Introduction to Computational Linguistics

Machine Translation

Jan-Philipp Söhn

jp.soehn@uni-tuebingen.de

November 21/28, 2007

Introduction – An example for a machine-translated text

First, the German text to be translated:

> *Der HERR ist mein Hirte, mir wird nichts mangeln. Er lagert mich auf grünen Auen, er führt mich zu stillen Wassern. Er erquickt meine Seele. Er leitet mich in Pfaden der Gerechtigkeit um seines Namens willen. (The Bible, Psalm 23/1–3, Rev. Elberfelder Translation)*

Second, the English reference:

> *The LORD is my shepherd, I shall not be in want. He makes me lie down in green pastures, he leads me beside quiet waters, he restores my soul. He guides me in paths of righteousness for his name's sake. (The Bible, Psalm 23/1–3, New International Version)*

(This is not a translation from German. Both texts are translated from the same Hebrew source.)

Introduction – An example for a machine-translated text

Third, the translation by Babelfish:

> *The GENTLEMAN is not my Hirte, me anything will lack. It stores me on green Auen, it leads me to quiet landing on water. He refreshes my soul. It leads me in paths of the justice around its name sake.*

Forth, the translation by Google:

> *The Lord is my shepherd, I will not pose a problem. He deposited me on the green woods, it leads me to quiet waters. It refreshed my soul. He guides me in paths of justice for his name's sake.*

Machine Translation

What makes Machine Translation an important application area to study:

- historically first application area, and for at least a decade the only application area, of computational linguistics
- requires all steps relevant to linguistic analysis of input sentences and linguistic generation of output sentences
- hence, machine translation is scientifically one of the most challenging and most comprehensive tasks in computational linguistics

The Purposes of Translation

- **Information Acquisition:**
  - e.g. Gather information on scientific articles or newspapers written in a foreign language.
- **Information Dissemination:**
  - e.g. Translation of technical manuals, legal texts, weather reports, etc.
- **Literary Translation:**
  - e.g. Translation of novels, poems, etc.

Relating Translation Purposes to MT

- **Information Acquisition:**
  - involves translation from a foreign to a native language
  - typically used by non-linguists with little or no linguistic competence in the source language
  - pre-processing of the input not feasible due to lack of linguistic competence by the user in the source language
  - may require special-purpose lexicon
  - low-quality translation is tolerable

Relating Translation Purposes to MT(2)

- **Information Dissemination:**
  - involves translation from a native to a foreign language
  - pre- and post-processing of the input feasible due to linguistic competence by the translator in the source language
  - may involve sublanguage with restricted vocabulary; e.g. translation of weather reports
  - often involves special terminologies stored in a terminology database; e.g. for translation of technical manuals
  - purely human translation for such tasks can be time-consuming, inconsistent, or tedious.

Relating Translation Purposes to MT(3)

- **Literary Translation**
  - requires stylistic elegance, often involves metaphorical and metonymic language
  - abundance of highly-trained human translators
  - task rarely performed by machine translation

What Makes Machine Translation Hard

- Lexical Ambiguity
- Lexical Gaps
- Syntactic Divergences between Source and Target Language

Problems: Word-to-Word Translations

English – German

The	ticket office	in the	train station
Der	Fahrkartenschalter	im	Bahnhof

öffnet	wieder	um	ein	Uhr.
re-opens		at	one	o'clock.

Lexical Ambiguity: Open (1)

English	**German**
in store door	Offen
on new building	Neu eröffnet
open door	Tür öffnen
open golf tourney	Golfspiel eröffnen
open question	offene Frage
open job	freie Stelle
open morning	freier Morgen
open football player	freier Fussballspieler

Lexical Ambiguity: Open (2)

English	**German**
loose ice	offenes Eis
blank endorsement	offenes Giro
private firm	offene Handelsgesellschaft
unfortified town	offene Stadt
blank cheque	offener Wechsel
to unbutton a coat	einen Mantel öffnen

Structural Divergence (1)

English – German

Max likes to swim.
NP VFIN INF

Max schwimmt gerne.
NP VFIN ADV

Structural Divergence (2)

Russian – English

Jego zovut Julian.
Him they call Julian.
They call him Julian.

Japanese – English

Kino ame ga futa.
Yesterday rain fell.
It was raining yesterday.

Differences in Word Order

English – German

Does it make sense to translate
Macht es Sinn

documents automatically ?
Dokumente automatisch zu übersetzen ?

Better German translation is even harder:
Ist es sinnvoll, Dokumente automatisch zu übersetzen?

MT: The Weaver Memo (1)

Translation and Context

> *If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words.*
>
> *But if one lengthens the slit in the opaque mask, until one see not only the central word in question but also say N words on either side, then if N is large enough one can unambiguously decide the meaning of the central word.*

MT: The Weaver Memo (2)

Translation and Context

> *The practical question is: "What minimum value of N will, at least, in a tolerable fraction of cases, lead to the correct choice of meaning for the central word?"*

Translation and Cryptography

> *... it is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the "Chinese code".*

MT: The Weaver Memo (3)

Translation and Language Universals (Invariants)

> *... there are certain invariant properties which are, again not precisely, but to some statistically useful degree, common to all languages.*
> *Thus may it be true that the way to translate Chinese to Arabic or from Russian to Portuguese, is not to attempt the direct route ... but down to the common base of human communication – the real but yet undiscovered universal language – and then to re-emerge by whatever particular route is convenient.*

Strategies for Machine Translation

Strategies for Machine Translation

- Word-to-Word (Direct) Translation
- Syntactic Transfer
- Semantic Transfer
- Interlingua Approach

Strategies for Machine Translation (2)

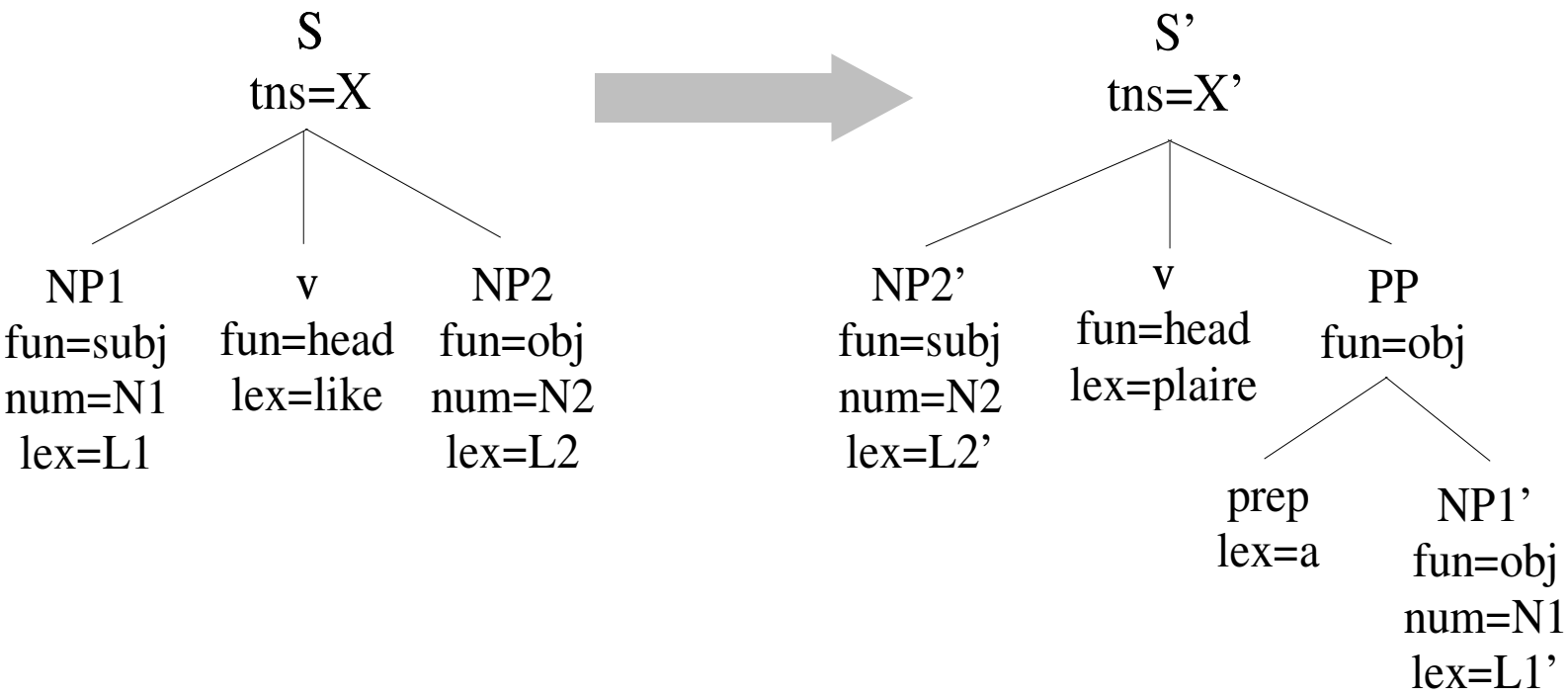
- Word-to-Word (Direct) Translation
  - simplest approach:
  - may require only an electronic, bi-lingual dictionary
  - depending on the source and target languages and the dictionary, minimal morphological analysis and generation may be required.
  - no use of syntactic or semantic knowledge

Strategies for Machine Translation (3)

- Syntactic Transfer
  - requires syntactic analysis of the source language
  - requires a syntactic parser
  - language-dependent rules that relate source language trees to target language trees.

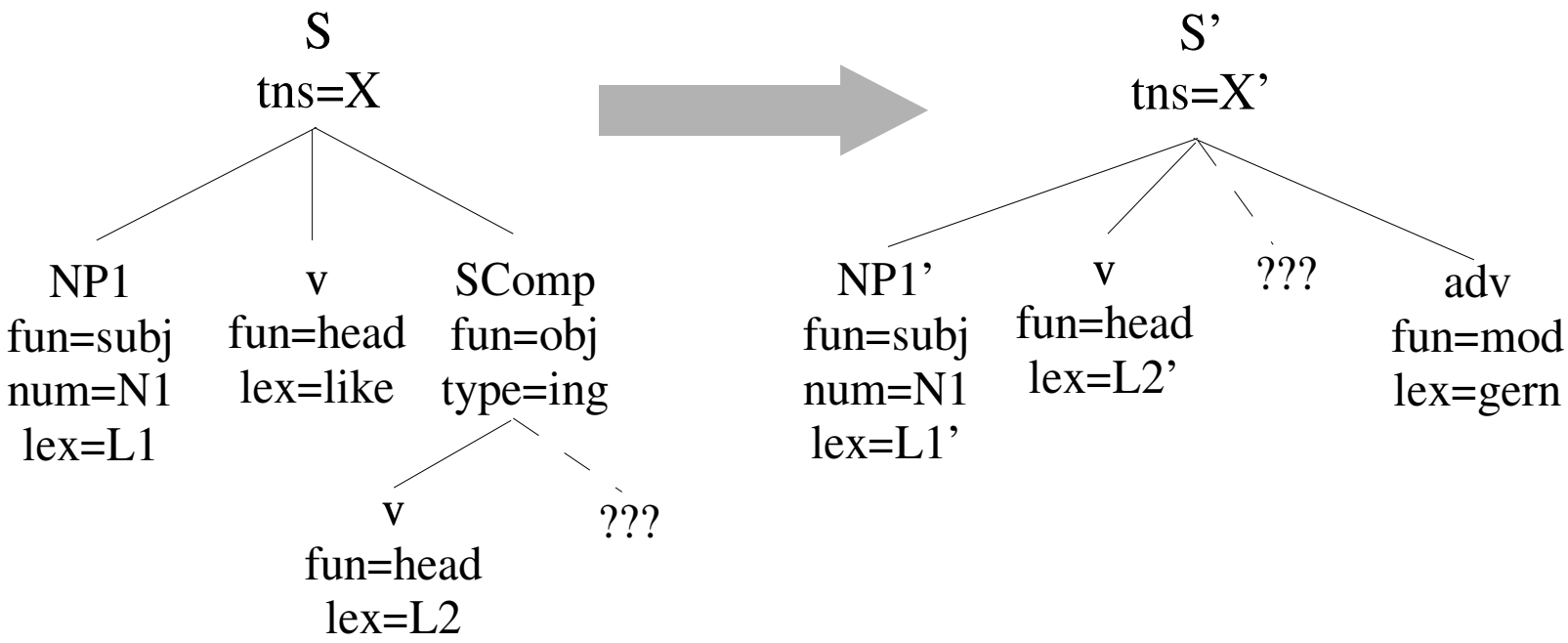
Syntactic Transfer Trees

An Example of a Transfer Tree for English *like* and French *plaire*



Syntactic Transfer Trees (2)

An Example of a Transfer Tree for English *like to* ⟨ *V* ⟩ and German ⟨ *V* ⟩ *gern*



Strategies for Machine Translation (4)

- Semantic Transfer
  - requires syntactic and semantic analysis of the source language
  - requires language-dependent meaning representation language
  - language-dependent rules that relate source language meaning representations to target language meaning representations
  - requires language generation component which maps target language meaning representations to output sentences

Strategies for Machine Translation (5)

- Semantic Transfer
  - synthesis typically performed in two stages: semantic synthesis (resulting in syntactic trees) and morphological synthesis (resulting in strings of inflected word forms).

Strategies for Machine Translation (5)

- Interlingua Approach
  - source language input is mapped to a language-neutral (quasi-universal) meaning representation language
  - requires syntactic and semantic analysis of the source language into interlingua
  - requires language generation component which maps interlingua to output sentences
  - synthesis typically performed in two stages: semantic synthesis from the interlingua (resulting in syntactic trees) and morphological synthesis (resulting in strings of inflected word forms).

Interlingua Representation for Motion Verbs

He walked across the road.
Il traversa la rue a pied.

$$\begin{bmatrix} \textsc{Pred} & = \langle \textsc{motion} \rangle \\ \textsc{Tense} & = \textit{past} \\ \textsc{Agent} & = \begin{bmatrix} \textsc{Pred} = \textsc{Pron} \\ \textsc{Num} = \textsc{sing} \\ \textsc{Pers} = 3 \\ \textsc{Sex} = \textsc{male} \end{bmatrix} \\ \textsc{Instr} & = \begin{bmatrix} \textsc{Pred} = \langle \textsc{foot} \rangle \end{bmatrix} \\ \textsc{Loc} & = \begin{bmatrix} \textsc{Pred} = \langle \textsc{cross} \rangle \\ \textsc{Obj} = \begin{bmatrix} \textsc{Pred} = \langle \textsc{road} \rangle \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Interlingua Representation for Motion Verbs (2)

They flew from Gatwick.
Ils partirent par avion de Gatwick.

$$\begin{bmatrix} \textsc{Pred} & = \langle \textsc{motion} \rangle \\ \textsc{Tense} & = \textit{past} \\ \textsc{Agent} & = \begin{bmatrix} \textsc{Pred} = \textsc{Pron} \\ \textsc{Num} = \textsc{plur} \\ \textsc{Pers} = 3 \end{bmatrix} \\ \textsc{Instr} & = \begin{bmatrix} \textsc{Pred} = \langle \textsc{plane} \rangle \end{bmatrix} \\ \textsc{Loc} & = \begin{bmatrix} \textsc{Pred} = \langle \textsc{leave} \rangle \\ \textsc{Obj} = \begin{bmatrix} \textsc{Pred} = \textsc{Gatwick} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Interlingua Representation for Verbs (1)

English	*wall*	German	*Wand* (inside a building) *Mauer* (outside)
English	*river*	French	*riviere* (general term) *fleuve* (major river, flowing into sea)

Interlingua Representation for Verbs (2)

English	*leg*	Spanish	*pierna* (human)
			pata (animal,chair)
			pie (table)
			etapa (of a journey)
		French	*jambe* (human)
			patte (animal,insect)
			pied (chair,table)
			étape (journey)

Interlingua Representation for Verbs (3)

English	*blue*	Russian	*goluboi* (pale blue) *sinii* (dark blue)
French	*louer*	English	*hire* or *rent*
French German	*colombe* *Taube*	English	*pigeon* or *dove*
German	*leihen*	English	*borrow* or *lend*

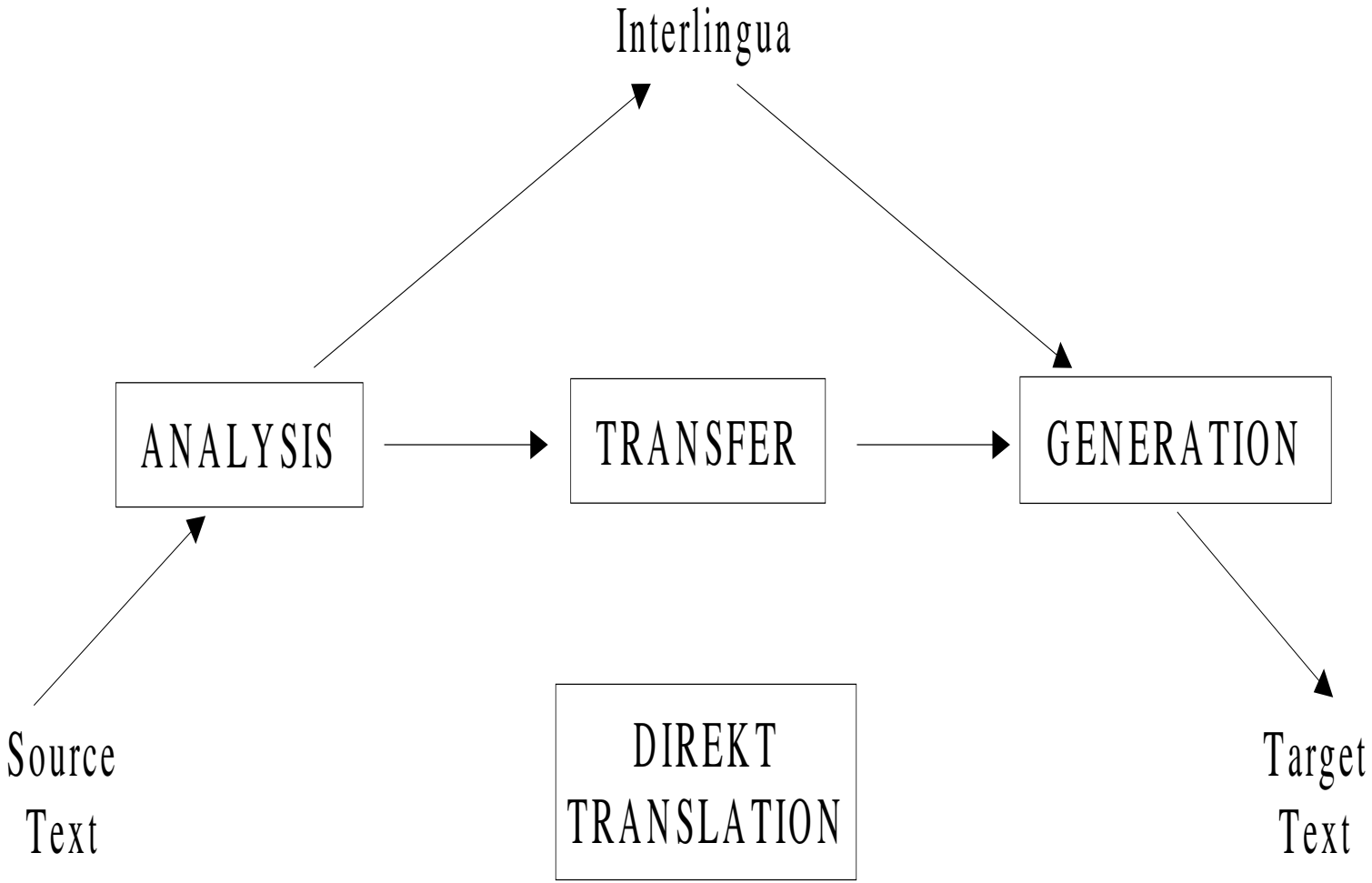
Interlingua Representation for Verbs (4)

English *rice* Malay *padi* (unharvested grain)
beras (uncooked)
nasi (cooked)
emping (mashed)
pulut (glutinous)
bubor (cooked as a gruel)

Interlingua Representation for Verbs (5)

English	*wear*	Japanese	*kiru* (generic)
			haoru (coat or jacket)
			haku (shoes or trousers)
			kaburu (hat)
			hameru (ring or gloves)
			shimeru (belt or tie or scarf)
			tsukeru (brooch or clip)
			kakeru (glasses or necklace)

The Vauquois Triangle



Strategies for Machine Translation

Modules required in an all-pairs MTS

Number of languages	Analysis modules	Generation modules	Transfer modules	Total modules
2	2	2	2	6
3	3	3	6	12
4	4	4	12	20
5	5	5	20	30
...				
9	9	9	72	90
n	n	n	n(n-1)	n(n+1)

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

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
- 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
- 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

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'Université de Montréal
  - 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 Traduction 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 machine translation tool
  `http://www.google.com/language_tools`
- Altavista and Yahoo offer the Babel Fish 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.

Conclusion

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 *hydraulic 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.
- **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.

Some Misconceptions about MT (3)

- **False:** There is a amazing South American Indian language with a structure of such logical perfection that it solves the problem of designing MT systems.
- **False:** MT systems are machines, and buying an MT system should be very much like buying a car.