Integrating Encyclopedic Knowledge into Neural Language Models

Yang Zhang, Jan Niehues, Alex Waibel
Polysemy problem in translation task, e.g.

“The author does research for his column.”

“Der Autor betreibt Forschung für seine Säule” →✗

“Der Autor betreibt Recherche für seine Kolumne” ✓

Why does this happen?

High translation score

Language model fails to capture contextual meaning

Problem for low-resource languages
Approach

Motivation

- Context can be used to disambiguate
  - author → column/Kolumne
- Can we use other resources to disambiguate?
  - Encyclopedia
Approach

Idea

- Add topic-related, external information into neural language models
  - Long context
  - Add arbitrary information

Sources

- Wikipedia
- zdic.net

Word level: Wikipedia Categories

Sentence level: Topic modelling of encyclopaedia documents

⇒ Use this to improve rescoring of a SMT system
Contents

- Motivation
- Integration
- Model
  - Word-level integration
  - Sentence-level integration
- Evaluation
- Conclusion
Rescoring

- Create **n-best lists** for dev/test set
- Retrieve information for encyclopedia
- Score these n-best lists with **new model**
- Add scores to previous n-best scores
- Find model weights (MERT, ListNet)
Neural Network Language Model

- LSTM-based LM
  - 2 layers
- Factored word representation
  - Input
  - Output
Approach 1 – Word level information

- Idea:
  - Annotate word with topic
  - Kolumne (Jornalism)
  - Säule (Building)

- Wikipedia:
  - category for each article
Approach 1 – Word level information
Approach 1 – Word level information

- Algorithm:
  - Find word in Wikipedia title
  - Determine Wikipedia category of page
  - Number of pages in category > N
    - Select category
  - Else
    - Use parent category
### Approach 1 – Word level information

- **Example: label nouns**

<table>
<thead>
<tr>
<th></th>
<th>The</th>
<th>author</th>
<th>does</th>
<th>research</th>
<th>for</th>
<th>his</th>
<th>column</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Der Autor</td>
<td>betreibt</td>
<td><strong>Forschung</strong></td>
<td>für</td>
<td>seine</td>
<td>Säule</td>
<td></td>
</tr>
<tr>
<td>ART</td>
<td>Person (Literatur)</td>
<td>VVFIN</td>
<td><strong>Wissenschafts-praxis</strong></td>
<td>ADJD</td>
<td>PPOSAT</td>
<td>Bauwesen nach Bauteil</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>Der Autor</td>
<td>betreibt</td>
<td><strong>Recherche</strong></td>
<td>für</td>
<td>seine</td>
<td>Kolumne</td>
<td></td>
</tr>
<tr>
<td>ART</td>
<td>Person (Literatur)</td>
<td>VVFIN</td>
<td><strong>Journalismus</strong></td>
<td>ADJD</td>
<td>PPOSAT</td>
<td>Journalismus</td>
<td></td>
</tr>
</tbody>
</table>
Approach 2 – Sentence level information

- Neural network can handle any input
  - Add topic information as additional input
- Model the topic of the sentence
  - Use topic model (TF-IDF/LDA/LSA/…) to represent sentence
- Use topic vector as additional input to NNLM
Approach 2 – Sentence level information

- Given translation W
- Find topic related web documents
- Represent documents in vector
  - Tf-Idf (10K dimension)
  - LSA, 300 dimension
  - LDA, 100 dimension
- Use vector in neural network
Approach 2 – sentence level information

\[ D_n = \{ d \in D \mid \text{score}(w, d) > c \} \]

\[ |D_n| = n \]

\[ f = \frac{\sum_{d \in D_n} h(d)}{|D_n|} \]
Approach 2 – sentence level information
Experiment Setup

- **Baseline**
  - EN-ZH
    - 2010 TED, UN;
    - 3000-best list;
    - 12 features
  - EN-RO
    - WMT 2015
    - 300-best list
    - 22-23 features;
Experiment Setup

- RNN Models
  - Voc (10K (ZH) / 5K (RO)); embedding (100), LSTMs (100, 200)
  - WikiCat classes: 3000-4000
  - Number of similar documents: 10
  - SGD NLLCriterion
## Evaluation – EN-ZH

<table>
<thead>
<tr>
<th>Model</th>
<th>Devdata[BLUE]</th>
<th>Testdata[BLEU]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>14.70</td>
<td>17.02</td>
</tr>
<tr>
<td>+ Word-based WikiCat</td>
<td>14.89 (+0.19)</td>
<td>17.63 (+0.61)</td>
</tr>
<tr>
<td>+ Word-based WikiCat + POS</td>
<td>14.75 (+0.05)</td>
<td>17.81 (+0.79)</td>
</tr>
<tr>
<td>+ Sentence Wiki (Tf-idf)</td>
<td>14.78 (+0.08)</td>
<td>17.68 (+0.66)</td>
</tr>
<tr>
<td>+ Sentence Wiki 2Conn</td>
<td>14.74 (+0.04)</td>
<td>17.81 (+0.79)</td>
</tr>
<tr>
<td>+ Sentence ZDICT</td>
<td>14.91 (+0.21)</td>
<td>17.58 (+0.56)</td>
</tr>
</tbody>
</table>
## Evaluation – EN-ZH

<table>
<thead>
<tr>
<th>Rank</th>
<th>Vect</th>
<th>Devdata [BLEU]</th>
<th>Testdata [BLEU]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td>14.70</td>
<td>17.02</td>
</tr>
<tr>
<td>TFIDF</td>
<td>TFIDF</td>
<td>14.78 (+0.08)</td>
<td>17.68 (+0.66)</td>
</tr>
<tr>
<td>LSA</td>
<td>TFIDF</td>
<td>14.78 (+0.08)</td>
<td>17.31 (+0.29)</td>
</tr>
<tr>
<td>LSA</td>
<td>LSA</td>
<td>14.83 (+0.13)</td>
<td>17.80 (+0.78)</td>
</tr>
<tr>
<td>LDA</td>
<td>TFIDF</td>
<td>14.79 (+0.09)</td>
<td>17.41 (+0.39)</td>
</tr>
<tr>
<td>LDA</td>
<td>LDA</td>
<td>14.79 (+0.09)</td>
<td>17.27 (+0.25)</td>
</tr>
</tbody>
</table>

**Motivation**

**Background**

**Word level**

**Sentence level**

**Evaluation**

**Conclusion**
## Evaluation – EN-RO – Single Score

<table>
<thead>
<tr>
<th>Input</th>
<th>Prediction</th>
<th>Single</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>Word</td>
<td>27.88</td>
</tr>
<tr>
<td>Word, POS, 2x clusters (4F)</td>
<td>Word, POS, 2x clusters (4F)</td>
<td>28.54</td>
</tr>
<tr>
<td>+ WikiCat (N)</td>
<td>+ WikiCat (N)</td>
<td>28.71 (+0.17)</td>
</tr>
<tr>
<td>+ WikiCat (All)</td>
<td>+ WikiCat (All)</td>
<td>28.84 (+0.30)</td>
</tr>
</tbody>
</table>
# Evaluation – EN-RO

## Word + Sentence

<table>
<thead>
<tr>
<th>Model</th>
<th>Conf1</th>
<th>Conf2</th>
<th>Conf3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>29.86</td>
<td>30.00</td>
<td>29.75</td>
</tr>
<tr>
<td>FRNNLM 4F</td>
<td>29.94</td>
<td>30.01</td>
<td>30.01</td>
</tr>
<tr>
<td>+Sentence</td>
<td>29.99 (+0.05)</td>
<td>30.19 (+0.18)</td>
<td>29.99 (-0.02)</td>
</tr>
<tr>
<td>+WikiCat (N)</td>
<td>29.90 (-0.04)</td>
<td>30.29 (+0.28)</td>
<td>30.23 (+0.22)</td>
</tr>
<tr>
<td>+WikiCat (All)</td>
<td>30.00 (+0.06)</td>
<td>30.20 (+0.19)</td>
<td>30.21 (+0.20)</td>
</tr>
</tbody>
</table>
Conclusion

- Integrate external information (from encyclopedia) into neural language models
  - use Wikipedia categories as word factors
  - Feature vector of topic-related articles for each sentence
- Improved EN-ZH by 0.79 BLEU, EN-RO by 0.2 BLEU
- Low-resource languages
- Future
  - Other web resources, topic models, network architecture
  - Use idea on other tasks, e.g. neural machine translation
THANK YOU!