

Current developments in machine translation

An overview of progress and future prospects

John Hutchins

Outline

- description of different approaches
 - Rule-based MT
 - Example based MT
 - Statistical MT
- speech translation
- evaluation
- problems of MT in general
- reducing problems: restriction, adaptation and control
- uses of MT
- translation tools, translation memory
- online MT, integration with other NLP systems
- quality and future developments

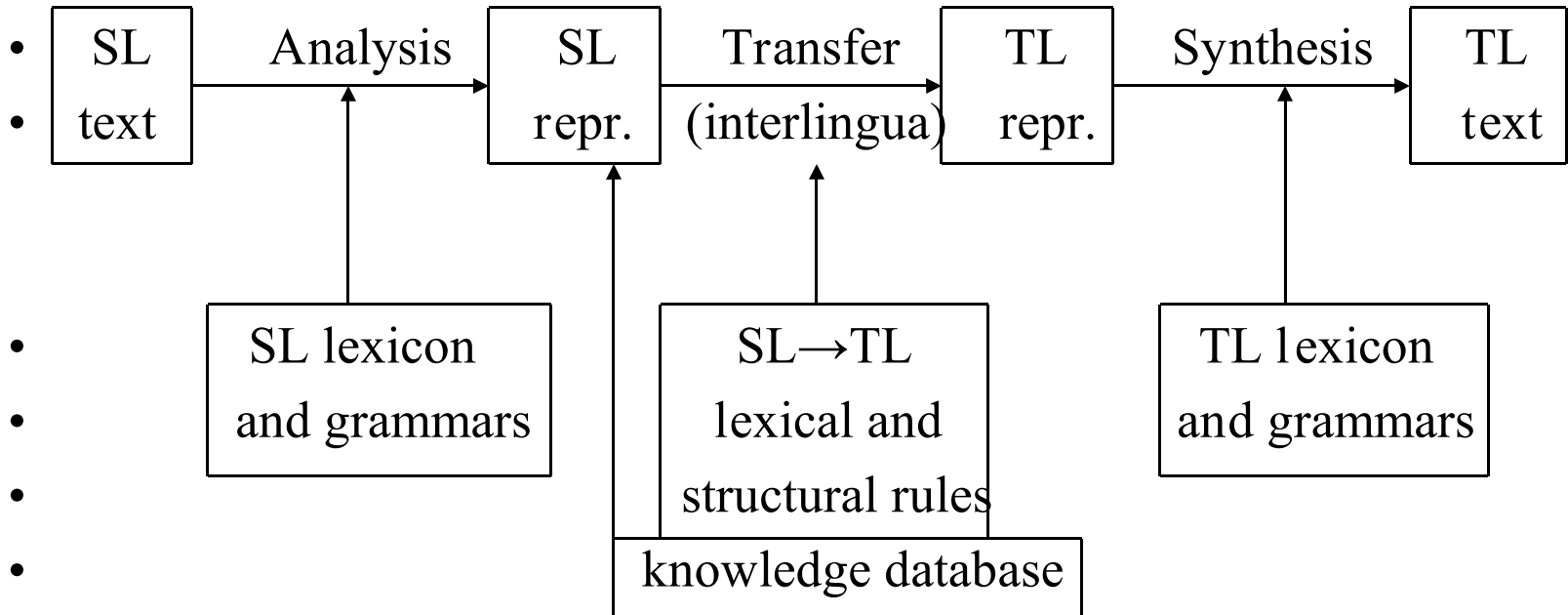
MT as it was in 1990

- Rule-based systems
 - latest: Knowledge-based MT
 - Transfer-based approaches: Eurotra, Ariane, etc.
- Use on Mainframes
 - still few PC systems
- Limitations
 - dictionaries
 - languages

Developments since 1990

- Availability of large text corpora
 - monolingual texts
 - bilingual texts (originals and translations)
- Corpus-based approaches
 - statistical machine translation
 - example-based machine translation
- Translation workstations
 - translation memory
- great increase in use of MT systems (both companies and individuals, online, etc.)

'Traditional' RBMT system



RBMT problems

- complexity of grammar rules
 - interactivity unpredictable, incomplete coverage
- complexity of dictionaries
 - incomplete coverage of meanings, selectional restrictions
- collocations, phrasal verbs, verb/noun phrases, etc.
- complex structures
 - long sentences, embeddings, discontinuities
- pronouns, anaphora
- semantic problems overcome (to some extent) by use of knowledge bases (in KBMT)
 - but knowledge bases hugely complex
- also overcome (to some extent) in domain restricted and/or controlled language systems

Simplistic RBMT (‘direct translation’)

- analysis of SL only as much as necessary for conversion into particular TL
- dictionary lookup followed by TL word-for-word output, then TL rearrangement
- dictionary entries include TL rearrangement rules
- as far as possible: one TL form for each SL word
- no analysis of SL syntax or semantics
- output close to SL structure
- many current MT systems (e.g. online) virtually word-word ‘direct’ systems

RBMT still continues

- direct approach common in commercial systems
- research on interlinguas still popular (e.g. Universal Network Language)
- knowledge-based systems being developed (e.g. Caterpillar)
- sublanguage systems usually ‘transfer-based’ (e.g. PaTrans)
- used in adaptation of systems for new (minority) languages
- enterprise and company-oriented systems are RBMT

Example-based MT

- origins: Nagao (1981)
- first motivation: collocations, bilingual differences of syntactic structures
- basic idea:
 - human translators search for analogies (similar phrases) in previous translations
 - MT should seek matching fragment in bilingual database, extract translations
- aim to have less complex dictionaries, grammars, and procedures (shallow parsing, no semantic analysis)
- improved generation (using actual examples of TL sentences)

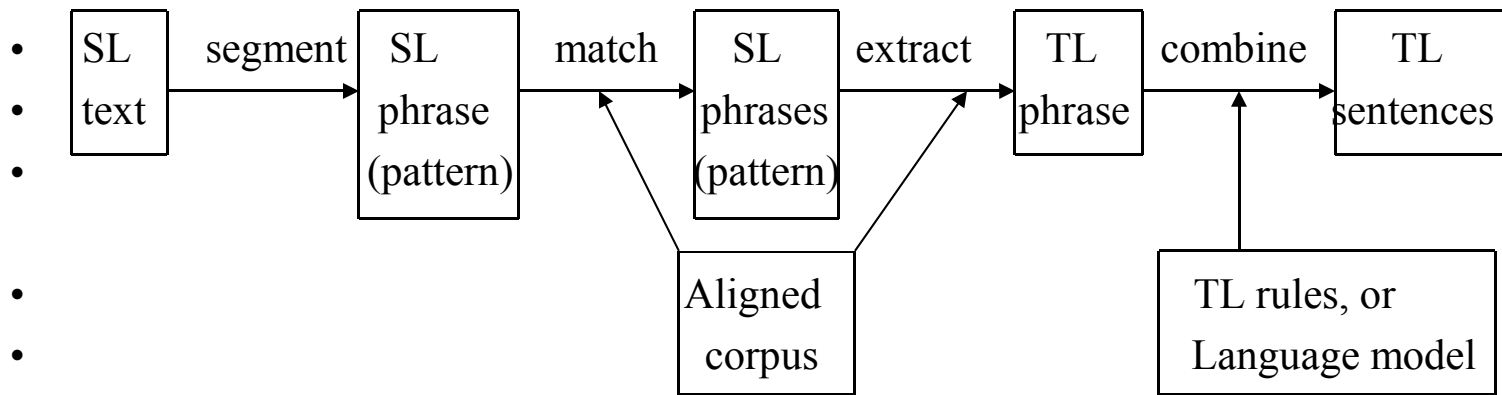
Two tendencies in EBMT

- From beginning two tendencies:
- EBMT as supplement to RBMT systems
 - as continuation of RBMT tradition
 - leading to ‘hybrid’ systems
- EBMT as a discrete approach
 - either as a new ‘paradigm’ (complete break with the past)
 - or rather as a new ‘framework’ (since EBMT researchers acknowledge and use work of predecessors)

‘pure’ EBMT processes

- core process is selection and extraction of TL elements (fragments) corresponding to SL fragments
- preceding stage is decomposition of input into fragments (or templates with or without variables) and matching against SL fragments in the database
- succeeding stage of synthesis (‘recombination’) adapts extracted TL fragments and combines them as output sentences
- preparatory stage for alignment of SL and TL sentences in the database, and/or for deriving templates and patterns used in matching and extraction
- alternatively, templates and patterns may be derived during run-time

EBMT schema



Example-based MT: some problems and issues (1)

- bilingual aligned corpora
 - size: adding examples may improve performance or may degrade performance
 - repetition of same or similar examples may reinforce selection or may be unnecessary clutter
 - suitability of examples: automatically compiled or manually compiled
 - need: phrases/clauses aligned (not sentences), length is open issue
 - stored: as word strings or as annotated trees(e.g. dependency or case grammar trees)
- 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.)
- analysis of corpus at run-time or in advance

Example-based MT: some problems (2)

- matching by characters:
 - This is shown as A in the diagram ↔ This is shown as B in the diagram
 - The large paper tray holds up to 400 sheets <≠> The small paper tray holds up to 300 sheets
 - (because system does not know that *large* and *small* are similar/substitutable)
- matching by words via thesaurus (close in meaning)
 - English *eat* → Japanese *taberu* or *okasu*
 - A man eats vegetables ↔ Hito wa yasai o taberu
 - Acid eats metal ↔ San wa kinzoku o okasu
- problem of large thesaurus (similar to RBMT problem of knowledge base)

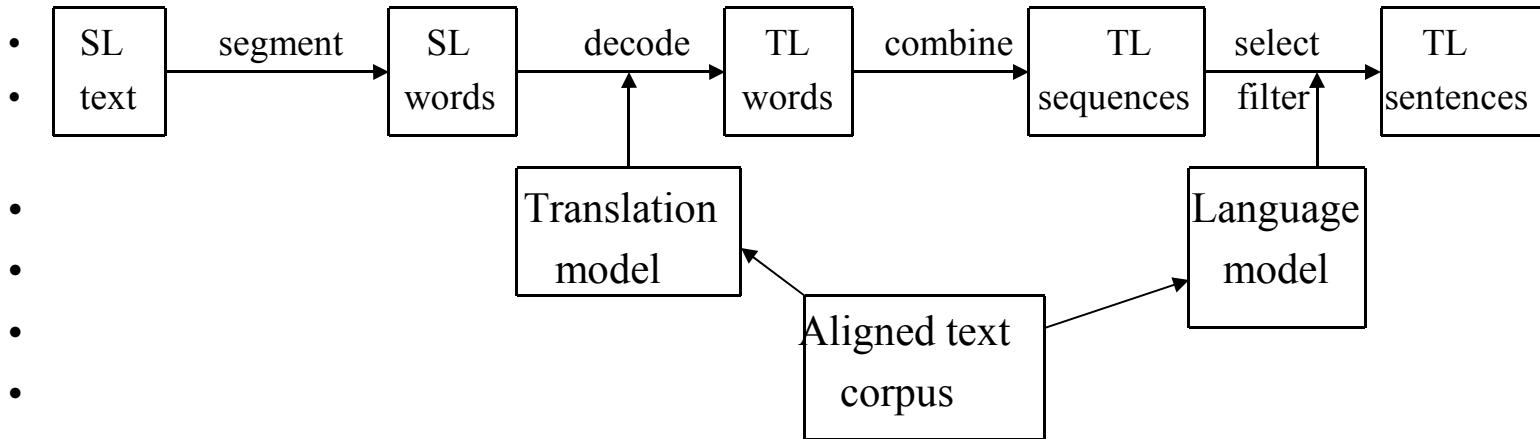
Example-based MT: some problems and issues (3)

- examples in database:
 - (1e) The obstinate man refused all help
 - (1g) Der hartnäckige Mann hat alle Hilfe verweigert
 - (2e) Help was rejected by the stubborn man
 - (2g) Hilfe wurde von dem starrköpfigen Mann abgewiesen
- sentence to be translated:
 - (3e) Help was rejected by the obstinate man
- but failure to relate *verweigern* and *abweisen*, and *hartnäckig* and *starrköpfig*
- therefore no matching
- but if example for ‘obstinate man’ (1g) inserted into (2g), wrong morphological form:
 - (3g) * Hilfe wurde von der hartnäckige Mann abgewiesen.
- **In general, morphological variation handled more easily by rule-based systems than by corpus-based systems**

Statistical MT basics

- based on assumption that translations observe statistical regularities
 - origins: Warren Weaver (1949): “When I look at an article in Russian, I say: ‘This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode’”; Shannon’s information theory; speech recognition
- core process is the probabilistic ‘translation model’ taking SL words or phrases as input, and producing TL words or phrases as output
- succeeding stage involves a probabilistic ‘language model’ which synthesizes TL words as ‘meaningful’ TL sentences
- preceding stage (matching) locates input words and phrases against entries in translation model
 - involving segmentation and matching processes
- vital pre-processing stage is the creation of the (bilingual) translation models and (monolingual) language models based on statistical analyses of the corpus (or corpora)

Statistical MT schema



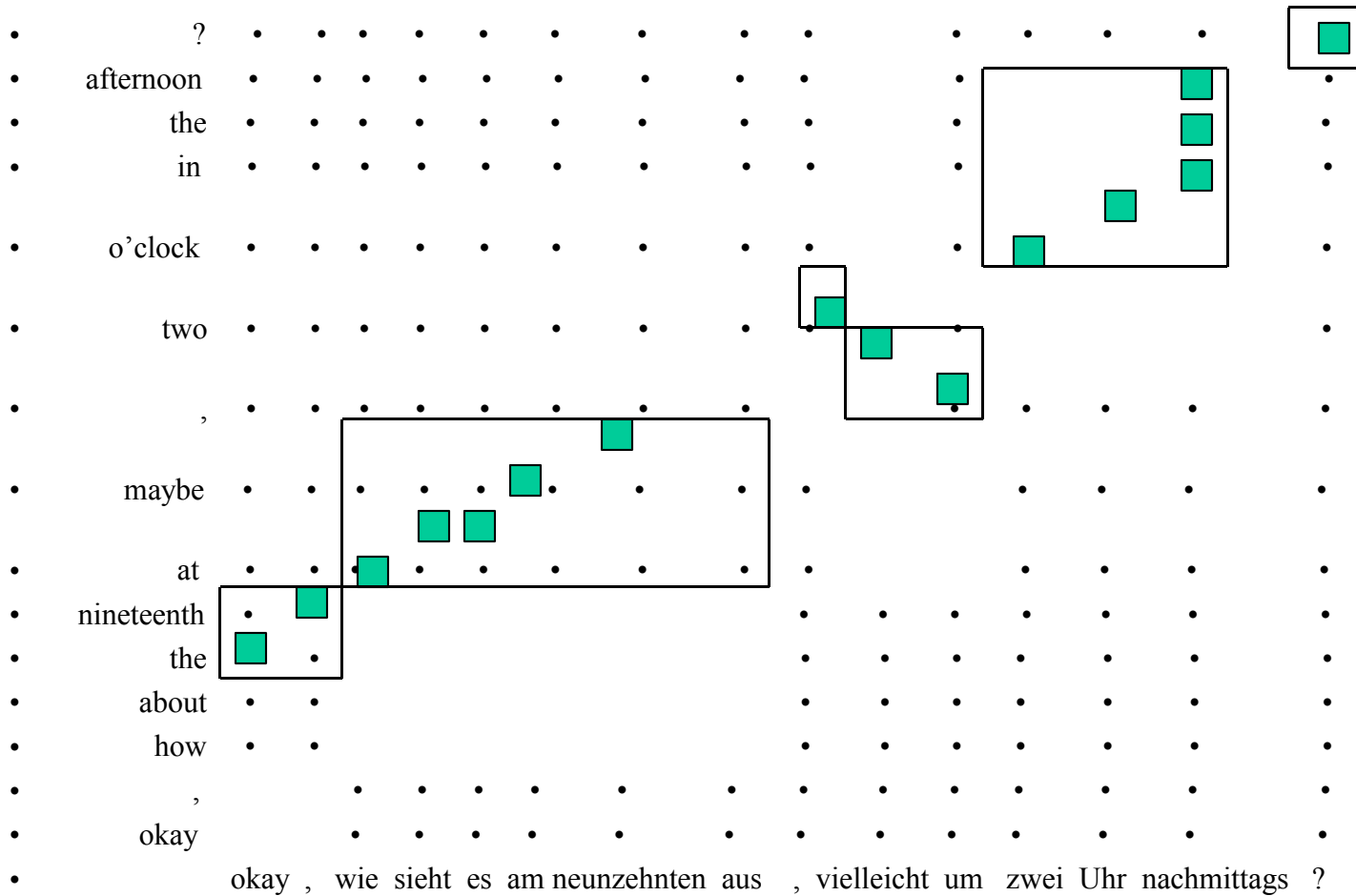
Statistical MT processes

- Bilingual corpora: original and translation
- 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
- Alignment: sentences aligned statistically (according to sentence length and position)
- Decoding: compute probability that a TL string is the translation of a SL string (‘translation model’), based on:
 - frequency of co-occurrence in aligned texts of corpus
 - position of SL words in SL string
- Adjustment: compute probability that a TL string is a valid TL sentence (based on a ‘language model’ of allowable bigrams and trigrams)
- search for TL string that maximizes these probabilities
- $\operatorname{argmax}_e P(e/f) = \operatorname{argmax}_e P(f/e) P(e)$

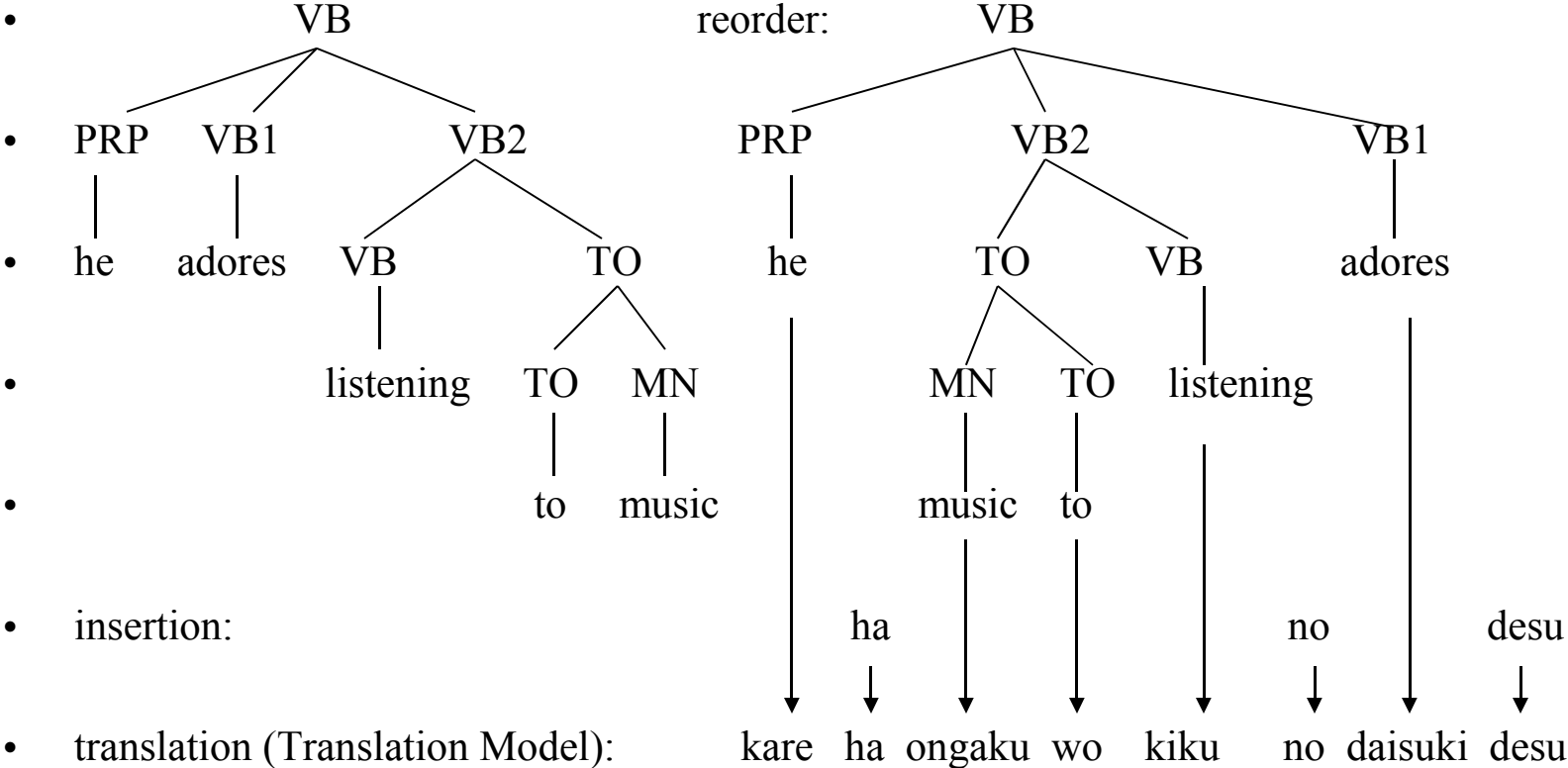
SMT issues

- ignores previous MT research (new start, new ‘paradigm’)
 - basically ‘direct’ approach: (a) replaces SL word by most probable TL word, (b) reorders TL words
 - decoding is effectively kind of ‘back translation’
- originally wholly word-based (IBM ‘Candide’ 1988) ; now predominantly phrase-based (i.e. alignment of word groups); some research on syntax-based
- mathematically simple, but huge amount of training (large databases)
- problems for SMT:
 - translation is not just selecting the most frequent ‘equivalent’ (wider context)
 - no quality control of corpora (unlike EBMT)
 - lack of monolingual data for some languages
 - insufficient bilingual data (Internet as resource)
- merit of SMT: evaluation as integral process of system development

SMT phrase alignment: example



Syntax-based SMT translation model



Spoken language MT

- probably most desired translation technology of all
- but MT with most intractable problems
- many potential applications
- interim ‘solutions’: voice input, text translation, voice output
- research using variety of methods (RBMT, EBMT and statistical analysis, speech recognition, etc.)
 - Japanese groups first to develop EBMT methods
- research concentrating on narrow domains
 - hotel booking, conference registration (since late 1980s)
 - military: phrasebook-type
 - medical consultations (doctor: questions; patient: simple answers)

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

Evaluation

- manual: EAGLES, FEMTI, DARPA
 - comprehensibility, readability, fidelity, adequacy, usability (e.g. amount of editing), appropriateness (in usage context)
 - ‘surprising’ DARPA results (1994) showing Candide (SMT) almost as good as Systran (RBMT): better fluency, although worse adequacy/fidelity
- automatic evaluation: testing of alternative versions of SMT systems
 - e.g. with/without morphology, structure, domain vocabulary
- BLEU (use of ‘reference texts’, human translations)
 - percentage of word sequences (n-grams) occurring in reference texts
- NIST (as BLEU with weighted n-grams)
- mWER (multi-reference word error rate)
 - number of edit operations (insertion, deletion, substitution) to transform MT output into one of reference texts
 - penalizes differences of word order
- mPER (as mWER but ignoring word order)

BLEU problems

- Large test corpora needed (20K words for 1% differences to be statistically significant)
- low correlation between sentence-level BLEU scores and sentence-level human judgements
- problematic to compare
 - different language pairs, different number of ‘reference texts’, different n-gram sizes, and different test corpora
- scoring depends on agreement between human translators
- counter-intuitive results:
 - SMT system 51% BLEU; commercial (RBMT) system 34% BLEU; HT 63% BLEU
 - MT output may be better (in BLEU scores) than HT -- but subjectively still worse
- i.e. BLEU cannot rank human quality translation; only useful for comparing SMT systems

Comparison of EBMT and RBMT

- both EBMT and RBMT represent SL input as strings, templates, patterns, structures
 - usually parsing in EBMT is shallow, while in RBMT usually ‘deep’
 - and EBMT usually little semantic analysis
- in some EBMT, structured representations of SL and corresponding TL (e.g. dependency trees) similar to RBMT representations
 - if decomposition, matching, extraction, recombination based on dependency (sub)trees, then:
 - EBMT processes are identical to RBMT tree transduction and comparison processes, and such EBMT systems are in effect RBMT systems
- core difference: EBMT representations are derived from example databases
 - however: RBMT rules can also be derived from bilingual databases

Comparison of EBMT and SMT

- initially distinct: SMT decomposition, matching and extraction based on individual SL words; while EBMT decomposition, matching and extraction based on strings (word sequences, fragments, examples)
- recent ‘phrase-based’ and ‘syntax-based’ SMT blurs distinction
 - methods introduced primarily to improve alignment and matching processes
- SMT is closest to EBMT when input is parsed, matching based on parsed representations in database and output to ‘language model’ also as parsed representations
 - in such cases the only remaining difference: SMT works exclusively with statistical methods, EBMT works mainly with symbolic (linguistic) fragments and text examples during the ‘core’ process
- many SMT researchers regard EBMT as a type of SMT

SMT and RBMT

- SMT black box: no way of finding how it works in particular cases, why it succeeds sometimes and not others
- RBMT/EBMT: rules and procedures can be examined
- RBMT and SMT are apparent polar opposites, but gradually ‘rules’ incorporated in SMT models
 - first, morphology (even in versions of first IBM model)
 - then, ‘phrases’ (with some similarity to linguistic phrases)
 - now also, syntactic parsing

Bilingual lexical differences

- bilingual lexical ambiguity (more than one equivalent, whether ambiguous in SL or not):
 - river: fleuve/rivière; Fluss/Strom
 - Taube: dove/pigeon
 - Schraube: screw/bolt/propeller
 - 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
 - kenner van het Turks: *knower of Turkish, someone who knows Turkish
- **Bilingual lexical ambiguity solved (?) by contextual rules (RBMT), or by examples (EBMT), or by word-word frequencies and ‘language models’ (SMT), but this is a perennial difficulty whatever the method**
- **lexical gaps are virtually unsolvable by any method**

Structural ambiguity

- (1) Peter mentioned the book I sent to Mary
 - 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
 - ... the man you told us about yesterday
- (2b) We will meet the man you told us about tomorrow
 - we will meet tomorrow the man...
- (3) pregnant women and children
 - des femmes et des enfants enceintes
- (4a) Smog and pollution control are important factors
- (4b) Smog and pollution control is under consideration
- (4c) The authorities encouraged smog and pollution control
- (5a) Old men and women receive a state pension
- (5b) Tickets were refunded for children, old men and women
- **Problems (1), (2), (3), and (5a) may be ‘solved’ by SMT ‘language model’ and by EBMT databases. But problems (4c) and (5b) require ‘knowledge’ (i.e. rule-based KBMT)**

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 rules (tree-transduction in RBMT) to cover all circumstances and contexts; example-based (EBMT) and statistics-based (SMT) yet to prove any better; possibly examples like the one in no.4 are inherently unsolvable**

Anaphora

- Die Europäische Gemeinschaft und ihre Mitglieder
 - The European Community and its members (*ihr* usually translated by *her*)
- The monkey ate the banana because it was hungry
 - Der Affe ass die Banane weil er Hunger hat
- The monkey ate the banana because it was ripe
 - Der Affe ass die Banane weil sie reif war
- The monkey ate the banana because it was lunch-time
 - Der Affe ass die Banane weil es Mittagessen war
- Particular problem when translating from Japanese where it is good style to omit the subjects of verbs and to avoid repetition.
- **Sentence-orientation of all types of systems makes most anaphora unresolvable; possibly only a discourse-oriented ‘language model’ is the only chance (no sign of one yet!)**

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 (RBMT) approaches**
- **No solutions with corpus-based (SMT and EBMT) approaches**
- **Perhaps only solution using Artificial Intelligence approaches (Knowledge-based machine translation)**
- However, perhaps this aspect is sometimes 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 [nominalization style]
- The engineers discussed whether it was possible to rectify the fault by inserting a valve [preference for verb forms]
- [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 could be more TL-oriented.**

Reducing problems

- Domain restriction (sublanguages)
 - e.g. weather reports, patents, hotel reservations, doctor consultations
- Style limitations (e.g. only complete sentences, no interrogatives)
- Adaptation of input
- Controlled language input
- Multiple MT engines: selecting the ‘best’ output

Adaptation of input

- MT-ese
 - writing with MT in mind (i.e. to avoid ambiguities)
- pre-editing
 - marking words for grammatical category
 - e.g. *convict* as noun or verb
 - indicating proper names
 - e.g. to ensure that *John White* is not translated as *Johann Weiss*
 - indicating compound nouns
 - e.g. to translate *light bulb* as *ampoule* and not *bulbe léger* or *oignon léger*
 - marking parenthetical phrases
 - e.g. *There are he says two options...* as *There are (he says) two options...*
 - dividing sentences into shorter clauses
 - in theory, need not know target language(s)

Controlled language

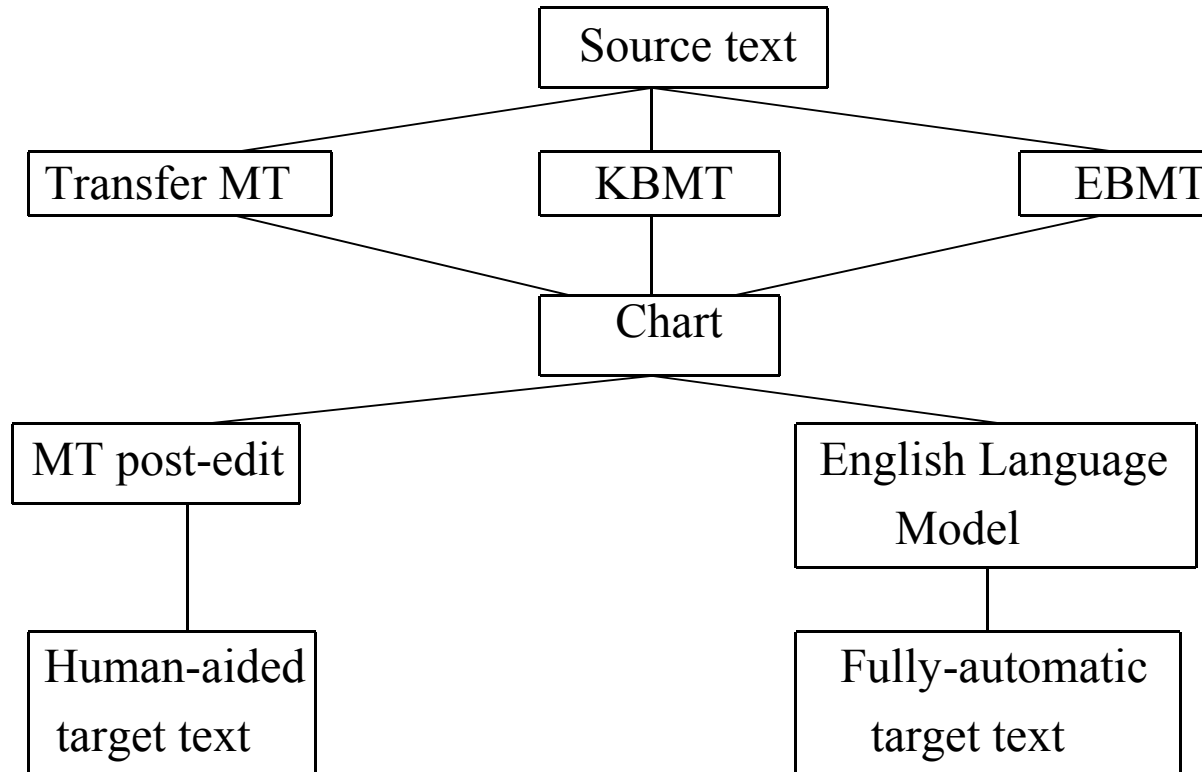
- Controlled authoring of the source text in standard manner, suitable for unambiguous translation
- Typical rules:
 - use only approved terminology, e.g. *windscreen* rather than *windshield*
 - use only approved sense: *follow* only as ‘come after’, not as ‘obey’
 - avoid ambiguous words: *replace*,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

Multi-engine and hybrid systems

- multi-engine: many systems, aiming for ‘best’ output (algorithms for comparison/integration, e.g. TransBooster, DEMOCRAT)
- hybrid: combining different approaches
- integration of SMT and RBMT (e.g. ArchTran)
- integration of SMT and EBMT
 - many SMT researchers regard EBMT as a subtype of SMT
- integration of EBMT and RBMT (e.g. Microsoft)

- EBMT may be most adaptable framework for incorporating methods from variety of approaches

Multi-engine system: an example (Pangloss Mark III)



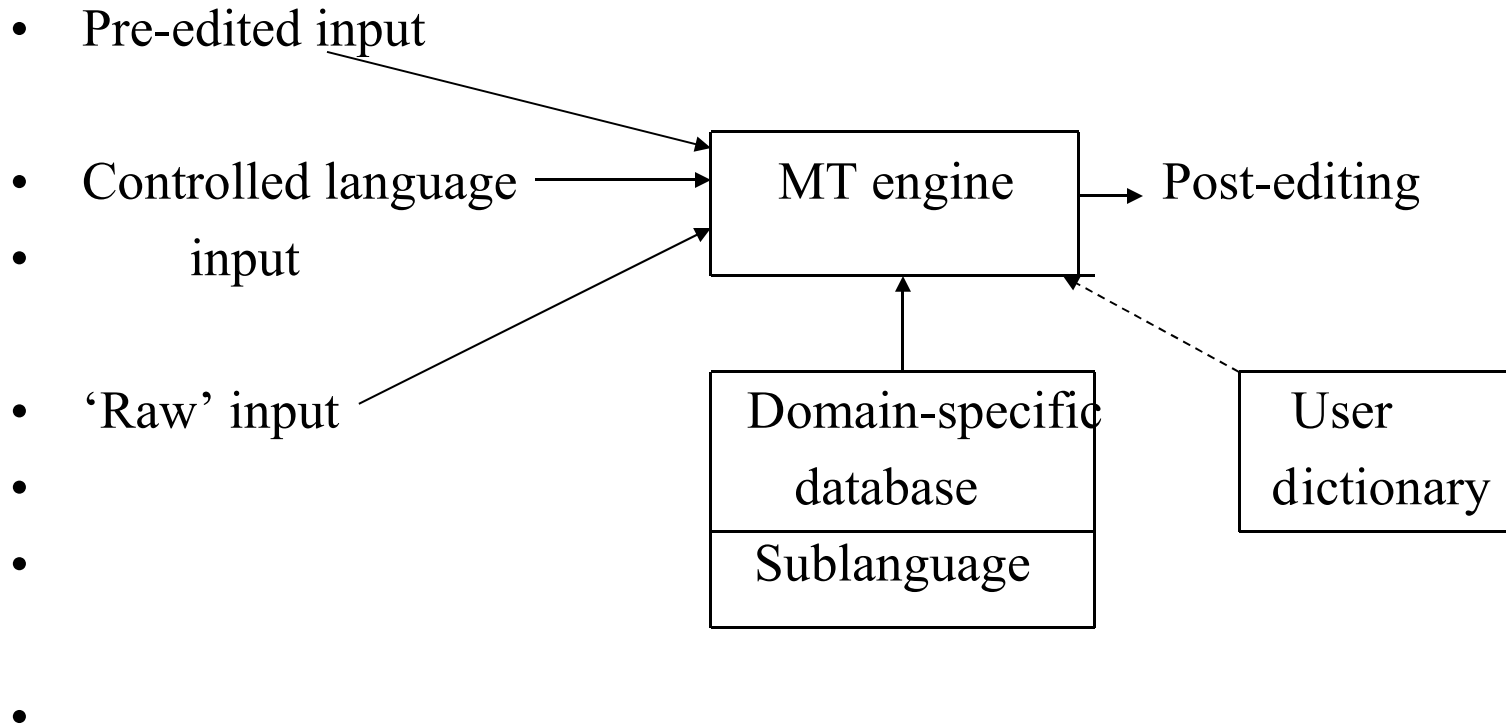
MT in use: issues

- MAHT (machine aids for human translation) and HAMT (human assisted MT)
- types of demand: dissemination (good quality) vs assimilation
- types of users: companies, government bodies, translators, individuals
- printed input vs electronic input
- platforms: mainframe, PC, Internet, PDA, cell phones (mobiles)

The translation demand

- dissemination: production of ‘publishable quality’ texts
 - but, since raw output is 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
- categories of systems: home (personal), professional, enterprise

Human-assisted MT



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.)

Issues for corporations

- MT or translation aids (TM)
- General-purpose system or specialised/customized system
- Controlled language (existing or developed in-house)
- Lexical resources (creation, maintenance)
- Translation memories (creation, use, maintenance)
- Control of terminology
- Quality control
- Standards; exchange formats
- Compatibility (hardware, software)
- Integration: technical authoring, publishing
- Management/staffing implications

Software (enterprises, government)

- requirements: client-server (intranet) systems, customizable
- facilities: large basic dictionary, technical dictionaries, user dictionaries
- platforms: Windows NT, Unix, Sun Solaris; or browser (client) access to server
- languages:
 - English, French, German, Italian, Portuguese, Spanish
 - LogoMedia Enterprise Solutions, PeTra Enterprise, Reverso Intranet, SDL Enterprise Translator, Systran Enterprise, WebSphere Translation Server (IBM)
 - English, Japanese, Korean, Chinese
 - ATLAS (Fujitsu), EWTranslate, Systran Enterprise, TranSphere (AppTek), WebSphere (IBM)
- other languages
 - TranSmart [Finnish], TranSphere [Arabic]
 - and one SMT system: LanguageWeaver [Arabic, Chinese, Hindi, Somali]

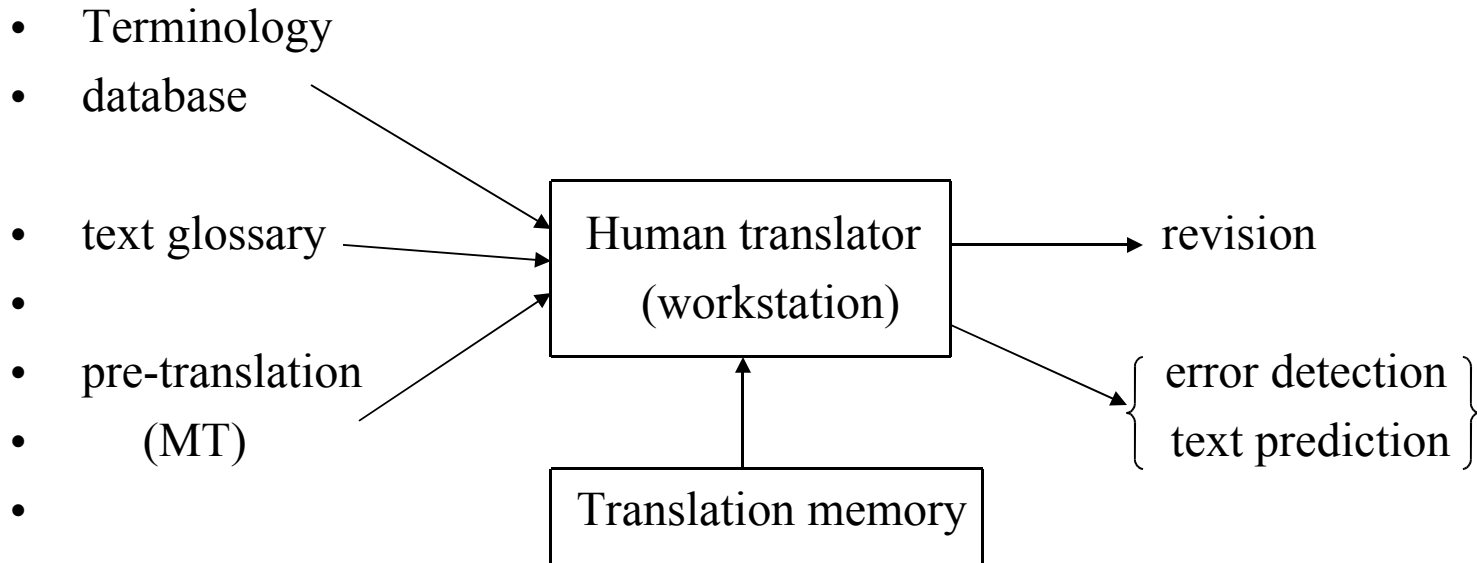
In-house systems: examples

- Pan American Health Organization [medical, social, welfare]
 - Spanish→English (SPANAM), English→Spanish (ENGSPAN), Portuguese→English (PORTENG)
- Japan Center for Science and Technology [abstracts]
 - English→Japanese
- NHK [news broadcasts]
 - Japanese→English
- IBM Japan
- CSK (Japan)
- PaTrans:[patents]
 - English→Danish
- GSI Erli
- Hook and Hatton [chemistry texts]
 - Dutch→English

Post-editing

- Why needed?
 - 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: longer sentences more difficult for MT
- Always necessary?
 - more standardised, more jargon-full documents mean less correction
- Can it be avoided?
 - if rough version acceptable

Machine-aided human translation



Computer-aided translation tools

- based on recognition that fully automatic translation not appropriate for professional translators; desire of *translators to be 'in control'*
- development (availability), mainly since 1985, of :
 - PCs and multilingual word processing, desk top publishing
 - dictionaries (monolingual, bilingual): on-line access
 - user glossary, terminology management, 'authorised' terms, 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 memory

- based on sets of original texts and their ‘authorized’ translations
- particularly suitable for translation of revisions and for translating standardized documents
- most suitable for large (organizational) translation agencies/departments
- major gains (time saving, etc.) from retrieving already translated text
- revised texts (i.e. updated documents) are checked against corpus for any changes; for unchanged source sentences, the ‘authorized’ translation is retained
- processes:
 - alignment of bilingual text corpora
 - search of exact matches or ‘fuzzy’ matches
 - extract target phrase for insertion and/or amendment (by human translator)

Translation memories: weaknesses

- sentence-based comparisons restrict potential use (no phrase matching)
- any TM likely to contain redundant, ambiguous versions; unnecessary (redundant), untypical (misleading); and 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)
 - newly translated texts added after the process
- fuzzy matching often too complex, and translators opt not to use the facility
- TM systems do not help in combining extracted phrases
- developments needed:
 - finding phrases (retrieval, fuzzy matching)
 - searching for words in combination (e.g. ...*take*... + ...*a swipe at*...)
 - re-combining phrases to produce sentences
- example-based MT research (integrated in: Déjà Vu)

Localization

- Internationalisation, globalisation (e.g. software and Web pages)
 - estimated market (end 2006) is \$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 (days, not weeks)
- Repetitive (translation memory)
- Graphics, formatting, layout, etc. (to be preserved)
- **companies use both translation tools (workstations, translation memories) and MT systems**
- own association: Localization Industry Standards Association

Translation prediction, error correction

- Text prediction
 - interactive drafting of TL text: anticipating (suggesting) continuations
 - using bilingual and monolingual databases
 - based on bigram and trigram frequencies (translation and language models)
 - TransType (Montreal), still under development
- Error correction
 - using aligned texts (original and translation)
 - using external resources (dictionaries, grammar rules, terminology)
 - to identify omissions (sentences), morphological errors, deceptive cognates (*faux amis*), names
 - for use by experts (translators) in revision process
 - TransCheck (Montreal) under development since early 1990s

Convergence of HAMT and MAHT

- increasingly, systems straddle different categories
- workstations (TM systems) include MT components (e.g. Trados, Atril, TransType)
- MT systems include Translation Memories
- enterprise and localization systems share 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 merging rule-based and database methods
- future: full integration (no distinctions)

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, PDAs, mobile phones
- systems can be any of those primarily designed for dissemination, e.g. as Systran (at EC) and earlier systems; and any PC system
- online systems

Online and PC translation: why so bad?

- old models (word for word, simplistic RBMT architecture)
 - often single equivalents, no morphological analysis or target adjustment; some no more than electronic dictionaries
- dictionaries too small, insufficient information, and difficult (or impossible) to update
- weak syntactic analysis/transfer -- only simple clauses possible
- 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
- **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**
- **potential improvements by use of SMT 'language model' for TL output?**

MT in the marketplace

- retail availability
 - many only purchased direct from manufacturer
- confusion of terms:
 - ‘translation systems’ no more than dictionaries
 - ‘computer aided translation’ either HAMT or MAHT
 - combination 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
- bench marks, consumer reports/reviews (evaluations)
- risks of marketplace (many systems have failed)

Quality improvements (of core systems)

- RBMT -- marginal improvements for French and German to English in last 10 years, none for Russian to English in last 20 years [Hutchins at MT Summit 2003]
- EBMT -- no evidence
- SMT -- claim that more data (i.e. larger aligned corpora) means better results:
 - [Och tutorial at MT Summit 2005]
 - BLEU 54%
 - 53
 - 52
 - 51
 - 50
 - 49
 - 48
 - 47
 - 46
 - 75M 150M 300M 600M 1.2B 2.5B 5B 10B 18B
- doubling test corpus (Arabic-English) produces about 0.5% higher BLEU score
- comparisons of SMT with RBMT inconclusive (different measures), but general agreement that ‘intuitively’ RBMT better for fidelity and comprehensibility, and RBMT always better on untrained corpora

Evaluation

- 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)
- Task suitability
 - dissemination/assimilation: publishing, gisting, extraction, triage, detection, filtering
- Resources evaluation
 - suitability and quality of dictionaries, terminology resources, translation memories (databases)
- Methods
 - Black box vs. glass box; test suites (set of 'standard' texts); interviews

Requirements for future development of domain-specific systems

- domain-specific systems for dissemination developed for specific companies/associations
- [for assimilation: may become generally available online, at cost]
- RBMT: expansion/adjustment of dictionaries, changes in grammars
- KBMT: expansion of RBMT by access to domain knowledge
- EBMT (ideally developed from scratch): specific database, specific parsing, templates, etc.
- SMT (developed from scratch): specific database, domain-oriented ‘translation model’ and ‘language model’.
- [domain-specific systems are obvious ‘niches’ for EBMT and SMT]
- spoken language MT: inevitably domain-specific

Rapid development for ‘minor’ languages

- RBMT: limited reusability or adaptation of grammars and dictionaries (mainly feasible for closely related languages)
- EBMT: faster than RBMT if algorithms for matching and extraction are transferable; however, the more parsing, preprocessing of templates, etc. the less easily developed
- SMT: fastest since (ideally) only change of bilingual database; however, the more phrase-based, the more complex preprocessing and matching

Further applications: MT for interchange

- [interchange does not demand exact translation]
- in principle, any systems can be used for written interchange (correspondence, emails):
 - 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)
- 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 and other LT applications

- document drafting
 - Japanese researchers, EC administrators, school essays
- 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)
 - 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

Future of online MT

- general-purpose ‘assimilation’ systems for major languages: downloaded or used online (not PC packages) -- free for short texts, unformatted
- general-purpose systems for emails and ‘text-messaging’
- ‘added value’ systems: retain text formats, webpage graphics (?), special characters; also longer texts and post-editing facilities [already available online]
- general-purpose ‘assimilation’ systems for ‘minor’ languages: online only
- special-purpose (domain-specific) ‘assimilation’ systems -- e.g. for medical, legal, sports, etc. texts: online only, with extra fees
- ‘dissemination’ systems downloaded (fewer packages); maintenance and trouble-shooting online
- text only [no spoken language MT online in foreseeable future]

Summary of strengths and weaknesses

- Relative strengths (potentially) of different approaches
- SMT: for poor resources, domain restriction, rapid development, ‘minor’ languages; but not good for online (slow running)
- RBMT: for general-purpose (or domain-specific) ‘major’ languages, where quality important (controlled input, post-edited output); otherwise information-only ‘assimilation’ online
- EBMT: good for controlled, domain restricted, quality output; but not good for online
- Hybrid (multi-engine) approaches: general-purpose and specialised; but probably not good for immediate results

Summary of future developments and expectations

- merging of MT and TM for enterprise dissemination systems
- Internet as major (chief) resource
- 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

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, LISA Forums
- Journals:
 - *Multilingual Computing and Technology*
 - *MT News International*
- *Compendium of translation software* [directory of current commercial systems on EAMT website]
- *Machine Translation Archive* (<http://www.mt-archive.info>)
- my website:
 - <http://ourworld.compuserve.com/homepages/WJHutchins>