Example-Based Machine Translation

Kevin Duh
April 21, 2005
UW Machine Translation Reading Group

Outline

1. Basics of Example-based MT
2. System example:
   ATR system (1991): E. Sumita & H. Iida paper
3. General EBMT Issues
4. Different flavors of EBMT
5. Connections to Statistical MT, Rule-based MT, Speech synthesis, Case-base Reasoning…

EBMT Basic Philosophy

- "Man does not translate a simple sentence by doing deep linguistic analysis, rather, man does translation, first, by properly decomposing an input sentence into certain fragmental phrases, and finally by properly composing these fragmental translations into one long sentence. The translation of each fragmental phrase will be done by the analogy translation principle with proper examples as its reference." -- Nagao (1984)

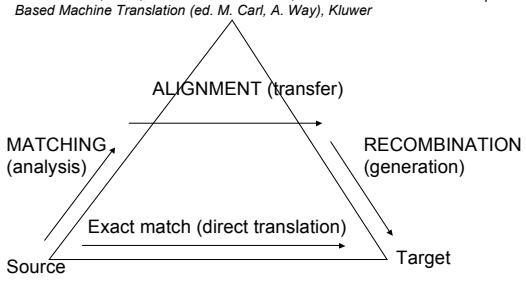
Example run of EBMT

- Example:
  Input: He buys a book on international politics
  1. He buys a notebook -- Kare wa noto o kau
  2. I read a book on international politics -- Watashi wa kokusai seiji nitsuite kakareta hon o yomu

  Output: Kare wa kokusai seiji nitsuite kakareta hon o kau
- 3 Main Components:
  - Matching input to a database of real examples
  - Identifying corresponding translation fragments
  - Recombining fragments into target text

EBMT "Pyramid"

From: H. Somers, 2003, "An Overview of EBMT," in Recent Advances in Example-Based Machine Translation (ed. M. Carl, A. Way), Kluwer



ATR System (1991)

- "Experiments and Prospects of Example-based Machine Translation," Eiichiro Sumita and Hitoshi Iida. In *29th Annual Meeting of the Association for Computational Linguistics*, 1991.
- Overview:
  1. Japanese-English translation: "N1 no N2" problem
  2. When EBMT is better suited than Rule-based MT
  3. EBMT in action: distance calculation, etc.
  4. Evaluation

Translating "N1 no N2"

- "no" の is an adnominal particle
- Variants: "deno" での, "madeno" までの, etc.
- "Noun no Noun" => "Noun of Noun"

Youka **no** gogo	The afternoon **of** the 8th
Kaigi **no** mokuteki	The objective **of** the conference
Kaigi **no** sankaryou	The application fee **for** the conference
Kyouto **deno** kaigi	The conference **in** Kyoto
Isshukan **no** kyuka	A week**'s** holiday
Mittsu **no** hoteru	Three hotels

Difficult linguistic phenomena

- It is difficult to hand-craft linguistic rules for "N1 no N2" translation phenomenon
  - Requires deep semantic analysis for each word
- Other difficult phenomena:
  - Optional case particles ("-de", "-ni")
  - Sentences lacking main verb ("-onegaishimasu")
  - Fragmental expressions ("hai", "soudesu")
  - "N1 wa N2 da" ("N1 be N2")
  - Spanish "de"
  - German compound nouns

When EBMT works better than Rule-based MT

1. Translation rule is difficult to formulate
2. General rule cannot accurately describe phenomena due to special cases (e.g. idioms)
3. Translation cannot be made by a compositional way using target words
4. When sentence to be translated has a close match in the database.

- How about when does Statistical MT work well?

EBMT in action

- Required resources:
  - Sentence-aligned parallel corpora
  - (Hierarchical) Thesaurus
    - for calculating semantic distance between content words of input and example sentences
- Distance calculation:
  - Input and example sentences (I, E) are represented by a set of attributes
  - For "N1 no N2":
    - For N1/N2: Lexical subcategory of noun, existence of prefix/suffix, semantic class in thesaurus
    - For No: "no", "deno", "madeno" binary variables

$$\text{distance}(I,E) = \sum_{j=1}^{J} d(I_j, E_j) * w_j$$

EBMT in action : Attribute distance & weight

$$\text{distance}(I,E) = \sum_{j=1}^{J} d(I_j, E_j) * w_j$$

- Attribute Distance:
  - For "no": d("no","deno")=1, d("no","no") = 0
  - For Noun1, Noun2, use thesaurus.
- Weight for each attribute: $w_j = \sqrt{\sum_{tp} (freq(tp) \text{ when } E_j = I_j)^2}$

Timei 地名[place]	Deno での[in]	Soudan 相談 [meeting]
B in A (freq: 12/27)	B in A (3/3)	B (9/24)
AB (4/27)		A's B (1/24)
B from A (2/27)		…
BA (2/27)		B on A (1/24)
B to A (1/27)		

Evaluation

- Corpus (Conversations re. Conference registration)
  - 3000 words, 2550 examples
- Jacknife evaluation
  - Ave success rate 78% (min 70%, max 89%)
  - Success rate improves as examples are added
  - Success rate for low-distance sentences are higher
- Failures due to:
  - Lack of similar examples
  - Retrieval of dissimilar examples due to current distance metric
- In practice:
  - EBMT is used as a subsystem within Rule-based MT to handle special cases like "N1 no N2"

General EBMT Issues: Granularity for locating matches

- Sentence or sub-sentence?
  - Sentence:
    - Better quality translation
    - Boundaries are easy to determine
    - Harder to find a match
  - Sub-sentence:
    - Studies suggest this is how humans translate
    - "Boundary friction"
      - The handsome boy ate his breakfast -- Der schone Junge as seinen Fruhstuck
      - I saw the handsome boy -- Ich sah den schonen Jungen

The following slides are based from: H. Somers, 2003, "An Overview of EBMT," in Recent Advances in Example-Based Machine Translation (ed. M. Carl, A. Way), Kluwer

General EBMT Issues: Suitability of Examples

- Some EBMT systems do not use raw corpus directly, but use manually-constructed examples or carefully-filtered set of real-world examples
- Real-world examples may contain:
  - Examples that mutually reinforce each other (overgeneration)
  - Examples that conflict
  - Examples that mislead the distance metric
    - Watashi wa kompyuta o kyoyosuru -- I **share the use** of a computer
    - Watashi wa kuruma o tsukau -- I use a car
    - Watashi wa dentaku o shiyosuru --> * I **share the use** of a calculator

General EBMT Issues: Matching

- String matching / IR-style matching
  - "This is shown as A in the diagram" = "This is shown as B in the diagram"
  - "The large paper tray holds 400 sheets of paper" =? "The small paper tray holds 300 sheets of paper"
- Matching by meaning:
  - use thesaurus and distance based on semantic similarity
- Matching by structure:
  - Tree edit distance, etc.

General EBMT Issues: Alignment and Recombination

- Once a set of examples close to the input are found, we need to:
  1. Identify which portion of the associated translation corresponds to input (alignment)
  2. Stitch together these fragments to create smooth ouput (recombination)
- Some interesting solutions:
  - "Adaptation-guided retrieval":
    - scores an example based on both sentence similarity and ability to align/recombine well
  - Statistical Language Model solution:
    - Pangloss EBMT (RD Brown, 1996)
  - Post-processing (c.f. der schone Jungen example)

Flavors of EBMT

- Different uses of EBMT:
  - Full automatic translation system
  - System component for handling difficult cases (ATR)
  - EBMT as one engine in Multi-engine MT (Pangloss)
- Different approaches to EBMT:
  - Run-time approach:
    - Sub-sentential alignment of input sentence and mapping on translation examples are computed at run-time (translation knowledge lies implicit in the corpus)
  - "Compile-time" approach:
    - Explicity extracts translation patterns from corpora.
    - Template-driven: uses taggers, analyzers, etc.
    - Structural Representations: abstracts sentences into more structured representation, such as LF, dependency tree

Connections to other areas

- Statistical MT
- Rule-based MT
- Unit Selection Speech Synthesis
- Case-base Reasoning
- *What do you think?*