Experiments with Surface Forms in EBMT (Matching and Alignment)

Monica Gavrila

NATS Oberseminar

03 May 2007

Motivation

- Thesis on Recombination in Example Based Machine Translation (EBMT)
  - "most difficult step in EBMT process" (H. Somers 2003, Kit et al. 2002)
  - "area that has received little attention" (McTait, 2001)
  - step dependent on the EBMT approach
  - not formalized
  - ***Romanian***, ***German***, English
    - language specific characteristics
    - less-resourced language

Contents

- **EBMT**
- Recombination in EBMT
- Previous Steps
  - Resources, Matching, Alignment
- Conclusions
- Further Work

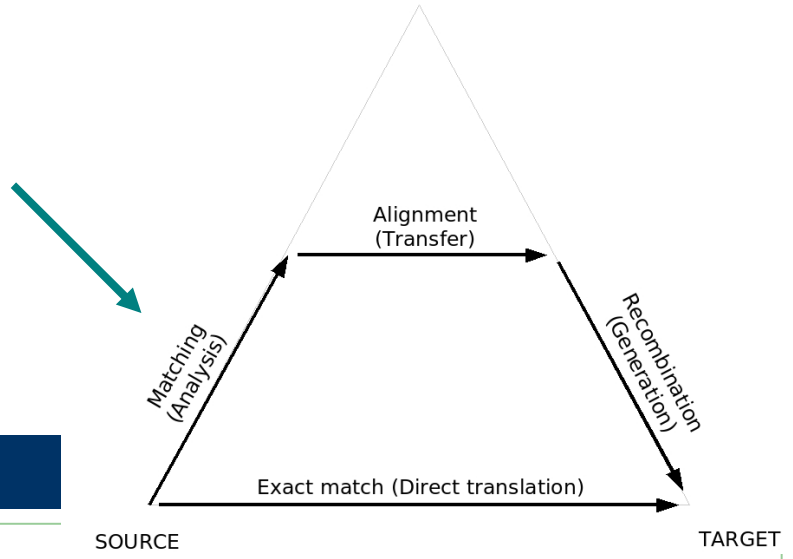
Example-Based Machine Translation

- Starting point: Makoto Nagao's work

"*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 principle with proper examples as its reference*" (Nagao, 1984)

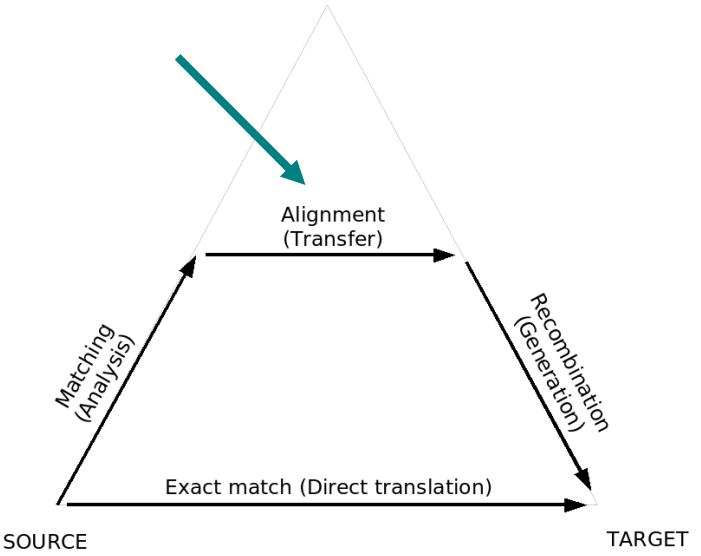
- Types
  - structure-based (trees), pattern-based etc.
- Resources
  - parallel aligned corpus, dictionary etc.

EBMT Steps



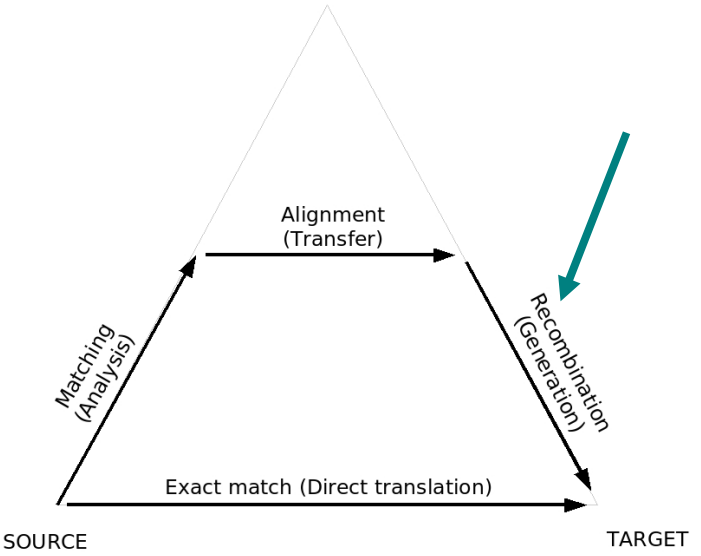
- ***Matching***
  - finds examples for the input
- Alignment
  - identifies corresponding translation fragments
- Recombination
  - recombines translation fragments into target texts

EBMT Steps



- Matching
  - finds examples for the input
- ***Alignment***
  - identifies corresponding translation fragments
- Recombination
  - recombines translation fragments into target texts

EBMT Steps



- Matching
  - finds examples for the input
- Alignment
  - identifies corresponding translation fragments
- ***Recombination***
  - recombines translation fragments into target texts

Contents

- EBMT
- **Recombination in EBMT**
- Previous Steps
  - Resources, Matching, Alignment
- Conclusions
- Further Work

Recombination in EBMT

- Depending on
  - previous steps
  - the EBMT approach
- ***Input***: SL input sentence, TL fragments aligned to SL fragments
- ***Output***: formed TL sentence (in case more results are considered: ranking scores)

Interesting Aspects in the Recombination Step

- boundary friction
- word order
- polysemy / homonymy
- data sparseness
- language dependent characteristics (e.g. „lampă/lamp“ – Romanian/English – 6 inflected forms)
- …. etc.

Contents

- EBMT
- Recombination in EBMT
- **Previous Steps**
  - Resources, Matching, Alignment
- Conclusions
- Further Work

Resources – The Corpus

- Technical corpus
  - parallel aligned at sentence level
  - 4 languages: Romanian, German, English, Russian
  - 2333 sentences, approx. 25.000 words

- Other possibility: JRC-Acquis
  (http://langtech.jrc.it/JRC-Acquis.html)
  - 22 languages
  - 9 million words (Romanian)

Matching in EBMT

- Different approaches
  - character-based
  - token-based
  - structure-based

- Preferred: token-based

Longest Common Subsequence Similarity - 1

- Surface-form matching similarity
- Based on *Longest Common Subsequence algorithm*
- *Token-based similarity*
- *Output*
  - *similarity value: between 0 and 1*
  - *the common subsequence*

Longest Common Subsequence Similarity - 2

$$1.\ LCS_{TokenString}(input,s) = LCS_{String}$$

$$2.\ LCS_{Tokens}(input,s) = Length_{token}\frac{(LCS_{String})}{Length_{token}}(input)$$

$$3.\ LCS_{Penalties} = Substracting\ a\ penalty\ of\ 0.1\ for\ each\ word\ distance - LCS_{TokenString}$$

$$4.\ LCS_{Chars}(input,s) = Length_{chars}\frac{(LCS_{String})}{Length_{chars}}(input)$$

“Simple” Matching with LCSS

- The input is tokenised
- Word index used to reduce the space

1. Find the sentence in the corpus that best matches the input (maximum value of the LCSS). Keep it part of the solution.
2. If the input is not fully covered, eliminate what was found. For the rest of the input repeat step 1.

“Simple” Matching with LCSS

- The input is tokenised
- Word index used to reduce the space

1. Find the sentence in the corpus that best matches the input (maximum value of the LCSS). Keep it part of the solution.
2. If the input is not fully covered, eliminate what was found. For the rest of the input repeat step 1.

“Simple” Matching with LCSS

- The input is tokenised
- Word index used to reduce the space

1. Find the sentence in the corpus that best matches the input (maximum value of the LCSS). Keep it part of the solution.
2. If the input is not fully covered, eliminate what was found. For the rest of the input repeat step 1.

“Simple” Matching with LCSS - Questions

- Is the highest LCSS score best solution?
- In case of the appearance of a word from the input several times in the example, which one to choose
- Case-sensitivity?
- … etc.

Matching - Results

- The coverage of the input
  - depends on data sparseness
    - solution: cognates in SL
- Same sentence, different language – not the same results
- Bad results in recombination
  - Improvement: considering the alignment

Alignmentq

- Tools: UPLUG, Kvec++, Moses, Twente, GIZA++
- Alignment Score based on surface forms – bilingual distribution and cognates

Alignment Score (starting point)

$$Score_{Align}(w1,w2)=\frac{Score_{BD}(w1,w2)+Score_{Cognates}(w1,w2)}{2}$$

$$Score_{BD}(w1,w2)=\frac{2*contains_{both}(w1,w2)}{contains(w1)+contains(w2)}$$

$$Score_{Cognates}(w1,w2)=1-\frac{LD(w1,w2)}{MaxLength(w1,w2)}$$

Alignment Example

- Romanian / English – only maximum considered

  Alte nume de produse si de firme mentionate aici pot fi nume comerciale sau *marci comerciale* apartinand proprietarilor respectivi

  Other product and company names mentioned herein may be *trademarks* or tradenames of their respective owners

- Improvement:
  - using a matrix - maximum on lines and on columns
  - more than word alignment needed

Alignment Example

- Romanian / English

  Alte nume de produse si de firme mentionate aici pot fi nume comerciale sau marci comerciale apartinand proprietarilor respectivi

  Other product and company names mentioned herein may be trademarks or tradenames of their respective owners

Alignment Example

- Romanian / English – only maximum considered

  Alte nume de produse si de firme mentionate aici pot fi nume comerciale sau marci comerciale apartinand proprietarilor respectivi

  Other product and company names mentioned herein may be trademarks or tradenames of their respective owners

Cognates – 1

- Cognate Score
- Calculated for
  - 1 language – matching
    - Clustering (Romanian: respectivi, respective, respectiva, respectiv): the higher the threshold, the more correct the results are – loss of information (e.g. No cluster: sesiune, sesiunii, sesiunea)
  - 2 languages - alignment

$$Score_{Cognates}(w1, w2) = 1 - \frac{LD(w1, w2)}{MaxLength(w1, w2)}$$

Cognates – 2

- Possible score improvement
  - Using penalties, according to the place where the differences in the words are

Cognates: Experimental Results - 1

	Threshold	En - Ro	Ge - Ro	En – Ge
L=0	0.5	10223	6707	6734
	0.66	2839	1534	1714
	0.75	1267	598	707
L=3	0.5	10143	6604	6643
	0.66	2759	1431	1623
	0.75	1250	582	689
L=5	0.5	8492	5598	5324
	0.66	2211	1097	1191
	0.75	930	424	460

Cognates: Experimental Results - 2

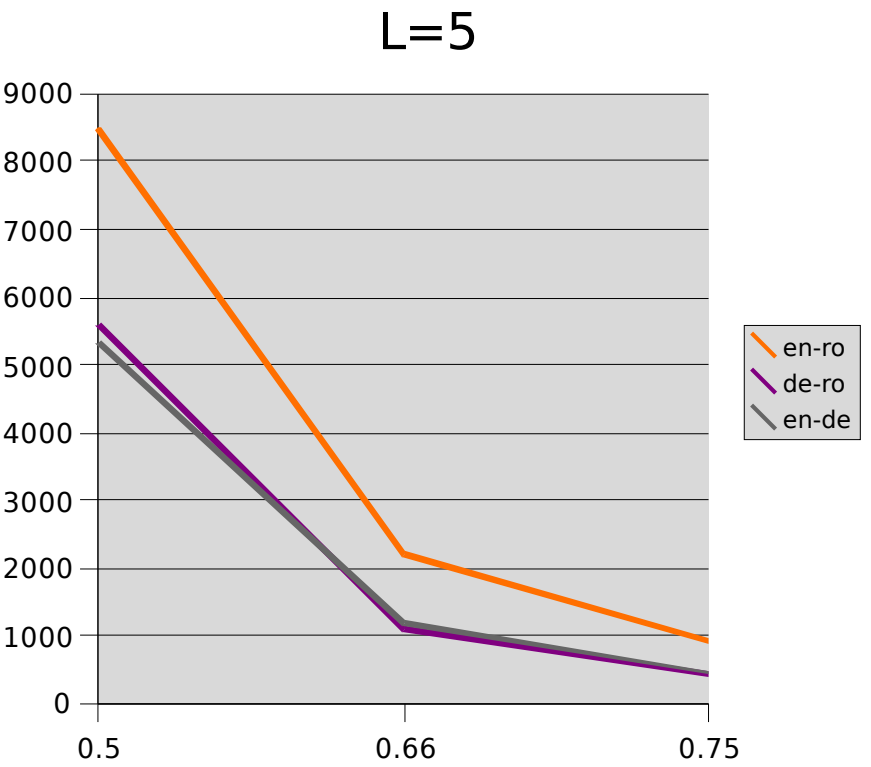
Length=5
Threshold=0.75

Length=4
Threshold=1

	En - Ro	Ge - Ro	En – Ge
No constraint	930	424	460
Common first letter	896	400	427
Correct	699	340	377
Different word	465 (51.9%)	240 (60%)	291 (68.15%)
Wrong	197 (22%)	60 (15%)	50 (11.7%)
	134	100	118

Cognates: Experimental Results - 3

- Different language family
- Flexion forms for Romanian
- Latin influence on English bigger than in German?!?!
- Compound words in German
- Corpus dependency: foreign words in the language
- Character transformations:
  - ü – ue (German)
  - ă – a (Romanian)



Again Alignment

- As now GIZA++ is compiled…
  - how it is working for Romanian, German, English
  - a way of improving it
  - more than word alignment possible?

Example - 1

TO TRANSLATE: Pentru a termina sesiunea chat apasati OK si selectati Renuntati

Results:

**** 1997:pentru a termina apasati si selectati renuntati

Values: 0.7 / 0.734375 / 0.6699999999999999

MaxC: true / MaxP: true

Example - 2

TO TRANSLATE*: Pentru a termina* sesiunea chat *apasati* OK *si selectati Renuntati*

Results:

**** 1048:chat ok

Values: 0.6666666666666666 / 0.4375 / 0.6666666666666666

MaxC: true / MaxP: true

Example - 3

TO TRANSLATE*: Pentru a termina* sesiunea *chat apasati OK si selectati* Renuntati

Results:

****

Example - 4

"Pentru a termina sesiunea chat apasati OK si selectati Renuntati"

-Sentence 1997

SL=**Pentru a** opri navigarea si a **termina** conexiunea , **apasati** Optiuni **si selectati Renuntati** .

TL=***To*** quit browsing and *to* ***end*** the connection , ***press*** Options ***and select Quit*** .

-Sentence 1048

SL=Tastati sau cautati in agenda telefonica numarul de telefon al persoanei cu care doriti sa schimbati mesaje in cadrul unei sesiuni **chat** , apoi apasati **OK** .

TL=Key in or search in the phone book the phone number of the person with whom you want to start the ***chat*** session and press ***OK*** .

-Sentence: SL= / TL=

Example - 5

"Pentru a termina sesiunea chat apasati OK si selectati Renuntati"

-Sentences 1997

SL=**Pentru a** opri navigarea si a **termina** conexiunea , **apasati** Optiuni **si selectati Renuntati** .

TL=***To*** quit browsing and *to* ***end*** the connection , ***press*** Options ***and select Quit*** .

Output: To end - - - press - - - and select Quit

Example - 6

"Pentru a termina sesiunea chat apasati OK si selectati Renuntati"

-Sentence 1048

SL=Tastati sau cautati in agenda telefonica numarul de telefon al persoanei cu care doriti sa schimbati mesaje in cadrul unei sesiuni **chat** , apoi apasati **OK** .

TL=Key in or search in the phone book the phone number of the person with whom you want to start the ***chat*** session and press ***OK*** .

Output: To end --- chat press OK and select Quit

Example - 7

"Pentru a termina sesiunea chat apasati OK si selectati Renuntati"

Output: To end --- chat press OK and select Quit

Problems: „sesiunea“ not found, word order

Solution: cluster for Romanian: {sesiunii, sesiunea} -> translation „session“.

Example - 8

„To end the chat session press Ok and select Quit“

Matching on 1057, 1997, 459

Output: Pentru a termina SESIUNII chat apasati OK si selectati Renuntati.

Problem: „Sesiunii“ is incorrect. Sesiunea needed.

Solution: cluster for „sesiunea“, and ranking (on www.google.ro).

Solution: Pentru a termina sesiunea chat apasati Ok si selectati Renuntati.

Contents

- EBMT
- Recombination in EBMT
- Previous Steps
  - Resources, Matching, Alignment
- **Conclusions**
- Further Work

Conclusions

- Surface form algorithms are helpful for
  - low-resourced languages (e.g. Romanian)
  - forming an idea about a phenomenon
- Surface form algorithms usually give worse results that more complex ones

Contents

- EBMT
- Recombination in EBMT
- Previous Steps
  - Resources, Matching, Alignment
- Conclusions
- **Further Work**

Further Work

- Finish the alignment
- Start the **recombination**

Thank you!

QUESTIONS?

Demo

(Some) References

- H. Somers, *An Overview of EBMT*, vol. recent Advances in EBMT, pp. 3-57, Kluwer Acad. Publ. 2003
- C. Kit et al. *EBMT: A New Paradigm*, vol Translation and Information technology, pp. 57-78, Chinese U of HK Press, 2002
- K. McTait, *Translation Pattern Extraction and Recombination for EBMT*, PhD Thesis, UMIST, 2001
- M. Nagao, *A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, NATO Symposium on AI and Human Intelligence, pp. 173-180, Elsevier North-Holland Inc., 1984*

(Some) References -2

- Vauquois, B. *A Survey of Formal Grammars and Algorithms for Recognition and Transformation in Machine Translation*. IFIP Congress-68, Edinburgh, pp. 254–260, 1968