Towards a definition of example-based machine translation

John Hutchins
[Email: wjhutchins@compuserve.com; jhutchins@beeb.net]
[web: http://ourworld.compuserve.com/homepages/WJHutchins/]
Why a definition is needed

• Not required if satisfied with a vague definition
  – e.g. “MT using examples of actual translations”

• Not required if interest/research in EBMT can be maintained (and increased) without one
  – lack of definition does not hold back research (e.g. AI)

• But it is required if EBMT has to be distinguished from other approaches to MT (previous and contemporary)
  – EBMT less clearly defined than SMT
  – plethora of different methods, many also used in other approaches

• And it is required if ‘outsiders’ want to know what is distinctive about EBMT [my own position]
Original conception of EBMT

• Based on observation that translators try to find similar SL phrases and sentences and their TL equivalents in previously translated texts
  – seek sets of analogies and examples from bilingual corpora
• as means of overcoming deficiencies of RBMT, particularly collocations and wide differences of word order
• less complex procedures (e.g. no phrase structure analysis, semantic analysis)
• potential to improve generation of TL sentences, since based on actual translations rather than ‘created’ by grammars and lexica
• originally (Nagao 1981) as means of augmenting RBMT
Two tendencies

• From beginning two tendencies:
  • EBMT as supplement to RBMT systems
    – as continuation of RBMT tradition
    – leading to ‘hybrid’ systems
  • EBMT as 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)
Definitions by Somers and by Turcato/Popowich

• Somers: “the main knowledge base stems from examples” and “the examples are used at run-time”
• Turcato/Popowich: it does not matter how system knowledge is acquired or expressed, what matters is how it is used
• crucial test is treatment of non-compositional translation
• if this knowledge is derived explicitly from example database and/or bilingual corpora, then EBMT no different from RBMT
• but if system makes direct reference to the example database during the process then EBMT system is clearly distinct from RBMT, it is based on (un-processed) examples as ‘analogies’
• therefore true EBMT is ‘run-time’ EBMT
• but both definitions exclude many systems considered to be EBMT
Basic definition of *any* MT system

• Conversion of SL elements (entities, text, sentences, words, phrase structures, etc.) into ‘equivalent’ TL representations (= the ‘core’ process)
  – using information about SL-TL correspondences (lexical and structural)
  – preserving ‘meaning equivalences’
  – producing usable unedited output (for gisting or use as pre-translation)
Ancillary processes

• Processing of input (sentences) preparatory to ‘core’ process of conversion
  – e.g. segmentation, morphological and syntactic analysis, semantic analysis, matching

• Processing of output of ‘core’ process to produce appropriate TL sentences
  – e.g. syntactic and morphological generation, recombination

• Pre-processing of database
  – e.g. alignment, parsing, templates, frequency analyses (for ‘translation models’ and ‘language models’)

Rule-based MT (transfer)

- SL text
- Analysis → SL repr.
- Transfer → TL repr.
- Synthesis → TL text

- SL lexicon and grammars
- SL→TL lexical and structural rules
- TL lexicon and grammars
RBMT defined

- Core process mediated by bilingual dictionaries and rules for converting SL structures into TL structures
- and/or by dictionaries and rules for deriving ‘intermediary’ (interlingual) representations from which output is generated
- preceding stage of ‘analysis’ interprets SL input as abstract SL representations
- succeeding stage of ‘synthesis’ derives TL texts from TL representations produced by the core (‘transfer’ or ‘interlingual’) process
- [NB. ‘direct translation’ (‘transformer’) architecture converts SL lexical items into TL items with minimal (or no) intermediary representations.]
Statistical MT

- SL text
- segment
- SL words
- decode
- TL words
- combine
- TL sequences
- select
- filter
- TL sentences

Translation model

Language model

Aligned text database
Statistical MT defined

- core process is the ‘translation model’ taking SL words or phrases as input, and producing TL words or phrases as output
- succeeding stage involves a ‘language model’ which synthesizes TL words as ‘meaningful’ TL sentences
- preceding stage locates input words and phrases against entries in translation model
  - involving segmentation and matching processes
  - (may involve morphological and word-class rules)
- important pre-processing stage is the creation of the (bilingual) translation models and (monolingual) language models based on statistical analyses of the corpus (or corpora)
- [NB. SMT essentially lexical substitution and rearrangement (as old ‘direct translation’ model)]
Example-based MT

- **SL text**
- **decompose**
- **SL phrase (pattern)**
- **match**
- **SL phrases (pattern)**
- **extract**
- **TL phrases**
- **combine**
- **TL sentences**

(alang.model)
EBMT defined (preliminarily)

- 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.
- Pre-processing stage: for alignment of SL and TL sentences in the database, and/or for deriving templates and patterns used in matching and extraction.
- [Secondary factors]: deriving templates and patterns in advance or during run-time.
Bilingual corpora and database

• essential database of information about SL and TL correspondences
  – should distinguish between source corpora and derived database
• a distinctive feature, but not unique to EBMT
  – also found in RBMT and SMT
• bilingual corpora/database in EBMT and SMT replaces RBMT
dictionaries, grammars - largely:
• since a bilingual database is not the only knowledge source in EBMT
  – also use dictionaries, thesauri, grammars
• differences between EBMT, SMT and RBMT are not located in the
  use of bilingual text corpora (but how they are used in the core process)
Grammatical information

- RBMT: essential information for analysis and synthesis is used explicitly
- EBMT/SMT: information about well-formedness and lexical correspondences is contained implicitly in databases
- and implicitly ‘extracted’ for matching and conversion
- and implicitly utilised in synthesis (in SMT by ‘language model’ and in EBMT with the extraction of well-formed TL fragments)
Core processes as defining essences of all MT systems

• RBMT: interlingua, transfer, ‘direct’
• SMT: statistical SL-TL word (and ‘phrase’) probabilities (in a ‘translation model’)
• EBMT: matching SL fragments, extracting corresponding TL fragments
• secondary distinctions are:
  • run-time processing vs. preparatory processing
    – what matters is how SL fragments are converted into TL fragments
  • use of templates (patterns) vs use of phrase structures
  • use of SMT-like statistical methods (to derive templates/patterns) vs use of RBMT-like parsing (to derive representations)
EBMT and RBMT

- EBMT representations of SL: strings, templates, patterns, structures
- ‘problematic’ are 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
  - except: EBMT representations are derived from example databases - whereas RBMT representations are derived by rules
- however: RBMT rules may also be derived from bilingual databases
- therefore, some convergence of EBMT and RBMT (‘hybrid’)

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
- introduced to improve alignment and matching processes
- closest to EBMT when input is parsed, matching based on parsed representations in database, and output to ‘language model’ also as parsed representations
- only remaining difference in the ‘core’ process: SMT works exclusively with statistical methods, EBMT works mainly with symbolic (linguistic) fragments and text examples
- so, convergence of EBMT and SMT (‘hybrid’?)
Summary

- essence of EBMT is the matching of SL fragments/strings (from input text) against SL fragments/strings (in a database) and the extraction of equivalent TL fragments/strings (as partial potential translations), whether the matching is against pre-compiled representations or whether it is against fragments/strings in the whole database (at run-time)

- the essential knowledge database is derived from a corpus (or corpora) of SL-TL examples [although such a database is not unique to EBMT]

- the characteristic feature of EBMT is the assumption (or hypothesis) that translation involves the finding of ‘analogues’ of SL sentences in existing TL texts [neither SMT nor RBMT work with analogues]

- EBMT stands in an intermediary position between RBMT and SMT, using both statistical and symbolic methods - perhaps a ‘true’ hybrid MT approach