IWSLT 2004 Workshop

The ISL EDTRL System

Jürgen Reichert
University of Karlsruhe
Overview

1. Introduction
2. Basic Ideas of EDTRL
3. Training and Translating
4. Experiments and Results
1. Introduction

Source Text

Syntax-based Translation

Interlingua

Semantics-based Translation

Target Text

(Semantics)

(Syntax)

Source Text

Target Text

Word-to-Word Translation

level of abstraction

Interactive Systems Labs
Usage of Knowledge

• Knowledge-based Approaches
  – Grammar writing
  – Expert system
  – Frame-based Machine Translation

• Data-driven Approaches
  – Example-based Machine Translation
  – Statistical Machine Translation
  – Grammar learning
2. Basic Ideas of EDTRL

Combine the Approaches of

- Interlingua Systems
  (2n Modules, NLP, Paraphrases)

- Data-Driven Systems
  (only depends from corpus)

- Knowledge-based Approaches
  (further semantic and morphologic knowledge)
What do we get?

⇒ Domain only depends from training corpus
⇒ No handcrafted work
⇒ Easily add new language
⇒ Reduce Parallel Data Sparseness Problem
  - Very few
    Spanish – Chinese
  - Large amount of
    Chinese – English
    English – Spanish
⇒ use Chinese – English – Spanish
Enriched and Formalized English as Interlingua

1. Cascaded translation
2. Preserve translation alternatives
3. Confidence Measures / Probabilities
4. Standardized and Simplified English
5. Linguistically enriched English
Formalizations & Enrichment

- Simplified English (common used alternative)
- Standardized word order (please give me … -> give me … please)
- Added Attributes
  - Morphological knowledge
  - Sense
  - Synonym Generator
  - Part-of-Speech Tags
  - Named Entity Tags
  - (sentence type, active/passive, politeness, domain, category, …)
Statistical Translation Rules

• Transfer knowledge from English to the foreign language using the statistical Alignment
• Error-Driven Learning (learn from errors)
• Interactive learning modus
• Translation error tracking and correction
• Small memory/time footprint (scalable)
3. EDTRL Training

- Parallel training data
  - Word / Phrase Align → Statistical Alignment
  - Chunks
    - Morphology
      - Base form + type
    - Sense Guesser
      - Word sense
    - Synonyms Gen.
      - Synonyms
    - POS Tagger
      - POS-Tags
    - Named E.Tagger
      - Named-Entity-Tags

- Knowledge sources
  - Translation rules
    - Rule Selection
      - Rules
    - Rule Generation
      - Meta rules, Dictionary
Weight functions for Alignment

1. Weight Position factor

\[
\frac{1}{\text{WordPos1} - \left( \text{WordPos2} \cdot \frac{\#\text{Words1}}{\#\text{Words2}} \right)}
\]

```
for a while it was very painful but now it's all right
```

```
for a while it was very painful but now it's all right
```

*
Weight functions for Alignment

2. Length penalty \( \frac{k}{\log(len)} \)

3. Matching Length factor (prefer same length)
\[
\frac{#LenA + #LenB}{2 \cdot \max(#LenA, #LenB)}
\]

4. Frequency Weight (prefer alignment between words with similar frequency)
\[
\frac{#wordsA + #wordsB}{2 \cdot \max(#wordsA, #wordsB)}
\]
Weight functions for Alignment

Alignment Distance

Interactive Systems Labs
Align Dependence = (align weights + align weights) / (align weights + align weights)

LM Dependence = \( \frac{\text{Sum} \ P(W_n \| W_{n-1})}{\text{Sum} \ P(W_n \| W_{n-1})} \)
Template generation

1. Cluster similar Sentence pairs
2. Generate Phrase Alignment
3. Build templates with classes for the different words

我孩子说她 <0:bodypart> 痛
我孩子说他 头痛
我孩子说他 牙齿痛

我孩子说她 <0:bodypart> 痛
Rule Generation/Selection

- Rules
  Cond1 | Cond2 | ... → Templ1 | Templ2 | ...
  Build form: Word, Phrase, Attribute Class Scores (Probabilities) for each Template
- Find ‘optimal’ rules
- Evaluate rule on verification set
- Using a class hierarchy
- Using meta-rules for the construction
Learning reorder rules

- Search reorders with a high alignment confidence
- Generalize or specialize the reorder rules by introducing classes and conditions
Translation

- Left to Right
- Find Matching rules -> probability
- Instantiate rules -> probability
- Beam-Search weighted by a trigram
- Pruning
Translation

A) Chinese -> IL (Tagged English):

我想我从某人那传染上感冒了

我从某人那<1>上<2>了

-> I've<1:VB>a<2:Disease>fromsomeone

传染: infection<NN>0.3, transmission<NN>0.1, infect<VB>0.1, catch<VB>0.2
感冒: cold<Disease>0.3, rheum<Body Substance>0.2, to catch cold<Change>0.4

Instantiation => I think I've caught a cold from someone
I think I've caught a cold from someone

I've a Disease from someone

→ 我从某人那 上了

catch VB: 捕捉 0.4, 逮 0.3, 传染 0.1 ...
catch NN: 陷阱 0.1, ...
cold Temperature attribute: 冷 0.4, 凉 0.4
cold Disease: 感冒 1.0

Instantiation => 我想我从某人那 捕捉上 感冒了
4. Experiments and Results

Preprocessing
- IWSLT
- New Segmentation (for Chinese)

Post processing
- a -> an
- removing duplicates
- Some verb form adaption
### Database

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># English Phrases</td>
<td># Chinese Phrases</td>
</tr>
<tr>
<td>BTEC</td>
<td>162314</td>
<td>162314</td>
</tr>
<tr>
<td>Medical</td>
<td>7634</td>
<td>7634</td>
</tr>
<tr>
<td>Tourism</td>
<td>2003</td>
<td>2003</td>
</tr>
<tr>
<td><strong>∑</strong></td>
<td>171951</td>
<td>171951</td>
</tr>
</tbody>
</table>
### Experiment: mixing Domains

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th>NIST-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTEC (162000)</td>
<td>BTEC (506)</td>
<td>4,7109</td>
</tr>
<tr>
<td>medical (6500)</td>
<td>medical (200)</td>
<td>2,8434</td>
</tr>
<tr>
<td>tourism (2000)</td>
<td>tourism (200)</td>
<td>3,1706</td>
</tr>
<tr>
<td>btec+medical+tourism</td>
<td>btec</td>
<td>4,7617</td>
</tr>
<tr>
<td>btec+medical+tourism</td>
<td>medical</td>
<td>2,8952</td>
</tr>
<tr>
<td>btec+medical+tourism</td>
<td>tourism</td>
<td>3,1735</td>
</tr>
<tr>
<td>btec+medical+tourism</td>
<td>btec (large system)</td>
<td>4,8383</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Systems</th>
<th>EDTRL</th>
<th>Systran</th>
</tr>
</thead>
<tbody>
<tr>
<td>C → E</td>
<td>7,34</td>
<td>5,74</td>
</tr>
<tr>
<td>E → S</td>
<td>5,17</td>
<td>6,06</td>
</tr>
<tr>
<td>C → S</td>
<td>3,17</td>
<td>-</td>
</tr>
<tr>
<td>C → E → S</td>
<td>3,41</td>
<td>2,84</td>
</tr>
<tr>
<td>C → E IL → S</td>
<td>3,69</td>
<td>-</td>
</tr>
</tbody>
</table>
## IWSLT 2004 evaluation
Chinese-English unrestricted

<table>
<thead>
<tr>
<th>Method</th>
<th>Score</th>
<th>Rank (of 9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>fluency</td>
<td>2.93</td>
<td>6</td>
</tr>
<tr>
<td>adequacy</td>
<td>3.25</td>
<td>3</td>
</tr>
<tr>
<td>BLEU</td>
<td>0.27</td>
<td>5</td>
</tr>
<tr>
<td>GTM</td>
<td>0.66</td>
<td>4</td>
</tr>
<tr>
<td>NIST</td>
<td>7.50</td>
<td>2</td>
</tr>
<tr>
<td>PER</td>
<td>0.42</td>
<td>3</td>
</tr>
</tbody>
</table>
Conclusion

• The EDTRL System has a better performances than simple cascaded multiple MT systems.

• The use of formulized, enriched English as Interlingua can reduce the Parallel Data Sparseness Problem form many languages pairs

• Results from IWSLT 2004 evaluation campaign lie behind the best systems