Machine translation history and current state

John Hutchins

(Email: WJHutchins@compuserve.com)

[http://ourworld.compuserve.com/homepages/WJHutchins]

Imperial College London, Tuesday 5 February 2002

The beginnings, to 1966

- 1933: Troyanskii's patent proposal
- 1949: Weaver's memorandum
- 1952: First conference, MIT in June 1952
- 1954: First MT system demonstrated, January 1954
- 1955 onwards: Many groups in US, Russia, West Europe, etc.
- 1958: first Soviet MT conference (Moscow): 340 participants from 50 institutions
- 1959: IBM system begins operations for USAF
- 1959: Unesco conference on Information Processing (Paris); first public demo of Georgetown system
- 1960: National Symposium on Machine Translation, UCLA; first conference of federally supported groups (Princeton, US)
- 1960: US House of Representatives hearings
- 1960: Bar-Hillel's critical review of MT (Advances in Computers): non-feasibility of FAHQT
- 1961: International conference, NPL, September 1961
- 1962: NATO summer school
- 1963: Georgetown system installed, Oak Ridge Nat.Lab. (US Atomic Energy) and Euratom (Italy)
- 1964: "semantic barrier", ALPAC set up

ALPAC report 1966

- Automatic Language Processing Advisory Committee
  - set up by NSF for government sponsors
- conclusions
  - “no machine translation of general scientific text, and none is in immediate prospect”
  - uneconomic, and “supply of translators greatly exceeds the demand”
  - no further funding of MT
  - development of translation aids
  - fund basic research in computational linguistics
- consequences
  - end of US government funding of MT research for more than 20 years
  - world-wide reductions (including USSR)
  - MT no longer ‘respectable’ occupation
  - AMTCL becomes ACL

‘Perfectionist’ tendency

- Characteristics of first decades
  - Military and intelligence funding
  - Russian/ English concentration
  - Direct translation systems
  - Interlinguas
  - Formal syntax
  - Fully automatic high quality translation (FAHQT)
- Neglect of
  - operational requirements (effective use of ‘less-than-perfect’ MT)
  - expertise of translators
  - machine aids for translators
- Three strands of MT since ALPAC
  - translation tools (translator workstations)
  - operational MT systems (pre- and post-editing, domain-specific, controlled language, sublanguage)
  - research (new theories, new methods)

From 1967 to 1979

- 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) begins

From 1980 to 1989

- first PC systems
  - ALPS, Weidner, PC-Translator, Globalink
- commercial systems
  - Logos, ATLAS, HICATS, ASTRANSAC, Duet, Tovna, etc.
- dominance of transfer-based approach
  - GETA-Ariane, SUSY, Eurotra
- Japanese research
  - Mu (Kyoto)
- knowledge-based systems
  - Carnegie-Mellon, New Mexico
- interlingua-based systems
  - DLT, Rosetta
- controlled language applications
  - Xerox use of Systran; Smart systems
- translation tools
  - termbanks, terminology management

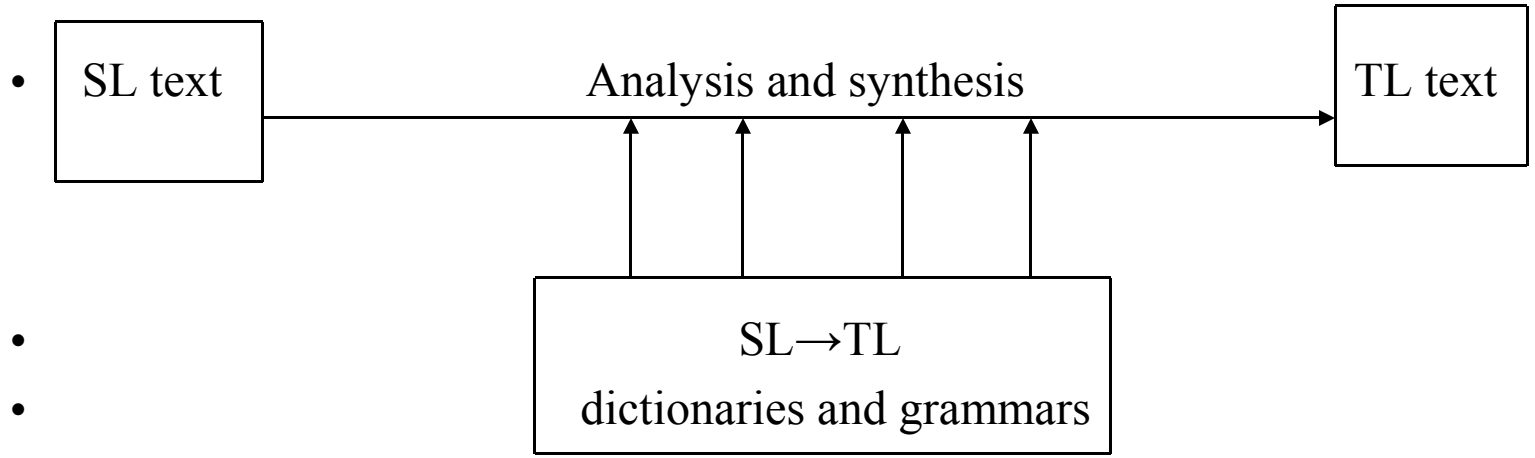
Since 1990

- corpus-based (statistical) approaches
  - IBM Candide
- example-based methods, translation memory
- revival of US research: Carnegie-Mellon, U.S.California, New Mexico
- systems for the Internet
- localization
- lexical resources: acquisition, reusability
- translation workstations

System architectures and strategies

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

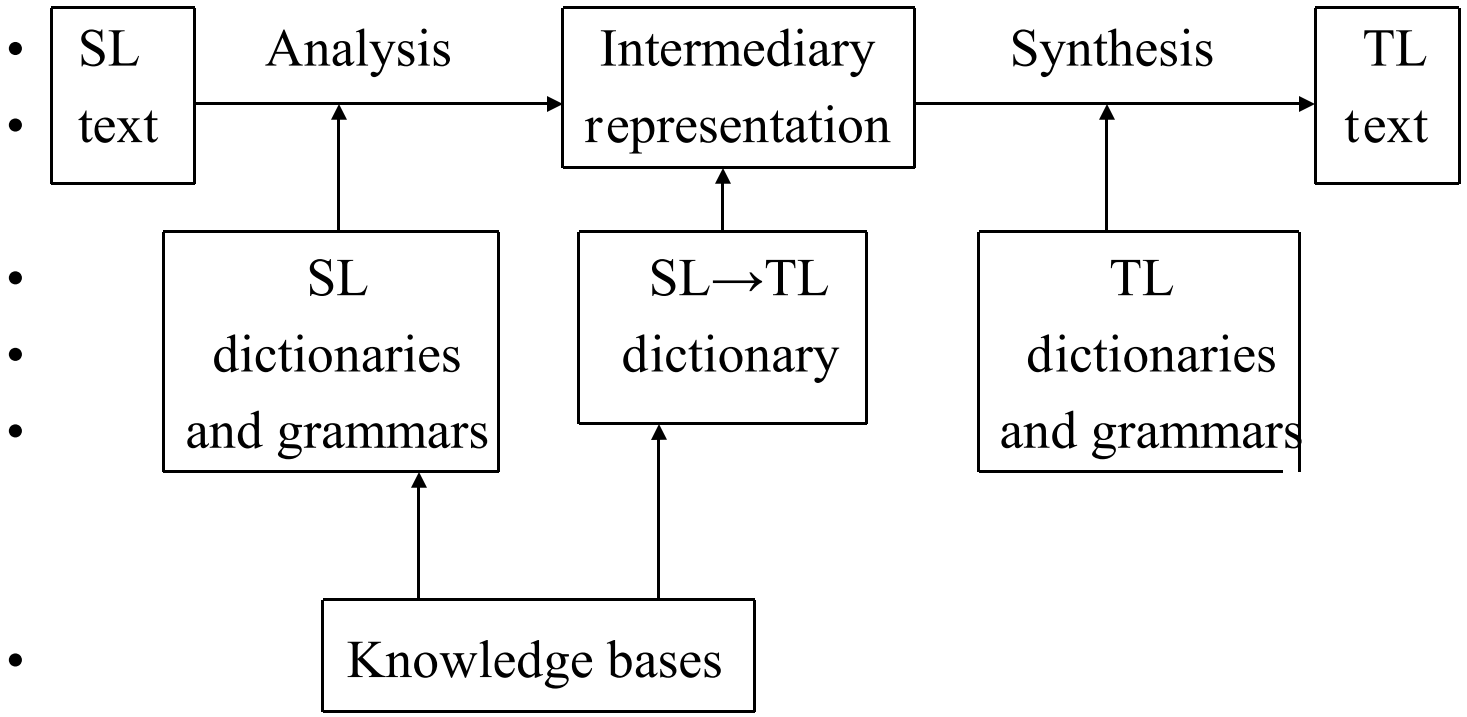
Direct translation



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
- Use of ‘cover’ words
- no analysis of SL syntax or semantics
- output too close to SL structure
- example (Russian to English):
  - On dopisal stranitsu i otložil ručku v storonu.
  - It wrote a page and put off a knob to the side
  - (i.e.) “He finished writing the page and laid his pen aside”
- systems:
  - Univ.Washington, IBM (US)
  - Georgetown University (US)
  - Ramo-Wooldridge (US)
  - Institute for Precision Mechanics and Computer Technology (USSR)
  - National Physical Laboratory (UK)

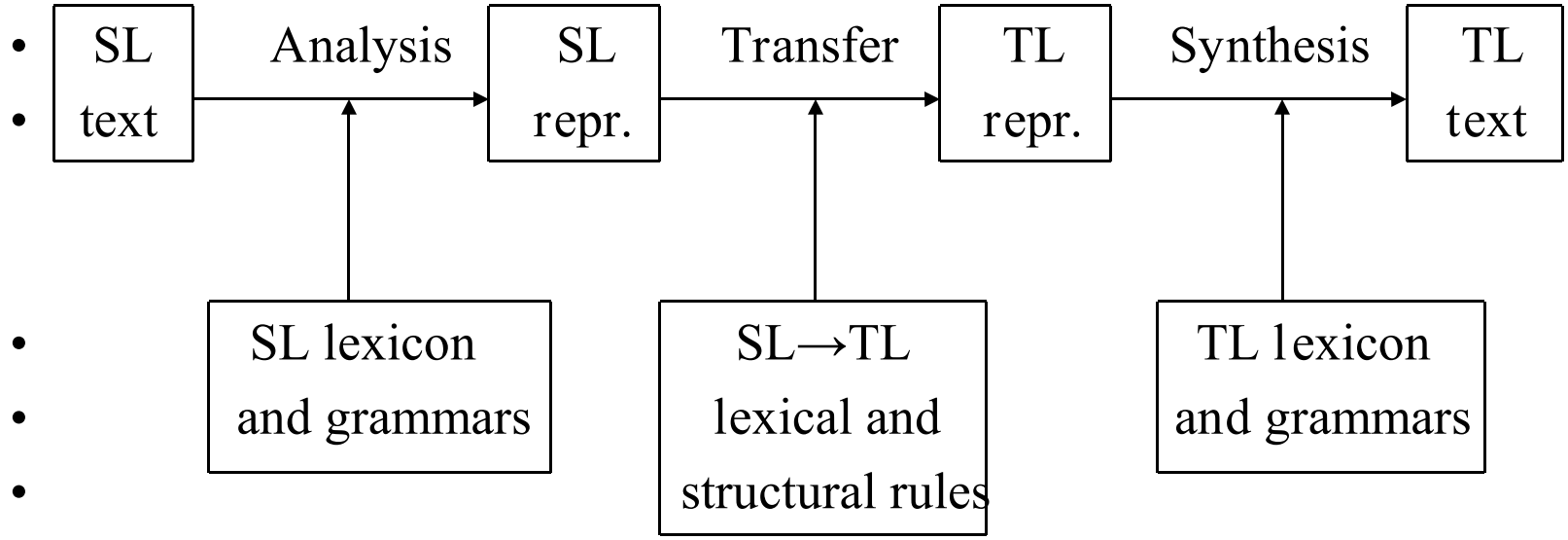
‘Interlingual’ system



Interlingua-based MT

- two independent stages: analysis, synthesis
- abstract language-neutral representation
- multistratal: morphology, syntax, semantics
- semantics-oriented (‘understanding’)
- domain-specific ‘knowledge bases’ (AI-oriented)
- projects:
  - Grenoble (CETA), Texas (METAL)
  - DLT, Rosetta, Pivot (NEC)
  - Carnegie-Mellon University (KBMT, KANT, CATALYST)
  - New Mexico State University (ULTRA, Pangloss)
  - Univ. Maryland (UNITRAN)

‘Transfer’ system



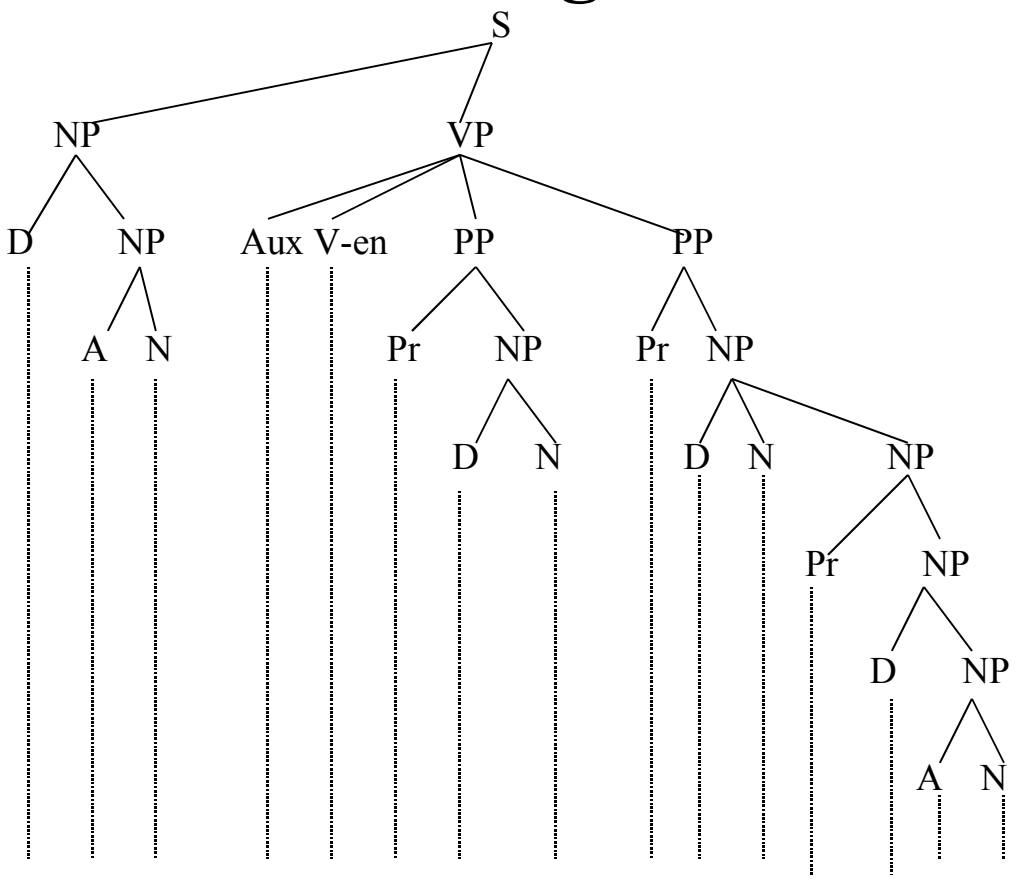
Transfer-based MT

- three stages: analysis, transfer, synthesis
- abstract semantico-syntactic interfaces/representations
- multiple level/strata: morphology, syntax, semantics
- syntax-oriented, tree-transduction
- batch processing, post-edited
- little/no discourse information (anaphora, etc.)
- projects/systems:
  - GETA-Ariane, Eurotra, LMT, Mu

Theories and formalisms

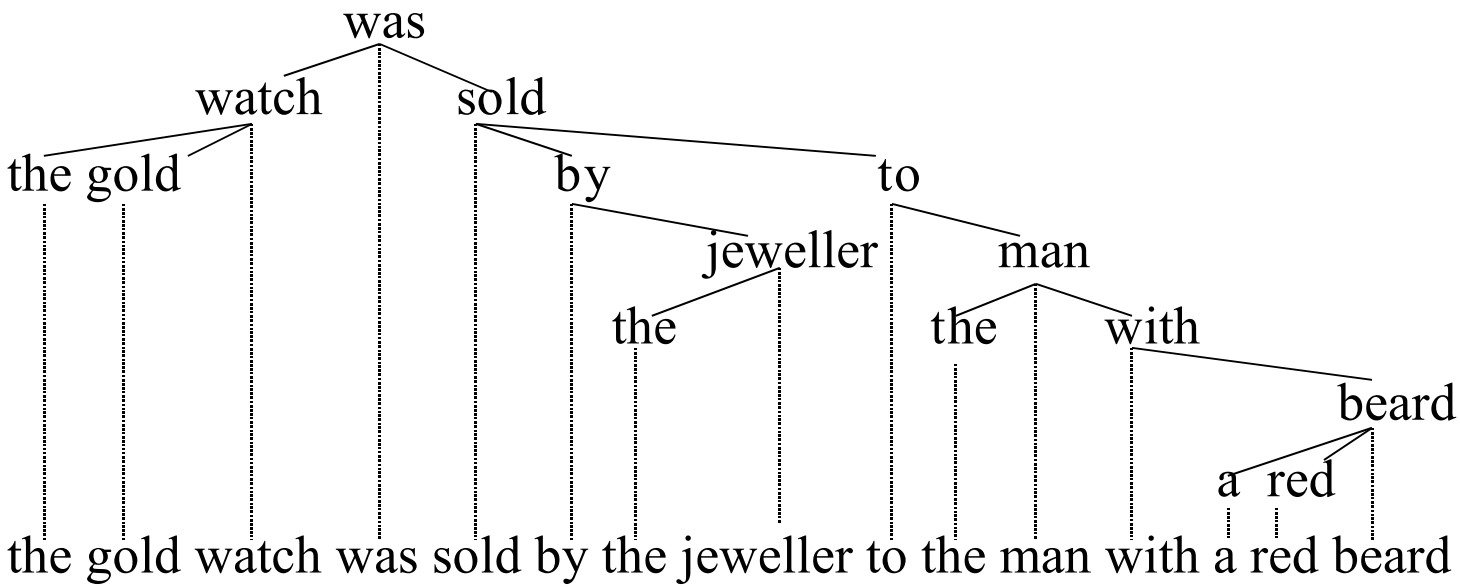
- Information theory
- Transformational-generative grammar
- Dependency grammar
- Stratificational grammar
- Artificial intelligence
- Lexical-functional grammar
- Generalized phrase-structure grammar
- Definite clause grammar
- Principles and parameters, Government-binding theory
- Categorial grammar
- Montague grammar
- Neural networks

Constituency ('phrase-structure') grammar

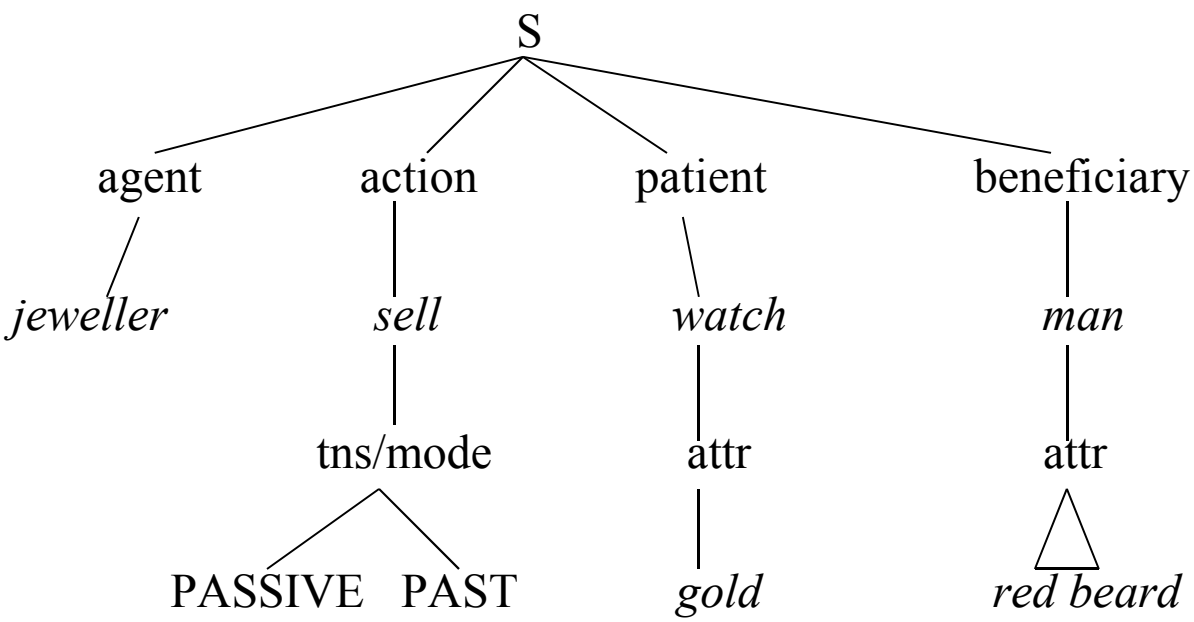


the gold watch was sold by the jeweller to the man with a red beard

Dependency grammar



Case grammar

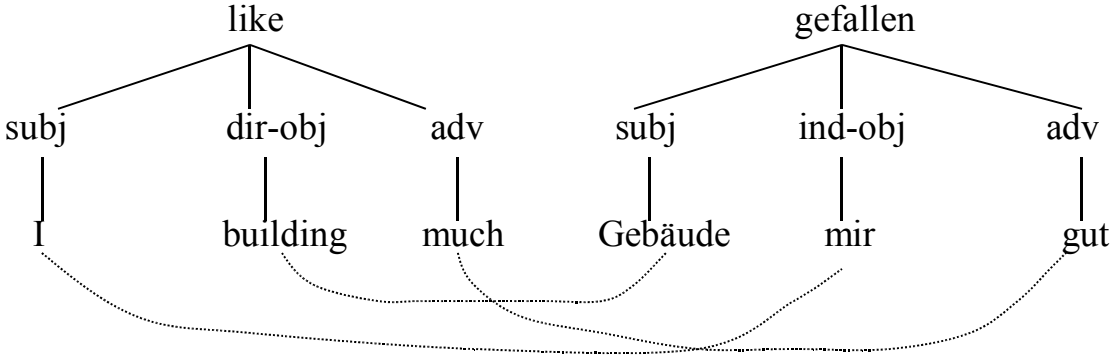


Unification grammar:
example (LFG)

- SL f-structure *John likes Mary*
- [PRED like
- SUBJ [PRED John]
- OBJ [PRED Mary]]
- like, V:
- (↑ PRED) = like < SUBJ, OBJ >
- (↑ PRED FR) = plaire < SUBJ, OBJ >
- (τ↑AOBJ OBJ) = τ (SUBJ)
- (τ↑ SUBJ) = τ (OBJ)
- TL f-structure *Marie plaît à Jean*
- [PRED plaire
- SUBJ [PRED Marie]
- AOBJ [OBJ [PRED Jean]]]

Tree transduction

- I like the new building very much ↔ Das neue Gebäude gefällt mir gut



- I like coffee ↔ ich trinke gern Kaffee
- He has just broken his leg ↔ il vient de se casser la jombe

Corpus-based systems

- Not rule-based: grammar rules (analysis, transfer, synthesis), multiple strata, ‘deep’ semantic analysis; complex dictionary entries
- based on bilingual text resources, e.g.
  - have a direct effect on… ont une influence directe sur…
  - have a direct effect on… intéressent directement
  - have a direct effect on… ont eu une répercussion directe sur…
  - has had a marked effect on… a largement influencé…
  - had a positive effect on… s’est avérée positive dans...
- Extraction of phrases for re-combination [Example-based MT]
- Statistical alignment, word-word frequencies, word co-occurrences [Statistics-based MT]

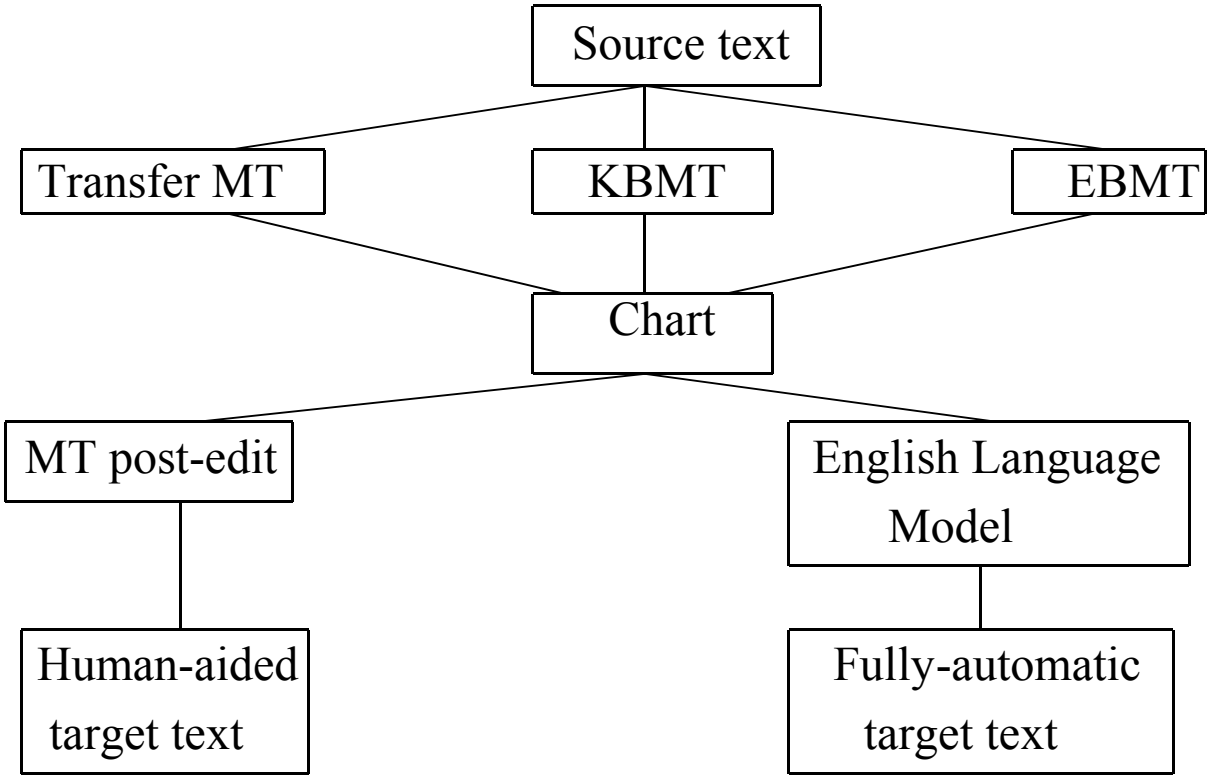
Statistics-based

- 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
- Sentences aligned statistically (according to sentence length and position)
- compute probability that a TL string is the translation of a SL string (based on frequency of co-occurrence in aligned texts of corpus)
  - including ‘fertility’ (how many TL words correspond to SL word)
  - position of SL words in SL string
- 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
- example:
  - IBM Candide (1988) on Canadian Hansard (English and French)

Example-based

- 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)
  - not trivial
- example (translating *no* from Japanese):
  - hoteru no juusho — address of the hotel
  - eigo no panfuretto — pamphlet in English
  - asu no tenkou — tomorrow's weather

Hybrid systems:
an example (Pangloss Mark III)



Speech translation

- 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)
- systems: ATR (Japan), JANUS (US, Germany), SLT (SRI, Cambridge), Verbmobil (Germany)
- examples:

– Der Montag, der passt mir gut	The Monday suits me fine
– Sie haben Zeit?	(do) you have time?
– Ich habe am Freitag bloss keine Zeit	But I don’t have time on Friday
– Ich habe bloss am Freitag Zeit	I have time only on Friday
– Ich muss nach Hannover	I have to go to Hanover
– klingt gut!	sounds OK!
– bin einverstanden!	agreed!

Types of translation demand

- dissemination
  - external publications
  - internal reports
  - operational manuals
  - localization
- assimilation
  - internal ‘monitoring’, information ‘filtering’
- interchange
  - correspondence, email
  - telephone
- information access
  - information retrieval/extraction, database searching, Web pages

Large-scale translation and MT

- Technical documentation (e.g. operating manuals)
- repetitive, frequent updates
- multilingual output (e.g. English to French, German, Japanese, Portuguese, Spanish)
- in-house terminological database
- controlled language input
- post-editing
- technical writing
- in-house printing/publishing
- technical expertise

Translation workstations

- Components and facilities
  - Terminology management
  - Translation memory, and alignment
- Products available
  - Trados: Translation Solution
  - STAR: Transit
  - IBM: TranslationManager/2
  - Atril: Déjà Vu
  - SDL: SDLX
  - EC [non-commercial]: EURAMIS

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
- alignment of bilingual text corpora
- revised texts (i.e. updated documents) are checked against corpus for any changes; for unchanged source sentences, the ‘authorized’ translation is retained
- search of exact matches or ‘fuzzy’ matches
- extract target phrase for insertion and/or amendment
- still much post-editing, and there is need for programs to ‘meld’ or conflate extracted phrases
- sometimes augmented by MT systems

Localization

- Internationalisation, globalisation (e.g. software and Web pages)
  - estimated market (end 2006) is $3.5 billion and $3 billion resp. (ABI this year)
- Cultural and linguistic adaptation (not just translation)
- Screen commands and help files
- Large scale
- Multiple language output
- Repetitive (translation memory)
- Fast results
- Organisation:
  - Localization Industry Standards Association
- Software companies (many in Ireland):
  - ALPNET; Berlitz; Compaq; Corel; IBM; Lotus; Microsoft; Oracle; SAP; Symantec

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
- examples:
  - ***not***: After agitation, allow the solution to stand for one hour
  - ***but***: If you shake the solution, do not use it for one hour.
  - ***not***: It is very important that you keep all of the engine parts clean and free of corrosion.
  - ***but***: Keep all of the engine parts clean. Do not let corrosion occur.
- controlled languages:
  - AECMA, MCE (Xerox), PACE (Perkins Engines), Caterpillar English, Smart

Management implications

- Terminology database: acquisition, consistency, management
- Translation memory: inclusion/exclusion policy, quality, access
- Text alignment: quality control
- Documentation flow
- Technical authoring: interaction with translation systems
- Publishing, formatting: graphics, layout
- Personnel training: project manager, translators
- Technical assistance: language engineer, computer technician
- Translators and post-editors
- Customer contact
- Management control systems
  - e.g. LTC Organiser, PASSOLO

Personal translation

- PC systems, e.g.
  - Al-Wafi (ATA Software): English/Arabic
  - Crossroad (NEC): English/Japanese
  - Easy Translator (Transparent Language): English/French, English/German, English→Italian, English→Portuguese, English/Spanish, Japanese→English
  - ESI Standard (WordMagic): English/Spanish
  - Instant Spanish (Bilingual Software): English→Spanish
  - Korya Eiwa (LogoVista): English/Japanese
  - LogoMedia Translate (LogoMedia): Chinese/English, English/French, English/German, English/Italian, English/Japanese, English/Korean, English/Portuguese, English/Russian, English/Spanish
  - LogoVista Personal (LEC): English/Japanese
  - NeuroTran (Translation Experts): Bosnian/English, Croatian/English, English/French, English/German, English/Hungarian, English/Polish, English/Serbian, English/Spanish
  - PC Translator 2002: Czech/English, Czech/German, English/Slovak, German/Slovak
  - Personal Translator PT Home (Linguatec): English/German
  - PeTra Word (Synthema): English/Italian
  - Pocket Transer (Nova): English/Japanese
  - PROMT Express (ProMT): English/Russian
  - Reverso Perso (Softissimo): English/French, English/Spanish
  - Systran Personal (Systran): English/French, English/German, English/Greek, English/Italian, English/Portuguese, English/Spanish

Personal translation (contd.)

- Special devices and mobile
  - Partner (Ectaco) (hand held): English/French, English/German, English/Italian, English/Portuguese, English/Spanish; and Gold Partner for English/Russian and English/Ukrainian; and Universal Translator
  - Language Teacher (Ectaco), and Quicktionary (Seiko): dictionaries
- Web page translation
  - Amiweb (Amikai): Chinese/English, English/French, English/German, English/Italian, English/Japanese, English/Korean, English/Portuguese, English/Spanish
  - CITAC: Chinese to English
  - LogoMedia Passport (LogoMedia): Chinese/English, English/French, English/German, English/Italian, English/Japanese, English/Korean, English/Portuguese, English/Russian, English/Spanish
  - LogoVista Internet Plus: (LEC): English to Japanese
  - Reverso Perso (Softissimo): English/French, English/Spanish
  - Systranet (Systran): English/French, English/German, English/Italian, English/Portuguese, English/Spanish
  - Translingo (Fujitsu), Transpad (AILogic): English/Japanese
  - WebTransSmart: Finnish/English
- Emails
  - Amimail (Amikai); CITAC; LogoMedia; Reverso Perso; T-Mail; Translingo
- Text messages
  - MobileTran; Petra-SMS; PT-SMS
- MT on the Internet: translation services
  - Ajeeb; Babelfish (AltaVista); Babylon; Cybertrans; Free Translation (and PlusTranslation); Hypertrans; InterTran; iTranslator; JICST; PROMT-Online/Reverso-Online; PT-Online; Systran

Professional translation

- PC systems, e.g.
  - Comprendium cX (Sail Labs): English/German, French/German, Russian/German
  - ENGSPAN (PAHO): English→Spanish
  - ESI Professional (WordMagic): English/Spanish
  - HICATS (Hitachi): English/Japanese
  - Honyaku Office (Toshiba): English/Japanese
  - Hypertrans (D'Agostini): English/French, English/German, English/Italian, English/Spanish, French/German, French/Italian, French/Spanish, German/Italian, German/Spanish, Italian/Russian, Italian/Spanish, Portuguese/Spanish [patents]
  - LogoVista X Pro (LEC): English/Japanese
  - Pensee (Oki): English↔Japanese
  - Personal Translator PT Office Plus (Linguatec): English/German
  - PeTra Expert (Synthema): English/Italian
  - ProMT Translation Office (ProMT): English/Russian, French/Russian, German/Russian, Italian/Russian
  - Reverso Expert (Softissimo): English/French, English/German, English/Spanish, French/German
  - SPANAM (PAHO): Spanish→English
  - Systran Professional Premium/Standard (Systran): Chinese→English, English/French (S), English/German (S), English/Italian (S), English/Japanese, English/Korean, English/Portuguese (S), English/Spanish (S), Russian→English
  - Transcend (SDL International): English/French, English/German, English→Italian, English/Portuguese, English/Spanish
- Translation workstations
  - still too expensive for independent translator

Human versus machine translation: Dissemination

•	HT	MT	TM
• Literary, legal	costly	no	
• Technical, scientific	v.costly	post-edited	?
• Weather reports	costly	yes	yes
• Localization	?	post-edited	yes
• Advertisements	?	adequate	yes
• Document drafting	no	yes	yes

Human versus machine translation: Assimilation

–	HT	MT		TM	
• Scientific, technical	rare	adequate		no	
• Non-literary (occasional)	rare	poor		no	
• Information monitoring		costly	adequate		no

Human versus machine translation: Interchange and information access

–	HT	MT	TM
• Business correspondence	yes	yes	yes
• Personal correspondence	?	adequate	no
• Electronic mail	no	poor	no
• Web pages	yes!	adequate?	no
• Database searching	no	adequate	no
• Summarising (with translation)	rare	possible?	no
• TV captions	no?	adequate	no?
• informal conversation	yes	no	no
• formal interpreting	yes	no	no
• telephone enquiries	rare	possible?	no

Where human (and machine) translation can fail

- insufficient knowledge of (data covering) source language
- insufficient knowledge of (data covering) subject matter
- lack of knowledge of specialist vocabulary (access to specialist lexica)
- inadequate familiarity with cultural background (no background)
- inadequate knowledge of (data for) target language (in relevant domain)
- lack of translation experience (no ‘understanding’ or ‘learning’)

Aids for translation and/or communication

- computer-produced draft translation (traditional post-edited MT)
- computer-based translation aids (dictionaries, terminology, translation memories, translator workstations)
- text assimilation aids (traditional use of ‘rough’ MT output)
- text production aids (multilingual generation, authoring aids)
- message dissemination aids (TV captions, public announcement, police messages)
- cross-language information access (information retrieval, information extraction, summarization)
- cross-language interchange (email, SMS, telephone, military ‘field’ communication, business negotiations, tourism, etc.)

New directions and challenges

- Spoken language
  - general-purpose?
- ‘Minor’ languages
  - languages of India, Africa, Asia
  - 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
  - word processing, Web pages, authoring systems
  - summarization, information extraction, information retrieval, data retrieval, question-answering, Internet search tools

Evaluation

- Who needs to know?
  - Potential purchasers, users (translators), service managers, system developers, researchers
- Quality control
  - fidelity, accuracy (of terminology), comprehensibility, intelligibility, readability, appropriate style
- Usability
  - adaptability, extendibility (e.g. to other languages and operating systems), compatibility (software and hardware), error levels (e.g. post-editing effort)
- Resources evaluation
  - dictionaries, terminology, translated texts

Conclusions

- MT is not translation as usually understood
- MT is ‘just’ symbol manipulation
- MT is a computer-based ***tool***
  - for translators
  - for cross-language communication
  - for access to information resources
- 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