DAMO-NLP at SemEval-2022 Task 11:
A Knowledge-based System for Multilingual Named Entity Recognition

Xinyu Wang*, Yongliang Shen*, Jiong Cai*, Tao Wang, Xiaobin Wang†, Pengjun Xie†
Fei Huang†, Weiming Lu*, Yueting Zhuang*, Kewei Tu*, Wei Lu†, Yong Jiang†*
†DAMO Academy, Alibaba Group
\(^{\circledR}\)School of Information Science and Technology, ShanghaiTech University
\(\blacklozenge\)College of Computer Science and Technology, Zhejiang University
\(\dagger\)StatNLP Research Group, Singapore University of Technology and Design
Dataset Analysis

- Multilingual task over 11 languages + 1 multilingual track + 1 codemixed track
- Low context, short sentences with ambiguous and complex entities
- Cross-domain ability is required:
  - Test Domain: Wikipedia + Web question + user query
Motivation

• Previous studies
  • Including document-level contexts of the target sentence in the input of contextual embeddings methods can boost the accuracy of NER models

• Problem
  • The datasets in the shared task are low-context short sentences

• Idea
  • Knowledge from knowledge base helps human annotators of NER
  • It should benefit machines too!
Motivating Example

Input Sentence: köpings is rate

Retrieve in KB

Predict without knowledge

köpings | LOC is rate

Retrieve Results:
1. **Kis:** Köpings IS, a Swedish sports club
2. **Köping, Sweden:** Köpings IS, association club, bandy and handball
3. **Kalle Samuelsson:** Kalle Samuelsson (born February 15, 1986) is a Swedish Bandy player who plays for Västerås SK as a goalkeeper. Kalle was a youth product of Köpings IS.
4. ...

Predict with knowledge

köpings is | GRP rate
Ours: A general Knowledge-based System

- Based on Wikipedia of the 11 languages, we build a multilingual knowledge base to search for the related knowledge of the input sentence.
- We then concatenate the input sentence and the text of the retrieved text together and feed the concatenated string into the NER model.
- Continue Fine-tuning: first train a unified multilingual model, then train the monolingual models for each language.
Building the knowledge base

• Wikipedia dumps of 11 languages from Wikimedia

• Add hyperlinks into the retrieval results

• Example: “Steve Jobs founded Apple”
  • “Steve Jobs" is linked to the wiki entry Steve_Jobs
  • "Apple" is linked to the wiki entry Apple_Inc

• mark the anchors marked in the wiki with a special marker:
  • Example: (Apple, Apple_inc is marked in the text as “<e:Apple_inc>Apple</e>”
Sentence Retrieval

- Search which knowledge base: by its language
- For codemixed: In all language knowledge base
- For multilingual: Decide the language and search
- ElasticSearch for searching
Coarse-to-Fine Entity Retrieval

• Motivation:
  • Retrieval at the sentence level tends to overlook the key entities in the sentences

• Our approach
  • Concatenate the entities in the sentences with "|" and then retrieve them on the title field of the index

• Training:
  • Use gold entities

• Testing:
  • Use predicted entities
Post-processing

• Concatenate retrieved knowledge at the end of input sentence, until 512 subtokens.
Named Entity Recognition Module

\[ \psi(y', y, v_i) = \exp(W_y^T v_i + b_{y', y}) \]

\[ p_\theta(y|x, x') = \frac{\prod_{i=1}^{n} \psi(y_{i-1}, y_i, v_i)}{\sum_{y' \in \mathcal{Y}(x)} \prod_{i=1}^{n} \psi(y'_{i-1}, y'_i, v_i)} \]

Input: \( x \)

Retrieve \( x \) in KB

Retrieval Results \( \hat{x} \):
1. ------------
2. ----------------
3. -------------
4. ------------
5. -------------
6. …

\( \hat{y}_\theta \)

Majority Voting Ensemble

Vote

NER Model 1

\( \hat{y}_{\theta_1} \)

NER Model m

\( \hat{y}_{\theta_m} \)

Transformer-Based Embedding

CRF

\( y_1, y_2, \ldots, y_n \)
Ensemble Module

• Span Based Voting

• Majority Voting

• Ranks all spans in the predictions by the number of votes and selects the spans with more than 50% votes into the final prediction.
Data

- MultiCoNER dataset
  - There are mainly three domains in the dataset:
    - LOWNER: low-context sentences from Wikipedia.
    - MSQ: a lot of natural language questions.
    - ORCAS: user queries from Microsoft Bing.
  - Test: from 100k to 500k

<table>
<thead>
<tr>
<th>Track</th>
<th>Training</th>
<th>Development</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multilingual</td>
<td>168,300</td>
<td>8,800</td>
</tr>
<tr>
<td>Monolingual</td>
<td>15,300</td>
<td>800</td>
</tr>
<tr>
<td>Code-mixed</td>
<td>1,500</td>
<td>500</td>
</tr>
</tbody>
</table>

Table 1: Statistics about the MultiCoNER dataset.
Results

- Our system performs the best on 10 out of 13 tracks and is competitive on the other 3 tracks.
- Our system outperforms our baseline by 14.39 F1 on average.

<table>
<thead>
<tr>
<th>System</th>
<th>en</th>
<th>es</th>
<th>nl</th>
<th>ru</th>
<th>tr</th>
<th>ko</th>
<th>fa</th>
<th>de</th>
<th>zh</th>
<th>hi</th>
<th>bn</th>
<th>multi</th>
<th>mix</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours: Baseline</td>
<td>77.81</td>
<td>76.80</td>
<td>80.51</td>
<td>74.65</td>
<td>72.83</td>
<td>70.81</td>
<td>72.68</td>
<td>81.92</td>
<td>65.56</td>
<td>67.80</td>
<td>65.27</td>
<td>74.19</td>
<td>77.75</td>
<td>73.74</td>
</tr>
<tr>
<td>USTC-NELSLIP</td>
<td>85.47</td>
<td>85.44</td>
<td>87.67</td>
<td>83.82</td>
<td>85.52</td>
<td>86.36</td>
<td>87.05</td>
<td>89.05</td>
<td>81.69</td>
<td>84.64</td>
<td>84.24</td>
<td>85.30</td>
<td>92.90</td>
<td>86.09</td>
</tr>
<tr>
<td>Ours</td>
<td>91.22</td>
<td>89.94</td>
<td>90.50</td>
<td>91.50</td>
<td>88.69</td>
<td>88.59</td>
<td>89.70</td>
<td>90.65</td>
<td>78.06</td>
<td>86.23</td>
<td>83.51</td>
<td>85.31</td>
<td>91.79</td>
<td>88.13</td>
</tr>
</tbody>
</table>

Table 2: Official results of all systems.
Analysis: The relevance of the retrieval results and the query

• The IoU values are concentrated around 1.0 on the training and development sets of EN, ES, NL, RU, TR, KO, FA, which indicates that most of the samples were derived from Wikipedia.

• The source of the test set for TR is different from the training set. However, the model still performs strongly on this language.
### Analysis: Per-domain F1

<table>
<thead>
<tr>
<th></th>
<th>en</th>
<th>es</th>
<th>nl</th>
<th>ru</th>
<th>tr</th>
<th>ko</th>
<th>fa</th>
<th>de</th>
<th>zh</th>
<th>hi</th>
<th>bn</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LOWNER</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>88.70</td>
<td>86.54</td>
<td>89.92</td>
<td>81.52</td>
<td>88.52</td>
<td>86.25</td>
<td>81.85</td>
<td>91.71</td>
<td>85.43</td>
<td>83.13</td>
<td>82.69</td>
<td>86.02</td>
</tr>
<tr>
<td>Ours</td>
<td>96.78</td>
<td>96.19</td>
<td>97.96</td>
<td>96.60</td>
<td>96.43</td>
<td>96.83</td>
<td>96.48</td>
<td>94.89</td>
<td>88.66</td>
<td>84.18</td>
<td>86.31</td>
<td>93.76</td>
</tr>
<tr>
<td>Δ</td>
<td>8.07</td>
<td>9.66</td>
<td>8.05</td>
<td>15.08</td>
<td>7.91</td>
<td>10.59</td>
<td>14.63</td>
<td>3.18</td>
<td>3.23</td>
<td>1.05</td>
<td>3.62</td>
<td>7.73</td>
</tr>
<tr>
<td><strong>MSQ</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>70.49</td>
<td>71.86</td>
<td>72.63</td>
<td>72.31</td>
<td>75.49</td>
<td>68.57</td>
<td>71.54</td>
<td>74.63</td>
<td>67.38</td>
<td>73.57</td>
<td>58.66</td>
<td>70.65</td>
</tr>
<tr>
<td>Ours</td>
<td>83.50</td>
<td>83.10</td>
<td>83.34</td>
<td>87.03</td>
<td>88.76</td>
<td>81.96</td>
<td>87.36</td>
<td>86.18</td>
<td>79.80</td>
<td>89.20</td>
<td>72.00</td>
<td>83.84</td>
</tr>
<tr>
<td><strong>OSCAS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>62.07</td>
<td>62.71</td>
<td>67.39</td>
<td>64.83</td>
<td>66.92</td>
<td>56.08</td>
<td>65.52</td>
<td>67.52</td>
<td>55.34</td>
<td>62.03</td>
<td>60.68</td>
<td>62.83</td>
</tr>
<tr>
<td>Ours</td>
<td>83.72</td>
<td>81.33</td>
<td>80.29</td>
<td>85.00</td>
<td>85.85</td>
<td>81.06</td>
<td>84.84</td>
<td>84.40</td>
<td>72.11</td>
<td>85.75</td>
<td>82.13</td>
<td>82.41</td>
</tr>
</tbody>
</table>

Table 3: Official results of all systems.
Conclusion

• Knowledge-based system can help low context sentences to disambiguate complex entities.

• Our system achieves state-of-the-art accuracy over 10 tracks.

• We show that the knowledge-based system can also significantly improve cross-domain accuracy.