
Libo Qin¹, Qiguang Chen¹, Tianbao Xie¹, Qixin Li¹, Jian-Guang Lou², Wanxiang Che¹, Min-Yen Kan³

¹Research Center for Social Computing and Information Retrieval (SCIR), Harbin Institute of Technology
²Microsoft Research Asia, Beijing, China
³Department of Computer Science, National University of Singapore
Background
Architecture of Task-oriented Dialogue

“I want to find a Chinese Restaurant”

“Where do you want to eat”

Spoken Language Understanding (SLU)

Intent (Find_Restaurant)
Slots (food = Chinese)

Natural Language Generation (NLG)

Request (location)

Dialog Manager

Knowledge Base

Query
Architecture of Task-oriented Dialogue

“\(\text{I want to find a Chinese Restaurant}\)"

Spoken Language Understanding (SLU)

Intent (Find_Restaurant)

Slots (food = Chinese)
Spoken Language Understanding SLU

- SLU
  - Slot filling -> sequence labeling
  - Intent detection -> classification task
Motivation
Progress

• Progress
  • With the success of Deep Learning technique, remarkable progress has been made in spoken language understanding.

• Problems
  • Rely on a considerable amount of labeled data, which is only available on the English data set[1]
  • Hard to scale to other low-resource languages

Progress

- With the success of Deep Learning technique, remarkable progress has been made in spoken language understanding.

Problems

- Zero-shot SLU have gained increasing attention
- Rely on a considerable amount of labeled data, which is only available on the English data set\[^1\]
- Hard to scale to other low-resource languages

---

Zero-shot Cross-lingual Framework
**Multilingual BERT**

*English*

I want to watch sports movie

*Japanese*

スポーツ映画を見たい

*Chinese*

我想看体育电影

Training

Zero-shot Testing
Multilingual BERT

I want to watch sports movie

No Alignment Signal

English

Japanese

Chinese

CoSDA-ML

Automatically and implicitly align the replaced word vectors in the source and all target languages by mixing their context information.


CoSDA-ML

Implicit Alignment
Challenges of CoSDA-ML

• First Challenge
  • Such implicit alignment process seems to be a black box, which not only seriously affects the alignment representation but also makes it hard to analyze the alignment mechanism.

• Second Challenge
  • Simply relying on shared parameters does not distinguish between the varying granularities of the tasks: the intent detection is sentence-level and the slot filling is token-level, which does not offer fine-grained cross-lingual transfer for token-level slot filling.

• To solve the first challenge
  • We employ contrastive learning (CL) to explicitly align representations of similar sentences across different languages.

• The key insight in GL-CLEF is to encourage representations of similar sentences to be more similar than negative example pairs via contrastive learning.

• To solve the second challenge
  • We first introduce a *Local module* in GL-CLEF to learn different granularity alignment representations (i.e., sentence-level Local intent CL and token-level local slot CL).
    • To be specific, *sentence-level local intent CL* and *token-level local slot CL* are introduced for aligning similar sentence and token representations across different languages for intent detection and slot filling, respectively.

• We further propose a *Global module* named semantic-level global intent – slot CL to bring the representations of slot and intents within a sentence closer together.
  • We further argue that slot and intent are highly correlated and have similar semantic meanings in a sentence.
Positive and Negatives samples of CL

- **Positive Samples Generation**
  - We employ CoSDA-ML to generate multi-lingual code-switched data, which is considered as the positive samples.

- **Negatives Samples Generation**
  - Other different queries in a batch can be considered as negative samples.
Models
Overall Framework

Sentence-level Intent Local CL  
Token-level Slot Local CL  
Semantic-level Intent-Slot Global CL

mBERT

Multi-lingual Code-Switching Generator

Watch  Sports  Movie

Positive Sample Generation  
Negative Sample Generation

Negative Sample Queue
A sentence-level local intent CL loss is introduced to explicitly encourage the model to align similar sentence representations into the same local space across languages for intent detection.
We propose a **token-level local slot CL loss** to help the model to consider token alignment for slot filling, achieving fine-grained cross-lingual transfer. In this situation, token-level CL is applied to all tokens in the query.
Further, we introduce a **semantic-level global intent-slot CL loss** to model the semantic interaction between slots and intent, which may further enhance cross-lingual transfer between them.
Training

- The overall objective in GL-CLEF is a tuned linear combination of the individual losses:

\[
\mathcal{L} = \lambda_I \mathcal{L}_I + \lambda_S \mathcal{L}_S + \lambda_{LI} \mathcal{L}_{LI} + \lambda_{LS} \mathcal{L}_{LS} + \lambda_{GIS} \mathcal{L}_{GIS},
\]

where \( \mathcal{L}_I \) denotes intent detection loss; \( \mathcal{L}_S \) is slot filling loss; \( \mathcal{L}_{LI} \) denotes sentence-level local intent CL loss; \( \mathcal{L}_{LS} \) denotes token-level local slot CL loss; \( \mathcal{L}_{GIS} \) denotes semantic-level global intent-slot CL loss and \( \lambda_i \) are tuning parameters for each loss component.
Experiments
Datasets

- Multi-ATIS++
  - 9 languages including English (en), Spanish (es), Portuguese (pt), German (de), French (fr), Chinese (zh), Japanese (ja), Hindi (hi), and Turkish (tr).

<table>
<thead>
<tr>
<th>Language</th>
<th>Utterances</th>
<th>Intents</th>
<th>Slots</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Dev</td>
<td>Test</td>
</tr>
<tr>
<td>English</td>
<td>4,488</td>
<td>490</td>
<td>893</td>
</tr>
<tr>
<td>Spanish</td>
<td>4,488</td>
<td>490</td>
<td>893</td>
</tr>
<tr>
<td>Portuguese</td>
<td>4,488</td>
<td>490</td>
<td>893</td>
</tr>
<tr>
<td>German</td>
<td>4,488</td>
<td>490</td>
<td>893</td>
</tr>
<tr>
<td>French</td>
<td>4,488</td>
<td>490</td>
<td>893</td>
</tr>
<tr>
<td>Chinese</td>
<td>4,488</td>
<td>490</td>
<td>893</td>
</tr>
<tr>
<td>Japanese</td>
<td>4,488</td>
<td>490</td>
<td>893</td>
</tr>
<tr>
<td>Hindi</td>
<td>1440</td>
<td>160</td>
<td>893</td>
</tr>
<tr>
<td>Turkish</td>
<td>578</td>
<td>60</td>
<td>715</td>
</tr>
</tbody>
</table>

Table 1: MultiATIS++ datasets and statistics.
Main Results

<table>
<thead>
<tr>
<th>Intent Accuracy.</th>
<th>en</th>
<th>de</th>
<th>es</th>
<th>fr</th>
<th>hi</th>
<th>ja</th>
<th>pt</th>
<th>tr</th>
<th>zh</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>mBERT* (Xu et al., 2020)</td>
<td>-</td>
<td>95.27</td>
<td>96.35</td>
<td>95.92</td>
<td>80.96</td>
<td>79.42</td>
<td>94.96</td>
<td>69.59</td>
<td>86.27</td>
<td>-</td>
</tr>
<tr>
<td>mBERT† (Devlin et al., 2019)</td>
<td>98.54</td>
<td>95.40</td>
<td>96.30</td>
<td>94.31</td>
<td>82.41</td>
<td>76.18</td>
<td>94.95</td>
<td>75.10</td>
<td>82.53</td>
<td>88.42</td>
</tr>
<tr>
<td>Ensemble-Net* (Razumovskaia et al., 2021)</td>
<td>92.50</td>
<td>90.26</td>
<td>96.64</td>
<td>95.18</td>
<td>77.88</td>
<td>77.04</td>
<td>95.30</td>
<td>75.40</td>
<td>84.99</td>
<td>87.20</td>
</tr>
<tr>
<td>CoSDA† (Qin et al., 2020)</td>
<td>95.74</td>
<td>94.06</td>
<td>92.29</td>
<td>77.04</td>
<td>82.75</td>
<td>73.25</td>
<td>93.05</td>
<td>80.42</td>
<td>78.95</td>
<td>87.32</td>
</tr>
<tr>
<td><strong>GL-CLEF</strong></td>
<td><strong>98.77</strong></td>
<td><strong>97.53</strong></td>
<td><strong>97.05</strong></td>
<td><strong>97.72</strong></td>
<td><strong>86.00</strong></td>
<td><strong>82.84</strong></td>
<td><strong>96.08</strong></td>
<td><strong>83.92</strong></td>
<td><strong>87.68</strong></td>
<td><strong>91.95</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Slot F1</th>
<th>en</th>
<th>de</th>
<th>es</th>
<th>fr</th>
<th>hi</th>
<th>ja</th>
<th>pt</th>
<th>tr</th>
<th>zh</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble-Net* (Razumovskaia et al., 2021)</td>
<td>85.05</td>
<td>82.75</td>
<td>77.56</td>
<td>76.19</td>
<td>14.14</td>
<td>9.44</td>
<td>74.00</td>
<td>45.63</td>
<td>37.29</td>
<td>55.78</td>
</tr>
<tr>
<td>mBERT* (Xu et al., 2020)</td>
<td>-</td>
<td>82.61</td>
<td>74.98</td>
<td>75.71</td>
<td>31.21</td>
<td>35.75</td>
<td>74.05</td>
<td>23.75</td>
<td>62.27</td>
<td>-</td>
</tr>
<tr>
<td>mBERT† (Devlin et al., 2019)</td>
<td>95.11</td>
<td>80.11</td>
<td>78.22</td>
<td>82.25</td>
<td>26.71</td>
<td>25.40</td>
<td>72.37</td>
<td>41.49</td>
<td>53.22</td>
<td>61.66</td>
</tr>
<tr>
<td>CoSDA† (Qin et al., 2020)</td>
<td>92.29</td>
<td>81.37</td>
<td>76.94</td>
<td>79.36</td>
<td>64.06</td>
<td>66.62</td>
<td>75.05</td>
<td>48.77</td>
<td>77.32</td>
<td>73.47</td>
</tr>
<tr>
<td><strong>GL-CLEF</strong></td>
<td><strong>95.39</strong></td>
<td><strong>86.30</strong></td>
<td><strong>85.22</strong></td>
<td><strong>84.32</strong></td>
<td><strong>70.34</strong></td>
<td><strong>73.12</strong></td>
<td><strong>81.83</strong></td>
<td><strong>65.85</strong></td>
<td><strong>77.61</strong></td>
<td><strong>80.00</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Overall Accuracy.</th>
<th>en</th>
<th>de</th>
<th>es</th>
<th>fr</th>
<th>hi</th>
<th>ja</th>
<th>pt</th>
<th>tr</th>
<th>zh</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR-S2S-PR* (Zhu et al., 2020)</td>
<td>86.83</td>
<td>34.00</td>
<td>40.72</td>
<td>17.22</td>
<td>7.45</td>
<td>10.04</td>
<td>33.38</td>
<td>-</td>
<td>23.74</td>
<td>23.77</td>
</tr>
<tr>
<td>IT-S2S-PR* (Zhu et al., 2020)</td>
<td>87.23</td>
<td>39.46</td>
<td>50.06</td>
<td>46.78</td>
<td>11.42</td>
<td>12.60</td>
<td>39.30</td>
<td>-</td>
<td>28.72</td>
<td>32.69</td>
</tr>
<tr>
<td>mBERT† (Devlin et al., 2019)</td>
<td>87.12</td>
<td>52.69</td>
<td>52.02</td>
<td>37.29</td>
<td>4.92</td>
<td>7.11</td>
<td>43.49</td>
<td>4.33</td>
<td>18.58</td>
<td>36.29</td>
</tr>
<tr>
<td>CoSDA† (Qin et al., 2020)</td>
<td>77.04</td>
<td>57.06</td>
<td>46.62</td>
<td>50.06</td>
<td>26.20</td>
<td>28.89</td>
<td>48.77</td>
<td>15.24</td>
<td>46.36</td>
<td>44.03</td>
</tr>
<tr>
<td><strong>GL-CLEF</strong></td>
<td><strong>88.02</strong></td>
<td><strong>66.03</strong></td>
<td><strong>59.53</strong></td>
<td><strong>57.02</strong></td>
<td><strong>34.83</strong></td>
<td><strong>41.42</strong></td>
<td><strong>60.43</strong></td>
<td><strong>28.95</strong></td>
<td><strong>50.62</strong></td>
<td><strong>54.09</strong></td>
</tr>
</tbody>
</table>

Our framework achieves the state-of-the-art performance by beating CoSDA-ML with 10.06% performance
Ablation Analysis

All components contribute a lot for the final performance
Visualization Analysis

Figure 5: t-SNE visualization of sentence vectors from (a) mBERT and (b) GL-CLEF. Different colors represent different languages.

GL-CLEF successfully pulls representations closer across different languages.
Conclusion

- We introduced a global-local contrastive learning (CL) framework (GL-CLEF) to explicitly align representations across languages for zero-shot cross-lingual SLU.

- Besides, the proposed Local CL module and Global CL module achieves to learn different granularity alignment (i.e., sentence-level local intent alignment, token-level local slot alignment, semantic-level global intent-slot alignment).

- Experiments on MultiATIS++ show that GL-CLEF obtains best performance and extensive analysis indicate GL-CLEF successfully pulls closer the representations of similar sentence across languages.
Thanks & QA

Paper

Code