Machine translation: past imperfect, future indefinite

John Hutchins

(Email: WJHutchins@compuserve.com)
[http://ourworld.compuserve.com/homepages/WJHutchins]

University of Leeds

November 2003

Outline of system types and uses

- Computer-based translation aids
  - terminology management
  - translation memories
  - translation workstations
- MT systems with human assistance, for good quality
  - pre-editing, controlled language input
  - post-editing
- MT systems for rough translation
  - PC systems, online services
  - translation for assimilation, from unknown languages
- MT systems with other language technologies

The development of MT: 1950s and 1960s

- Sponsored by government bodies in USA and USSR (also CIA and KGB)
  - assumed goal was fully automatic quality output (i.e. of publishable quality) [dissemination]
  - actual need was translation for information gathering [assimilation]
- Methods: word for word, lexicographic, direct translation, interlingua, syntactic analysis, transfer systems, use of statistics (distribution analysis, redundancy), pre-editing, controlled language, restricted language, post-editing
- Survey by Bar-Hillel of MT research (1960):
  - criticised assumption of fully automatic high quality translation (FAHQT) as goal
  - demonstrated 'non-feasibility' of FAHQT (without 'unrealisable' encyclopedic knowledge bases)
  - advocated "man-machine symbiosis", i.e. HAMT and MAHT
- ALPAC 1966, set up by disillusioned funding agencies

Consequences of ALPAC

- ALPAC recommendations, and shortcomings
  - compared latest systems with early unedited MT output (IBM-GU demo, 1954), criticised for still needing post-editing
  - advocated machine aids, and no further support of MT research
  - but failed to identify the actual needs of funders [assimilation]
  - therefore failed to see that output of systems were used and appreciated
- MT research virtually ended in US
- henceforth: identification of actual needs: assimilation vs. dissemination
- recognition of different needs/purposes of full automation vs. HAMT and MAHT
- recognition that 'perfectionism' (FAHQT) had neglected: operational factors and requirements, expertise of translators, machine aids for translators
- henceforth three strands of MT:
  - translation tools
  - operational systems (post-editing, controlled languages, domain-specific systems)
  - research (new approaches, new methods)

From 1967 to 1989

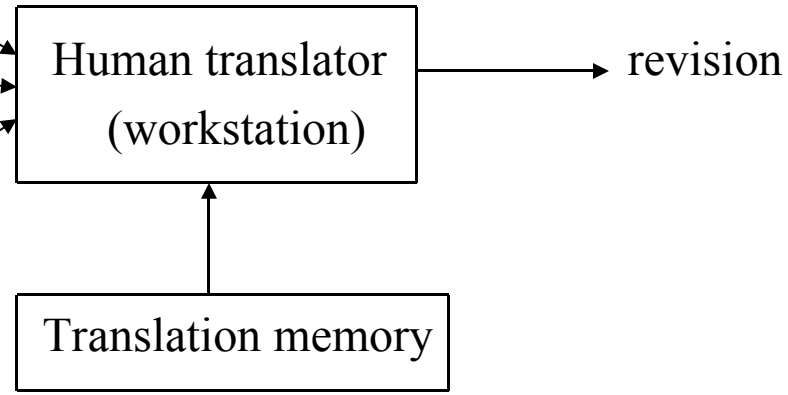
- Continuation of research in US (Texas, Wayne State), Soviet Union, UK, Canada, France
- rule-based approaches: interlingua and transfer
- 1970: Systran installed at USAF (Foreign Technology Division)
- 1975: Météo 'sublanguage' English-French system (weather broadcasts)
- 1976: European Commission acquires Systran
- 1979: Pan American Health Organization system (SPANAM)
- Rule-based systems: involving long-term efforts compiling grammar rules (interlocking) and creating dictionaries
- Interlingua systems: DLT, Rosetta, Carnegie Mellon
- Transfer-based systems: GETA (Ariane), SUSY, Eurotra, Mu (Kyoto)
- Knowledge-based systems: Carnegie Mellon, New Mexico, Pangloss
- Speech translation: ATR, C-STAR, Verbmobil
- **Computer-based tools**

Changes since late 1980s

- Increasing use of MT by large enterprises
- Translation memory and translation workstations
- Localization
- Growth in PC systems
- The impact of the Internet
- Online translation
- MT and other language activities
- Research on corpus-based MT methods

Machine-aided human translation

- Terminology
- database
- text glossary
- pre-translation
- (MT)



Computer-aided translation and translation tools

- recognition that fully automatic translation not appropriate for professional translators
- PCs and multilingual word processing, desk top publishing
- Translator ‘in control’
- dictionaries (monolingual, bilingual): on-line access
- grammar aids, spelling checkers
- user glossary, terminology management, ‘authorised’ terms, standards, specialist glossaries
- input, output, transmission (OCR, pre-editing, controlled language)
- translation memory, alignment
- management support tools (project control, budgeting, workflow)
- previous antagonism of translators to MT diminished

Translation memories: weaknesses

- Expensive to build (in time and money)
- sentence-based comparisons restrict potential use (no phrase matching); whole sentence repetition is rare (except with revised texts)
- loss of context beyond sentence
- any TM likely to contain redundant, ambiguous versions
- any TM likely to contain conflicting translations (with little or no guidance)
- sentences are edited by translators outside TM environment and therefore not included in the database
- TM systems do not 'learn' decisions/choices made by users (e.g. which potential translations are preferred, which rejected)
- fuzzy matching often fails (hidden tags) and too complex, and translators opt not to use the facility; prefer translating from scratch
- combining extracted translation segments left entirely to user/translator
- developments needed:
  - **finding phrases (retrieval, fuzzy matching)**
  - **combining phrases; searching for words in combination**
  - **phrase repetition**

Machine translation
System architectures and strategies

- Rule-based
  - Direct translation
  - Interlingua-based MT
  - Transfer-based MT
- Corpus-based MT
  - Statistics-based
  - Example-based
- Hybrid systems

Direct translation

- analysis of source language (SL) text only as much as necessary for conversion into particular target language (TL) text
- SL dictionary lookup followed by TL word-for-word output, then TL rearrangement based on dictionary entries
- use of ‘cover’ words (i.e. single most frequent equivalent; therefore often not most appropriate)
- no deep analysis of SL syntax, no semantic analysis
- TL output too close to SL structure

- problems of direct translation systems:
  - too complex for modification and enhancehancement (not just computationally)
  - mixture of lexical rules and syntactic rules (no linguistic or translation ‘theory’)
- systems:
  - Univ.Washington, IBM (US), Georgetown University (US), Ramo-Wooldridge (US), Institute for Precision Mechanics and Computer Technology (USSR), National Physical Laboratory (UK)
  - some modern PC systems

Interlingua-based MT

- two independent stages: analysis, synthesis
- abstract language-neutral representation
- multistratal: morphology, syntax, semantics
- semantics-oriented ('understanding'), hence domain-specific 'knowledge bases' (AI-oriented)
- problems:
  - nature of interlingua: natural, artificial, logical?; language-neutral or language-universal? (latter not feasible, most are 'neutral' for a few specific languages)
  - few 'pure' interlinguas: most only syntax, retain bilingual SL-TL lexicon
  - complexity of representations, complexity of fully disambiguating analysis
- projects:
  - [Trojanskij, 1933], Milan (Ceccato), Cambridge (CLRU), Grenoble (CETA), Texas (pre-METAL), [Mel'chuk (MMT)], Utrecht (DLT), Eindhoven (Rosetta), NEC (Pivot), Carnegie-Mellon University (KBMT, KANT, CATALYST), New Mexico State University (ULTRA, Pangloss), Univ. Maryland (UNITRAN), United Nations University (UNL)

Transfer-based MT

- three stages: analysis, transfer, synthesis
- abstract semantico-syntactic interfaces/representationsl but basically syntax-oriented
- multiple level/strata: morphology, syntax, semantics
- problems:
  - failure at one analysis stage may mean no output
  - separation of morphological and syntactic analysis may not be relevant (in fact, many did not), similarly, distinction between syntax and semantics may not be helpful
  - distinction between interlingua-based and transfer-based often unclear (many combined features of both, e.g. Eurotra)
  - little/no discourse information (anaphora, etc.)
  - complexity of tree transduction rules
- projects/systems:
  - Georgetown University, MIT -- METAL (Texas), GETA-Ariane (Grenoble), SUSY (Saarbrücken) -- Eurotra, LMT (IBM), Mu (Japan), and many Japanese systems (JICST, Fujitsu, Toshiba…) -- many current PC systems

Statistics-based MT

- bilingual corpora: original and translation (not available for some languages)
- little or no linguistic 'knowledge', based on word co-occurrences in SL and TL texts (of a corpus), relative positions of words within sentences, length of sentences
- SL and TL sentences aligned statistically (according to sentence length and position)
- 'translation model': probability that a TL string is the translation of a SL string, based on:
  - frequency of SL/TL co-occurrence in aligned texts of corpus
  - position of SL words in SL string, and TL words in TL string
- 'language model': probability that a TL string is a valid TL sentence (based on frequencies of bigrams and trigrams), search for TL string that maximizes these probabilities
- first example: IBM Candide (1988) on Canadian Hansard (English and French)

Example-based MT

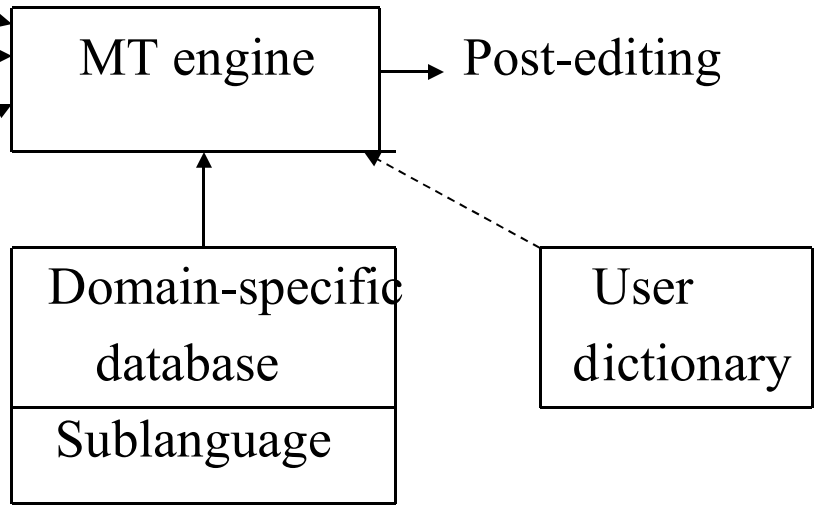
- Use of already translated sentences or phrases either from actual translations (corpus) or from data supplied by user or developer
- Sentences/phrases aligned in database (either by rule-based parser or statistically)
- matching algorithm (exact and close) of SL input and TL examples
- combination algorithm (for generating a TL sentence from extracted examples)
- problems: adding examples may not improve performance; repetition of same or similar examples may introduce unnecessary clutter; 'boundary friction'(*that old man has died* ↔ ce vieil homme est mort; *that old woman has died* ↔ (**not simple substitution**: ce viel femme est mort), **but**: cette vieille femme est morte); use of grammatical categories (patterns):
  - templates (e.g. <1st name><family name> flew to <city> on <date>)
  - X [pron] eats Y [noun/NP] ↔ X [pron] ga Y [noun/NP] o taberu
  - X o onegai shimasu → may I speak to the X (if X=jimukyoku 'office', … etc.); or: please give me the X (if X=bangō 'number', … etc.)

Hybrid systems

- **clearly, none of the current MT 'models' are capable of solving all problems**
- **hence search for hybrid architectures**
- **in theory, it would seem that (on average):**
  - **RBMT better for SL analysis**
  - **EBMT better for transfer**
  - **SMT best for TL generation**
- **Problem is that different approaches not easily compatible:**
  - **there are however research prototypes combining:**
  - **EBMT with statistical methods**
  - **EBMT using rules similar to those in RBMT systems**
  - **perhaps a version of EBMT will be the answer**
- **Currently 'hybrid' systems are parallel systems with a selection mechanism, as in:**

Human-assisted MT

- Pre-edited input
- Controlled language input
- 'Raw' input



MT for dissemination

- General-purpose system or specialised system
- Controlled language
- Lexical resources
- Management implications
- Control of terminology
- Consistency
- Standards; exchange formats
- Compatibility (hardware, software)
- Integration: technical authoring, publishing

Large-scale translation and MT

- accurate, good quality, publishable (dissemination)
- publicity, marketing, reports, operational manuals, localization
- technical documentation; large volumes
- repetitive, frequent updates; saving costs (and staffing?)
- multilingual output (e.g. English to French, German, Japanese, Portuguese, Spanish)
- available in-house terminological database; user (company) dictionaries
- backup resources (translated texts, personnel for dictionaries, etc.)
- human assistance for quality (controlled language input, post-editing)
- integrate with technical writing and publishing
- availability of in-house printing/publishing
- technical expertise (computers, printers, etc.)

Controlled language

- Controlled authoring of the source text in standard manner, suitable for unambiguous translation
- Typical rules:
  - use only approved terminology, e.g. *windscreen* rather than *windshield*
  - use only approved sense: *follow* only as ‘come after, not ‘obey’
  - avoid ambiguous words: *replace*, either (a) remove and put back, or (b) remove and put something else in place; not *appear* but: come into view, be possible, show, think
  - only one ‘topic’ per sentence, e.g. one instruction, command
  - do not omit articles; use relative pronouns (which, in order that); avoid post office-nominally gerundive form (*wires connecting*…→ *wires that connect*…)
  - do not use pronouns instead of nouns if possible
  - do not use phrasal verbs, such as *pour out*
  - do not omit implied nouns
  - use short sentences, e.g. maximum 20 words
  - avoid co-ordination of phrases and clauses

- **advantage of controlled language is improvement of original SL text; sometimes translation no longer necessary; later revision can be faster**

Controlled language and special-purpose systems: requirements and issues

- system developed by external agency (e.g. Smart, LANT) or in-house?
- special dictionaries (domain, company): existing, or to develop?
- terminology databases
- new or adapted from existing controlled languages
  - despite previous models, SAP developed own language (SKATE)
- grammar and style analysis (usual grammar checkers inadequate)
- lexicon
  - internal (company) and external (standard terminology)
- grammar
  - to be recommendations or to be obligations

Convergence of HAMT and MAHT

- increasingly, systems straddle different categories
- workstations (TM systems) include MT components (e.g. Trados, Atril)
- MT systems include TM components (e.g. globalwords)
- localization systems embracing, or as components of, either TM or MT systems
- common facilities:
  - terminology management; integration with authoring and publishing systems; project management; quality control; Internet access and downloading; Lexical acquisition; Web translation
- common aim: production of quality translations for **dissemination;** utilization of translator skills
- at present: both approaches in parallel rather than integrated
- in research: EBMT investigates merging of rule-based and database methods
- future: full integration (no distinctions)

MT for translators (office systems): issues

- translation database -- ownership, copyright
- terminology management -- acquisition
- integration with other IT equipment
- translation workstations still too expensive for individual translators
- insufficient functionality in downsizing systems for large organizations onto stand-alone (PC) systems
- suitable project management tools (currently most for large agencies and companies)

- **currently downsized versions of ‘enterprise’ systems, or upgraded version of ‘home’ systems - not yet well defined category**

MT for assimilation (home use, online)

- must be fast, immediate, real-time
- must be readable, but accept poor quality
- more languages
- webpage compatible (translate graphics)
- translate electronic mail
- steady improvement

Online and PC translation: why so bad?

- old models (word for word, simple transformer architecture)
  - often single equivalents, no morphological analysis or target adjustment
- dictionaries too small, insufficient information, and difficult (or impossible) to update
- weak syntactic analysis/transfer
- poor disambiguation (little semantic information)
- general-purpose (not domain restricted)
- not designed for language/style of emails
- web page translations: graphics not translated, distorted, ignored; format lost
- need special functions if used as aid for writing in foreign language
- language coverage uneven; many languages of Africa and Asia are lacking
- translation from English often poorer than into English

- **conclusion: of use/value only if source language unknown or known only poorly and if essence and not full information is adequate**
- **the less the user knows the source language, the more useful becomes automatic translation**

MT in the marketplace

- retail availability
  - many only purchased direct from manufacturer
- promotion by vendors
  - confusion of terminology:
    - some ‘translation systems’ are no more than dictionaries
    - ‘computer aided translation’ either HAMT or MAHT
    - combination of MT and support tools
    - translation memories either independent or components
- low profits, slow quality improvement, few differences between rivals
- categorisation (enterprise, professional, home, workstations) unclear
- expectations of users
  - steady (faster) quality improvement
  - more languages
- suitability of system to expected use
- bench marks, consumer reports/reviews

MT for interchange: what's needed?

- correspondence, emails, face-to-face, etc.
- in principle, any systems can be used for written interchange
  - many PC systems have specific facilities for email translation
- in future there may be special-purpose systems for business correspondence (e.g. with interactive authoring in controlled language)
  - has been subject of research (e.g. UMIST)
- interchange in military ('field') situations
  - e.g. systems for translating standard phrases (Diplomat, Phraselator)
- interchange in tourist situations
  - so far only dictionaries of words and phrases (hand-held devices)
- interchange with deaf and hearing impaired
  - translation into sign languages [mainly research so far]
- interchange by telephone or in business oral communication
  - still at research stage (speech translation)

MT and other LT applications

- document drafting
  - Japanese researchers, EC administrators, school essays
- information retrieval (CLIR): translation of search terms
- information filtering (intelligence):
  - for human analysis of foreign language texts
  - document detection (texts of interest); triage (ranking in order of interest)
  - deciding whether text worth translating (discard irrelevant ones)
- information extraction: retrieving specific items of information (domain-tuned, captured by key words/phrases)
  - e.g. specific events, named people or organizations
- summarization: producing summaries of foreign language texts
- multilingual generation from (structured) databases
- localization of interactive commands (computers, mobile phones)
- television subtitling

Speech translation: problems

- speech recognition, speech synthesis
- highly context dependent, use of ‘knowledge databases’
- discourse semantics, ‘ill-formed’ utterances
- ellipsis, use of stress, intonation, modality markers
- restricted domain (e.g. hotel booking by telephone)
- colloquial usage not yet investigated sufficiently (even in linguistics)

- half-way ‘solutions’ available with voice input/output
  - Word processing add-ons (Dragon Naturally Speaking, IBM ViaVoice)
  - PC translation systems with voice input/output (Al-Wafi, CITAC, ESI, Korya Eiwa, Personal Translator PT, Reverso Voice, TranSphere, Vocal PeTra, ViaVoice Translator)
  - Online translation with voice output (Translation Wave)

Has MT improved?

- In what respect?
  - translation quality: general-purpose vs. domain-specific
  - usability (ease of use)
  - adaptability (integration with other software)
- Since when?
  - quality perhaps not in last ten years, but since 1980 it has
- Why not?
  - inherent problems of language
  - inherent problems of ‘cultural’ differences

MT: when it works and when it doesn't

- cannot be both fully automatic (no pre- or post-editing) and general-purpose
- beyond its scope:
  - literature, philosophy, sociology, law
- large corporations, cost-effective if:
  - controlled input, standardised terminology, multilingual output, repetitive documentation, restricted domain
- occasional (information-only)
  - rough, not for publication; immediate (fast) production
- small-scale MT
  - 'formulaic' documents (business correspondence), restricted domain
  - interactive assistance

MT and HT in complementation

- Dissemination
  - HT: single item, context/culture-sensitive,
  - HT with TM: repetitive (e.g. localization, web localization)
  - MT only: restricted language, repetitive (e.g. Meteo); document drafting
  - MT with post-editing/controlled language: large scale, technical, localization
- Assimilation
  - MT with (rapid) post-editing:scientific, technical
  - MT only (PC or online): single item (non-literary), general purpose; information monitoring/filtering
  - MT domain-specific (online)
- Interchange
  - HT: business correspondence; interpreting
  - MT: email, personal correspondence; database searching; TV captions
  - MT domain-specific: business correspondence
  - MT (speech) domain-specific: telephone enquiries

General comments

- MT is not *translation* as usually understood, it is merely a computer-based tool
  - for translators
  - for cross-language communication
  - for access to information resources
- Perfectionism is not necessary or essential
  - publishable quality will always require human editing/revision
  - assimilation/interchange can always tolerate imperfect communication
- MT should be used only as required to save costs/effort in appropriate circumstances
- Judgement should be based
  - ***not*** on whether system produces 'real' translations
  - and particularly not whether it produces 'good' translations
  - ***but***: whether the output can be *used*
  - and: whether its use will save time or money

New directions and challenges

- Spoken language translation
  - general-purpose?
- 'Minor' languages
  - languages of India, Africa, Asia
  - non-national ('official') languages (e.g. Welsh, Basque, Catalan)
  - languages of minorities (e.g. non-indigenous languages in Britain)
- Systems for monolinguals
  - from unknown source language
  - to unknown target language
- Improvement expectations
  - particularly PC commercial and Internet systems
- Reusability of resources (particularly dictionaries and translation memories)
- Integration
  - MT as option in word processing packages, on Web pages
  - MT as option with summarization, information extraction, information retrieval, data retrieval, question-answering, Internet search tools

Sources of information

- EAMT website (www.eamt.org) with links to other IAMT sites, etc.
- LISA website (www.lisa.org)
- Conferences:
  - MT Summit, EAMT workshops, AMTA conferences, LISA Forums
- Journals (non-research):
  - *Multilingual Computing and Technology*
  - *MT News International*
- *Compendium of translation software*
- my website:
  - http://ourworld.compuserve.com/homepages/WJHutchins