HISTORY OF MACHINE TRANSLATION

LTI MT Graduate Class

Jaime Carbonell

January-22-2007

OUTLINE:

- Origins of MT
- MIT and Georgetown Experiments
- ALPAC Report
- The MT Winter
- MT in Europe and Japan
- Resurgence of MT
- Current approaches to MT

Origins of MT:Early "Successes"

- 1933 – Smirnov-Troyanskii Patent for a *word translation & printing machine*
- 1939-1941 – Troyanskii added memory (first Russian computer)
- 1946 – MT as code-braking (ENIAC in US), Weaver et al
- 1946-1947 – Weaver, Booth, Weiner… Weaver realizes complexity
- 1949 – Weaver Memorandum (what it would take for MT)

Origins of MT: Early "Successes"

- 1951 – Bar Hillel survey → Human/machine is best
- 1952 – MIT Conference on MT (first small scale E-F, F-E mostly)
- 1954 – *Mechanical Translation Journal* (Yngve)
- 1954 – Georgetown-IBM Experiment (50 sentences R-E) → massive US funding

Origins of MT: Early "Successes"

- 1956-1962 – Massive MT efforts at U of Washington, IBM, Georgetown, MIT, Harvard, Oakridge, Rand, using any and all hardware including Mark II, ILIAC, …
- 1960-1964 – Kuno (Harvard) and Oettinger (Georgetown) parser
- 1955-1967 – UK active in MT (Booth, Cambridge group)
- 1956-1965 – MT in Japan starts (Wada at ETL, Fukuoka at Kyushu, …)
- 1960's → on – GETA in Grenoble (Vauquois)

Origins of MT: End of Optimism

- 1960 – Bar-Hillel report and the FAHQT Myth
- 1964, April – ALPAC Report

The MIT Early History: Bar-Hillel

- Philosopher & Mathematician, but turned Linguist & MT booster
- First-ever full-time MT researcher (MIT: 1951-1953)
- Recognized lexical ambiguity as largest challenge for MT
- Identified other MT challenges

Ambiguity Makes MT Hard

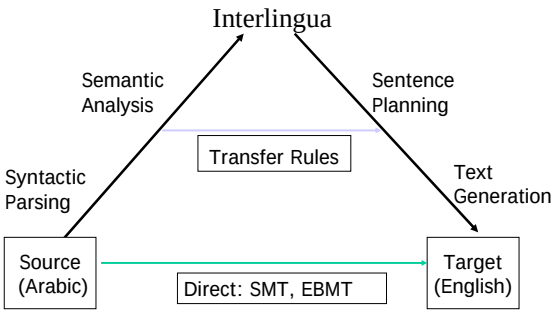
(not Bar Hillel's examples)

- Syntactic
  I saw the Grand Canyon flying to New York.
  Observe the man with the telescope with care.
- Word Sense (i.e., "polysemy")
  Power **line** (cable)
  Subway **line** (track)
  Be on **line** (be connected to internet)
  Be on the **line** (be on telephone)
  Line up (*verb*: to form a straight line)
  Line one's pockets (*verb*: to get rich)
  Line one's jacket (*verb*: add layer)
  Actor's **line** (what an actor says)
  Get a **line** on someone (verb: get info)

Ambiguity Makes MT Hard

- Word Sense (even more senses in multiple English-Japanese Dictionaries)
  Power **line** – densen (電線)
  Subway **line** – chikatetsu (地下鉄)
  (Be) on **line** – onrain (オンライン)
  (Be) on the **line** – denwachuu (電話中)
  Line up – narabu (並ぶ)
  Line one's pockets – kanemochi ni naru (金持ちになる)
  Line one's jacket – uwagi o nijuu ni suru (上着を二重にする)
  Actor's **line** – serifu (セリフ)
  Get a **line** on – joho o eru (情報を得る)

Types of Machine Translation



The MIT Early History: Victor Yngve

- High-Energy Physicist turned Linguist
- 2nd-ever full-time MT researcher (MIT: 1953-1961)
- Word-for-word MT => syntax matters (for resolving homonyms e.g. "block" and for word-order inversion)
- Recognized phrasal lexicon

The MIT Early History: Victor Yngve

- Invented analysis-transfer-generation method
- Invented COMIT (operational grammar encoding)
- Implemented Chomsky's TG in COMIT (which proved a dismal failure for analysis)

The Georgetown Early History: Leon Dosert

- Linguist & Interpreter during WWII
- Attracted most MT funding (military)
- Focused on Russian => English
- Strongest advocate for MT research

The Georgetown Early History: Other Contributors

- ***Peter Toma*** – system builder
- ***Murial Vasconcellos*** – later PanAm MT
- ***M Zarechnak*** -- Linguist

The Georgetown Early History: First "large-scale" MT

- About 100,000-word Russian Text MTed in demo adding out-of-dictionary words (1958)
- System scaled further in next 5 years
- GAT (Georgetown Automated Translator) → Well-known SYSTRAN in later years

The ALPAC Report: Members

- Pierce (Chair) Bell Labs
- Several discouraged MT researchers (Oettinger, Hays)
- Linguists (Hamp, Hockett)
- Token Computer Scientist (Alan Perlis from Carnegie Tech)

The ALPAC Report: Findings

- ***Myth*** – MT does not and cannot work
- ***Reality*** – MT is more difficult than originally envisioned
- ***Reality*** – Basic Research in NLP should be done before doing MT
- ***Reality*** – MT is too expensive (computers cost more than people)

The ALPAC Report: Net Effect

- The end of Government-funded MT research in US for 10+ years
- Continuation of private MT (e.g. Systran, Logos) in US
- Not much effect on Japan or France (efforts continued)
- USSR and UK followed US example, it appears

MT: 1967-1985
ALPAC Myth Fades Away in US

- SYSTRAN quite successful in E-R (Air Force at Wright-Patterson etc.)
- Partial success E-S, E-F, E-G (SYSTRAN, Logos, Weidner)
- SYSTRAN → use in Europe (later by EC)
- Knowledge-Based MT (KBMT) concept advanced (Carbonell, Nirenburg, …)

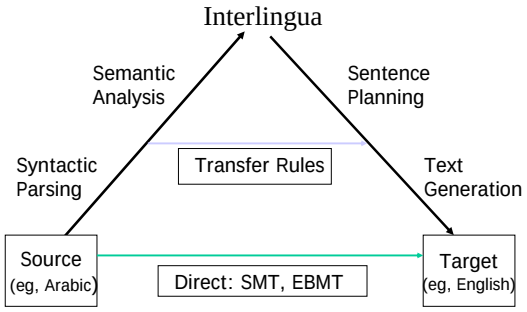
MT: 1967-1985 (II)
ALPAC Myth Fades Away in US

- "Underground MT" in US Universities dares to seek funding again
- Machine-aided Translation (MAT) concept advanced (Kay, …)
- Very-narrow-domain MT demonstrated (Kittredge et al, METEO)

MT: 1975-1985
Golden-Age of MT in Japan:1980's

- Nagao proposes Example-Based MT (not taken seriously then)
- Nagao proposes Transfer-Based MT for E-J (Mu project)
- Mu's success triggers MT-mania in giant Japanese companies, e.g., ATLAS in Fujitsu, PIVOT in NEC, HICATS in Hitachi, …
- Japanese MT Research budgets soar, US and Europe take note
- JEIDA Report paints upbeat future for MT

Types of Machine Translation



MT: 1975-1985
MT in Europe, not as Rosy

- "Interlingua" approach tried (ROSETTA, DLT)
- First language-neutral Interlingua (Yale-MT, Carbonell & Cullingford 1979, 1981)
- Eurotra proposed and started to build ultimate collaborative MT system, but later tanks due to incompatible transfer paradigms
- …but SYSTRAN adopted by EC for volume internal translations

MT Matures 1985-1995:
MT Spring in US

- Center for Machine Translation at CMU opens in 1986
- Interlingual KBMT success at CMU for domain-oriented MT (KANT) with controlled-language input, but did not generalize to open-ended and uncontrolled domains (PANGLOSS)
- Resurgence of statistical corpus MT at IBM (Brown et al), which also succeeds for E-F but needs huge training corpus

MT Matures 1985-1995: MT Spring in US

- Speech-to-Speech MT launched at CMU (first JANUS, the DIPLOMAT)
- CSTAR launched (International consortium for speech-speech MT)
- SYSTRAN, LOGOS, GLOBAL-LINK (formerly Weidner), … survive
- Conferences: MT-Summit, TMI, … (MT regains respectability)

MT Matures 1985-1995: MT Summer and Fall in Japan

- Japanese systems reach performance plateau, typical for transfer-MT
- Funding reduced, especially when economic difficulties intrude
- MT useful with extensive post-editing (e.g. ATLAS-II MT bureau)
- ATR Successful in speech-speech MT for limited domains
- Example-based MT re-emerges (Iida at ATR, Nagao at Kyoto)

MT Matures 1985-1995: MT Mostly Sub-Rosa in Europe

- EUROTRA a massively distributed un-collaborative failure
- Companies abandon MT efforts (DLT, Rosetta, Metal)
- SYSTRAN in large-scale deployment and use in EU shines through
- Vermobil speech-speech MT in Germany concluded with reasonable large-scale success for speech-MT

The Modern Period: MT post 1995 Technological Trends

- Transfer MT works with high development & post editing costs
- Interlingual KBMT works well in technical domains (but requires high development cost)
- Speech-to-Speech MT increasing in popularity, but not yet robust
- Example-Based MT => Generalized EBMT

The Modern Period: MT post 1995 Technological Trends

- New-wave of Statistical MT (CMU, ISI, JHU)
- Example-Based MT (Kyoto U, CMU)
- MT research ongoing and respectable, but with modest funding (in US, Japan, and Europe)
- Rapid-development MT becomes hot topic (US Govt., CMU, NMSU, internet)

The Modern Period: MT post 1995 Application Trends

- SYSTRAN, LOGOS, L&H, IBM, Fujitsu, remain steady MT suppliers
- Interlingual KBMT in first massive use (at Caterpillar)
- PC-based MT Systems explode (Fujitsu, IBM, Globalink, L&H)

The Modern Period: 1995-Present

- Internet MT off to a good start (Babblefish, Google)
- Translingual IR + MT hot (CMU, IBM, Google, …)
- Speech-speech MT reinvigorated
- New DARPA MT initiative
  - Statistical MT dominates
  - Evaluation centric (NIST, BLEU, …)
  - Focus on non-European languages (Arabic, Chinese)
- Japan & Europe → MT slidelines
- India, China, Russia become serious MT players

MT: Present & Future Trends

- Evaluation is here to stay
  - New, better methods (e.g. METEOR at CMU)
- New paradigms for MT flourish
  - Transfer-rule learning (CMU)
  - CMBT = EBMT without parallel text (Meaningful M.)
  - Hybrid methods EBMT/SMT/RuleMT
  - Multi-Engine MT
- Biggest challenge: Breaking the Accuracy Bottleneck
  - MT with accuracy comparable to Human Translators
  - Huge translation market (20+ billion/year)

Lessons from MT History

- Translation ≠ Transduction
- MT is a paradigm task for NLP
- Context, context, context
  - word-for-word
  - transfer grammars + lexical substitution
  - KBMT with semantic interpretation rules
  - statistical MT with bi-grams & trigrams
  - phrases (bigger n-grams) matter (EBMT, SMT)
  - new methods are based on yet longer n-grams
- Machine learning enters MT, more and more
- In MT perseverance and longevity matter