Neural semi-Markov CRF for Monolingual Word Alignment

Wuwei Lan*, Chao Jiang* and Wei Xu
Monolingual Word Alignment is **Challenging**

- Aims to align words or phrases with similar meaning in two sentences in the same language.
- **Spans may not follow linguistic boundaries**, and annotation is expensive and time-consuming.
Utility of Monolingual Word Alignment

- Support the analysis of human editing operations.

Some of Russian space agency's launching missions used hydrazine as fuel for the initial few stages.

Hydrazine is used to power the early stages of some Russian launchers.
Utility of Monolingual Word Alignment

- Support the analysis of human editing operations.

Re-ordering: Some of Russian space agency’s launching missions used hydrazine as fuel for the initial few stages.

Deletion: space agency

Rephrase: launching missions

Keep: used hydrazine as fuel for the initial few stages.

Hydrazine is used to power the early stages of some Russian launchers.
Utility of Monolingual Word Alignment

• Support the analysis of human editing operations.

• Improve the performance of text-to-text generation tasks (this work).

• Improve the interpretability of NLU tasks (SemEval 2016 on iSTS).

• Used for label projection and data argumentation (Culkin et al., 2021).
Our Solution for Monolingual Word Alignment

- Neural semi-markov CRF Alignment Model
  - Unify word and phrase alignment by variable-length spans.
  - 92.4 F1 on in-domain evaluation.

- A Multi-genre Monolingual Word Alignment Benchmark
  - Covers 4 different genres (MTRef, Wiki, news, scientific writing).
  - 9k sentence pairs annotated by in-house annotators.
## Semi-CRF Alignment Model

Tokens in target sentence are labels. [NULL] label means no alignment.

<table>
<thead>
<tr>
<th>Source sentence</th>
<th>Target sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wall street stocks fell sharply.</td>
<td>Stocks slump on wall street.</td>
</tr>
</tbody>
</table>

**Source:** Wall street stocks fell sharply.

**Target:** Stocks slump on wall street.

**Formulate it as a sequence tagging problem.**
Semi-CRF Alignment Model

Source: Wall street stocks fell sharply.

Target: Stocks slump on wall street.

Formulate it as a sequence tagging problem.

Tokens in target sentence are labels. [NULL] label means no alignment.

Utilize semi-markov property to handle source spans.

Expand label space to bigram and trigram to handle target spans.
Semi-CRF Alignment Model

Span Interaction Matrix

Span representation based on SpanBERT

\[ h_i^s = (e_{\text{start}(i)}; e_{\text{end}(i)}; \text{att}_i) \]

\[ \text{score}(s_i, t_j) = \text{FFNN}(h_i^s; h_j^t; |h_i^s - h_j^t|; h_i^s \circ h_j^t) \]

2-layer FFNN to capture semantic similarity between \((s_i, t_j)\)
Semi-CRF Alignment Model

Alignment Label Transition

semi-Markov Conditional Random Fields for span alignment

$$\Psi(a, s, t) = \sum_{i} score(s_i, t_{a_i}) + T(a_{i-1}, a_i) + cost(a, a^*)$$

Negative Log-likelihood Loss  Hamming Loss

$$P(a | s, t) = \frac{exp(\Psi(a, s, t))}{\sum_{a \in A} exp(\Psi(a, s, t))}$$

all possible alignments

stocks_slump ⋯

stocks  fell sharply

stocked  on  wall  street

Wall  street  stocks  fell  sharply

on_wall_street ⋯

semi-Markov Conditional Random Fields for span alignment

Source  s

Target  t
Semi-CRF Alignment Model

Bi-directional Training / Decoding

Training objective:

\[
\sum_{s,t,a} - \log P(a_{s2t} | s, t) - \log P(a_{t2s} | t, s)
\]

Source-to-target  Target-to-source

Decoding:

Viterbi-like Algorithm + Intersect + Expand

To handle spans longer than 3
## Multi-Genre Monolingual Word Alignment Benchmark

Annotated by in-house annotators, covers four different domains and the largest to date.

<table>
<thead>
<tr>
<th>MultiMWA</th>
<th>Size</th>
<th>Length</th>
<th>%word/phrase</th>
<th>%identical/non-id</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTRRef</td>
<td>3,998</td>
<td>22 / 17</td>
<td>62.0 / 38.0</td>
<td>52.6 / 47.3</td>
<td>News</td>
</tr>
<tr>
<td>Wiki</td>
<td>4,099</td>
<td>30 / 29</td>
<td>95.6 / 4.4</td>
<td>94.1 / 5.9</td>
<td>Wikipedia</td>
</tr>
<tr>
<td>Newsela</td>
<td>500</td>
<td>27 / 23</td>
<td>74.6 / 25.4</td>
<td>67.1 / 32.9</td>
<td>News</td>
</tr>
<tr>
<td>arXiv</td>
<td>200</td>
<td>29 / 28</td>
<td>96.6 / 3.4</td>
<td>93.4 / 6.6</td>
<td>Scientific writing</td>
</tr>
<tr>
<td>Total</td>
<td>8,797</td>
<td>26 / 23</td>
<td>79.3 / 20.7</td>
<td>73.8 / 26.2</td>
<td>All above</td>
</tr>
</tbody>
</table>
Experiments on **MultiMWA Benchmark**

Achieves SOTA performance on both in-domain and out-of-domain evaluation.

<table>
<thead>
<tr>
<th></th>
<th>In-domain</th>
<th>Out-of-domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTReference</td>
<td>Newsela</td>
<td>arXiv</td>
</tr>
</tbody>
</table>
Experiments on **MultiMWA Benchmark**

Achieves SOTA performance on both in-domain and out-of-domain evaluation.

<table>
<thead>
<tr>
<th></th>
<th>In-domain</th>
<th></th>
<th>Out-of-domain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MTReference</td>
<td>Newsela</td>
<td>arXiv</td>
</tr>
<tr>
<td>JacanaToken (Yao et al. 2013a)</td>
<td>76.2</td>
<td>79.8</td>
<td>95.8</td>
</tr>
<tr>
<td>JacanaPhrase (Yao et al. 2013b)</td>
<td>75.8</td>
<td>79.4</td>
<td>93.7</td>
</tr>
<tr>
<td>PipelineAligner (Sultan et al. 2014)</td>
<td>74.8</td>
<td>80.3</td>
<td>96.5</td>
</tr>
</tbody>
</table>
# Experiments on MultiMWA Benchmark

Achieves SOTA performance on both in-domain and out-of-domain evaluation.

<table>
<thead>
<tr>
<th></th>
<th>In-domain</th>
<th></th>
<th>Out-of-domain</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MTReference</td>
<td>Newsela</td>
<td>arXiv</td>
<td>Wikipedia</td>
</tr>
<tr>
<td>JacanaToken (Yao et al. 2013a)</td>
<td>76.2</td>
<td>79.8</td>
<td>95.8</td>
<td>95.8</td>
</tr>
<tr>
<td>JacanaPhrase (Yao et al. 2013b)</td>
<td>75.8</td>
<td>79.4</td>
<td>93.7</td>
<td>94.9</td>
</tr>
<tr>
<td>PipelineAligner (Sultan et al. 2014)</td>
<td>74.8</td>
<td>80.3</td>
<td>96.5</td>
<td>97.1</td>
</tr>
<tr>
<td>Our Neural CRF aligner</td>
<td>90.8</td>
<td>86.6</td>
<td>95.7</td>
<td>97.0</td>
</tr>
</tbody>
</table>
# Experiments on MultiMWA Benchmark

Achieves SOTA performance on both in-domain and out-of-domain evaluation.

<table>
<thead>
<tr>
<th></th>
<th>In-domain</th>
<th>Out-of-domain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MTReference</td>
<td>Newsela</td>
</tr>
<tr>
<td>JacanaToken (Yao et al. 2013a)</td>
<td>76.2</td>
<td>79.8</td>
</tr>
<tr>
<td>JacanaPhrase (Yao et al. 2013b)</td>
<td>75.8</td>
<td>79.4</td>
</tr>
<tr>
<td>PipelineAligner (Sultan et al. 2014)</td>
<td>74.8</td>
<td>80.3</td>
</tr>
<tr>
<td>Our Neural CRF aligner</td>
<td>90.8</td>
<td>86.6</td>
</tr>
<tr>
<td>Our Neural semi-CRF aligner</td>
<td>92.4</td>
<td>87.2</td>
</tr>
</tbody>
</table>

🚀 16.2 F1  🚀 6.9 F1  🚀 0.8 F1  🚀 0.3 F1
Text Simplification: EditNTS* + Our Aligner

Word alignment can help to explicitly learn edit operations (addition, deletion and paraphrase).

**Deletion**
With Canadian collaborators, Lloyd performed laboratory simulations of his model.

**Paraphrase**
Lloyd performed successful laboratory experiments of his model.

**Addition**

Used edit labels derived from our aligner for training:

- DEL, DEL, DEL, KEEP, KEEP, ADD(‘successful’), KEEP, REPLACE-S, ADD(‘experiments’), REPLACE-E, KEEP, KEEP, KEEP

* EditNTS: An Neural Programmer-Interpreter Model for Sentence Simplification through Explicit Editing, Dong et al., ACL 2019
## Text Simplification Experiments

Our aligner improves the SOTA text simplification model - EditNTS on two benchmark datasets.

<table>
<thead>
<tr>
<th></th>
<th>Newsela-Auto</th>
<th>Wiki-Auto</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SARI</td>
<td>add</td>
</tr>
<tr>
<td>Complex (input)</td>
<td>11.8</td>
<td>0.0</td>
</tr>
<tr>
<td>Simple (reference)</td>
<td>86.9</td>
<td>84.7</td>
</tr>
<tr>
<td>EditNTS (Dong et al. 2019)</td>
<td>36.6</td>
<td>1.1</td>
</tr>
<tr>
<td>EditNTS + Aligner</td>
<td>37.5</td>
<td>1.3</td>
</tr>
</tbody>
</table>
Take Away

- Neural semi-markov CRF Alignment Model
  - Unify word and phrase alignment by variable-length spans.
  - 92.4 F1 on in-domain evaluation.
- A Multi-genre Monolingual Word Alignment Benchmark
  - Covers 4 different genres (MTRef, Wiki, news, scientific writing).
  - 9k sentence pairs annotated by in-house annotators.

Trained model / code / data available at github.com/chaojiang06/neural-Jacana