Multiple Uses and Applications of Machine Translation and Computerised Translation Tools

John Hutchins

Presentation on 2 July 2009
ISMTCL conference, Besançon

Outline

- General features of MT and MAT
- Use by companies and large organizations
- Tools for translators, translation memories, localization
- Use for assimilation, interchange
- Online MT, webpages, email, mobiles
- Special purpose systems: speech, minorities, embedding
- Conclusions

Categories of systems

- Machine translation – for enterprises
- Machine translation – for professional translators
- Machine translation – for casual/home use
- Machine translation for bilingual communication
- Translation memory systems
- Translation workstations – for professional translators
- Electronic dictionaries

Basic types of use

- Dissemination (for publication)
  - Enterprise systems (corporations, organizations)
  - Free-lance translators and agencies
- Assimilation
  - Acceptable lower quality (information purposes)
- Bilingual communication
  - Interchange, with feedback and elucidation
- Translation aids
  - Drafts, dictionaries

General-purpose vs Subject-specific

- General purpose systems
  - General dictionaries with all translation options (or selection of most common only); wide grammatical coverage
- Subject-specific systems
  - Fewer ambiguities within subjects (sublanguages)
  - Subject-specific system dictionaries
  - User dictionaries, terminology
  - Controlled language input
    - restrict vocabulary choice and syntactic complexity; avoid ambiguity (articles, pronouns, conjuctions, prepositions, etc.)

Basic architectures

- Rule-based
  - Direct translation (dictionary-based) - *segment, substitute, rearrange*
  - Interlingual approach: two stages - *analyse, abstract representation, generate*
  - Transfer approach: three stages - *analyse, transfer representations, generate*
- Corpus-based
  - Example-based MT - *segment, select TL phrases, combine*
  - Statistical MT - *segment, select TL forms, rearrange*
  - Translation memory - *search, extract, combine*
- Combinations: hybrid and multi-engine

System types from the users' viewpoint

- The differences between MT system architectures and methods are largely irrelevant.
- Users are normally only concerned with
  - compiling and/or augmenting dictionaries
  - storing texts for translation memory systems (preparing corpora)
  - controlling (adapting) text input (pre-editing)
  - interactive disambiguation
  - editing text output (post-editing)
- In theory any MT systems can be used for any of the functions (dissemination, assimilation, interchange, information access)
- Overall quality of translation is less important than whether output is good enough to be useful (usable) in particular context of use

MT for dissemination: companies and government organisations

- Dissemination originally only use (e.g. US Atomic Energy, Euratom, USAF)
- usually general-purpose systems (Systran, SDL)
  - adapted with subject-specific terminology (JobBank, GHIN, GM, SAP, etc.)
  - system dictionaries (general vocabulary) usually unalterable
- often with controlled language input (earliest: Xerox in late 1970s)
  - closely integrated with authoring software
- usually with post-editing
    - the less post-editing the more cost-effective
  - processing closely integrated with publishing software
- subject-specific systems:
  - PAHO, JAPIO, ProLingua

Dissemination: Translators' computer-based tools

- (since 1966) recognition that fully automatic translation not appropriate for professional translators
- Term banks (since 1970): TEAM, LEXIS, TERMIUM, Dicautom, Eurodicautom
- Text-related glossaries (since 1970s: Bundeswehr, ALPS)
- Terminology management (Mercury/Termex)
- Electronic dictionaries (software, CDs, etc.)
- Translation databases ('translation memory')
  - first: Arthern (1978), Kay (1980), ALPS
- Melby's three levels (early 1980s)
  - word processor with integrated terminology aids, manual insertion of words
  - machine-readable input texts, concordance (to find occurrences of words in text), local term bank, automatic insertion of terms
  - integrated 'workstation' with MT system, and automatic 'quality' evaluation

Computer-aided translation tools since 1980s

- PCs and multilingual word processing, desk top publishing
- dictionaries (monolingual, bilingual): on-line access
- grammar aids, spelling checkers, concordances
- user glossaries, terminology management, ‘authorised’ terms, standards, specialist glossaries, text-related glossaries
- input, output, transmission (OCR, pre-editing, controlled language)
- translation memory, alignment
- text prediction (TransType)
- management support tools (project control, budgeting, workflow)
- translation workstations (combining tools [and MT], compatible with authoring/publishing software)
  - examples: Trados, Déjà Vu, MultiTrans, WordFast, ProMemoria, MetaTexis, etc
- translators ‘in control’, previous antagonism of translators to MT has gradually diminished

Translation for dissemination: using translation memories

- based on sets of original texts and their ‘authorized’ translations
- particularly suitable for translation of revisions and for translating standardized documents; with major gains (time saving, etc.)
- most suitable for large (organizational) translation agencies/departments
- any TM likely to contain redundant, ambiguous versions, untypical, rare, conflicting translations (with little or no guidance)
- TM systems do not ‘learn’ decisions/choices made by users (e.g. which potential translations are preferred, which rejected) - weak feedback
- sentence-based comparisons restrict potential use (no phrase matching)
- fuzzy matching often too complex, e.g. without linguistic information such as morphology, and translators opt not to use the facility
- combining extracted translation segments left entirely to user/translator; sentences edited by translators not automatically added to the database
- still much post-editing

Localization

- Internationalisation, globalisation (e.g. software and Web pages)
  - estimated market (end 2006) $3.5 billion and $3 billion resp. (ABI, 2001)
- Cultural and linguistic adaptation (not just translation): currency, measurements, power supplies
- Screen commands and help files; users' guides; warranties; publicity, marketing; packaging; workshop manuals
- Large scale, multiple language output, fast results (within days, not weeks)
- Repetitive (translation memory)
- Graphics, formatting, layout, etc. (to be preserved)
- **companies use both translation tools (workstations, translation memories) and MT systems**
- Software companies (many in Ireland):
  - ALPNET; Berlitz; Compaq; Corel; Eastman-Kodak; IBM; Lotus; Microsoft; Oracle; SAP; Symantec

MT for dissemination: individual translators

- translation workstations still too expensive or not appropriate for individual translators
- PC sysetms offer easier integration with other IT equipment
- cost-saving, easy post-editing (familiar word processors)
- commercial ‘professional’ systems with functions as for large organizations
  - i.e. include terminology management and use of translation database (own or shared)
- vendors either downsize client-server systems or upgrade cheaper (home) PC systems
- other users of such systems?:
  - companies not able to afford (or without facilities for) client-server systems
  - smaller translation agencies
  - occasional translators (perhaps)

MT for assimilation

- publication-level quality not necessary
- fast/immediate; translation (service) not otherwise available
- readable (intelligible), for information use
  - intelligence services (e.g. NAIC)
  - occasional translation (home use)
- as draft for translation
- aid for writing in foreign language
  - as used by EC administrators
- emails, Web pages
- any system type can be used
  - in early (mainframe) MT (e.g. by USAF), a usage reluctantly conceded [but not by ALPAC]
  - PC systems [perhaps principal use]
  - online MT [undoubtedly the principal use]
- but generally no facilities for adding (or changing) dictionaries)

MT for interchange: examples

- correspondence, emails, 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)
- 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 by telephone or in business oral communication; still research only (speech translation)
- interpreting ex tempore (unlikely ever to be even semi-automated) , but:
  - interpreters (at EC etc.) do use rough MT of technical speeches to aid them

MT in the marketplace

- retail availability
  - most products only purchasable direct from manufacturer (online ordering)
- promotion by vendors can be misleading by confusion of terms:
  - ‘translation systems’ no more than dictionaries
  - ‘computer aided translation’ (either human-aided MT or translation tools)
  - various mixtures of MT and support tools
  - translation memories either independent or components
- expectations of users
  - steady quality improvement
  - more languages
  - suitability of system to expected use (difficult for users to assess)
- need for bench marks, consumer reports/reviews

Risks of marketplace

- Failures of previous products, e.g.:
  - ALPS Transactive, Weidner and Bravice
  - Intergraph and Transparent Language
  - Globalink (Microtac)
  - Lernout & Hauspie
  - Logos Corporation
  - Winger
- current system categories used by vendors - are they understood?
  - Enterprise systems, i.e. Client-server (intranet)
  - Workstations (TM systems)
  - Professional systems
  - Home systems
- low profits, slow quality improvement, few differences between rivals
  - not helped by free online services

Free Online MT

- First systems: 1988 - Minitel (Systran), Niftyserve (ATLAS)
- 1992 CompuServe, 1994 Globalink
- 1997 Babelfish (Altavista, Systran)
- FreeTranslation, Gist-in-Time, ProMT, Google, etc.
- Limited lengths of text input (e.g. 100 words)
- No user dictionaries, but can be restricted to subject areas
- Free, vendors hope for sales of products
- 'Value-added' post-editing services (charged)
- Raised profile of MT, but at a cost...

Online MT

- For many users:
  - First use of MT
  - Unaware of PC products
  - Unaware of limitations
  - Test with 'inappropriate' texts, back-translation
  - Produce howlers of ‘first generation’ MT
    - The spirit is willing, but the flesh is weak; Out of sight, out of mind
  - Often disappointed with results

Online MT usage

- No data on users: ages, background knowledge, types of texts, etc.
- Used by translators as rough drafts?
- Average length 20 words; 50% of submitted ‘texts’ just one or two words
- Very few webpages (unexpected!)
- Overall usage continues to grow exponentially
- The less the language knowledge of users , the more useful the output!

- Quality improvements?
  - Desirable but not commercially attractive
  - mainly rule-based systems (Babelfish), now some statistical (Google)

Webpages

- Colloquial, culture-dependent language
- Texts in graphic format cannot be translated (very common in Japanese webpages)
- But website developers often recommend users to online MT services – do they know the dangers to their reputations?
- Website localisation systems for companies, etc. (IBM Websphere)

Electronic mail

- On PCs
  - initially mainly Japanese systems, now standard
- On intranets
  - basic facility of company (‘enterprise’) systems
- Commercial systems specifically for emails (e.g. Translution)
  - access online or via intranets
  - adapted to company terminology

Hand-held devices

- 'Pocket translators' (special equipment)
  - Ectaco, Phraselator
  - Mechanised phrase books for military, tourism
  - often no more than word dictionaries
  - Sold in large numbers (but how successful?)
- Mobile (PDA) devices
  - Text messaging (SMS)
  - Only for common languages
  - Direct access to online MT services

Spoken Language

- PC systems with voice input/output, i.e. speech-text-text-speech
  - first?: Globalink (1995)
- Genuine speech translation
  - only research systems: ATR, CMU, JANUS, C-STAR, Verbmobil, NESPOLE
- ‘bottleneck’ is speech recognition: often very limited range of speakers
- Domain restriction
  - telephone, hotel booking, business communication (ATR, Verbmobil)
  - military (DIPLOMAT, Phraselator)
  - medical, doctor-patient, etc. (MedSLT)
  - tourism (ATR) - BTEC (for SMT evaluation)

MT for minorities

- No clear definition: language may be widespread globally, but minor in particular country (e.g. Hindi in UK)
- European examples: Basque, Catalan, Galician, Estonian, Latvian, etc.
- Not commercially significant market
- Poor resources (dictionaries, grammars)
- Often not even word-processing (alphabets)
- Lack of bilingual corpora
  - even SMT rapid development not an option
- instead of MT: other 'low-level' (NLP) aids more suitable?

Rapid development of MT systems

- For languages poorly covered
- For languages of interest to ‘intelligence’ services
- Rule-based systems: not feasible because of:
  - Complex grammar, large dictionaries
  - Slow costly development
- Statistical MT
  - Based on large corpora (but not always available): Internet as resource
  - Little additional data required (e.g. grammars, thesauri)
  - Open source systems and components
    - GIZA, Moses, Apertium, GPL, etc.
  - Commercialisation, e.g. Language Weaver

Embedding MT

- Information retrieval
  - multilingual access to document information (cross-language information retrieval)
- Information extraction
  - data mining, text mining
- Intelligence
  - languages: Arabic, Chinese, Farsi, ...
- Summarization
- Transliteration (names)
- Question answering
- Authoring software

Subject-specific MT systems

Sublanguage systems (few successors of Météo)

e.g. police, drug enforcement, news

Commercial PC systems for medicine/patents (Japanese)

Availability of special glossaries (ranked for preference)

e.g. medicine, law, Bible, business

Wide range of dictionaries and glossaries available (but how many sold/used?)

Other applications: actual and possible

- subtitles, broadcast transcripts, syndicated feeds
- chatrooms, social networking (Facebook, etc.)
  - problems comparable to spoken language translation
- distance education, language teaching
- emergency services
- MT for the deaf
- Photocopier-MT; Scanner-MT ('pen' scanner)
- Camera-MT (menus, road signs?)
- Surround MT
- MT for robots (spoken?)
- decipherment (back to MT origins!)

Current usages of MT: summary

- Systems for dissemination (publication)
  - traditional use by corporations, agencies, localisation
  - rough drafts for authors
- Systems for assimilation (information acquisition)
  - 'unedited' MT, intelligence/analysis, online MT
- Systems for interchange
  - electronic mail, correspondence, Web pages, tourism
- Language coverage
  - good (usable) for English, French, German, Spanish, Japanese, Chinese, Korean, Arabic
  - poor for: African, Indian, S.E.Asian, E.European, UK minorities

Future expectations: summary

- merging of MT and TM for enterprise dissemination systems
- internet as major (chief) data resource - not only SMT
- integration of semantic annotations (Semantic Web)
- rapid development of systems (SMT)
- reuse of MT components (for closely related languages)
- improvements in quality of MT
  - hybrid, multi-engine systems
- minor (and minority) languages
  - i.e. languages not of major commercial or military interest
- special-purpose systems (domain and function) - also online
- rapid updating of dictionaries (special and general), of terminology databases
- spoken language MT, domain-specific only [not general-purpose]
- much greater embedding of MT in other LT systems
- bilingual (multilingual) communication as much as translation

Resources

- associations: European Association for Machine Translation (www.eamt.org); Localization Industry Standards Association (www.lisa.org); Translation Automation Users Society (translationautomation.com)
- conferences: MT Summit, AMTA conferences, EAMT conferences, Aslib Translating and the Computer
- Compendium of translation software (www.eamt.org/soft_comp.php)
  - conversion to searchable database in preparation
- Machine Translation Archive (www.mt-archive.info)
- My website for *history of MT (*www.hutchinsweb.me.uk)