

Future prospects in machine translation usage and research

John Hutchins

Leeds, February 2006

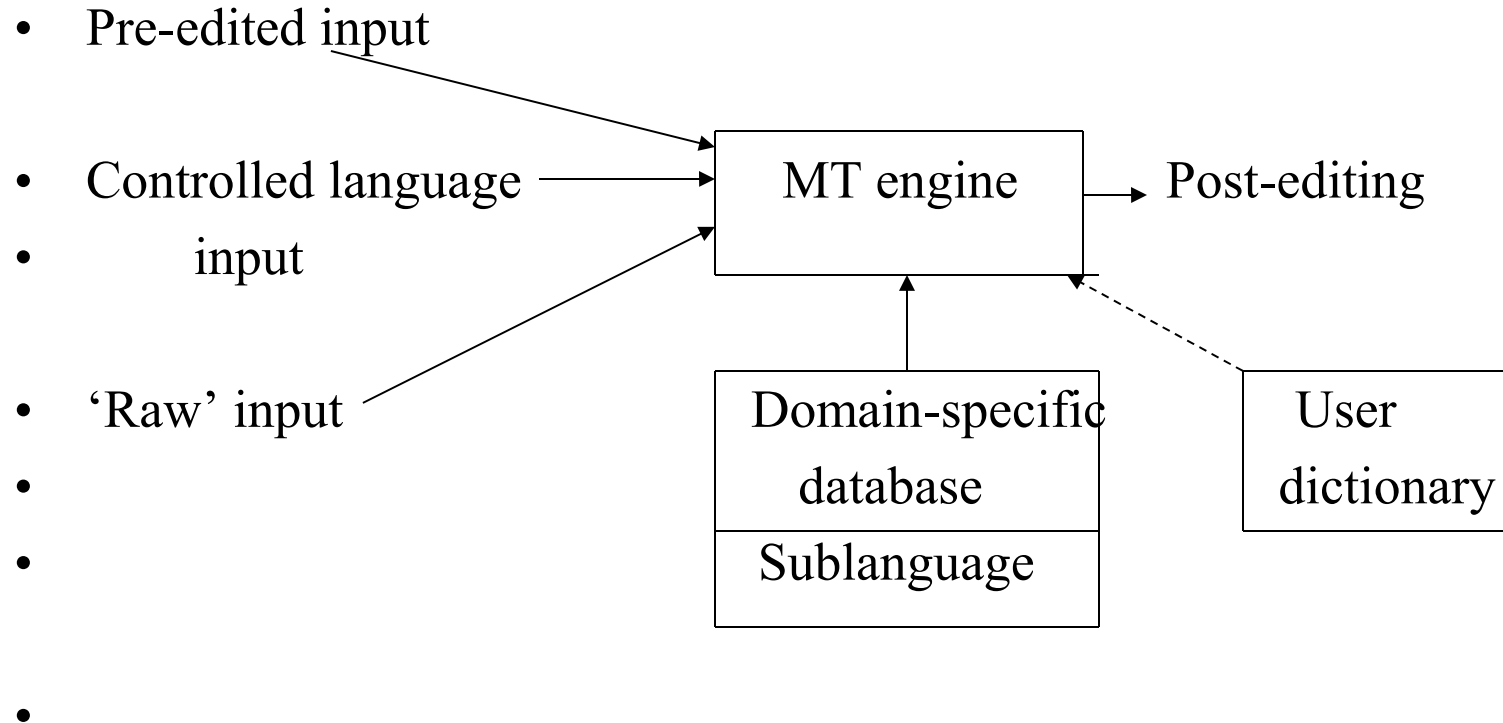
Overview

- Large organisations
- Individual translators
- Other users
- Research prospects
- Problems and expectations

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
- types of users: companies (managers), translators, individuals
- different platforms: PC, internet, PDA, mainframes (intranets)

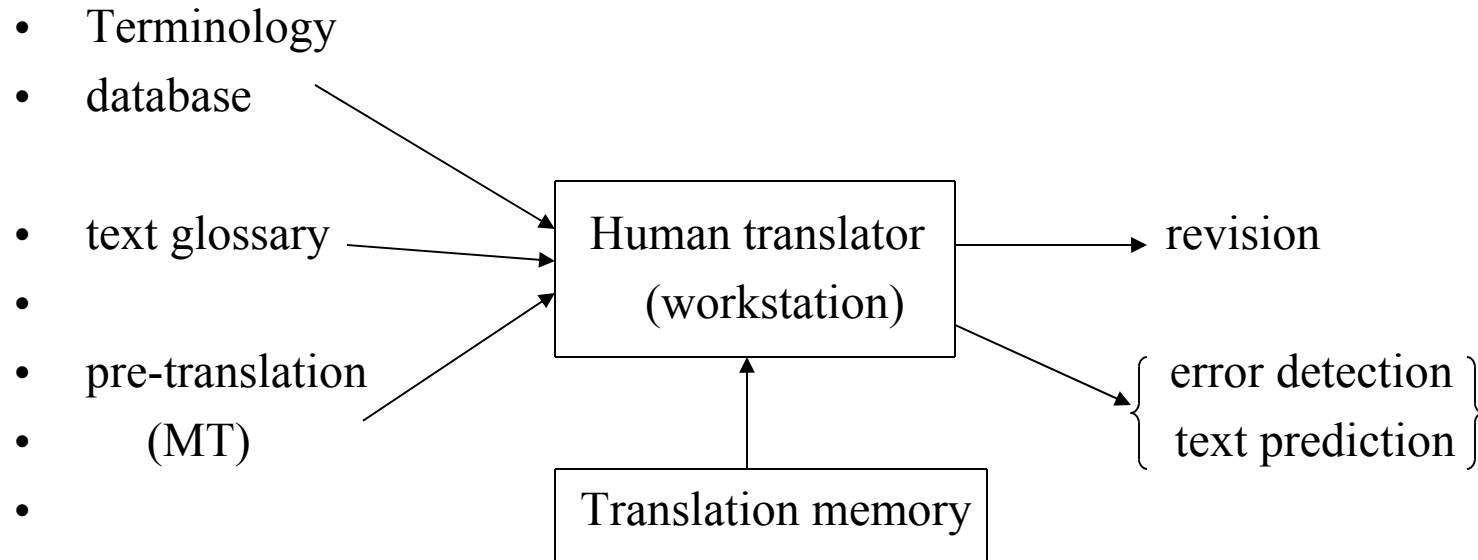
Human-assisted MT



Large-scale translation: needs

- accurate, good quality, publishable (dissemination)
- publicity, marketing, reports, operational manuals, localization, website localization
- technical documentation; large volumes, repetitive, frequent updates
- multilingual output (e.g. English to French, German, Japanese, Portuguese, Spanish)
- system choices: MT or translation aids (TM), or both; general-purpose system or specialised/customized system; commercial system or in-house system
- resources: controlled language (existing or developed in-house); lexical and terminological resources (creation, maintenance, control); translation memories (creation, use, maintenance)
- management: standards; exchange formats; compatibility (hardware, software); integration with technical authoring, publishing

Machine-aided human translation



Translation memories and their weaknesses

- based on sets of original texts and their ‘authorized’ translations
- particularly suitable for translation of revisions and for translating standardized documents
- **Problems:** 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
- **Future** developments needed:
 - finding phrases (retrieval, fuzzy matching)
 - searching for words in combination (e.g. ...*take*... + ...*a swipe at*...)
 - re-combining phrases to produce sentences
 - benefits from example-based MT research (already partly integrated in Déjà Vu)

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

Large organisations: future needs

- Post-editing: immediate (online) updating of TMs and MT systems
- Authoring: controlled language designed for translation
- Terminology: immediate updating (inclusion in TM and MT dictionaries)
- Text prediction, error detection (for missing text, incorrect/unauthorized usages)
- Full integration of MT and TM [already some workstations (TM systems) include MT components (e.g. Trados, Atril) and many MT systems include Translation Memories]
- Multiple facilities: terminology management; integration with authoring and publishing systems; project management; quality control; Internet access and downloading; Lexical acquisition; Web translation
- ‘Database to document’ production
- Full integration into company-wide documentation and (global) content management systems
- Customer services (help desks)

Individual translators

- Working for agencies or subcontracting for organisations, and freelancers
 - Many MT systems designed for ‘professional use’
 - Few TM systems suitable for individual translators (and too expensive), except specialist translators with regular clients
- Problems: post-editing, terminology control, project management
- Future?: --
 - Affordable software
 - Integrated MT and TM, terminology, controlled language, etc.
 - Customers will expect translators to use computer tools
 - Must emphasise quality (comparisons with online MT inevitable)
 - Must specialise (to make good use of TM)

MT for assimilation

- publication quality not necessary
- fast/immediate
- readable (intelligible), for information use
 - intelligence services (e.g. NAIC)
 - internal company information, monitoring competition
 - 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) [but some PC systems now available]
- 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
- **online MT is 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**

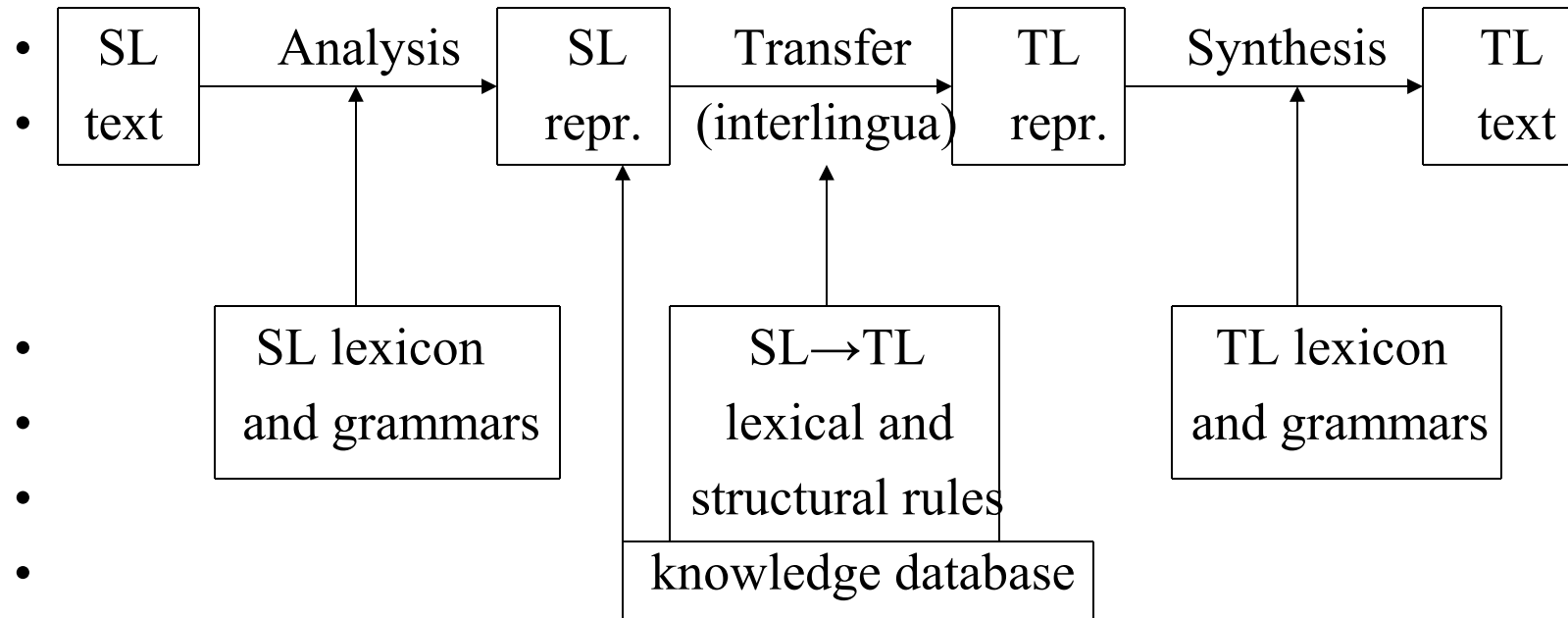
MT in the marketplace

- retail availability
 - many only purchased direct from manufacturer
- confusion of terms:
 - ‘translation systems’ no more than dictionaries
 - ‘computer aided translation’ means either HAMT or MAHT
 - combination of MT and support tools
 - translation memories either independent or components
- expectations of users
 - steady quality improvement; more languages
- unknown suitability of system to expected use
- bench marks, consumer reports/reviews (evaluations) - [urgent need]
- risks of marketplace (many systems/companies have failed)

Future prospects from MT research

- ‘Traditional’ rule-based systems
- Corpus-based approaches
 - statistical machine translation
 - example-based machine translation
- Multi-engine and hybrid systems
- Spoken language translation
- Integration with other NLP applications

'Traditional' RBMT schema



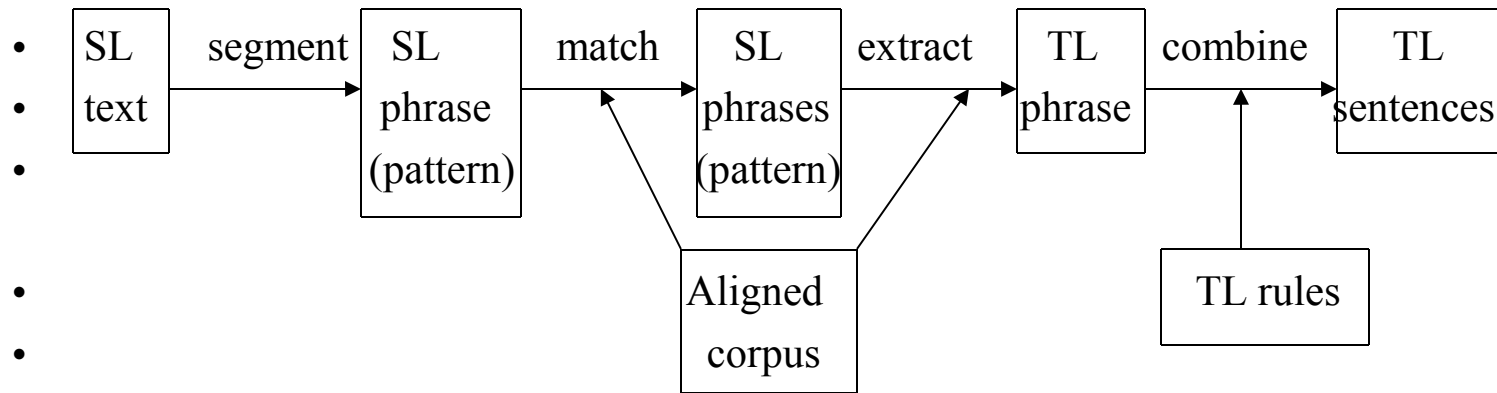
RBMT problems

- complexity of grammar rules
 - interactivity unpredictable, incomplete coverage, multi-level representation (artificial distinctions?, failure at one stage means no result)
- complexity of dictionaries
 - incomplete coverage of meanings, selectional restrictions
- collocations, phrasal verbs, verb/noun phrases, etc.
- complex structures
 - complex tree transduction rules; long sentences, embeddings, discontinuities, coordination
- pronouns, anaphora, ‘discourse’
- semantic problems
 - overcome (to some extent) by use of knowledge bases (in KBMT), but knowledge bases hugely complex
- many difficulties overcome/minimized in domain restricted and/or controlled language systems

RBMT still continues

- ‘direct’ approach common in commercial systems
 - 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
- 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, closely-related languages
- enterprise and company-oriented systems are RBMT

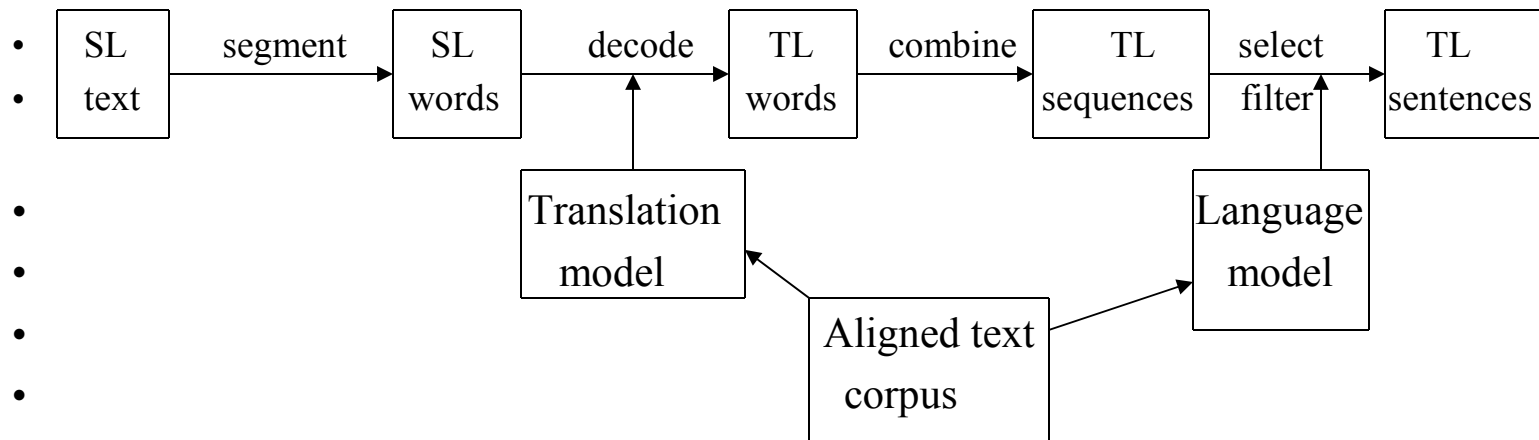
EBMT schema



EBMT basics

- two tendencies: EBMT methods to augment RBMT; **or**: EBMT as new ‘paradigm’
- 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 database examples
- extraction of TL examples (fragments) aligned to SL fragments
- 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.)

Statistical MT schema



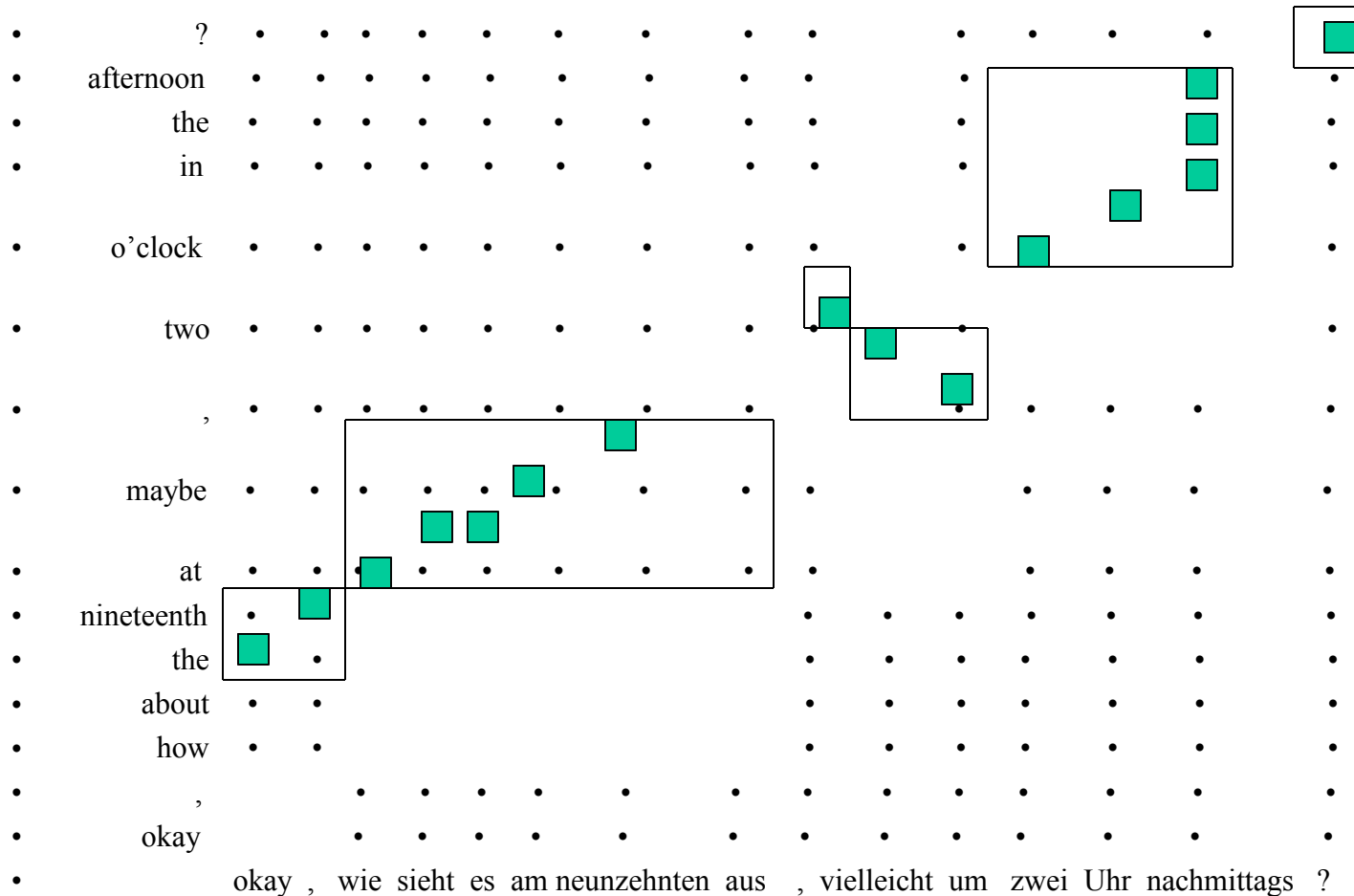
SMT basics

- 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’: probabilities that a TL string (word/phrase/fragment) is the translation of a SL string, based on:
 - frequencies of SL/TL co-occurrences in aligned texts of corpus
- ‘language model’: probabilities that a TL string is a valid TL fragment (based on frequencies of bigrams and trigrams), searches for TL string that maximizes these probabilities

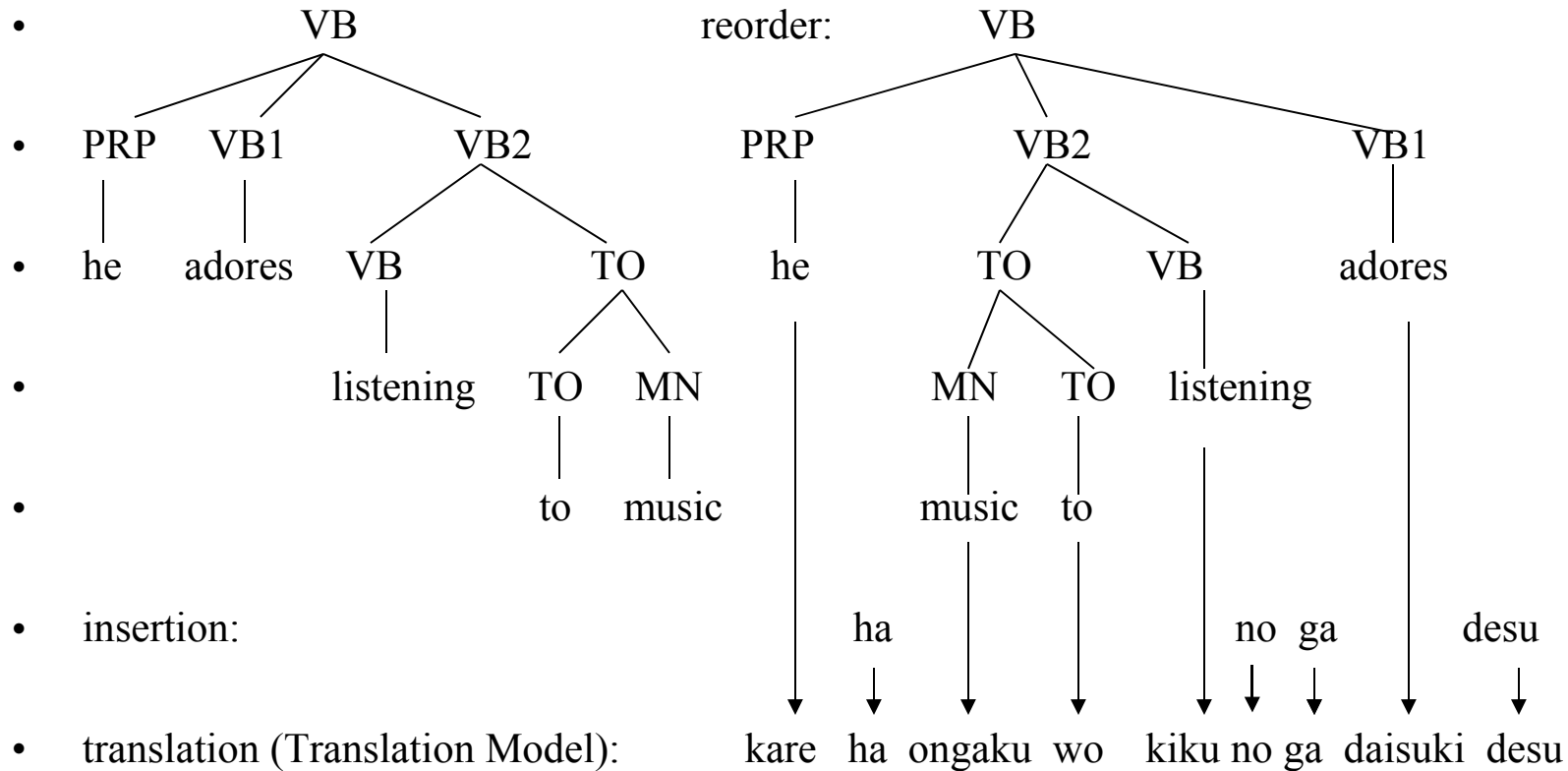
SMT issues

- 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:
 - no quality control of corpora (unlike EBMT)
 - lack of monolingual data for some languages
 - insufficient bilingual data, even for some major languages (Internet as resource)
- merits of SMT: evaluation as integral process of system development
- rapid development, but slow in operation

SMT phrase alignment: example



Syntax-based SMT translation model



Basic processes of RBMT, EBMT, SMT

	RBMT	EBMT	SMT
• pre-processing • (database)	dictionaries grammars	alignment parsing	alignment (statistical parsing)
• analysis	morphology deep parsing semantic analysis	morphology shallow parsing	(morphology)
• interface	structured repres.	templates (trees)	words/phrases
• transfer • (core)	lexical transfer tree transduction	matching (tree conversion) extraction	matching 'translation model'
• synthesis	structured repres. generation	strings (trees)	strings 'language model'

Comparison of SMT, EBMT and RBMT

- SMT black box: no way of finding how it works in particular cases, why it succeeds on some occasions and not on others - [test by full run of system version]
- RBMT/EBMT: rules and procedures can be inspected - [changes may have unpredictable consequences]
- EBMT closer to RBMT when input, matching, extraction are based on structured representations (database also structured) - [parsing not trivial]
- EBMT/SMT: matching, extraction, re-combination equivalent processes (statistics-based); EBMT as subtype of SMT
- EBMT: its distinctive feature is example strings (analogy-based approach)
- SMT closest to EBMT when corpora and input (statistically) parsed: syntax-based SMT virtually identical to syntax-based EBMT
- RBMT and SMT began as 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

Persistent problems of MT

- bilingual ambiguity (lexical transfer)
 - RBMT contextual rules, SMT phrases, EBMT examples
- collocations, non-compositional constructions, discontinuous structures
 - RBMT lexicons, SMT phrases and ‘language models’, EBMT examples
- bilingual structural divergences (cases, dependencies)
 - RBMT tree transduction, syntax-based SMT, EBMT examples
- semantic constraints/choice
 - RBMT rules, thesauri, knowledge bases, WordNet, Semantic Web
- complex (long) sentences, embedding, ellipses, coordination
- reference, anaphora, articles
 - KBMT, ‘discourse-oriented’ systems
- stylistic differences (instead of retention of SL structures)
 - SMT language model
- unpredictable interactions
- better for domain-specific than general-purpose (for special language than for ‘ordinary language’)
- ‘inherent’ limitations of computational linguistics and MT (?)

Hybrid and multi-engine systems

- Hybrid: take ‘best’ methods from each type, e.g.
 - RBMT for analysis (morphology, phrase dependencies)
 - EBMT for collocations, non-compositional constructions
 - RBMT and EBMT for bilingual structural divergence
 - SMT for selection of most frequent TL forms (in domain-specific systems)
 - EBMT for selection of actual TL examples (idiomatic selection)
 - KBMT and thesauri for problems of semantics and reference
 - SMT ‘language model’ for smoothing TL output
- Multi-engine: take ‘best’ outcome from each system
 - e.g. output from RBMT, SMT and EBMT systems run through a ‘language model’
 - and tested for semantic coherence (a discourse model?)

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 better on untrained corpora

Spoken language MT

- probably most desired translation technology of all
- most problems (ellipses, intonation, modality, discourse forms, ‘ill-formed’ utterances, dialect, etc.)
- many potential applications
- interim ‘solutions’: voice input, text translation, voice output
- research using variety of methods (RBMT, EBMT and statistical analysis, speech recognition, knowledge databases, etc.) -- typically ‘hybrid’
 - Japanese groups first to develop EBMT methods
 - systems: ATR (Japan), JANUS (US, Germany), SLT (SRI, Cambridge), Verbmobil (Germany), DIPLOMAT (Carnegie-Mellon), MedSLT
- wide-coverage systems not realistic objective
- research concentrating on narrow domains
 - hotel booking, conference registration (since late 1980s);
 - military (phrasebook-type);
 - medical consultations (doctor: questions; patient: simple answers)

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]

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]

Summary of future developments and expectations

- merging of MT and TM for enterprise dissemination/assimilation systems
- Internet as major (chief) resource
- rapid development of systems (SMT)
- greater (easier) reuse of MT components (for closely related languages)
- improvements in quality (evaluation, hybrid, multi-engine systems)
- minor (and minority) languages
- special-purpose systems (domain and function) - also online
- spoken language MT, domain-specific only [not general-purpose]
- wider access (wireless), new devices [spectacles! etc.]
- 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>