Towards integrating (Sense, Structure and Entity) Disambiguation into Statistical Machine Translation

Presented by:
Sudha (114050002)
Abhijit (114056002)
Rahul (11305R013)

Under the guidance of :
Prof. Pushpak Bhattacharyya
Ambiguities and Translation

- Ambiguities: Inherent property of a language.

- Kinds of ambiguity
  - Sense ambiguity
    (Arises due to polysemy)
  - Structural ambiguity
    (Arises due to different possible structures)
  - NE ambiguity
    (Whether a word is a Named entity)

She could not **bear** children.
1. Tom and Susan or John will go.
2. Mary likes Tom more than Susan.

पूजा के लिए फूल
Examples

S1: Bank of the river is sandy.
   Google: नदी के तट रेतीला है
   <nadl ke taT retIia hai>

S2: The bank of the West Tapti is dry.
   Google: पश्चिमी ताप्ती की बैंक शुष्क है.
   <pashchimi taApTI ki beIN k shushka hai>

S3: आकाश आकाश में पतंग उड़ा रहा था
   <AkAsh AkAsh meIM pataMg udA rha thA>
   Google: The sky was flying kite in the sky
More Examples

S4: I washed the shirt with soap.
Google: मैं साबुन के साथ शर्ट धोया
<meiM sAbun ke sAth shaRt dhoyA>

S5: I washed the shirt with green buttons.
Google: मैं हरे रंग की बटन के साथ शर्ट धोया.
<meiM hare raMg kl batan ke sAth shaRt dhoyA>

- Tourism Domain:
  Party: उत्सव <utshav>
- Judicial Domain:
  Party: पक्ष <paksha>
Motivation

- Simple Phrase Based/Hierarchical SMT systems make little use of context/syntactic properties to resolve cases of ambiguity due to the complexity involved in implementing this type of strategy.

- **Accuracy-efficiency tradeoff**
  - Introduction of disambiguation modules are expensive in terms of translation time
  - Significant increment of translation accuracy in terms of fluency and adequacy is desirable.

- **Challenge**: Finding appropriate places for these modules in the translation pipeline.
Sense Disambiguation and SMT
State of the art

• Various researchers tried integrating WSD with SMT
• They focus on single word rather than full phrases
• Phrasal SMT + Single-word WSD leads to unpredictable and inconsistent effects on translation quality
• Dekai Wu and Carpuat (2007) introduced the concept of phrase sense disambiguation which generalizes word sense disambiguation to multi-word targets
• Phrasal SMT + Multi-word PSD reliably and consistently improves translation quality
## PSD vs. WSD

<table>
<thead>
<tr>
<th>Phrase Sense Disambiguation (PSD)</th>
<th>Word Sense Disambiguation (WSD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Targets are phrases</td>
<td>Targets are single words</td>
</tr>
<tr>
<td>Phrases are defined as any sequence of words up to a given length</td>
<td></td>
</tr>
<tr>
<td>Target phrase need not necessarily be syntactically well formed, needs only collocations defined by their surface form</td>
<td>The grammatical category of the target word is known and eventually helps in finding the sense of the word</td>
</tr>
<tr>
<td>PSD classifiers are trained for every multi-word phrase</td>
<td>WSD classifiers are trained for every word</td>
</tr>
<tr>
<td>Sense candidates are defined by the SMT translation lexicon</td>
<td>Sense candidates are defined by sense inventories</td>
</tr>
</tbody>
</table>
Building multi-word PSD models for SMT

• There is no concept of sense IDs since we are not dealing with sense repositories.
• Instead, we use the target phrases as sense indicators.
  Example:

  1: Piece of cake: सरल काम <saral kAm> <Simple Task> (Sense 1)
  2: Piece of cake: केक का टुकड़ा <kek kA tukdA> <piece of cake> (Sense 2)

• Training is done using the same parallel corpora used for SMT.
• Integrating PSD in phrase-based SMT architectures
Integration of multi-word PSD into SMT

Parallel Corpora

Training (Phrase table generation)

Phrase Sense Disambiguation model

Decoding using PSD

Translation output

Language model
PSD helps SMT

- The rich context features help rank the correct translation first with PSD while it is ranked lower according to baseline translation probability scores.

- Even when PSD and baseline translation probabilities agree on the top translation candidate, the stronger PSD scores help override wrong language model predictions.

- The strong PSD scores for phrases help the decoder pick longer phrase translations, while using baseline translation probabilities often translate those phrases in smaller chunks that include a frequent (and incorrect) translation candidate.
Insight

• Consider a Phrase table:
  Piece of cake : सरल काम ( Prob = 0.7 )
  Piece of cake : केक का टुकड़ा (Prob = 0.3 )

For a test sentence
I ate a piece of cake for breakfast.

The presence of contextual information related to eating (e.g. “ate” and “breakfast”) suggests that केक का टुकड़ा is the appropriate translation even of the translation probability is low.
Comparative evaluation results

• Dataset used:
  • multilingual BTEC corpus (travel domain)
    – Training: 40000 sentences
    – Testing: 500 sentences in each test set
  • NIST test set
• Baseline: Phrase-based decoder Pharaoh (Koehn, 2004) trained on IWSLT training set
## Evaluation on IWSLT-06 test sets

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Experiment</th>
<th>BLEU</th>
<th>NIST</th>
<th>METEOR</th>
<th>METEOR (no syn)</th>
<th>TER</th>
<th>WER</th>
<th>PER</th>
<th>CDER</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>Baseline</td>
<td>42.21</td>
<td>7.888</td>
<td>65.40</td>
<td>63.24</td>
<td>40.45</td>
<td>45.58</td>
<td>37.80</td>
<td>40.09</td>
</tr>
<tr>
<td></td>
<td>+WSD (all words)</td>
<td>41.94</td>
<td>7.911</td>
<td>65.55</td>
<td>63.52</td>
<td>40.59</td>
<td>45.61</td>
<td>37.75</td>
<td>40.09</td>
</tr>
<tr>
<td></td>
<td>+PSD (all phrases)</td>
<td>42.38</td>
<td>7.902</td>
<td>65.73</td>
<td>63.64</td>
<td>39.98</td>
<td>45.30</td>
<td>37.60</td>
<td>39.91</td>
</tr>
<tr>
<td>#2</td>
<td>Baseline</td>
<td>41.49</td>
<td>8.167</td>
<td>66.25</td>
<td>63.85</td>
<td>40.95</td>
<td>46.42</td>
<td>37.52</td>
<td>40.35</td>
</tr>
<tr>
<td></td>
<td>+WSD (all words)</td>
<td>41.31</td>
<td>8.161</td>
<td>66.23</td>
<td>63.72</td>
<td>41.34</td>
<td>46.82</td>
<td>37.98</td>
<td>40.69</td>
</tr>
<tr>
<td></td>
<td>+PSD (all phrases)</td>
<td>41.97</td>
<td>8.244</td>
<td>66.35</td>
<td>63.86</td>
<td>40.63</td>
<td>46.14</td>
<td>37.25</td>
<td>40.10</td>
</tr>
<tr>
<td>#3</td>
<td>Baseline</td>
<td>49.91</td>
<td>9.016</td>
<td>73.36</td>
<td>70.70</td>
<td>35.60</td>
<td>40.60</td>
<td>32.30</td>
<td>35.46</td>
</tr>
<tr>
<td></td>
<td>+WSD (all words)</td>
<td>49.73</td>
<td>9.017</td>
<td>73.32</td>
<td>70.82</td>
<td>35.72</td>
<td>40.61</td>
<td>32.10</td>
<td>35.30</td>
</tr>
<tr>
<td></td>
<td>+PSD (all phrases)</td>
<td>51.05</td>
<td>9.142</td>
<td>74.13</td>
<td>71.44</td>
<td>34.68</td>
<td>39.75</td>
<td>31.71</td>
<td>34.58</td>
</tr>
</tbody>
</table>
Evaluation on NIST test sets

<table>
<thead>
<tr>
<th>Experiment</th>
<th>BLEU</th>
<th>NIST</th>
<th>METEOR</th>
<th>METEOR (no syn)</th>
<th>TER</th>
<th>WER</th>
<th>PER</th>
<th>CDER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>20.20</td>
<td>7.198</td>
<td>59.45</td>
<td>56.05</td>
<td>75.59</td>
<td>87.61</td>
<td>60.86</td>
<td>72.06</td>
</tr>
<tr>
<td>+PSD</td>
<td>20.62</td>
<td>7.538</td>
<td>59.99</td>
<td>56.38</td>
<td>72.53</td>
<td>85.09</td>
<td>58.62</td>
<td>68.54</td>
</tr>
</tbody>
</table>
Comparative evaluation results contd..

• Single-word WSD results are inconsistent across different test sets and depend on which evaluation metric is chosen

• PSD shows consistent improvement in translation quality across all metrics and on all test sets
Multi-word PSD helping translation

• Example 1:

Reference: I’d like to call Tokyo, Japan. What time is it now in Tokyo?

WSD: I want to make a call to Tokyo, Japan is Tokyo time now?

PSD: I want to make a call to Tokyo, Japan what time is it now in Tokyo?
Multi-word PSD helping translation

• Example 2:

Input: 请转乘中央线。
Reference: You should transfer to the Central Line.
WSD: Please turn to the Central Line.
PSD: Please transfer to Central Line.
Multi-word PSD helping translation

• Example 3:

**Input** 咖啡 和 红茶，您要哪个？

**Reference:** Which would you like, coffee or tea?

**WSD:** Which would you like, and coffee black tea?

**PSD:** Which would you like, black tea or coffee?
When WSD/PSD predictions go wrong?

Example 1:

**Input**: 我要送餐服务。

**WSD**: I will take meal service.

**PSD**: I want to eat service.

**Reference**: Room service, please.
When WSD/PSD predictions go wrong?

Example 2:

Input: 啊。给我帐单。
Reference: Uhh. Give me a Tab.
WSD: Oh. I have the bill.
PSD: Well, let me check.
Summary

- First attempt which fully integrates state-of-art WSD with phrase based SMT
- WSD targets are not only single words but multi-word phrases
- Single-word WSD+ SMT leads to unpredictable effects on the translation quality
- PSD + SMT consistently improves translation quality on all test sets for all evaluation metrics.
- Therefore, it tends to confirm that generalization from WSD to PSD is indeed necessary for SMT
Named Entity and SMT
What's in a name? That which we call a rose
By any other name would smell as sweet.

By: William Shakespeare
Romeo and Juliet
Named Entity

- Entities which have rigid designators
- Rigid designator includes proper names like Ford Motor company
- Also includes biological species and substances like Pheidole barbata (ants)
- Includes rigid temporal expressions like year 2001 (some relaxation in definition of rigid designators)
NER for MT

- Can we achieve improvement in translation quality using only phrase tables?
- Does NER improve performance of MT system?
Named Entities and phrase table

- Phrase table associates the phrases between two languages.
- Phrase table contains only words which it has seen during training.
- Vast majority of named entities are very rare and never seen during training.
- Just using phrase table for translation is not enough.
Named Entities and translation

- NE translation involves identifying NE boundaries and their type

- Different type of NE require different translation strategies

- Usually named entities are classified into PERSON, ORGANISATION, LOCATION
NE identification and translation

- Identifying boundaries of NE is important to isolate for different translation strategy

- PERSON and LOCATION are usually transliteration of each other across languages

- Certain components of ORGANISATION name might need translation.
Example

- **Eg. 1**
  - Hindi - “मंदिर की शांति कहा है?”
  - English - Where is *Shanti* of temple?
  - English - Where is temple’s peace?

- **Eg. 2**
  - Hindi - “*Indian railways* is the biggest railway network
  - English - “भारतीय रेलवे दुनिया का सबसे बड़ा रेल नेटवर्क है”

- **Eg. 3**
  - Hindi - “*पुणे* जाने वाली गाड़ी १० बजे निकलेगी “
  - English - Bus for *Pune* will leave at 10.
Integration of NE in MT

- Several possibilities exist to use NE information in SMT.
  - Maintain DNT translate list
  - Adding named entities to help the aligner
  - Modifying language model
  - Modifying the ranking function using NE information
Do Not Translate

- DNT (Do not translate)
- Most of the existing translation system provide the feature of accepting a do not translate list
- If NEs are provided to the system as DNT then it has been shown that morpho-syntactic and lexical well-formedness improve. (Babych et. al, 2003)
Improving aligner

- Add a list of NEs to bi-text training data.

- This will increase the probability of matching these NEs from source language to their counterpart in target language.

Results

<table>
<thead>
<tr>
<th></th>
<th>Dev Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Case</td>
<td>31.968</td>
<td>31.349</td>
</tr>
<tr>
<td>With NER Data</td>
<td>32.276</td>
<td>30.427</td>
</tr>
</tbody>
</table>
Modifying language model

1. “रिचर्ड टहलने के लिए जा रहा है”
2. “राहुल टहलने के लिए जा रहा है”

If we have Hindi monolingual corpus, finding Richard would be very rare but Rahul would be abundant.

So LM score of 1 will be lower than 2
Modifying language model Contd.

- **Solution**
  - Change named entity in a sentence to their categories
    - raj went for fishing – PERSON went for fishing
  - Learn language model on modified monolingual data

- **Results**

<table>
<thead>
<tr>
<th></th>
<th>Dev Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Case</td>
<td>31.968</td>
<td>31.349</td>
</tr>
<tr>
<td>With class based LM</td>
<td>26.935</td>
<td>25.532</td>
</tr>
</tbody>
</table>

Due to large occurrences of NE tags in training, language model is distorted
Adding Extra Feature

- Correct translation of a sentence should have equal number of NEs
- Define extra feature function in Decoding step

Feature Value = \exp(\text{abs}(\text{Difference in number of proper nouns in source and target}))

- Eg.
  - Hindi - चीनी प्रधानमंत्री
  - Base case - Chinese prime minister
  - EF - Chinese premier

Results

<table>
<thead>
<tr>
<th></th>
<th>Dev Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Case</td>
<td>31.968</td>
<td>31.349</td>
</tr>
<tr>
<td>With extra feature</td>
<td>32.155</td>
<td>30.914</td>
</tr>
</tbody>
</table>
Summary

- Current studies show that adding NE to MT system does not always result in better BLUE score.

- But the tangibility of the sentence increases.

- There is considerable reduction in dropping and mistranslation of NEs.
Structure Disambiguation and SMT
A sentence having two different underlying structures is structurally ambiguous.

Example:
- The boy saw the lady with glasses. (PP attachment)
- Old men and women escaped. (Scope)

Sentences may not be ambiguous for humans but computers may not resolve the ambiguities properly.
Phrase Based and Hierarchical SMT may not always take care of structure disambiguation.

The quality of output in Hierarchical SMT depends on parse trees at source side.

Example:
मैं दौड़ते हुए पतंग को देखा
Moses: I saw the running kite
Joshua: I saw the kite running
Reference: While running, I saw the kite.

Not much work on integrating Structural Disambiguation modules in SMT systems.

The KANT Pipeline

- Source Sentence
- Extraction Heuristics
- PP-attachment tuples
  - Rule based disambiguation
  - Correct attachment information
- KBMT System
- Target Sentence
The KANT Pipeline

- Morphological analysis is performed, and the set of possible lexical entries for each input token is retrieved;

- A unification grammar is used to produce the legal set of grammatical functional structures for the input tokens.

- PP attachment tuples are extracted and disambiguated using a rule based technique.

- When the text is eventually translated, the information stored in the processing instruction is used to automatically select the desired prepositional attachment.
PP attachment Disambiguation

- The disambiguation module encodes attachment preferences in the form of tuples
  \[\langle \text{Verb}, \text{N}1, \text{PP}, \text{N}2 \rangle \langle \text{attachment} \rangle\]

  Example:
  \[\langle \text{Use, chain, for, pulling} \rangle \langle \text{Verb} \rangle\]

- For a new tuple extracted from the test sentence, the attachment is decided by checking the closest relationship of the tuple with those present in the Rule Base.

  Example:
  \[\langle \text{employ, machines, for, dragging} \rangle\]
  Is closer to \[\langle \text{use, chain, for, pulling} \rangle \langle \text{Verb} \rangle\]
  Than \[\langle \text{introduce, plan, for, advertisers} \rangle \langle \text{Noun} \rangle\]
Experiments and Results

- Test corpus contains 12000 sentences with 11,607 instances of PP attachment ambiguity.

- The KANT system was able to disambiguate 8209 of these PPs.

- The overall accuracy of the translations from English to Spanish increased from 83% to 89% by introducing the PP disambiguation modules and pre-defined patterns.
Discussion

- We can use state-of-the-art PP disambiguation techniques such as corpus based methods to avoid sparse data problem (Nadh and Christian, 2009)
- In phrase based MT, attachment disambiguation steps can be carried out as a pre processing step and the attachment information can be used as for grammar correction at target side.
- In Hierarchical SMT, correction of source parse-trees can be carried as a preprocessing step of decoding.
Conclusion

- The accuracy of simple SMT systems may not go beyond a certain level without adding disambiguation modules.

- Integration of such modules in the SMT pipeline is desirable but should not affect the overall efficiency of the system.

- Future work includes introducing less resource intensive state-of-the-art modules in SMT pipeline.
References

Thank you