Continuous Space Language Models for the IWSLT 2006 Task

Holger Schwenk
Marta R. Costa-jussà and José A. R. Fonollosa

LIMSI-CNRS, France
schwenk@limsi.fr

UPC, Spain
mruiz, adrian@gps.tsc.upc.edu

November, 27 2006
Plan

1. Context and motivation
2. Continuous space language model
3. Baseline SMT systems
4. Experimental evaluation on the IWSLT’06 tasks
5. Conclusion and perspectives
Introduction

Context of this work

- **BTEC task of IWSLT 2006**
- Statistical MT systems rely on representative resources
- Resources to train SMT systems are very limited (40k sentences bitexts, 320k words for LM)
  ⇒ Need for techniques to take better advantage of the available resources

Language modeling for SMT

- Most systems use $n$-gram word or class back-off LMs
- Language model adaptation [CMU, IWSLT’05]
- Factored LMs [Kirchoff, ACL wshop’05], syntax-based LMs [Charniak, MT Summit’03]
Introduction

Context of this work

- BTEC task of IWSLT 2006
- Statistical MT systems rely on representative resources
- Resources to train SMT systems are very limited (40k sentences bitexts, 320k words for LM)

⇒ Need for techniques to take better advantage of the available resources

Language modeling for SMT

- Most systems use $n$-gram word or class back-off LMs
- Language model adaptation [CMU, IWSLT’05]
- Factored LMs [Kirchoff, ACL wshop’05], syntax-based LMs [Charniak, MT Summit’03]
Continuous Space Language Models

Introduction

Theoretical Drawbacks of Back-off LM

- Words are represented in a high-dimensional discrete space
- Probability distributions are not smooth functions
- Any change of the word indices can result in an arbitrary change of LM probability

⇒ True generalization is difficult to obtain

New Approach [Bengio, NIPS’01]:

- Project word indices onto a continuous space and use a probability estimator operating on this space
- Probability functions are smooth functions and better generalization can be expected
Continuous Space Language Models
Introduction

Theoretical Drawbacks of Back-off LM

- Words are represented in a high-dimensional discrete space
- Probability distributions are not smooth functions
- Any change of the word indices can result in an arbitrary change of LM probability

⇒ True generalization is difficult to obtain

New Approach [Bengio, NIPS’01]:

- Project word indices onto a continuous space and use a probability estimator operating on this space
- Probability functions are smooth functions and better generalization can be expected
Continuous Space Language Models

Introduction

Application of Continuous Space Language Model
- Very successful in LVCSR
- Initial experiments with a word-based SMT system [Schwenk, ACL’06]

Cooperation with UPC
- First application of the CSLM to a state-of-the-art SMT system
- \( n \)-best list rescoring of UPC’s phrase and Ngram-based system
- All four languages are considered (translation of Mandarin, Japanese, Arabic and Italian to English)
## Continuous Space Language Models

### Introduction

<table>
<thead>
<tr>
<th>CSLM for IWSLT 2006</th>
<th>LIMSI-UPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td></td>
</tr>
<tr>
<td>Architecture of the CSLM</td>
<td></td>
</tr>
<tr>
<td>Baseline SMT systems</td>
<td></td>
</tr>
<tr>
<td>Evaluation</td>
<td></td>
</tr>
<tr>
<td>Dev data</td>
<td></td>
</tr>
<tr>
<td>Eval data</td>
<td></td>
</tr>
<tr>
<td>Conclusion</td>
<td></td>
</tr>
</tbody>
</table>

### Application of Continuous Space Language Model

- Very successful in LVCSR
- Initial experiments with a word-based SMT system [Schwenk, ACL’06]

### Cooperation with UPC

- First application of the CSLM to a state-of-the-art SMT system
- $n$-best list rescoring of UPC’s phrase and Ngram-based system
- All four languages are considered (translation of Mandarin, Japanese, Arabic and Italian to English)
Probability Calculation

- Outputs = LM posterior probabilities of all words: 
  \[ P(w_j = i | h_j) \quad \forall i \in [1, N] \]
- Context \( h_j \) = sequence of \( n-1 \) points in this space
- Word = point in the \( P \) dimensional space
- Projection onto a continuous space
- Inputs = indices of the \( n-1 \) previous words

\[ h_j = w_{j-n+1}, \ldots, w_{j-2}, w_{j-1} \]
Continuous Space Language Models
Architecture - Probability Calculation

Probability Calculation
- Outputs = LM posterior probabilities of all words: $P(w_j = i|h_j) \ \forall i \in [1, N]$
- Context $h_j = \text{sequence of } n-1 \text{ points in this space}$
- Word = point in the $P$ dimensional space
- Projection onto a continuous space
- Inputs = indices of the $n-1$ previous words

$h_j = w_{j-n+1}, \ldots, w_{j-2}, w_{j-1}$
Continuous Space Language Models
Architecture - Probability Calculation

Probability Calculation
- Outputs = LM posterior probabilities of all words: \( P(w_j = i | h_j) \ \forall i \in [1, N] \)
- Context \( h_j = \) sequence of \( n-1 \) points in this space
- Word = point in the \( P \) dimensional space
- Projection onto a continuous space
- Inputs = indices of the \( n-1 \) previous words

\[ h_j = w_{j-n+1}, \ldots, w_{j-2}, w_{j-1} \]
Continuous Space Language Models
Architecture - Probability Calculation

**Introduction**

**Architecture of the CSLM**

**Baseline SMT systems**

**Evaluation**

**Dev data**

**Eval data**

**Conclusion**

**Probability Calculation**

- Outputs = LM posterior probabilities of all words:
  \[ P(w_j = i|h_j) \quad \forall i \in [1, N] \]
- Context \( h_j = \text{sequence of } n-1 \text{ points in this space} \)
- Word = point in the \( P \) dimensional space
- Projection onto a continuous space
- Inputs = indices of the \( n-1 \) previous words

**Diagram:**

A neural network diagram illustrating the architecture of the CSLM. The network takes as input the indices of the \( n-1 \) previous words and outputs the probabilities of all words after the current one.

**Equation:**

\[
h_j = w_{j-n+1}, \ldots, w_{j-2}, w_{j-1}
\]
Continuous Space Language Models

Architecture - Training

- Backprop training, cross-entropy error
  \[ E = \sum_{i=1}^{N} d_i \log p_i \]
  + weight decay
  ⇒ NN minimizes perplexity on training data
- Continuous word codes are also learned (random initialization)
Continuous Space Language Models

Architecture - Training

**Training**

- Backprop training, cross-entropy error

\[
E = \sum_{i=1}^{N} d_i \log p_i + \text{weight decay}
\]

⇒ NN minimizes perplexity on training data

- Continuous word codes are also learned (random initialization)
Continuous Space Language Models
Architecture - Training

Training

- Backprop training, cross-entropy error
  \[ E = \sum_{i=1}^{N} d_i \log p_i \]

+ weight decay

⇒ NN minimizes perplexity on training data

- Continuous word codes are also learned (random initialization)
Continuous Space Language Models
Architecture - Practical Issues

Interpolation

- Back-off LM (modified Kneser-Ney smoothing, SRILM) and CSLM trained on 326k words,
- Both LM seem to be complementary
  → interpolated together
- Several neural networks are trained independently using different sizes of the continuous representation
- EM optimization of the interpolation coefficients: minimize perplexity on the Dev data (0.33 for LM)
- Replace the original LM scores with those of this interpolated LM
- Alternatively we could use several feature functions and tune the coefficients on the BLEU score
Baseline SMT systems

Incorporation into UPC’s SMT systems

- Use of UPC’s phrase-based and Ngram-based system
- Both systems were described in detail just before the break
- Slight difference with respect to official evaluation systems (most of them achieve better results)
- 1000-best list rescoring
  + re-optimization of feature function weights

Phrase-based system

- Standard phrase extraction algorithm
- Translation model probabilities in both directions are estimated using relative frequencies
Baseline SMT systems

Incorporation into UPC’s SMT systems

- Use of UPC’s phrase-based and Ngram-based system
- Both systems were described in detail just before the break
- Slight difference with respect to official evaluation systems (most of them achieve better results)
- 1000-best list rescoring
- + re-optimization of feature function weights

Phrase-based system

- Standard phrase extraction algorithm
- Translation model probabilities in both directions are estimated using relative frequencies
N-gram-based system

- Monotonic segmentation of each sentence pair
- Translation model probabilities are estimated as a bilingual LM

\[ p(e, f) = Pr(t^K_1) = \prod_{k=1}^{K} p(t_k | t_{k-2}, t_{k-1}) \]

- This translation model includes an implicit target language model

→ Is an improved target LM still helpful?
Baseline SMT systems

N-gram-based system

- Monotonic segmentation of each sentence pair
- Translation model probabilities are estimated as a bilingual LM

\[ p(e, f) = Pr(t^K_1) = \prod_{k=1}^{K} p(t_k \mid t_{k-2}, t_{k-1}) \]

- This translation model includes an implicit target language model

→ Is an improved target LM still helpful?
Baseline SMT systems

Additional Features

Log-linear combination of feature functions

\[ \tilde{e}_1 = \arg\max_{e_1} \left\{ \sum_{m=1}^{M} \lambda_m h_m(f'_1, e'_1) \right\} \] (1)

- Phrase translation probabilities 
or Ngram translation language model
- Word bonus model (and phrase bonus model)
- Source → target lexicon model (IBM1 probabilities)
- Target → source lexicon model (IBM1 probabilities)
- Target language model 
  (4-gram back-off or continuous space LM)
Experimental Evaluation

Data sets

BTEC Open data track

- Open data track of the 2006 IWSLT evaluation
- Only the supplied subset of the full BTEC corpus was used
- Results on the supplied Dev corpus of 489 sentences (<6k words) and the official test set (evaluation server)
- Scoring is case insensitive and punctuations are ignored
### BLEU scores

<table>
<thead>
<tr>
<th>Language</th>
<th>Phrase-based system</th>
<th></th>
<th>N-gram-based system</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Oracle</td>
<td>Ref.</td>
<td>CSLM</td>
<td>Oracle</td>
</tr>
<tr>
<td>Mand.</td>
<td>33.1</td>
<td>20.68</td>
<td>21.97</td>
<td>32.0</td>
</tr>
<tr>
<td>Japan.</td>
<td>26.9</td>
<td>17.29</td>
<td>18.27</td>
<td>28.6</td>
</tr>
<tr>
<td>Arabic</td>
<td>40.1</td>
<td>27.92</td>
<td>30.28</td>
<td>41.6</td>
</tr>
<tr>
<td>Italian</td>
<td>56.2</td>
<td>41.66</td>
<td>44.03</td>
<td>58.1</td>
</tr>
</tbody>
</table>

- Oracle scores calculated using cheating Dev-LM
- Improvements between 1 and 3 points BLEU
- Slightly better gains for Ngram-based systems
- Notable differences between the languages (also lower oracle BLEU scores)
### BLEU scores

<table>
<thead>
<tr>
<th></th>
<th>Phrase-based system</th>
<th>N-gram-based system</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Oracle</td>
<td>Ref.</td>
</tr>
<tr>
<td>Mand.</td>
<td>33.1</td>
<td>20.68</td>
</tr>
<tr>
<td>Japan.</td>
<td>26.9</td>
<td>17.29</td>
</tr>
<tr>
<td>Arabic</td>
<td>40.1</td>
<td>27.92</td>
</tr>
<tr>
<td>Italian</td>
<td>56.2</td>
<td>41.66</td>
</tr>
</tbody>
</table>

- Oracle scores calculated using cheating Dev-LM
- Improvements between 1 and 3 points BLEU
- Slightly better gains for Ngram-based systems
- Notable differences between the languages (also lower oracle BLEU scores)
## Experimental Evaluation

Results on Development Data (1)

### BLEU scores

<table>
<thead>
<tr>
<th>Language</th>
<th>Phrase-based system</th>
<th></th>
<th>NLG-based system</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Oracle</td>
<td>Ref.</td>
<td>CSLM</td>
<td>Oracle</td>
</tr>
<tr>
<td>Mand.</td>
<td>33.1</td>
<td>20.68</td>
<td>21.97</td>
<td>32.0</td>
</tr>
<tr>
<td>Japan.</td>
<td>26.9</td>
<td>17.29</td>
<td>18.27</td>
<td>28.6</td>
</tr>
<tr>
<td>Arabic</td>
<td>40.1</td>
<td>27.92</td>
<td>30.28</td>
<td>41.6</td>
</tr>
<tr>
<td>Italian</td>
<td>56.2</td>
<td>41.66</td>
<td>44.03</td>
<td>58.1</td>
</tr>
</tbody>
</table>

- Oracle scores calculated using cheating Dev-LM
- Improvements between 1 and 3 points BLEU
- Slightly better gains for Ngram-based systems
- Notable differences between the languages (also lower oracle BLEU scores)
## Experimental Evaluation

Results on Development Data (1)

<table>
<thead>
<tr>
<th>Language</th>
<th>Phrase-based system</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Oracle</td>
<td>Ref.</td>
<td>CSLM</td>
<td>Oracle</td>
<td>Ref.</td>
</tr>
<tr>
<td>Mand.</td>
<td>33.1</td>
<td>20.68</td>
<td>21.97</td>
<td>32.0</td>
<td>20.84</td>
</tr>
<tr>
<td>Japan.</td>
<td>26.9</td>
<td>17.29</td>
<td>18.27</td>
<td>28.6</td>
<td>18.34</td>
</tr>
<tr>
<td>Arabic</td>
<td>40.1</td>
<td>27.92</td>
<td>30.28</td>
<td>41.6</td>
<td>29.09</td>
</tr>
<tr>
<td>Italian</td>
<td>56.2</td>
<td>41.66</td>
<td>44.03</td>
<td>58.1</td>
<td>41.65</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N-gram-based system</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ref.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSLM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oracle</td>
<td>32.0</td>
<td>20.84</td>
<td>21.83</td>
<td></td>
</tr>
<tr>
<td>Ref.</td>
<td>28.6</td>
<td>18.34</td>
<td>19.77</td>
<td></td>
</tr>
<tr>
<td>CSLM</td>
<td>41.6</td>
<td>29.09</td>
<td>30.89</td>
<td></td>
</tr>
</tbody>
</table>

- Oracle scores calculated using cheating Dev-LM
- Improvements between 1 and 3 points BLEU
- Slightly better gains for Ngram-based systems
- Notable differences between the languages (also lower oracle BLEU scores)
Experimental Evaluation
Results on Development Data (2)

<table>
<thead>
<tr>
<th></th>
<th>Phrase-based</th>
<th></th>
<th></th>
<th>N-gram-based</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Oracle</td>
<td>Ref.</td>
<td>CSLM</td>
<td>Oracle</td>
<td>Ref.</td>
<td>CSLM</td>
</tr>
<tr>
<td>Ma/En</td>
<td>mWER</td>
<td>59.1</td>
<td>67.4</td>
<td>66.5</td>
<td>58.1</td>
<td>67.8</td>
</tr>
<tr>
<td></td>
<td>mPER</td>
<td>44.5</td>
<td>50.8</td>
<td>50.1</td>
<td>45.3</td>
<td>51.5</td>
</tr>
<tr>
<td>Ja/En</td>
<td>mWER</td>
<td>70.8</td>
<td>74.6</td>
<td>77.0</td>
<td>63.5</td>
<td>73.0</td>
</tr>
<tr>
<td></td>
<td>mPER</td>
<td>48.5</td>
<td>52.2</td>
<td>54.6</td>
<td>46.1</td>
<td>53.4</td>
</tr>
<tr>
<td>Ar/En</td>
<td>mWER</td>
<td>49.1</td>
<td>56.0</td>
<td>52.7</td>
<td>48.1</td>
<td>55.7</td>
</tr>
<tr>
<td></td>
<td>mPER</td>
<td>40.3</td>
<td>45.7</td>
<td>43.3</td>
<td>39.6</td>
<td>44.0</td>
</tr>
<tr>
<td>It/En</td>
<td>mWER</td>
<td>34.1</td>
<td>42.3</td>
<td>40.7</td>
<td>33.1</td>
<td>42.8</td>
</tr>
<tr>
<td></td>
<td>mPER</td>
<td>26.6</td>
<td>31.6</td>
<td>30.5</td>
<td>26.0</td>
<td>31.9</td>
</tr>
</tbody>
</table>

- Nice gains for the Arabic/English system
- Problem with the phrase-based system for Japanese
### Experimental Evaluation

Results on Development Data (2)

#### Word Error rates

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Phrase-based</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Oracle</td>
<td>Ref.</td>
<td>CSLM</td>
<td>Oracle</td>
</tr>
<tr>
<td>Ma/En</td>
<td>mWER</td>
<td>59.1</td>
<td>67.4</td>
<td>66.5</td>
</tr>
<tr>
<td></td>
<td>mPER</td>
<td>44.5</td>
<td>50.8</td>
<td>50.1</td>
</tr>
<tr>
<td>Ja/En</td>
<td>mWER</td>
<td>70.8</td>
<td>74.6</td>
<td>77.0</td>
</tr>
<tr>
<td></td>
<td>mPER</td>
<td>48.5</td>
<td>52.2</td>
<td>54.6</td>
</tr>
<tr>
<td>Ar/En</td>
<td>mWER</td>
<td>49.1</td>
<td>56.0</td>
<td>52.7</td>
</tr>
<tr>
<td></td>
<td>mPER</td>
<td>40.3</td>
<td>45.7</td>
<td>43.3</td>
</tr>
<tr>
<td>It/En</td>
<td>mWER</td>
<td>34.1</td>
<td>42.3</td>
<td>40.7</td>
</tr>
<tr>
<td></td>
<td>mPER</td>
<td>26.6</td>
<td>31.6</td>
<td>30.5</td>
</tr>
</tbody>
</table>

- Nice gains for the Arabic/English system
- Problem with the phrase-based system for Japanese
Experimental Evaluation
Example Translations

Phrase-based system

Zh: could you we arrive time is two thirty departure time is two five ten
→ you can the time we arrive at two thirty departure time is two fifty
Ar: information your will we arrive at two thirty and an appointment is two and the fifty minutes
→ information i’ll arrive at two thirty and time is two and fifty minutes
It: for your information we’ll be arriving at two o’clock and thirty and your departure time is at two o’clock and fifty
→ for your information we’ll arrive at two thirty and your departure time is at two fifty

Ngram-based system

Ja: we arrive at two thirty takeoff time is fifty two o’clock so you reference you please
→ we arrive at two thirty take off time is two o’clock in fifty so you your reference please
Ar: i’ll information you arrive at two thirty time and is two and fifty minutes
→ i’ll information you arrive at two thirty and time is two and fifty minutes
### Experimental Evaluation

**Results on Evaluation Data (1)**

<table>
<thead>
<tr>
<th></th>
<th>Phrase-based</th>
<th>N-gram-based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ref.</td>
<td>CSLM</td>
</tr>
<tr>
<td><strong>Mandarin/English:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLEU</td>
<td>19.74</td>
<td>21.01</td>
</tr>
<tr>
<td>mWER</td>
<td>67.95</td>
<td>68.16</td>
</tr>
<tr>
<td>mPER</td>
<td>52.46</td>
<td>51.87</td>
</tr>
<tr>
<td><strong>Japanese/English:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLEU</td>
<td>15.11</td>
<td>15.73</td>
</tr>
<tr>
<td>mWER</td>
<td>77.51</td>
<td>78.15</td>
</tr>
<tr>
<td>mPER</td>
<td>55.14</td>
<td>54.96</td>
</tr>
</tbody>
</table>

- Good generalization behavior for Mandarin (Dev +1.3/1.0)
- Small gain for Japanese
- mWER increases in most cases (but not mPER)
### Experimental Evaluation

Results on Evaluation Data (1)

<table>
<thead>
<tr>
<th></th>
<th>Phrase-based Ref.</th>
<th>CSLM</th>
<th>N-gram-based Ref.</th>
<th>CSLM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mandarin/English:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLEU</td>
<td>19.74</td>
<td>21.01</td>
<td>20.34</td>
<td>21.16</td>
</tr>
<tr>
<td>mWER</td>
<td>67.95</td>
<td>68.16</td>
<td>68.30</td>
<td>67.63</td>
</tr>
<tr>
<td>mPER</td>
<td>52.46</td>
<td>51.87</td>
<td>52.81</td>
<td>52.31</td>
</tr>
<tr>
<td><strong>Japanese/English:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLEU</td>
<td>15.11</td>
<td>15.73</td>
<td>16.14</td>
<td>16.35</td>
</tr>
<tr>
<td>mWER</td>
<td>77.51</td>
<td>78.15</td>
<td>75.45</td>
<td>75.59</td>
</tr>
<tr>
<td>mPER</td>
<td>55.14</td>
<td>54.96</td>
<td>55.52</td>
<td>55.29</td>
</tr>
</tbody>
</table>

- Good generalization behavior for Mandarin (Dev +1.3/1.0)
- Small gain for Japanese
- mWER increases in most cases (but not mPER)
Experimental Evaluation
Results on Evaluation Data (1)

<table>
<thead>
<tr>
<th></th>
<th>Phrase-based</th>
<th></th>
<th>N-gram-based</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ref.</td>
<td>CSLM</td>
<td>Ref.</td>
<td>CSLM</td>
</tr>
<tr>
<td><strong>Mandarin/English:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLEU</td>
<td>19.74</td>
<td>21.01</td>
<td>20.34</td>
<td>21.16</td>
</tr>
<tr>
<td>mWER</td>
<td>67.95</td>
<td>68.16</td>
<td>68.30</td>
<td>67.63</td>
</tr>
<tr>
<td>mPER</td>
<td>52.46</td>
<td>51.87</td>
<td>52.81</td>
<td>52.31</td>
</tr>
<tr>
<td><strong>Japanese/English:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLEU</td>
<td>15.11</td>
<td>15.73</td>
<td>16.14</td>
<td>16.35</td>
</tr>
<tr>
<td>mWER</td>
<td>77.51</td>
<td>78.15</td>
<td>75.45</td>
<td>75.59</td>
</tr>
<tr>
<td>mPER</td>
<td>55.14</td>
<td>54.96</td>
<td>55.52</td>
<td>55.29</td>
</tr>
</tbody>
</table>

- Good generalization behavior for Mandarin (Dev +1.3/1.0)
- Small gain for Japanese
- mWER increases in most cases (but not mPER)
Experimental Evaluation
Results on Evaluation Data (2)

<table>
<thead>
<tr>
<th></th>
<th>Phrase-based Ref.</th>
<th>Phrase-based CSLM</th>
<th>N-gram-based Ref.</th>
<th>N-gram-based CSLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic/English:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLEU</td>
<td>23.72</td>
<td>24.86</td>
<td>23.83</td>
<td>23.70</td>
</tr>
<tr>
<td>mWER</td>
<td>63.04</td>
<td>60.89</td>
<td>62.81</td>
<td>61.97</td>
</tr>
<tr>
<td>mPER</td>
<td>49.43</td>
<td>48.61</td>
<td>49.41</td>
<td>48.85</td>
</tr>
<tr>
<td>Italian/English:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLEU</td>
<td>35.55</td>
<td>37.41</td>
<td>35.95</td>
<td>37.65</td>
</tr>
<tr>
<td>mWER</td>
<td>49.12</td>
<td>47.22</td>
<td>48.78</td>
<td>47.59</td>
</tr>
<tr>
<td>mPER</td>
<td>38.17</td>
<td>36.62</td>
<td>38.12</td>
<td>37.26</td>
</tr>
</tbody>
</table>

- No improvement in BLEU score with Ngram-system for Arabic (BLEU decreases despite gain in mWER and mPER)
- Improvements of 1.8 point BLEU for Italian
### Experimental Evaluation
Results on Evaluation Data (2)

<table>
<thead>
<tr>
<th></th>
<th>Phrase-based</th>
<th></th>
<th>N-gram-based</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ref.</td>
<td>CSLM</td>
<td>Ref.</td>
<td>CSLM</td>
</tr>
<tr>
<td><strong>Arabic/English:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLEU</td>
<td>23.72</td>
<td>24.86</td>
<td>23.83</td>
<td>23.70</td>
</tr>
<tr>
<td>mWER</td>
<td>63.04</td>
<td>60.89</td>
<td>62.81</td>
<td>61.97</td>
</tr>
<tr>
<td>mPER</td>
<td>49.43</td>
<td>48.61</td>
<td>49.41</td>
<td>48.85</td>
</tr>
<tr>
<td><strong>Italian/English:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLEU</td>
<td>35.55</td>
<td>37.41</td>
<td>35.95</td>
<td>37.65</td>
</tr>
<tr>
<td>mWER</td>
<td>49.12</td>
<td>47.22</td>
<td>48.78</td>
<td>47.59</td>
</tr>
<tr>
<td>mPER</td>
<td>38.17</td>
<td>36.62</td>
<td>38.12</td>
<td>37.26</td>
</tr>
</tbody>
</table>

- No improvement in BLEU score with N-gram-system for Arabic (BLEU decreases despite gain in mWER and mPER)
- Improvements of 1.8 point BLEU for Italian
Discussion and Perspectives

Summary

- Continuous space LM on top of UPC’s evaluation systems
- Dev-data: gain between 1 and 3 points BLEU
- Eval data: up to 1.9 points BLEU

⇒ Promising approach for tasks with limited resources

Ongoing Work

- Further analysis of the improvements
- Interaction with word reordering?
- Usefulness of long span LMs
- Continuous space translation model (Ngram system)
Discussion and Perspectives

Summary

- Continuous space LM on top of UPC’s evaluation systems
- Dev-data: gain between 1 and 3 points BLEU
- Eval data: up to 1.9 points BLEU
- Promising approach for tasks with limited resources

Ongoing Work

- Further analysis of the improvements
- Interaction with word reordering?
- Usefulness of long span LMs
- Continuous space translation model (Ngram system)