Machine Translation
Alignment & Example-based MT

Jörg Tiedemann
jorg.tiedemann@lingfil.uu.se
Department of Linguistics and Philology
Uppsala University
October 2009

Introduction
Sentence alignment

Example-Based Machine Translation

The general idea: translation by analogy

- collect a large database of examples with translations
- when translating new sentences:
  - retrieve similar examples including partial matches
  - preference for longer matches
  - take partial matches and merge them together (recombination)
- advanced: generalize examples to templates

"The Amazing Utility of Parallel Corpora"

Example-Based Machine Translation

- strong relation to TM's and fuzzy matching
- alternative names: "analogy-based", "case-based", "memory-based", "experience-guided"

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 ..., then by translating these phrases into other language 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: 178f)
Example-Based Machine Translation

Famous example (Sato & Nagao, 1990)

- **translate**: “He buys a book on international politics.”
- **examples**:
  1. He buys a notebook.
     Kare wa nōto o kau.
     HE topic NOTEBOOK obj BUY.
  2. I read a book on international politics.
     Watashi wa kokusai seiji nitsuite kakareta hon o yomu.
     I topic INTERNATIONAL POLITICS ABOUT CONCERNED
     BOOK obj READ.
- **output**: “Kare wa kokusai seiji nitsuite kakareta hon o kau.”

Vauquois pyramid adapted for EBMT

(Von Somers:99)

Building an Example Database

What do we need?

- Database of aligned examples!
- Engine for retrieving matched fragments
- Algorithms for aligning and merging fragments

→ First of all: A sentence aligned parallel corpus!
Introduction
Sentence alignment

Automatic sentence alignment

Task: align corresponding sentences to each other (may be sequences of sentences)

Assumption: sentence alignment can be done monotonically (no crossing links!)

Challenges: non-1:1 alignments, insertions, deletions, incomplete translations

Many different ways to align sentences:
\[(s_1, t_1) (s_2, t_2) (s_3, t_3) \ldots\]
\[(s_1, t_1 + t_2) (s_2, t_3) (s_3, 0) \ldots\]
\[(s_1 + s_2, t_1) (s_3, t_2 + t_3) \ldots\]

Jörg Tiedemann 9/43

Sentence alignment approaches

Length-based methods: assumption = sentences (and sequences of sentences) that correspond to each other are also similar in length (characters or words) (more than others)

Lexical methods: assumption = corresponding sentences contain more corresponding words; use distribution of corresponding words in source and target language texts

Combined methods: use lexical cues in length-based settings

Jörg Tiedemann 10/43

Gale & Church: a length-based model

Strong correlation between paragraph lengths for various language pairs!

![Graph showing correlation between German and English paragraph lengths](chart.png)

(correlation = 0.991)

Jörg Tiedemann 11/43

Pirates of the Caribbean

1. Drink up me’ sailor yo-ho!
2. We kidnap and ravage and don’t give a hoot
3. Drink up me’ sailor yo-ho!
4. Yo ho yo ho a pirate’s life for me!
5. We extort, we pillage, we flinch and snack
6. Drink up
7. Omit, misery, cursed pirates and these waters.
8. You don’t want to bring them down on us now, do you?
9. - Mr. Gibbs that will do!
10. - She was singing about pirates!
11. Bad luck to be singing about pirates with as snared in this unnatural frog ...
12. Mark my words!
13. Consider them marked
14. On your way!
15. Aye, Lieutenant
16. It’s bad luck to have a woman on board, too, even a miniature one.
17. I think it’d be rather exciting to meet a pirate.
18. Time again, Miss Swan.
19. Vile and dimwitted creatures, the lot of them.
20. I intend to see to that any man who sails under a pirate flag or ... wears a pirate brand gets what he deserves.
21. A short drop and a sudden stop
22. Lieutenant Norrington, I appreciate your service.
23. But I’m ... I’m concerned about the effect this subject will have upon our daughter

PIRATES OF THE CARIBBEAN

1. Drink up me’ sailor yo-ho!
2. Trick up, yo ho ho ...
3. Trick up, yo ho ho! ...
4. Yo ho ho, on pirate villain’s trail ...
5. Vi plunderar, rövar och skäller ...
6. Tom dit gis ...
7. Tydligen, hirudinna piratens seger ockring här ...
8. Vi vill inte kalla på dem ?
9. - Mr. Gibbs, det räcker !
10. Hans sjöblod om piraten ...
11. Det beter att att sjunga om piratens när vi seger i den här dimman ...
12. Jom ihag det ...
13. Jom ska komma ihag det ...
14. Träg nu ...
15. Jämst, jämst ...
16. Även en i minustry storlek ...
17. Jag tror det skulle vara ganska upphördande att nöja en pirat ...
18. Tänk om igen miss Swan ...
19. Vilda och otredjeckna varelser ...
20. Piraflagget skrapp ...
21. Och här ett problem får vad hon börjar ...
22. I trycket och märklig ...

Jörg Tiedemann 12/43
Gale & Church: a length-based model

- define a distance based on the costs of aligning source to target sentences (for a fixed finite set of possible alignment types)
- minimize this distance by finding the best alignment using dynamic programming → recursive definition of

\[
D(i,j) = \min \begin{cases} 
D(i-1,j) + \text{cost}(\text{align}_1, 0, t_j) \\
D(i-1,j-1) + \text{cost}(\text{align}_1, s_i, s_j) \\
D(i-1,j-2) + \text{cost}(\text{align}_2, s_i, t_j) \\
D(i-2,j-1) + \text{cost}(\text{align}_2, 1, s_j) \\
D(i-2,j-2) + \text{cost}(\text{align}_2, 1, 1) 
\end{cases}
\]

- alignment costs are based on length correspondence

Gale & Church: parameters

- empirically find parameters \(c\) and \(s^2\) from example corpora (English/German and English/French)

\[
\delta = (l_2 - l_1 c) / \sqrt{l_1 s^2}
\]

parameters:
\[
c = 1 \\
s^2 = 6.8
\]

→ distribution: approximately normal
Gale & Church: parameters

- additional parameters: prior probabilities
  (from English/German and English/French corpora)

\[
\begin{align*}
P(\text{aligntype } = 1 : 1) & = 0.89 \\
P(\text{aligntype } = 1 : 0) & = 0.0099 \\
P(\text{aligntype } = 0 : 1) & = 0.0099 \\
P(\text{aligntype } = 2 : 1) & = 0.089 \\
P(\text{aligntype } = 1 : 2) & = 0.089 \\
P(\text{aligntype } = 2 : 2) & = 0.011
\end{align*}
\]

Gale & Church: dynamic programming

- applying this algorithm:
  - compute alignment costs for each sentence pair \((i, j)\)
  - start with 0.0 and fill the entire table
  - read the alignment path with minimal costs

→ more efficient on shorter text chunks
→ paragraph alignment to speed up the process!

Gale & Church: comments

- fixed parameters seem to be very ad-hoc ...
- ... but it works surprisingly well
  (>95% for most language pairs)

Possible issues:
- fixed set of alignment types not always adequate
- problems with incomplete translations
- follow-up errors
→ add synchronization points (hard boundaries)

Pirates of the Caribbean: misaligned

<table>
<thead>
<tr>
<th>English</th>
<th>Norwegian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drakk up me ' earties ye ho l</td>
<td>Sjørorve ar jag</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Yo ho yo ho a pirate’s life for me</td>
<td>Det fins sjørorve hår</td>
</tr>
<tr>
<td>We extort, we piller, we flich and sack</td>
<td>Du vil vi inte locka hit dem ?</td>
</tr>
<tr>
<td>Drakk up me ' earties ye ho l</td>
<td>Mr Gibbs, det räcker</td>
</tr>
<tr>
<td>Quiet, misy, cursed pirates and these waters</td>
<td>Tyst, låt trikke l</td>
</tr>
<tr>
<td>You don’t want to bring them down on us now, do you ?</td>
<td>Att sjunga om sjørorve i denne dimma gar otar , om</td>
</tr>
<tr>
<td>- Mr. Gibbs that will do</td>
<td>jag får sága det</td>
</tr>
<tr>
<td>She was singing about pirates</td>
<td>- i vág med er</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Bad luck to be singing about pirates with us mixed in</td>
<td>Att ha kvinnor ombord gør også otar , även om hon</td>
</tr>
<tr>
<td>this unnatural fog</td>
<td>är en miniatyr ...</td>
</tr>
<tr>
<td>Mark my words</td>
<td>...</td>
</tr>
<tr>
<td>Consider them marked</td>
<td>Att träffa en sjørorve vore spánande</td>
</tr>
<tr>
<td>On your way</td>
<td>...</td>
</tr>
<tr>
<td>Aye, Lieutenant</td>
<td>Tøn etter , miss Swann</td>
</tr>
<tr>
<td>You're bad luck to have a woman on board , too , even a</td>
<td>De er avskyværa og løstbarna sildløp</td>
</tr>
<tr>
<td>miniature one</td>
<td>...</td>
</tr>
</tbody>
</table>
**Introduction**

Sentence alignment

Pirates of the Caribbean: added hard boundaries

**Lexical methods**

- General idea: use lexical cues to find optimal sentence alignments
  - Kay and Röscheisen: partial alignment of lexical items
    - Iterate between alignment of lexical items and induced sentence alignment
    - Start with pillow shaped envelope of possible alignments
    - Choose pairs of co-occurring words with similar distributions and induce new anchor alignments with many possible lexical correspondences and repeat

→ Accurate but computationally expensive

---

Kay & Röscheisen’s method (intermediate step)
Lexical methods

other variants:

Chen: construct word-to-word translation models (used for cost estimation) to find best alignment like Gale&Church

Melamed: Geometric Mapping and Alignment (GMA): find true points of correspondence (cognates, dictionaries, etc) and search for local chains between them

Combined methods

▶ use cognates to find hard boundaries (anchor points) for length-based alignment methods (usually paragraph boundaries are used as hard boundaries)

▶ induce lexical (word) alignments from corpora initially aligned using simple length-based techniques and re-align using reliable lexical correspondences as additional anchor points (hunalign)
Subtitle alignment using time-slot overlaps

Sometimes other alignment techniques are more appropriate

<table>
<thead>
<tr>
<th>English</th>
<th>Dutch</th>
</tr>
</thead>
<tbody>
<tr>
<td>00:00:34,500 --&gt; 00:00:28,434</td>
<td>De wereld van Wayne 00:00:32,333 --&gt; 00:00:28,943</td>
</tr>
<tr>
<td>Spend all day with us.</td>
<td>Er zijn twee, excuseer me, twee van Zantar. 00:00:37,314 --&gt; 00:00:38,664</td>
</tr>
<tr>
<td>two of everything in</td>
<td>...gestoorde helicoperen. 00:00:42,434 --&gt; 00:00:45,367</td>
</tr>
<tr>
<td>Noah's arcade.</td>
<td>Het is goed om je weer te zien, Benjamin.</td>
</tr>
</tbody>
</table>

Available Tools

Some sentence alignment tools:

- Vanilla aligner (Gale&Church) (http://nl.ijs.si/telri/Vanilla/)
- hunalign (http://mokk.bme.hu/resources/hunalign)
- Melamed’s GMA (http://nlp.cs.nyu.edu/GMA/)
- Uplug toolbox (http://sourceforge.net/projects/uplug/)
  - pre-processing
  - automatic sentence alignment
  - interactive sentence alignment

Back to Example-Based Machine Translation

Now we have a large database of translation examples!
What else do we have to discuss?

Issues to be addressed:

- size & suitability of example database
- way of storing examples
- matching strategy
- alignment (adaptation) & recombination
Introduction

Example-Based Machine Translation

Issues to be addressed:
▶ size & suitability of example database
▶ way of storing examples
  ▶ plain text examples (purist approach)
  ▶ aligned tree structures
  ▶ morphologically analyzed examples
  ▶ generalized examples
▶ matching
▶ alignment (adaptation) & recombination

Morphologically analyzed examples

English - Turkish

without morphological analysis with morphological analysis
▶ I am coming
geliyorum
▶ I am going
gidiyorum
▶ I am come+ing
gel+Hyor+yHm
▶ I am go+ing
gid+Hyor+yHm

("H" captures Turkish vowel harmony)

Aligned Trees in Example Database

Kanojo wa kami ga nagai.
She has long hair.
SHE topic HAIR subj IS-LONG

Example-Based Machine Translation

Issues to be addressed:
▶ size & suitability of example database
▶ way of storing examples
▶ matching
  ▶ character based
  ▶ word based
  ▶ annotation/structure based
▶ alignment (adaptation) & recombination
Example-Based Machine Translation

Issues to be addressed:
- size & suitability of example database
- way of storing examples
- matching
- alignment (adaptation) & recombination
  - fragment alignment
  - boundary friction
  - statistical ranking

Alignment (Adaptation)

In case of partial matches:
need to align fragments with appropriate target language fragments
- requires word-level (or even character-level) alignment
- in case of annotated examples
  (for example syntactic trees): syntactic (tree) alignment

→ Difficult issue!

(More on word alignment next time!)

Recombination

Problem of “boundary friction”:
- syntactic function not always marked in the same way
- wrong agreement in recombined fragments
- problems with “filler” words, ambiguous/redundant function words, gaps in example translations ...

→ Ranking candidates with target language model helps!

Automatic Generalization: Templates

- replace named entities (PER, LOC, ORG), numbers, dates with place-holder variables:
  - John Hancock was in Philadelphia on July 4th.
  - PERSON was in CITY on DATE.
  - PERSON war am DATE in CITY.
- generalization by syntactic category (e.g. NP)
  - rekodo no nagasa wa saidai 512 baito de aru.
  - The maximum length of records is 512 bytes.
  - X[NP] no nagasa wa saidai Y[N] baito de aru.
  - The maximum length of X[NP] is Y[N] bytes.
- generalization by semantic role/feature
  - play X[NP/sport]
  - X[NP] o suru
  - play X[NP/instrument]
  - X[NP] o hiku
Related Topics

Many of the ideas above move away from the pure analogy-based approach

- relation to transfer based approaches
  - derive transfer rules from examples
  - recombination & target language generation
- relation to statistical approaches
  - word alignment & statistical preference
  - output ranking using language models

Summary EBMT's

- no clear definition of EBMT
- required: bilingually aligned examples
- data-driven rather than theory-driven
- many (open) issues/problems
- relations to both, transfer-based MT & statistical MT

What's next?

Another (more popular) data-driven approach: statistical machine translation (SMT)

- basic concepts
- word alignment
- language modeling
- decoding