Using wordnet-based word sense disambiguation to improve MT performance

Nora Kumpikova

University of Tübingen

July 3, 2018
Overview

1 Introduction
- SMT approach
- RBMT approach

2 Experiments
- Description
- Setup
- Processing
- WSD
- Evaluation
- Discussion
- Conclusion
MT errors

→ ambiguity proves difficult for MT:
  - lexical items refer to more than a single concept
  - MT systems forced to choose between several translation equivalents (different senses)
  - wrong choices - result in grave errors

*striking* → beautiful, surprising, delivering a hard blow or indicating a certain time

*course* → something we give, take, teach, eat
WSD: Traditional approach

In SMT: lexical choice governed by target language model

**Problems:**

1. distant dependencies not taken into account
2. context used limited to several words

→ significant improvements when using context-dependent/dynamic lexicons
WSD: Traditional approach

In RBMT: rely on hand-crafted rules, semantic lexica or probabilistic models from corpora

**Problems:**

1. ambiguous words - number of equivalents/unrelated meanings, however similar syntactic behavior
2. sentences difficult to disambiguate without broader context

*Let us wait for the next course*
Experiments performed on English ←→ Slovene language pair
Slovene - less resourced language

**Goal:**
Assess performance of three MT systems. Check if wordnet-based WSD can improve performance and assist in avoiding translation errors.

**Statistical systems:**
1. Google
2. Bing

**Rule Based system:**
1. Presis
Word Sense Disambiguation

Using UKB (Universal Knowledge Base, http://ixa2.si.ehu.es/ukb/), providing a graph-based algorithm and Wordnet(sloWNet) to compute probability of each sense of polysemous word by considering the senses of context words.

Issues:
→ Wordnet sense too fine-grained for NLP tasks

Solution:
Word Sense Disambiguation

Using UKB (Universal Knowledge Base, http://ixa2.si.ehu.es/ukb/), providing a graph-based algorithm and Wordnet(sloWNet) to compute probability of each sense of polysemous word by considering the senses of context words.

Issues:
→ Wordnet sense too fine-grained for NLP tasks

Solution:
• using full sense inventory
Word Sense Disambiguation

Using UKB (Universal Knowledge Base, http://ixa2.si.ehu.es/ukb/), providing a graph-based algorithm and Wordnet(sloWNet) to compute probability of each sense of polysemous word by considering the senses of context words.

Issues:
→ Wordnet sense too fine-grained for NLP tasks

Solution:
- using full sense inventory
- using automatically induced coarser-grained sense clusters
Evaluation

1. WSD performance evaluated manually from MT perspective, both settings compared

2. Analyze agreement between each MT system and UKB/wordnet-derived translation

3. Generate raw translation and compare BLEU, NIST, METEOR scores achieved with each translation version
Corpus: "1984" novel by George Orwell, published in 1949 (English) → first translated and published in Slovene in 1967

Why literary language?

1. richer in ambiguity
2. provides more complex semantic space
MT systems

→ novel translated into Slovene by the 3 MT systems (Google Translate, Bing, Presis)
MT systems

→ novel translated into Slovene by the 3 MT systems (Google Translate, Bing, Presis)

- Google Translate and Bing:
  - lexical choice governed by combination of target language model
  - target language model contains probabilities of phrases
  - translation model - proposes various equivalents

- Presis:
  - choice of words in given context - governed by verb templates
  - verb templates: encode all possible combinations of verbs and objects together with their semantic properties (animate/inanimate)
  - semantic lexicon contains common patterns of adjective+nouns, prepositional phrases
  - in case of coordination - resolve ambiguity by searching for common hypernym
MT systems

→ novel translated into Slovene by the 3 MT systems (Google Translate, Bing, Presis)

- Google Translate and Bing:
  - lexical choice governed by combination of target language model
→ novel translated into Slovene by the 3 MT systems (Google Translate, Bing, Presis)

- Google Translate and Bing:
  1. lexical choice governed by combination of target language model
  2. target language model contains probabilities of phrases

- Presis:
  1. choice of words in given context - governed by verb templates
  2. verb templates: encode all possible combinations of verbs and objects together with their semantic properties (animate/inanimate)
  3. semantic lexicon contains common patterns of adjective+nouns, prepositional phrases
  4. in case of coordination - resolve ambiguity by searching for common hypernym
→ novel translated into Slovene by the 3 MT systems (Google Translate, Bing, Presis)

- Google Translate and Bing:
  1. lexical choice governed by combination of target language model
  2. target language model contains probabilities of phrases
  3. translation model - proposes various equivalents

- Presis:
  1. choice of words in given context - governed by verb templates
  2. verb templates: encode all possible combinations of verbs and objects together with their semantic properties (animate/inanimate)
  3. semantic lexicon contains common patterns of adjective+nouns, prepositional phrases
  4. in case of coordination - resolve ambiguity by searching for common hypernym
→ novel translated into Slovene by the 3 MT systems (Google Translate, Bing, Presis)

- Google Translate and Bing:
  1. lexical choice governed by combination of target language model
  2. target language model contains probabilities of phrases
  3. translation model - proposes various equivalents

- Presis:
  1. choice of words in given context - governed by verb templates
  2. verb templates: encode all possible combinations of verbs and objects together with their semantic properties (animate/inanimate)
  3. semantic lexicon contains common patterns of adjective + nouns, prepositional phrases
  4. in case of coordination - resolve ambiguity by searching for common hypernym
→ novel translated into Slovene by the 3 MT systems (Google Translate, Bing, Presis)

- **Google Translate and Bing:**
  1. lexical choice governed by combination of target language model
  2. target language model contains probabilities of phrases
  3. translation model - proposes various equivalents

- **Presis:**
  1. choice of words in given context - governed by verb templates
→ novel translated into Slovene by the 3 MT systems (Google Translate, Bing, Presis)

- **Google Translate and Bing:**
  1. lexical choice governed by combination of target language model
  2. target language model contains probabilities of phrases
  3. translation model - proposes various equivalents

- **Presis:**
  1. choice of words in given context - governed by verb templates
  2. verb templates: encode all possible combinations of verbs and objects together with their semantic properties (animate/inanimate)
→ novel translated into Slovene by the 3 MT systems (Google Translate, Bing, Presis)

- **Google Translate and Bing:**
  1. lexical choice governed by combination of target language model
  2. target language model contains probabilities of phrases
  3. translation model - proposes various equivalents

- **Presis:**
  1. choice of words in given context - governed by verb templates
  2. verb templates: encode all possible combinations of verbs and objects together with their semantic properties (animate/inanimate)
  3. semantic lexicon contains common patterns of adjective+nouns, prepositional phrases
MT systems

→ novel translated into Slovene by the 3 MT systems (Google Translate, Bing, Presis)

- Google Translate and Bing:
  1. lexical choice governed by combination of target language model
  2. target language model contains probabilities of phrases
  3. translation model - proposes various equivalents

- Presis:
  1. choice of words in given context - governed by verb templates
  2. verb templates: encode all possible combinations of verbs and objects together with their semantic properties (animate/inanimate)
  3. semantic lexicon contains common patterns of adjective+nouns, prepositional phrases
  4. in case of coordination - resolve ambiguity by searching for common hypernym
Experiment

**English:**
It was a bright cold day in April and the clocks were striking thirteen.

**English-preprocessed:**
be bright cold day April clock be strike

**Slovene-reference:**
Bil je jasan, mrzel aprilski dan in ure so bile trinajst.

**Slovene-reference-preprocessed:**
biti biti jasen mrzel aprilski dan ura biti biti

**Slovene-Google:**
Bilo je svetlo mrzel dan v aprilu, in ure so bile trinajst presenetljiv.

**Slovene-Google-preprocessed:**
biti biti svetlo mrzel dan april ura biti biti presenetljiv

**Slovene-Bing:**
Je bil svetlo hladne dan aprila in v ure so bili presenetljivo trinajst.

**Slovene-Bing-preprocessed:**
biti biti svetlo hladen dan april ura biti biti presenetljivo

**Slovene-Presis:**
Svetel hladen dan v aprilu je bilin so ure udarjale trinajst.

**Slovene-Presis-preprocessed:**
svetel hladen dan april biti bilin biti ura udarjati
Analysis

→ according to Wordnet - almost half of all tokens are ambiguous
→ Slovene net smaller than English net ⇒ almost half the words have no equivalent in sloWNet

→ restrict the research:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus size in tokens</td>
<td>103,769</td>
</tr>
<tr>
<td>Corpus size in types</td>
<td>10,982</td>
</tr>
<tr>
<td>Ambiguous tokens</td>
<td>48,632</td>
</tr>
<tr>
<td>Ambiguous types</td>
<td>7,627</td>
</tr>
<tr>
<td>Synsets with no equivalent in sloWNet</td>
<td>3,192</td>
</tr>
<tr>
<td>Contexts with no equivalent in sloWNet</td>
<td>8,073</td>
</tr>
<tr>
<td>Contexts with no cluster assignment</td>
<td>25,810</td>
</tr>
</tbody>
</table>

Table 1. Corpus size and number of ambiguous words
Word Sense Disambiguation: Process

1. disambiguating English corpus using UKB
1. disambiguating English corpus using UKB
2. compute probability of each graph based on the number and weight of edges between nodes representing semantic concepts
Word Sense Disambiguation: Process

1. disambiguating English corpus using UKB
2. compute probability of each graph based on the number and weight of edges between nodes representing semantic concepts
3. disambiguation performed on monolingual context for all words included in English wordnet
Disambiguation for word **face**, 13 possible senses in wordnet:

*WSD: ctx_Oen.1.1.2 24  !! face*

- **W**: 0.173463  **ID**: eng-30-05600637-n  **ENGWN**: face, human face, (the front of the human head from the forehead to the chin and ear to ear)  **SLOWN**: fris, obraz, faca, človeški obraz, (EMPTYDEF)
- **W**: 0.116604  **ID**: eng-30-08510666-n  **ENGWN**: side, face, (a surface forming part of the outside of an object)  **SLOWN**: stranica, ploskev, (EMPTYDEF)
- **W**: 0.0956895  **ID**: eng-30-03313602-n  **ENGWN**: face, (the side upon which the use of a thing depends (usually the most prominent surface of an object))  **SLOWN**: sprednja stran, prava stran, zgornja stran, lice, (EMPTYDEF)
- **W**: 0.0761554  **ID**: eng-30-04679738-n  **ENGWN**: expression, look, aspect, facial expression, face, (the feelings expressed on a person's face)  **SLOWN**: izraz, pogled, obraz, izraz na obrazu, (EMPTYDEF)
- **W**: 0.0709513  **ID**: eng-30-03313456-n  **ENGWN**: face, (a vertical surface of a building or cliff)  **SLOWN**: stena, fasada, (EMPTYDEF)
- **W**: 0.0653514  **ID**: eng-30-06825399-n  **ENGWN**: font, fount, typeface, face, case, (a specific size and style of type within a type family)  **SLOWN**: font, pisava, črkovna družina, vrsta črk, črkovna podoba, črkovni slog, (EMPTYDEF)
- **W**: 0.0629878  **ID**: eng-30-04838210-n  **ENGWN**: boldness, nerve, brass, face, cheek, (impudent aggressiveness)  **SLOWN**: predrznost, nesramnost, (EMPTYDEF)
WSD using sense clusters

Approach:
Each of the possible senses of an ambiguous lexical item is assigned a cluster with the aim of grouping similar meanings into the same cluster.

Example:
→ sentence containing ambiguous word *place*:

*People were leaping up and down in their places and shouting at the tops of their voices in an effort to drown the maddening bleating voice that came from the screen.*

→ UKB selects definition: a *point located with respect to surface features of some region*, literals: topographic point, place, spot
→ original 4 Slovene *kraj, mesto, prostor, točka*, adding 10 equivalents to the cluster: *dom, bivališ če, posest, stanovanje, domovanje, sedež, trg, posestvo, položaj, sediš če*
Manual evaluation

- UKB disambiguation tool performance in context of MT:
  Precision reported previously: unsupervised WSD 58% for all words, 72% for all nouns.

Precision evaluated manually:

<table>
<thead>
<tr>
<th></th>
<th>Correct</th>
<th>Incorrect</th>
<th>Borderline</th>
</tr>
</thead>
<tbody>
<tr>
<td>No sense clusters</td>
<td>345</td>
<td>126</td>
<td>29 (6%)</td>
</tr>
<tr>
<td></td>
<td>(69%)</td>
<td>(25%)</td>
<td></td>
</tr>
<tr>
<td>With sense clusters</td>
<td>420</td>
<td>55</td>
<td>35 (7%)</td>
</tr>
<tr>
<td></td>
<td>(82%)</td>
<td>(11%)</td>
<td></td>
</tr>
</tbody>
</table>

→ Auxiliary verbs and copula(be, have, do, can) are not disambiguated when occurring in purely functional role. Those cases are therefore disregarded during evaluation.
Comparison between WSD/wordnet-based equivalent and translations proposed by Presis, Google, Bing and reference translation, using and not using sense clusters

<table>
<thead>
<tr>
<th></th>
<th>No sense clusters</th>
<th>Using sense clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total no. of disambiguated tokens</td>
<td>40,560</td>
<td>42,157</td>
</tr>
<tr>
<td>Synsets with no sloWNet equivalent</td>
<td>8,073</td>
<td>6,476</td>
</tr>
<tr>
<td>WSD = reference</td>
<td>18,544</td>
<td>20,277</td>
</tr>
<tr>
<td>WSD = Presis</td>
<td>19,858</td>
<td>21,866</td>
</tr>
<tr>
<td>WSD = Google</td>
<td>20,522</td>
<td>22,471</td>
</tr>
<tr>
<td>WSD = Bing</td>
<td>20,112</td>
<td>21,963</td>
</tr>
<tr>
<td>WSD = ref = Presis = Google = Bing</td>
<td>12,815</td>
<td>14,126</td>
</tr>
<tr>
<td>WSD = ref ≠ Presis ≠ Google ≠ Bing</td>
<td>1,041</td>
<td>1,061</td>
</tr>
</tbody>
</table>
Evaluation with metrics

1. Generate a translation version consisting of content words in lemmatized form.
2. Translate disambiguated words with WordNet.
3. Proper names, hyphenated compounds and neologisms - translated using dictionaries/bilingual lexicon from Orwell corpus.
4. Pre-process the 5 version of the corpus (original, reference translation, Presis, Google and Bing):
   **Parameters:** lemmatization, removal of function words, sentence removal when alignment not 1:1, sentence removal where lexical items have no equivalent in sloWNet.
Evaluation with metrics

Evaluating the performance of WSD in the context of Machine Translation through metrics for automatic evaluation:

<table>
<thead>
<tr>
<th></th>
<th>BLEU (n=1)</th>
<th>NIST</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bing</td>
<td>0.506</td>
<td>3.594</td>
<td>0.455</td>
</tr>
<tr>
<td>Google</td>
<td>0.579</td>
<td>4.230</td>
<td>0.481</td>
</tr>
<tr>
<td>Presis</td>
<td>0.485</td>
<td>3.333</td>
<td>0.453</td>
</tr>
<tr>
<td>WSD</td>
<td>0.440</td>
<td>3.258</td>
<td>0.429</td>
</tr>
<tr>
<td>WSD-amend</td>
<td>0.410</td>
<td>3.308</td>
<td>0.430</td>
</tr>
<tr>
<td>WSD-dict</td>
<td>0.405</td>
<td>3.250</td>
<td>0.427</td>
</tr>
<tr>
<td>WSD-align</td>
<td>0.448</td>
<td>3.588</td>
<td>0.434</td>
</tr>
<tr>
<td>WSD-wikt</td>
<td>0.442</td>
<td>3.326</td>
<td>0.429</td>
</tr>
</tbody>
</table>

Corpus: 2,428 segments
using unsupervised WSD, comparing wordnet-based translation equivalents to ones proposed by MT systems - no significant improvement

using hyponyms and coarse-grained distinctions - higher potential

impact of WSD on fiction literature rather than domain specific texts

using sense clusters - significant improvement
Conclusion

1. Manual evaluation - correct equivalent proposed 69% of the cases
2. Sense clusters - 82% precision
3. Comparison with three MT systems:
   - High correspondence of WSD/wordnet-based equivalent with reference translation
   - WSD chooses correct equivalent where MT systems disagree

Future work:
- Redesign experiment to directly use WSD as a post-processing step instead of generating separate version
- Improve WSD precision by combining 2 different algorithms and only use cases when both agree
- Experiment with different text types and context lengths
Conclusion

1. Manual evaluation - correct equivalent proposed 69% of the cases
2. Sense clusters - 82% precision
3. Comparison with three MT systems:
   ▶ High correspondence of WSD/wordnet-based equivalent with reference translation
   ▶ WSD chooses correct equivalent where MT systems disagree
4. **Future work:** redesign experiment to directly use WSD as a post-processing step instead of generating separate version
Conclusion

1. Manual evaluation - correct equivalent proposed 69% of the cases
2. Sense clusters - 82% precision
3. Comparison with three MT systems:
   - High correspondence of WSD/wordnet-based equivalent with reference translation
   - WSD chooses correct equivalent where MT systems disagree
4. **Future work:** redesign experiment to directly use WSD as a post-processing step instead of generating separate version
5. **Future work:** improve WSD precision by combining 2 different algorithms and only use cases when both agree
Conclusion

1. Manual evaluation - correct equivalent proposed 69% of the cases
2. Sense clusters - 82% precision
3. Comparison with three MT systems:
   - High correspondence of WSD/wordnet-based equivalent with reference translation
   - WSD chooses correct equivalent where MT systems disagree
4. **Future work:** redesign experiment to directly use WSD as a post-processing step instead of generating separate version
5. **Future work:** improve WSD precision by combining 2 different algorithms and only use cases when both agree
6. **Future work:** experiment with different text types and context lengths