Example-based Machine Translation Systems

Cristina Vertan

University of Hamburg • Informatics Department

Natural Language Systems Group

WWW: http://nats-www.informatik.uni-hamburg.de/~cri/

E-Mail: vertan@informatik.uni-hamburg.de

These slides follow „EBMT - A tutorial“, R. Brown 2002“

Early EBMT Systems -I-

Satoshi Sato and Makoto Nagao (1990) (EN, JP)

- Operated on dependency trees
- Correspondence points between source- and target-language trees for an example provide the ability to replace portions of a sentence to match previously unseen text
- Hand-coded semantic network for computing semantic dstance to select among translation candidates

Early EBMT systems -II-

Eiichiro Sumita et al (1991, 1993) (JP, EN)

- Translated only Japanese phrases of the form:
  - NOUN1 no NOUN2
- In most contexts the English translation is
  - NOUN1 of NOUN2
- System used a commercial semantic network of everyday Japanese and calculated the semantic distance of the nouns, searching up the hierarchy for the most specific common abstraction

System: Gaijin -1-

(Veale & Way 1997) (DE, EN)

- PoS tagging in both languages
- Translation examples converted into templates consisting of PoS tags
- Matching performed at the level of complete tag sequences (no partial matching); phrases within the translation example can be templatized

System: Gaijin -II

Phrasal segmentation using Marker Hypothesis

- Psycholinguistic constraint on grammatical structure
- States that natural languages are marked for grammar by a closed set of lexemes and morphemes
- Gaijin exploits such markers as signals for beginning and end of a phrasal segment:
  - Prepositions: in, out, on, with,...
  - Determiners: the, those, a, an,....
  - Quantifiers: all, some, many,....
- Markers not considered to start a new segment if previous/next segment would consist entirely of marker words

System: Gaijin -III

Segment Alignement

- Possible segment correspondences between source and target are evaluated using segment length and word correspondence weights
- Bonus for having leading marker of the same category type (e.g. „with“ and „mit“)
- Many-to-one segment mappings are (partially) handled by merging contigous segments which all map to the same segment in the other language
- Non-contigous mappings are considered unusable

System: Gaijin -IV

Templates

- All well-formed segment mappings are converted into variables, generating a template for the translation example
- Non frequent marker words are removed from the variablized segment and retained in the template literally
- To simplify lookups, segment merging is represented in the target side only; when the source segments need to be merged the system uses a compound variable on the target side

System: Gaijin -V-

Template Example

E: Displays controls for coloring the extruded surfaces

G: Durch Klicken auf dieses Symbol lassen sich Optionen zum Kolorieren der extrudierten Flaechen anzeigen

Template:

E: {_A}{prep B}{det C}

G: Durch klicken auf {prep A}{prep B}{det C} anzeigen

Chunks

A: Displays Controls

dieses Symbol lassen sich Optionen

B: for coloring

zum kolorieren

C: the extrudede surfaces

der extrudierten Flaechen

System: Gaijin -VI

Retrieving Examples

- Examples indexed under both the phrasal chunks they contain and under the sequence of marker-word types
- Previous example would be indexed under
  - „displays controls“
  - „for coloring“
  - 2the extruded surfaces“
  - ?-prep-det

System: Gaijin VII

Adaptation

- Grafting: replacing one phrasal segment with another from a different example
- Keyhole surgery: replacing or morphologically fine-tuning individual words in a target segment
- Gaijin tries to minimize boundary friction during grafting by ensuring that the replacement is as compatible with the template position as possible
  - When multiple options are available, choose the one which shares the most words with the phrase that was in the original from which the template was formed

System: EDGAR

Michael Carl et al, University of Saarbrücken

- Applies morphological analysis to both languages
- Induces translation templates from analyzed reference translations
- Multiple levels of generalization
- Matched chunks from case base are re-specialized and refined in the target language

System: ReVerb -I-

(Brona Collins 1996, 1999)

English-German, irisch-english translation

- Explicitly uses Case-Based Reasoning
- Training examples are abstracted to syntactic dependency representation
  - Sgallower processing than original Nagao7Sato approach, usinf flet feature lists
- Retrieval criterion is combination of similarity and adaptability
- Retrieved examples are adapted to fit the text to be translated

System: ReVerb -II

Knowledge Representation

- Corpus is converted ito a Case Base
- Each sentence pair is stored as a case; cases refer to chunks, which may be replaced on adaptation
- Individual word types habe separate WORD objects indexing their occurences in cases and chunks
- A translation dictionary is generated from word-to-word correspondences in the case bae

System : Reverb -III

Template Creation

- Examples are generalized where chunks can „safely“ be replaced or otherwise adapted
- Heuristic determination:
  - Translation probability between SL and TL words in chunk
  - Functional equivalence on either side of chunk
- For restricted domains, „careful“ generalization is used, which merely masks the surface details of chunks and does not assume modularity between levels of linguistic description

System: ReVerb -IV-

Case Creation

- Bitext alignement and linking of possibly-corresponding words using a bilingual dictionary; chunks will be aligned using linkage pattern
- Case -based parsing to generate chunks
- Chunk-boundary adjustements
  - Fragmentation
  - Extending chunk to include an additional word not not otherwise covered
  - Statistics used to increase the likelyhood of a good chunk boundary

System: Guvenir & Cicekli

(1996 -)

- Training examples are abstracted into templates by replacing certin word stems and morphemes by co-indexed variables
- Generalization based on the heuristic that differences in mostly similar entences should correspond

System: D^3

D^3 : DP-match driven transDucer

Eiichiro Sumita (2001)

- Similarity metric includes edit and semantic distance
- Generates translation patterns on the fly, selects most commonly used pattern
- Adapts examples by substituting target words for variables
- 90% coverage for „travel conversation“ sentences with 200K training examples, about 80% good quality

System: HPA/HPAT

Kenji Imamura (2001) Hierarchical Phrase Alignment

- Works by finding equivalent phrases from bilingual text
  - Corresponding content words
  - Same syntactic category
- Parse failurescause problems; try to alleviate by combining partial trees

HPAT: HPA-based Translation

- Generate transfer patterns from HPA-processed corpus
- Parse source using source patterns, map to target patterns, then translate leaves of tree using a dictionary
- About 70% good quality translation of „travel“ sentences using 125K training examples