NLP for the translation class

Lars Ahrenberg and Ljuba Tarvi

2nd workshop on NLP for Computer-Assisted Language Learning
Oslo, May 22, 2013
Overview

- Background
- Our proposal
- The Token-Equivalence Method (TEM)
- Alignment experiments
- Conclusions
Process of a translation exercise

**Teacher Tasks:**
- Define exercise (text + brief)
- Preprocessing
- Distribution

**Student tasks:**
- Translation, Delivery
- Reading for assessment and grading, Preparation

**Before class**
- Classroom presentation and discussion

**In class**
- Final grading

**After class**
- Revision
Examples of computational aids for the translation exercise

- E-learning environments

- Corpora
  - Lopez-Rodriguez and Tercedor-Sanchez, 2008;
  - Pastor and Alcina, 2009

- CAT tools
  - Assessment of translations as literal or liberal
    - Shei and Pain, 2002
Our idea

- Computer-aided support for the Token-Equivalence Method (TEM; Tarvi, 2004)
- A new application area for alignment technology

Supporting
- teacher's assessment and grading
- discussion in class
Process of a translation exercise

**Before class**
- Define exercise (text + brief)
- Preprocessing Distribution

**In class**
- Reading for assessment and grading, Preparation
- Classroom presentation and discussion

**After class**
- Final grading
- Revision

**Student tasks:**
- Translation, Delivery

**Teacher Tasks:**
- Classroom presentation and discussion
- Before class
- After class
- Final grading

- Revision
Token alignment as a basis for instruction in class

- **Segment views**
  - display of different translations of the same source segment

- **Token views**
  - display of different translations of the same token(s)

- **Type views**
  - e.g. frequency tables of translations of words and phrases

- **Global views**
  - metrics and grades computed for the full text or parts thereof
The Token-Equivalence Method (TEM)

- Token correspondences, based on
  - content words
  - denotational meaning

- Frames
  - metrics that quantify relations between source and translation
  - combined to rank translations
An example (RU - EN)

Pushkin, Eugene Onegin, stanza LIX: 1-2

- Proshla lyubov, yavilas' muza, i projasnilsya tyomnyi um.

Translation (by Nabokov)

- Love passed, the Muse appeared, and the dark mind cleared up.

Indexing tokens


[passed] [love] [appeared] [muse] [and] [cleared up] [dark] [mind]


"Standard" alignment representation

1-2 2-1 3-5 4-4 5-6 6-10 6-11 7-8 8-9 0-3 0-7
An example (RU - EN)

Pushkin, Eugene Onegin, stanza LIX: 1-2

- Proshla lyubov, yavilas' muza, i projasnilsya tyomnyi um.

Translation (by Nabokov)

- Love passed, the Muse appeared, and the dark mind cleared up.

Indexing tokens

[passed] [love] [appeared] [muse] [and] [cleared up] [dark] [mind]


"Standard" alignment representation

1-2 2-1 3-5 4-4 5-6 6-10 6-11 7-8 8-9 0-3 0-7

multiword correspondence
An example (RU - EN)

Pushkin, Eugene Onegin, stanza LIX: 1-2

- Proshla lyubov, yavilas' muza, i projasnilsyva tyomnyi um.

Translation (by Nabokov)

- Love passed, the Muse appeared, and the dark mind cleared up.

Indexing tokens


[passed] [love] [appeared] [muse] [and] [cleared up] [dark] [mind]


"Standard" alignment representation

1-2 2-1 3-5 4-4 5-6 6-10 6-11 7-8 8-9 0-3 0-7

null links
# Differences TEM and standard SMT alignment

<table>
<thead>
<tr>
<th>Aspect</th>
<th>TEM</th>
<th>SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approach</td>
<td>manual</td>
<td>automatic</td>
</tr>
<tr>
<td>Punctuation</td>
<td>ignored</td>
<td>tokenized</td>
</tr>
<tr>
<td>Token types</td>
<td>two types</td>
<td>one type</td>
</tr>
<tr>
<td>Multiword units</td>
<td>single tokens</td>
<td>several tokens</td>
</tr>
</tbody>
</table>
Some definitions

- $t_s$: a source token
- $t_t$: a target token
- null(t): a token without correspondent
- nonnull(t): a token with at least one correspondent
- cont(t): a content token
- gram(t): a grammar token
TEM frames

- Basic content frame
  - "the percentage of source content tokens that have received a translation"
  - $\text{BCF} = 100 \times \frac{|\{t_s | \text{cont}(t_s) \land \text{nonnull}(t_s)\}|}{|t_s|}$

- Optional content frame
  - $\text{OCF} = |\{t_T | \text{cont}(t_T) \land \text{null}(t_T)\}|$
TEM frames (cont.)

- **Basic formal frame**
  - "the number of grammar tokens in the translation"
  - \[ \text{BFF} = | \{ t_T | \text{gram}(t_T) \land \text{null}(t_T) \} | \]

- **Optional formal frame 1**
  - "the percentage of source tokens that are translated by a token of the same part-of-speech"

- **Optional formal frame 2**
  - "the percentage of pairs of source tokens whose order and dependency relation is kept under translation"
The translation quotient (TQ)

- The TQ is defined as the average of all frames that are expressed as percentages:
  - \( TQ = \frac{BCF + OFF1 + OFF2}{3} \)
- All frames may be used to compute a rank for each translation
Word alignment for the translation class

- Source texts are short
- Translations, on the other hand, may be many
- Source texts are known beforehand
- Content tokens and grammar tokens should be treated differently

→ Statistical and rule-based methods may be combined
Alignment experiments

- **Russian-English data**
  - 8 translations of 17 stanzas from Eugene Onegin

- **English-Swedish data**
  - 5 translations of two small extracts of English prose text used as exercises in a course.

- **Systems used**
  - Giza++ (both corpora)
  - A ”pressure-aligner” (only EN-SE), using
    - a dictionary
    - part-of-speech patterns
    - alignment topology
Alignment results, RU-EN, Giza++ (model 4)

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 trl, all links</td>
<td>0.308</td>
<td>0.298</td>
<td>0.303</td>
</tr>
<tr>
<td>8 trls, no null links</td>
<td>0.434</td>
<td>0.467</td>
<td>0.450</td>
</tr>
<tr>
<td>8 trls, all links</td>
<td>0.482</td>
<td>0.480</td>
<td>0.481</td>
</tr>
</tbody>
</table>

**Note:** the gold standard used has some 40% added tokens, while Giza++ gives 20%.
Alignment results for EN-SE, Giza++ (model 4)

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 trl, no null links</td>
<td>0.751</td>
<td>0.652</td>
<td>0.698</td>
</tr>
<tr>
<td>1 trl, all links</td>
<td>0.681</td>
<td>0.681</td>
<td>0.681</td>
</tr>
<tr>
<td>5 trls, no null links</td>
<td>0.816</td>
<td>0.698</td>
<td>0.752</td>
</tr>
<tr>
<td>5 trls, all links</td>
<td>0.752</td>
<td>0.738</td>
<td>0.745</td>
</tr>
</tbody>
</table>
## Alignment results for EN-SE, rule-based aligner

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>PA1, no null links</td>
<td>0.815</td>
<td>0.492</td>
<td>0.614</td>
</tr>
<tr>
<td>PA1, all links</td>
<td>0.502</td>
<td>0.554</td>
<td>0.527</td>
</tr>
<tr>
<td>PA2, no null links</td>
<td>0.885</td>
<td>0.608</td>
<td>0.721</td>
</tr>
<tr>
<td>PA2, all links</td>
<td>0.606</td>
<td>0.664</td>
<td>0.633</td>
</tr>
</tbody>
</table>

PA1 has a small lexicon, while PA2 has a lexicon adapted for the corpus.
Alignment results for EN-SE, combinations of Giza++ and rule-based aligner

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Union, no null</td>
<td>0.775</td>
<td>0.777</td>
<td>0.776</td>
</tr>
<tr>
<td>Union, all</td>
<td>0.739</td>
<td>0.789</td>
<td>0.763</td>
</tr>
<tr>
<td>Intersection, no null</td>
<td>0.980</td>
<td>0.530</td>
<td>0.688</td>
</tr>
<tr>
<td>Intersection, all</td>
<td>0.875</td>
<td>0.543</td>
<td>0.670</td>
</tr>
<tr>
<td>Grown, no null</td>
<td>0.849</td>
<td>0.665</td>
<td>0.746</td>
</tr>
<tr>
<td>Grown, all</td>
<td>0.794</td>
<td>0.660</td>
<td>0.721</td>
</tr>
</tbody>
</table>
Observations on alignment performance

- As expected, adding more translations improves the results of the statistical aligner.
- Since the source text is known, and small, creating a dictionary for the source adapted for the task is not so much work and improves the results of the pressure aligner substantially.
- A combination of statistical and dictionary-based alignment can give very high precision.
- All possibilities have not been explored yet...
Conclusions

- There is much work ahead
  - implementation
  - trying it out
- Even with further improvements in the automatic tools, there will still be much to do for the teacher in reviewing and correcting token alignments
  - Need for good interactive tools!
Thank You