Machine translation: problems and issues

John Hutchins

panel at conference, 13 December 2007

Inherent (linguistic) problems: bilingual lexical differences

- bilingual lexical ambiguity (more than one equivalent, whether ambiguous in SL or not):
  - river: fleuve/rivière
  - Taube: dove/pigeon
  - Schraube: screw/bolt/propellor
  - corner: coin or angle; Ecke or Winkel
  - light: léger, clair, facile, allumer, lumière, lampe, feu
  - look: regarder, chercher, sembler
- lexical gaps
  - dacha, cottage, marmelade, vodka, etc.
  - snub: infliger un affront; verächtlich behandeln, or: derb zurückweisen
  - het Turks kennen: to know Turkish
  - kenner van het Turks: *knower of Turkish, someone who knows Turkish

- **Solved (?) by contextual rules (RBMT), or examples (EBMT), or frequencies and 'language models' (SMT)**

Inherent (linguistic) problems: structural ambiguity

- (1) Peter mentioned the book I sent to Mary [ambiguous for HT]
  - Peter mentioned the book which I sent to Mary
  - Peter mentioned to Mary the book which I sent [to Peter/David]
- (2a) We will meet the man you told us about yesterday [unambiguous for HT]
  - … the man you told us about yesterday
- (2b) We will meet the man you told us about tomorrow [unambiguous for HT]
  - we will meet tomorrow the man...
- (3) pregnant women and children [unambiguous for HT]
  - des femmes et des enfants enceintes [produced by MT system]
- (4a) Smog and pollution control are important factors
- (4b) Smog and pollution control is under consideration
- (4c) The authorities encouraged smog and pollution control

- **Often, problems such as (1), (2), and (3) are problematic for RBMT, but they may be 'solved' by SMT 'language model' and by EBMT databases. But problem (4c) requires 'knowledge' (i.e. rule-based KBMT)**

Inherent (linguistic) problems: bilingual structural differences

- (1) Young people like this music
  - Cette musique plaît aux jeunes gens
- (2) The boy likes to play tennis
  - Der Junge spielt gern Tennis
- (3) He happened to arrive in time
  - Er ist zufällig zur rechten Zeit angekommen
- (4) Le moment arrivé je serais prêt
  - When the time comes, I shall be ready

- **Difficult to specify transfer rules (RBMT) to cover all circumstances and contexts; but example-based (EBMT) and statistics-based (SMT) approaches yet to prove any better. Probably examples like no.4 are just unsolvable**

Non-linguistic problems of 'reality'

- The soldiers shot at the women and some of them fell
- The soldiers shot at the women and some of them missed
  - must know what 'them' refers to e.g. if translating into French (ils or elles)

- **No solutions with linguistic rule-based approaches**
- **No solutions with corpus-based approaches**
- **Perhaps only solution using Artificial Intelligence approaches (Knowledge-based machine translation, e.g. Carnegie-Mellon University)**

- However, perhaps this problem is exaggerated: no need to understand what AIDS and HIV are in order to translate:
  - The AIDS epidemic is sweeping rapidly through Southern Africa. It is estimated that more than half the population is now HIV positive.

Problems of stylistic difference

- The possibility of rectification of the fault by the insertion of a valve was discussed by the engineers
- The engineers discussed whether it was possible to rectify the fault by inserting a valve

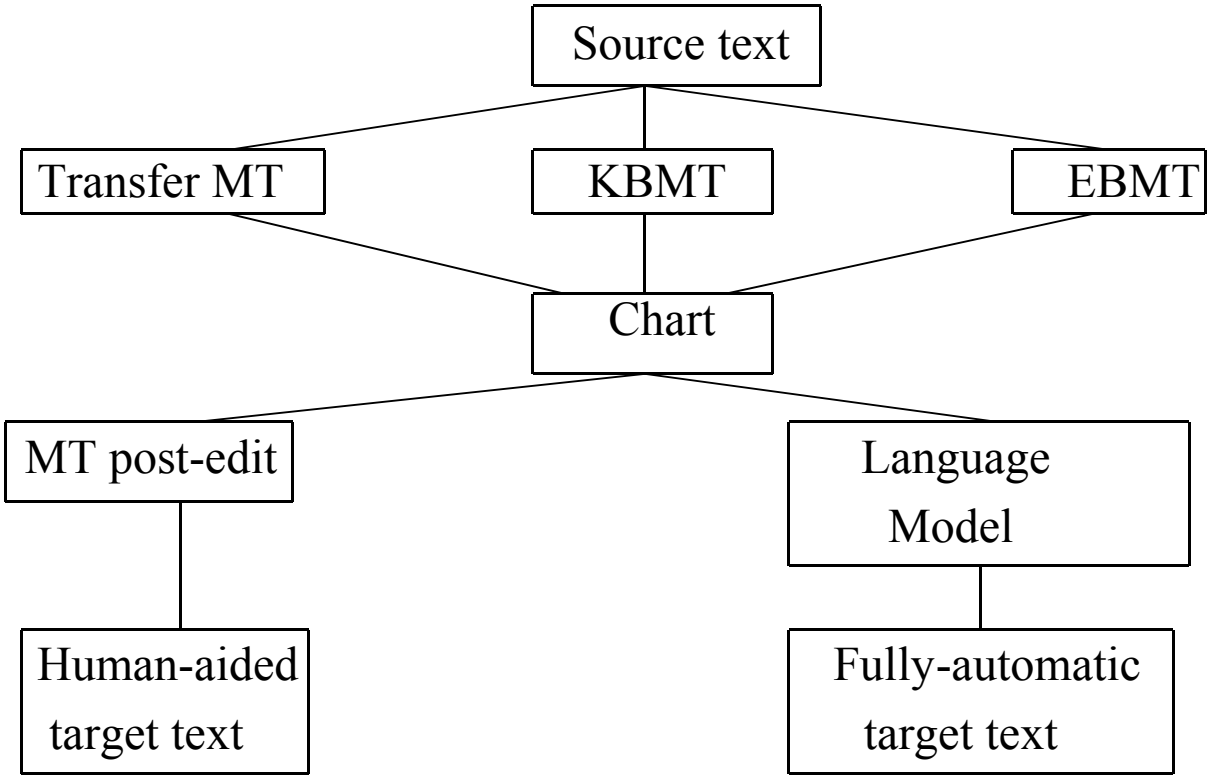
- [English] Advances in technology created new opportunities
- [Japanese] Because technology has advanced, opportunities have been created
- [or Japanese] Technology has advanced. There are new opportunities.

- **All methods of MT tend to retain SL structural features; however, theoretically SMT 'language model' approach should be more TL-oriented.**

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

Hybrid (multi-engine) systems

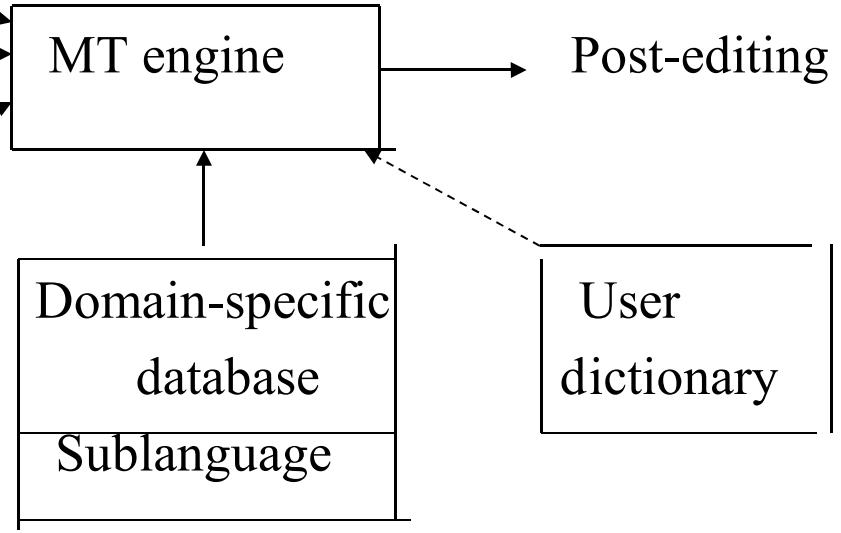


The translation demand

- Dissemination: production of ‘publishable quality’ texts
  - but, since raw output inadequate:
    - post-editing
    - control of input (pre-editing, controlled language)
    - domain restriction (reducing ambiguities)
- assimilation: for extracting essential information
  - use of raw output, with or without light editing
- interchange: for cross-language communication (correspondence, email, etc.)
  - if important: with post-editing; otherwise: without editing
- information access to databases and document collections
  - limited use before 1990

Operational MT (dissemination) typical configurations

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



Post-editing: types of errors for correction

- Misspelling in original not recognised, therefore not translated;
- missing punctuation
  - e.g. *The Commission vice president* translated as *Le président du vice de la Commission* (because no hyphen between *vice* and *president*)
- complex syntax
- prepositions:
  - …el desarrollo de programs de educación nutricional...
  - MT: …the development of programs of nutritional education
  - PE: …**in** nutritional education…
- verb phrases:
  - …el procedimento para registrar los hogares…
  - MT: the procedure in order to register the households
  - PE: …the procedure for registering households

Post-editing: types of errors (contd.)

- inversions:
  - …la inversión de la Argentina en las investigaciones de malaria
    - MT: …the investment of Argentina in the research of malaria
    - PE: Argentina's investment in malaria research
- reflexive verbs with inversions:
  - Se estudiarán todos los pacientes diagnostocados como…
    - MT: There will be studied all the patients diagnosed as…
    - PE: Studies will be done on all patients diagnosed as…
  - En 1972 se formuló el Plan Decenal de Salud para las Américas.
    - MT: In 1972 there was formulated the Ten-Year Health Plan for the Americas
    - PE: The year 1972 saw the formulation of the Ten-Year Health Plan for the Americas.

Translators and post-editors

- post-editing by translators:
  - not foreseen initially
  - skills acquired over time and practice in real working conditions
  - requires perseverance (initially post-editing takes longer than complete translation)
- advantages:
  - translators can maintain quality control
  - consistency of terminology
  - repetitive matter produced by MT, linguistic quality by HT
- disadvantages:
  - correction of ‘trivial’ mistakes; too often correcting same type of error
  - style too much SL oriented
  - translators as ‘slaves’ to machine
- need for special post-editing tools (not always provided)
- specially trained post-editors [still rare]

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
  - 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

MT for assimilation

- publication quality not necessary
- fast/immediate
- 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 (including those originally for mainframes and PCs
  - online MT has all types of rule-based systems - and now also SMT

MT and other LT applications

- document drafting (in poorly known languages)
- for tourists/shoppers: so far only dictionaries of words and phrases (hand-held devices).
- scanner-translator (scan/OCR/MT/print) - desktop, portable, online?
- interchange with deaf and hearing impaired: translation into sign languages [mainly research so far]
- information retrieval (CLIR): translation of search terms [very active field]
- information filtering (intelligence):
  - for human analysis of foreign language texts
  - document detection (texts of interest); triage (ranking in order of interest)
- information extraction: retrieving specific items of information (e.g. news analysts)
- summarization: producing summaries of foreign language texts [not yet feasible]
- multilingual generation from (structured) databases
- television subtitling [already available]

Evaluation of systems

- Who needs to know?
  - potential purchasers, potential users (translators), service managers, system developers, researchers
- Quality control
  - fidelity, accuracy (of terminology), comprehensibility, intelligibility, readability, appropriate style
- Usability
  - adaptability (e.g. to new domains), extendibility (e.g. to other languages and operating systems), compatibility (software and hardware), error levels (e.g. post-editing effort)
- Automatic evaluation
  - comparison of MT output and HT versions - emphasis on exact matches and close similarity of structures (statistical methods): tends to favour statistical MT systems

Some future developments and expectations

- merging of MT and TM for enterprise dissemination systems
- Internet as major (chief) resource - not only SMT
- rapid development of systems (SMT)
- reuse of MT components (for closely related languages)
- improvements in quality (evaluation, 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
- spoken language MT, domain-specific only [not general-purpose]
- embedding of MT in other LT systems
- bilingual (multilingual) communication as much as translation