

Carnegie Mellon University
Spring 2007

11-731 Machine Translation

Commercial MT

Additional Instructor: Bob Frederking

Phone: 268-6656

Email: ref@cs.cmu.edu **Office Hours:** By
appointment

MT Companies:

We will examine three in some detail (ordered oldest to newest):

- SYSTRAN
- METEO
- METAL

(Historically interesting ones; not even trying to keep up-to-date on current companies.)

We will not look at a number of others:

- IBM Machine Translation
- Pan American Health Organization (PAHO)
- AppTek
- Sakhr
- etc.

The commercial MT world

- What counts as “commercial”?
 - Selling a software product/service
 - In-house systems at major companies
 - Translation bureaus using software
 - PAHO? (Only licensed to non-commercial entities)
 - Free MT on web? (Google!)

Many different brands on web are really SYSTRAN under the hood!

- Commercial world has been highly volatile compared to academia
 - Normal free-market “creative destruction”
 - The L&H debacle

IP generally does not disappear, but may sit on a shelf for quite a while

Range of Approaches to Users' Needs

- Fully custom system (METEO, Catalyst)
- Base system + customization (SYSTRAN, METAL)
- Off-the-shelf (Langenscheidt T1)

SYSTRAN History

- Grew out of earliest MT work in the late 1950s (Georgetown Rus-Eng system)
- Principal architect, Peter Toma, left Georgetown in 1962 to set up his own company
- AUTOTRAN, TECHNOTRAN: atomic energy, medicine
- Development of SYSTRAN began in Germany in 1964
- New company Latsec Inc in La Jolla, CA, developed Rus-Eng for USAF starting 1968; tested early 1969 at Wright-Patterson AFB, adopted in 1970 by Foreign technology Division
- Used by NASA during Apollo-Soyuz space project, 1974-1975
- Eng-Frn demo for CEC (ancestor of EU) in 1975
- Funding by CEC for Eng-Frn, Frn-Eng, Eng-Ita; by March 1981 prod. center in Luxembourg
- CEC developed additional language pairs over time

SYSTRAN History (2)

- Numerous companies set up to develop/promote SYSTRAN:
 - Systran Institut (Germany)
 - World Translation Corp. (Canada)
 - Systran Corp. (Japan)
- In 1986, Gachot acquired the rights to all except Japan (IONA Co.)
- Many language pairs (as of 1992, see Table 10.1 in H&S)
- Many large users (as of 1992, see Section 10.5 in H&S)

SYSTRAN Architecture

- Relatively high degree of modularity
- Two main types of programs:
 - system programs written in assembler, independent of languages (control and utilities)
 - translation programs broken into several different stages, each with different modules
- Analysis/generation independence claimed
- Common Romance-language analysis components
- Modularity added in the Gachot era to support introduction of new techniques, languages

SYSTRAN Architecture (2)

- Main component: bilingual dictionaries
 - lexical equivalences
 - grammatical, semantic info
 - dictionaries contain algorithms, too
- Dependence on language pairs:
 - Bilingual dictionaries have some reusability when source/target are reversed or used with other sources/targets
 - Because the lexicons contain algorithms, they can be complex and hard to maintain/extend

SYSTRAN Dictionaries

- Main Stem dictionary
 - Bilingual single-word dictionary
 - Full source info (POS, morph, semantics)
 - Ambiguous homographs represented
 - Single translation only
- Various multi-word contextual dictionaries
 - (see next slide)

SYSTRAN Dictionaries (2)

Various multi-word contextual dictionaries

- Idiom dictionary: “in order to”
- Limited Semantics dictionary: “hydraulic brake” + “fluid”
- Homograph dictionary: “prendre un chapeau” vs. “prendre note”
- Analytic dictionary: exceptions to general syntactic rules; “nor could he see” vs. “or he could see”
- Conditional Semantics dictionary: intervenes at the final stage of transfer to make final target lexical selection: “grow” as “elever” (animate), “cultiver” (plant), or “grandir” (general)
 - can be quite large: 400 entries for “oil”, “huile” vs. “petrole”

SYSTRAN Data Representation

- “byte areas” corresponding to each word in the input
 - essentially, a vector of values for each word (POS, semantics, etc.)
 - some linking between byte areas (for rudimentary grammatical relations, such as agreement)

SYSTRAN Translation Process

- Multi-step process
 - Pre-processing, Analysis, Transfer, Synthesis
- Disambiguation
 - “by near context” homograph rules
 - if unresolved, choose POS with highest frequency
- Details: Sec. 10.2.3, Diagram: Fig 10.1.

SYSTRAN Characteristics

- A synthesis of Direct and Transfer MT approaches
- Architecture changed over the years
 - increasing modularity
 - less direct, more transfer
- Use of many bilingual lexicons: lingering trait of direct systems
- Lack of full linguistic analysis
 - “pieces” of linguistic analysis stored in the lexicon
- Inconsistent assignment of semantic features
- “First Generation” system
 - primitive data structures, low-level programming, lack of declarative rule-writing mechanisms
 - “methods are inconsistent, coverage and quality are uneven, and modifications of lexical information can often have unexpected consequences” (p. 186, H&S)

METEO

- Developed by TAUM group, Montreal
- Translated weather bulletins from English to French
- In operation 1977 to present
- Success! Completeness and accuracy achieved by sublanguage restriction

METEO Task

- Ideal task for MT:
 - boring, highly repetitive
 - low job satisfaction, high turnover
- Volume: 37k words per day!
- Format highly structured and invariant
- Vocabulary: fixed, predictable (unknown words = errors)
- Sample input: Fig. 12.1.

METEO Steps

- Pre-processing: segmentation, regularization (Fig. 12.2)
- Linguistic data: bilingual dictionaries for idioms, place names, and domain vocabulary
- Processing modules: Eng analysis, Frn generation, Frn morphology
- Direct MT: telegraphic style is identical in Eng/Frn
- Dictionary look-up
 - plural forms entered directly
 - semantic features: time-point, phenomenon-falling, etc. (Table 12.1)
- Syntactic analysis
 - only 5 kinds of tree structure!
 - simple bottom-up parsing
- Syntactic generation
 - surface ordering
 - preposition selection
- Morphological generation

METEO Architecture

- Production system: single data structure and rule-writing formalism for all modules
- Chart data structure
- Bottom-up, breadth-first parsing
- “Q-systems”: early prototype of Prolog by Colmerauer
- Details: Sections 12.4.1-3

METEO Characteristics

- Second Generation system
 - explicit formalisms, full analysis / generation
 - linguistic strategies “well-motivated and finely tuned”
- Domain-specific direct system
 - narrow domain
 - single language pair
 - single direction of translation (initially)
- “First successful sublanguage system”

METAL History

- Ger-Eng system introduced by Siemens in 1989
- Transfer system, in Lisp, based on research since 1960s at UT Austin
- In 1992, plans for Eng-Ger, Dch-Frn, Frn-Dch, Ger-Spn, Ger-Frn, Ger-Dan
- Spun of as Sietec in mid-1990s, then LANT
- Now “T1”, from Langenscheidt

METAL Steps

- Text Acquisition
- Deformatting (tables, diagrams, etc.)
- Pre-analysis (unknown words and compounds; known words)
- Dictionary lookup
- Translation
- Reformatting
- Post-editing
- Output

METAL Knowledge

- Dictionaries
 - arranged in an explicit hierarchy
 - function words, general, common, technical (standard)
 - user-specialized glossaries, in a specified order
 - country-specific glossaries
 - “full” lexical entries (incl. semantic role info)
 - Intercoder tool for lexicon development (Figs. 15.2-3)
- Grammar rules
 - unordered sets of context-free phrase structure rules, augmented by tests, conditions, and specified output structures
 - formulated as Lisp functions
 - Metalshop development tool (Fig. 15.4)

METAL Translation Process

- Analysis
 - lexical, morphological analysis
 - handling of unknown terms
 - prioritized chart parser, prunes unlikely paths
= “some paths” approach
- Transfer
 - rules identified for each node built
 - lex and struct rules applied top-down
- Generation
 - morphology only

METAL Characteristics

- “Modified Transfer” system
 - separation of dictionaries, linguistic analysis
 - full syntactic analysis
 - transfer and generation processes not separated
 - analysis and transfer rules interdependent
- Most sophisticated commercial system as of 1992, in approach and support tools

New developments in 2006

- Several Statistical systems going commercial:
 - LanguageWeaver (Knight/Marcu)
 - Google (some of its languages; free)
 - IBM (they claim, but ads still seem to feature transfer)
- Meaningful Machines “Context-Based MT”
(Carbonell)

The End

(What follows are descriptions of other systems, not covered this time.)

LOGOS

- Started with Vietnamese in 1972...
- Only other language mentioned is German.

LOGOS

Technology:

- 10ks of grammar and semantic rules (25 years of development)
- Extensible dictionaries, semantic foundation, homograph disambiguation
- Full morphological, syntactic, and semantic analysis.
- Now more interlingual than transfer, heavily semantic (originally transfer)
- Probably somewhere between SYSTRAN and METAL in sophistication.

LOGOS

Corporate strategy:

- Besides MT system, provides regular translation service and translation consulting
- Customers: OSRAM, Lexitech, Ericsson, translation bureaus

LOGOS

Practical details:

- ALEX and SEMANTHA acquisition tools;
- Dictionary and semantic db tool for finding words not in dictionary
- Import/export tools
- AMTA-94: experimental Translatability Index tool
- Integrated with several TM packages
- Many file formats handled; maintains document formatting
- PC or Unix; Java-based client/server system available

PAHO: SPANAM and ENGSPAN

- SPANAM started in 1976, operational 1980 (in PL/1).
- ENGSPAN begun in 1982, operational in 1985: ATN transfer-based system.
- SPANAM rewritten in 1990 to include parser and transfer.

PAHO

Technology:

- Developed over 20 years; originally an internal-use-only system.
- Originally SYSTRAN-like, now ATN transfer-based.
- Over 60k lexical items and phrases, 5k-10k rules (analysis, transfer, context-sensitive).
- Also probably somewhere between SYSTRAN and METAL in sophistication, but all of their energy has gone into one language pair.

PAHO

“Corporate strategy”:

- Not especially commercial...
- Aimed at institutions with translation staffs, for technical material.

PAHO

Practical details:

- Pre- and post-editing macros for Word.
- Dictionary maintenance by trained staff.
- Handles many word processor text formats.
- Runs under DOS, locally or over LAN.

Globalink

- More recent company: about ten years old?
- Spanish, French, German, Italian, Portuguese.
- Talked about Russian and Chinese in 1994, but not now.

Globalink Technology:

- Document, Sentence, and Interactive modes
- Apparently SYSTRAN-style technology:
rule-based transfer.
- Priority ordering of supplemental dictionaries

Globalink

Corporate strategy:

- Targeted at end-users: fast and cost-effective.
- Conversation mode, Clipboard utility
- Web-oriented products: Web Translator and now Compende
- In 1994, talked about voice output...

Globalink

Practical details:

- Dictionary editor: incl. phrases, morph, and some semantic info
- Proprietary Rule Editor for “Barcelona” language:
 - “Keyed rules” activated by dictionary headword
 - Lexical “Verb Frame” Rules: verbs not allowed in dictionary
- OLE and MTAPI interfaces
- Portrait/landscape translation windows
- “Formatting support”, formatting preservation
- Supports many file formats
- DOS, Windows, OS/2, Mac, Unix