What Matters Most
In Morphologically Segmented
SMT Models?

Mohammad Salameh
Colin Cherry
Greg Kondrak
Overview

• Determine what steps and components of phrase-based SMT pipeline benefits the most from segmenting target language.

• Testing several scenarios by changing the desegmentation point in the pipeline on English-Arabic SMT system

• Phrases with flexible boundaries are a crucial property to a successful segmentation approach

• Show impact of unsegmented LMs on generation of morphologically complex words
**Segmentation/Desegmentation**

Original Word

Segmentation: $t$ to $p$

Desegmentation

- **Morphological Segmentation** is the process of segmenting words into meaningful morphemes.
- **Desegmentation** is the process of converting segmented words into their original orthographically and morphologically correct surface form.
- Segmented vs Unsegmented vs Desegmented
Benefits and Complications of Segmentation

English to Arabic (Morphologically Complex Language)

Benefits segmentation bring to SMT

• Improves correspondence with morphologically simple languages
• Reduces data sparsity
• Increases expressive power by creating new lexical translations

Complications caused by segmentation

• Account for less context compared to word based models
• Less efficient statistically
• Introducing errors due to reversing the segmentation process at the end of the pipeline
Measuring Segmentation Benefits

Experimental study on English to Arabic

• Scenarios changing desegmentation point in pipeline:
  • Before evaluation
  • Before decoding
  • Before phrase extraction

• How these changes affect SMT component models:
  • Alignment model, lexical weights, LM and

• Introducing phrases with flexible boundaries
  • suffix start: +h m$AryE fy “his projects in”
  • Prefix end: jA’ b+ “arrived with”
  • Both: +hA AlAtHAd l+ “her union to”
Techniques for Morphological Segmentation/Desegmentation

Segmentation

• Penn Arabic Treebank Tokenization Scheme (El K holy et al.[2012]) using MADA tool

Desegmentation

• Table+Rule based for Arabic (Badr et al [2008])

<table>
<thead>
<tr>
<th>segmented</th>
<th>unsegmented</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>AbA' +km</td>
<td>AbAŷkm</td>
<td>22</td>
</tr>
<tr>
<td>AbA' +km</td>
<td>AbAWkm</td>
<td>19</td>
</tr>
<tr>
<td>DAŷqp +hm</td>
<td>DAŷqthm</td>
<td>9</td>
</tr>
<tr>
<td>kly +hA</td>
<td>klAhA</td>
<td>5</td>
</tr>
</tbody>
</table>
Unsegmented Baseline

- Suffers from data sparsity
- Poor correspondence
- All component models are based on words
- No desegmentation is required
One-best Desegmentation

- Alleviates data sparsity
- Improves correspondence
- All component models are based on morphemes
- LM spans shorter context
- Desegmentation is required at the end of the pipeline

<table>
<thead>
<tr>
<th>SMT components</th>
<th>Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desegment before</td>
<td>Evaluation</td>
</tr>
<tr>
<td>Alignment Model</td>
<td>Morph</td>
</tr>
<tr>
<td>Lexical weights</td>
<td>Morph</td>
</tr>
<tr>
<td>Language Model</td>
<td>Morph</td>
</tr>
<tr>
<td>Tuning</td>
<td>Morph</td>
</tr>
<tr>
<td>Flexible Boundaries?</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Alignment Desegmentation

- Segment
- Train
- Tune
- Decode

...}

Morpheme alignment

Morpheme desegmentation

Alignment desegmentation

Phrase extraction

...}

0 1 2 3 4
regarding the bank 's policies

SMT components | Scenario
---|---
Desegment before | Phrase extraction
Alignment Model | Morph
Lexical weights | Word
Language Model | Word
Tuning | Word
Flexible Boundaries? | No

0 1 2 3 4 5
[w+ b+ Alnsbp][l+ syAsp][Albnk]
Alignment Desegmentation

- **Segment**
  - ...
  - Morpheme alignment
  - Morpheme desegmentation
  - Alignment desegmentation
  - Phrase extraction
  - ...

- **Train**

- **Tune**

- **Decode**

**SMT components**

- Desegment before: Phrase extraction
- Alignment Model: Morph
- Lexical weights: Word
- Language Model: Word
- Tuning: Word
- Flexible Boundaries?: No

**Example:**

- **Regarding the bank's policies**
  - wbAlnsbp 0
  - IsyAsp 1
  - Albnk 2
Phrase Table Desegmentation

- Remove phrases with flexible boundaries from phrase table
- Desegment phrases in the phrase table
- Use word LM to score desegmented phrases

Similar to Lyong et al. 2010

<table>
<thead>
<tr>
<th>SMT components</th>
<th>Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desegment before</td>
<td>Decoding</td>
</tr>
<tr>
<td>Alignment Model</td>
<td>Morph</td>
</tr>
<tr>
<td>Lexical weights</td>
<td>Morph</td>
</tr>
<tr>
<td>Language Model</td>
<td>Word</td>
</tr>
<tr>
<td>Tuning</td>
<td>Word</td>
</tr>
<tr>
<td>Flexible Boundaries?</td>
<td>No</td>
</tr>
</tbody>
</table>

Phrase with flexible boundaries

- Suffix start: +h m$aryE fy “his projects in”
- Prefix end:  jA’ b+ “arrived with”
- Both: +hA AlAtHAd l+ “her union to”
Lattice Desegmentation
(Salameh et al)

Segment
- Train: segmented model
- Tune: using segmented reference
- Decode: generate lattice on tuning set [segmented output]
- Desegment Lattice
- Retune with added new features using unsegmented reference
- Decode on Desegmented Model

Benefits:
- gain access to a compact desegmented view of a large portion of the translation search space.
- Use features that reflect the desegmented target language
- Annotate with Unsegmented LM + Discontiguity features

<table>
<thead>
<tr>
<th>SMT components</th>
<th>Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desegment before</td>
<td>Evaluation</td>
</tr>
<tr>
<td>Alignment Model</td>
<td>Morph</td>
</tr>
<tr>
<td>Lexical weights</td>
<td>Morph</td>
</tr>
<tr>
<td>Language Model</td>
<td>Morph + Word</td>
</tr>
<tr>
<td>Tuning</td>
<td>Morph then Word</td>
</tr>
<tr>
<td>Flexible Boundaries?</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Flexible Boundaries? Yes
Segmented LM scoring in Desegmented Models

- Add additional LM feature that scores segmented form to:
  - Phrase table Desegmentation
  - Alignment Desegmentation

All our problems and conflicts

[kl m$AkInA] [wxlAfAtna]

[kl m$akl +nA] [w+ xlAfAt +nA]
Data

English-Arabic Data

- Train on NIST 2012 training set, excluding the UN data (1.49M sentence pairs)
- Tune on NIST 2004 (1353 pairs)
  Test on NIST 2005 (1056 pairs)
- Tune on NIST 2006 (1664 pairs)
  Test on NIST 2008 (1360 pairs)
  Test on NIST 2009 (1313 pairs)
System

- Train a 5-gram Language Model on target side using SRILM
- Align parallel data with GIZA++
- Decode using Moses
- Tune the decoder’s log-linear model with MERT
- Reranking Lattice desegmented model is tuned using a batch variant of hope-fear MIRA
- Evaluate the system using BLEU
Results on MT05

<table>
<thead>
<tr>
<th>Method</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unseg</td>
<td>32.8</td>
</tr>
<tr>
<td>Align. Deseg</td>
<td>33.4</td>
</tr>
<tr>
<td>Align. Deseg + seg.LM</td>
<td>33.7</td>
</tr>
<tr>
<td>PT Deseg</td>
<td>33.4</td>
</tr>
<tr>
<td>PT Deseg + seg.LM</td>
<td>33.6</td>
</tr>
<tr>
<td>1-best Deseg</td>
<td>33.7</td>
</tr>
<tr>
<td>1-best Deseg without flexible boundaries</td>
<td>33.7</td>
</tr>
<tr>
<td>Lattice Deseg</td>
<td>34.3</td>
</tr>
</tbody>
</table>
Decoder Integration: lattice desegmentation and 1-best are only systems without access to unsegmented information in the decoder
Results on MT05

Flexible Boundaries: PT Deseg and Align Deseg lack flexible phrase boundaries with respect to 1-best Deseg
Results on MT05

Flexible Boundaries: PT Deseg and Align Deseg. lack flexible phrase boundaries with respect to 1-best Deseg
Language Models: Align Deseg and Phrase Table Deseg show consistent but small, improvements from addition of a segmented LM.
Results on MT05

Language Models: Phrase Table Deseg with segmented LM and 1-best Deseg without flexible boundaries have exactly same output space.
Results on MT05

Language Models: main difference between 1-best Deseg. and Lattice Deseg. Is the unsegmented LM and discontiguity features.
Analysis

1. Flexible boundaries
   • Constitute 12% of phrases in final output of 1-best-deseg
   • Novel words: 3% of the desegmented types
     • Randomly selected 40 out of each set:
       • 64/120 violates morphological rules
       • 37/115 novel words from the reference could be constructed from morphemes

2. Impact of ngram order for segmented LM
   • No improvement seen over 5-gram LM with 6, 7 and 8-grams

3. Overall affix usage
# Overall affix usage

<table>
<thead>
<tr>
<th>Model</th>
<th>mt05</th>
<th>mt08</th>
<th>mt09</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>15.9</td>
<td>18.1</td>
<td>18.9</td>
</tr>
<tr>
<td>Unsegmented</td>
<td>12.0</td>
<td>12.2</td>
<td>12.6</td>
</tr>
<tr>
<td>Alignment Deseg.</td>
<td>11.6</td>
<td>11.0</td>
<td>11.8</td>
</tr>
<tr>
<td>Alignment Deseg. with Segmented LM</td>
<td>11.7</td>
<td>11.2</td>
<td>12.0</td>
</tr>
<tr>
<td>Phrase Table Deseg.</td>
<td>11.3</td>
<td>10.1</td>
<td>11.2</td>
</tr>
<tr>
<td>Phrase Table Deseg. with Segmented LM</td>
<td>11.6</td>
<td>10.5</td>
<td>11.4</td>
</tr>
<tr>
<td>1-best Deseg.</td>
<td>16.1</td>
<td>18.2</td>
<td>19.2</td>
</tr>
<tr>
<td>1-best Deseg. without flexible boundaries</td>
<td>14.2</td>
<td>14.7</td>
<td>15.4</td>
</tr>
<tr>
<td>Lattice Deseg.</td>
<td>10.0</td>
<td>11.5</td>
<td>12.2</td>
</tr>
</tbody>
</table>

Percentage of words in SMT output that have non-identity morphological segmentation
## Overall affix usage

<table>
<thead>
<tr>
<th>Model</th>
<th>mt05</th>
<th>mt08</th>
<th>mt09</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>15.9</td>
<td>18.1</td>
<td>18.9</td>
</tr>
<tr>
<td>Unsegmented</td>
<td>12.0</td>
<td>12.2</td>
<td>12.6</td>
</tr>
<tr>
<td>Alignment Deseg.</td>
<td>11.6</td>
<td>11.0</td>
<td>11.8</td>
</tr>
<tr>
<td>with Segmented LM</td>
<td>11.7</td>
<td>11.2</td>
<td>12.0</td>
</tr>
<tr>
<td>Phrase Table Deseg.</td>
<td>11.3</td>
<td>10.1</td>
<td>11.2</td>
</tr>
<tr>
<td>with Segmented LM</td>
<td>11.6</td>
<td>10.5</td>
<td>11.4</td>
</tr>
<tr>
<td>1-best Deseg.</td>
<td>16.1</td>
<td>18.2</td>
<td>19.2</td>
</tr>
<tr>
<td>without flexible boundaries</td>
<td>14.2</td>
<td>14.7</td>
<td>15.4</td>
</tr>
<tr>
<td>Lattice Deseg.</td>
<td>10.0</td>
<td>11.5</td>
<td>12.2</td>
</tr>
</tbody>
</table>

Percentage of words in SMT output that have non-identity morphological segmentation
Conclusion

• Presented experimental study on translation into segmented language by creating models that apply desegmentation at different points.

• *Flexible boundaries* are the most important factor in improving translation in segmented models.

• Although unsegmented LMs improve BLEU score, they hinder generation of morphologically complex words.
Thank You