a broad spectrum of corpus-based and rule-based methods.
  - combines the results of machine learning from large corpora with hand-crafted knowledge sources to achieve an adequate level of robustness and accuracy.

Langenscheidt's T1 Text Translator

- T1 is a commercial product that builds on the METAL system.
- T1 is bi-directional: translates from English into German and German into English; French into German and German into French; and German into Russian and Russian into German.
- T1 is flexible. It provides users with a number of different translation methods to choose from: batch translation and real-time on-screen translation.

T1's Resources and Functionality

- T1 has a big general purpose lexicon of 450 000 word forms; with domain-specific sublexica to choose from.
- T1 supports a dynamic system lexicon which can be enriched by the user, including grammatical information and multi-word expressions. Supported by an intelligent lexicon editor.
- Larger external dictionary for lexical lookup.

T1's Translation Options

- For individual sentences or short texts you can use the ScratchPad, and watch the actual translation process.
- For longer texts and RTF documents, you can translate from the Workspace. The draft translations retain the format of the original documents, and you can specify where you want the results to be stored. A useful feature here is the Translation Queue. This allows you to queue your documents for translation at a more convenient time.

T1's Translation Workspace

The advantages of translating in the Workspace are:

- you can translate RTF documents as well as ASCII and HTML documents.
- you can queue documents for translation at a more convenient time.
- you retain the layout and formatting of the original document.
- you can create a New Words List and add it to the lexicon.

Machine Translation on the Internet

Several search engines offer language support:

- Google offers a beta-version machine translation window
  `http://www.google.de/language_tools`
- Altavista offers Babelfish translator
  `http://de.altavista.com/babelfish`
  developed by Systran
  `http://www.systransoft.com`
- Both engines offer type-in windows for translation of short texts and translation of web sites.

MT: Performance Google/Altavista (1)

- Maria hat dem Kind ein Buch gegeben.
  Maria gave a book to the child.

MT: Performance Google/Altavista (1)

- Maria hat dem Kind ein Buch gegeben.
  Maria gave a book to the child.
- Ich glaube nicht, dass diese Maschine gute Übersetzungen liefern kann.
  I do not believe that this machine can supply good translations.

MT: Performance Google/Altavista (1)

- Maria hat dem Kind ein Buch gegeben.
  Maria gave a book to the child.
- Ich glaube nicht, dass diese Maschine gute Übersetzungen liefern kann.
  I do not believe that this machine can supply good translations.
- Wenn man einen Satz aus der Zeitung nimmt, dann müßte das Programm ihn übersetzen können.
  If one takes a sentence from the newspaper, then the program would have to be able to translate him.

MT: Performance Google/Altavista (2)

- Peter hat den Löffel abgegeben.
  Peter delivered the spoon.

MT: Performance Google/Altavista (2)

- Peter hat den Löffel abgegeben.
  Peter delivered the spoon.

- Das ist nicht der Grund dafür, dass ich ihm nicht traue.
  That is not the reason for the fact that I do not trust it.

Some Misconceptions about MT (1)

- **False:** MT is a waste of time because you will never make a machine that can translate Shakespeare.

Some Misconceptions about MT (1)

- **False:** MT is a waste of time because you will never make a machine that can translate Shakespeare.
- **False:** There was/is an MT system which translated *the spirit is willing, but the flesh is weak* into the Russian equivalen of *The vodka is good, but the steak is lousy*, and *hydrailic ram* into the French equivalent of *water goat*. MT is useless.

Some Misconceptions about MT (2)

- **False:** Generally, the quality of translation you can get from an MT system is very low. This makes them useless in practice.

Some Misconceptions about MT (2)

- **False:** Generally, the quality of translation you can get from an MT system is very low. This makes them useless in practice.
- **False:** MT threatens the jobs of translators.

Some Misconceptions about MT (2)

- **False:** Generally, the quality of translation you can get from an MT system is very low. This makes them useless in practice.
- **False:** MT threatens the jobs of translators.
- **False:** The Japanese have developed a system that you can talk to on the phone. It translates whatever you say into Japanese, and translates the other speaker's replies into English.

Incremental Linguistic Analysis

- tokenization
- morphological analysis (lemmatization)
- part-of-speech tagging
- named-entity recognition
- partial chunk parsing
- full syntactic parsing
- semantic and discourse processing