A Multi-lingual Multi-task Architecture for Low-resource Sequence Labeling

YING LIN¹, SHENQI YANG², VESELIN STOYANOV³, HENG JI¹

¹ Computer Science Department, Rensselaer Polytechnic Institute
² Intelligent Advertising Lab, JD.com
³ Applied Machine Learning, Facebook
MOTIVATION

- Most high-performance data-driven models rely on a large amount of labeled training data. However, a model trained on one language usually performs poorly on another language.

- Extend existing services to more languages:
  - Collect, select, and pre-process data
  - Compile guidelines for new languages
  - Train annotators to qualify for annotation tasks
  - Annotate data
  - Adjudicate annotations and assess the annotation quality and inter-annotator agreement
MOTIVATION

• Most high-performance data-driven models rely on a large amount of labeled training data. However, a model trained on one language usually performs poorly on another language.

• Extend existing services to more languages:
  • Collect, select, and pre-process data
  • Compile guidelines for new languages
  • Train annotators to qualify for annotation tasks
  • Annotate data
  • Adjudicate annotations and assess inter-annotator agreement

7,097 languages are spoken today

• Rapid and low-cost development of capabilities for low-resource languages.
  • Disaster response and recovery
Leverage existing data of related languages and tasks and transfer knowledge to our target task.

Multi-task Learning (MTL) is an effective solution for knowledge transfer across tasks.

In the context of neural network architectures, we usually perform MTL by sharing parameters across models.

Parameter Sharing: When optimizing model A, we update and hence \( \theta \). In this way, we can partially train model B as \( \theta \).
To illustrate our idea, we take sequence labeling as a case study. In the NLP context, the goal of sequence labeling is to assign a categorical label (e.g., Part-of-speech tag) to each token in a sentence. It underlies a range of fundamental NLP tasks, including POS Tagging, Name Tagging, and Chunking.

B-, I-, E-, S-: beginning of a mention, inside of a mention, the end of a mention and a single-token mention
O: not part of any mention

Although we only focus on sequence labeling in this work, our architecture can be adapted for many NLP tasks with slight modification.
BASE MODEL: LSTM-CRF (CHIU AND NICHOLS, 2016)

Each token in the given sentence is represented as the combination of its word embedding and character feature vector.

The Bidirectional LSTM (long-short term memory) processes the input sentence from both directional, encoding each token and its context into a vector (hidden states).

The linear layer projects hidden states to label space.

The CRF layer models the dependencies between labels.

The Character-level CNN extracts character-level features for each token.
Yang et al. (2017) proposed three transfer learning architectures for different use cases.

* Above figures are adapted from (Yang et al., 2017)
OUR MODEL: MULTI-LINGUAL MULTI-TASK ARCHITECTURE

- Our model combines multi-lingual transfer and multi-task transfer.
- It is able to transfer knowledge from multiple sources.
OUR MODEL: MULTI-LINGUAL MULTI-TASK MODEL

Part-of-speech Tagging

Cross-task Transfer
POS Tagging → Name Tagging

Name Tagging

Cross-lingual Transfer
English → Spanish

Spanish Linear Layer
Shared Linear Layer

Spanish Sentence

Cross-lingual Cross-task Shared Bidirectional LSTM

Spanish Word Embeddings

Character Embeddings

English Word Embeddings

Final tags
Shared across languages
Unary tag scores
Shared across languages
Language-specific
Encode each word and its context into a vector
OUR MODEL: MULTI-LINGUAL MULTI-TASK MODEL

- The bidirectional LSTM, character embeddings and character-level networks serve as the basis of the architecture. This level of parameter sharing aims to provide universal word representation and feature extraction capability for all tasks and languages.
For the same task, most components are shared between languages.

Although our architecture does not require aligned cross-lingual word embeddings, we also evaluate it with aligned embeddings generated using MUSE's unsupervised model (Conneau et al. 2017).
OUR MODEL: MULTI-LINGUAL MULTI-TASK MODEL - LINEAR LAYER

- We add a language-specific linear layer to allow the model to behave differently towards some features for different languages.

English: improvement, development, payment, ...
French: vraiment, complètement, immédiatement

We combine the output of the shared linear layer and the output of the language-specific linear layer using

$$ y = g \odot y^s + (1 - g) \odot y^u $$

where $g$ and $u$ are optimized during training. $s$ is the LSTM hidden states. As $s$ is a square matrix, $g$, $s$, and $u$ have the same dimension.
Linear layers and CRF layers are not shared between different tasks.

Tasks of the same language use the same embedding matrix: mutually enhance word representations.
ALTERNATING TRAINING

• To optimize multiple tasks within one model, we adopt the alternating training approach in (Luong et al., 2016).

![Diagram of alternating training]

- At each training step, we sample a task with probability:

  \[ p(d_i) = \frac{r_i}{\sum_j r_j} \]

- In our experiments, instead of tuning mixing rate \( r \), we estimate it by:

  \[ r_i = \mu_i \varsigma_i \sqrt{N_i} \]

where is the task coefficient, is the language coefficient, and is the number of training examples. (or ) takes the value 1 if the task (or language) of is the same as that of the target task; Otherwise it takes the value 0.1.
EXPERIMENTS - DATA SETS

- Name Tagging
  - English: CoNLL 2003
  - Spanish and Dutch: CoNLL 2002
  - Russian: LDC2016E95 (Russian Representative Language Pack)
  - Chechen: TAC KBP 2017 10-Language EDL Pilot Evaluation Source Corpus
- Part-of-speech Tagging: CoNLL 2017 (Universal Dependencies)
• 50-dimensional pre-trained word embeddings
  • English, Spanish and Dutch: Wikipedia
  • Russian: LDC2016E95
  • Chechen: TAC KBP 2017 10-Language EDL Pilot Evaluation Source Corpus
• Cross-lingual word embedding: we aligned mono-lingual pre-trained word embeddings with MUSE (https://github.com/facebookresearch/MUSE).
• 50-dimensional randomly initialized character embeddings
• Optimization: SGD with momentum (), gradient clipping (threshold: 5.0) and exponential learning rate decay.

<table>
<thead>
<tr>
<th>CharCNN Filter Number</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highway Layer Number</td>
<td>2</td>
</tr>
<tr>
<td>Highway Activation Function</td>
<td>SeLU</td>
</tr>
<tr>
<td>LSTM Hidden State Size</td>
<td>171</td>
</tr>
<tr>
<td>LSTM Dropout Rate</td>
<td>0.6</td>
</tr>
<tr>
<td>Learning Rate</td>
<td>0.02</td>
</tr>
<tr>
<td>Batch Size</td>
<td>19</td>
</tr>
</tbody>
</table>
EXPERIMENTS - COMPARISON OF DIFFERENT MODELS

- Target task: Dutch Name Tagging
- Auxiliary task: Dutch POS Tagging, English Name Tagging, English POS Tagging

11.9%-24.9% F-score Gain
18.2%-50.0% F-score Gain
EXPERIMENTS - COMPARISON OF DIFFERENT MODELS

- Target task: Spanish Name Tagging
- Auxiliary task: Spanish POS Tagging, English Name Tagging, English POS Tagging

13.5%-50.5% F-score Gain
11.6%-22.6% F-score Gain
EXPERIMENTS - COMPARISON OF DIFFERENT MODELS

- Target task: Chechen Name Tagging
- Auxiliary task: Russian POS Tagging + Name Tagging or English POS Tagging + Name Tagging

4.3%-15.9% F-score Gain
15.8%-25.4% F-score Gain

All training data:
Baseline: 78.9%
Our Model: 82.3%
We also compared our model with state-of-the-art models with all training data.
<table>
<thead>
<tr>
<th>#1 [DUTCH]: If a Palestinian State is, however, the first thing the Palestinians will do.</th>
</tr>
</thead>
<tbody>
<tr>
<td>* [B] Als er een Palestijnse staat komt, is dat echter het eerste wat de Palestijnen zullen doen</td>
</tr>
<tr>
<td>* [A] Als er een [S-MISC Palestijnse] staat komt, is dat echter het eerste wat de [S-MISC Palestijnen] zullen doen</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#2 [DUTCH]: That also frustrates the Muscovites, who still live in the proud capital of Russia but can not look at the soaps that the stupid farmers can see on the outside.</th>
</tr>
</thead>
<tbody>
<tr>
<td>* [B] Ook dat frustreert de Moskoviërs , die toch in de fiere hoofdstad van Rusland wonen maar niet naar de soaps kunnen kijken die de domme boeren op de buiten wel kunnen zien</td>
</tr>
<tr>
<td>* [A] Ook dat frustreert de [S-MISC Moskovietten] , die toch in de fiere hoofdstad van [S-LOC Rusland] wonen maar niet naar de soaps kunnen kijken die de domme boeren op de buiten wel kunnen zien</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#3 [DUTCH]: And the PMS centers are merging with the centers for school supervision, the MSTs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>* [B] En smelten de PMS-centra samen met de centra voor schooltoezicht, de MST’s.</td>
</tr>
<tr>
<td>* [A] En smelten de [S-MISC PMS-centra] samen met de centra voor schooltoezicht, de [S-MISC MST’s]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#4 [SPANISH]: The trade union section of CC.OO. in the Department of Justice has today denounced more attacks of students to educators in centers dependent on this department ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>* [B] La sección [I-ORG sindical] [I-ORG de] [S-ORG CC.OO.] en el Departamento [I-ORG de] [E-ORG Justicia] ha denunciado hoy ms agresiones de alumnos a educadores en centros dependientes de esta [S-ORG consellería] ...</td>
</tr>
<tr>
<td>* [A] La sección sindical de [S-ORG CC.OO.] en el Departamento [I-ORG de] [E-ORG Justicia] ha denunciado hoy ms agresiones de alumnos a educadores en centros dependientes de esta consellería ...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#5 [SPANISH]: ... and the Single Trade Union Confederation of Peasant Workers of Bolivia, agreed upon when the state of siege was ended last month.</th>
</tr>
</thead>
<tbody>
<tr>
<td>* [B] ... y la Confederación [I-ORG Sindical] [I-ORG Unica] [I-ORG de] [E-ORG Trabajadores Campesinos de Bolivianos] , pactadas cuando se dio fin al estado de sitio, el mes pasado .</td>
</tr>
<tr>
<td>* [A] .. y la [B-ORG Confederación] [I-ORG Sindical] [I-ORG Unica] [I-ORG de] [I-ORG Trabajadores] [I-ORG Campesinos] [I-ORG de] [E-ORG Bolivia] , pactadas cuando se dio fin al estado de sitio, el mes pasado .</td>
</tr>
</tbody>
</table>
EXPERIMENTS - CROSS-TASK TRANSFER VS CROSS-LINGUAL TRANSFER

With 100 Dutch training sentences:

- The baseline model misses the name “Ingeborg Marx”.
- The cross-task transfer model finds the name but assigns a wrong tag to “Marx”.
- The cross-lingual transfer model correctly identifies the whole name.
- The task-specific knowledge that B-PER $\rightarrow$ S-PER is an invalid transition will not be learned in the POS Tagging model.
- The cross-lingual transfer model transfers such knowledge through the shared CRF layer.

[DUTCH] ... Ingeborg Marx is her name, a formidable heavy weight to high above her head!

★ [B] ... Zag ik zelfs onlangs niet dat een lief, mooi vrouwje, **Ingeborg Marx** is haar naam, een formidabel zwaar gewicht tot hoog boven haar hoofd stak!

★ [CROSS-TASK] ... Zag ik zelfs onlangs niet dat een lief, mooi vrouwje, **[B-PER Ingeborg] [S-PER Marx]** is haar naam, een formidabel zwaar gewicht tot hoog boven haar hoofd stak!

★ [CROSS-LINGUAL] ... Zag ik zelfs onlangs niet dat een lief, mooi vrouwje, **[B-PER Ingeborg] [E-PER Marx]** is haar naam, een formidabel zwaar gewicht tot hoog boven haar hoofd stak!
### EXPERIMENTS - ABLATION STUDIES

<table>
<thead>
<tr>
<th>Model</th>
<th>0</th>
<th>10</th>
<th>100</th>
<th>200</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>2.06</td>
<td>20.03</td>
<td>47.98</td>
<td>51.52</td>
<td>77.63</td>
</tr>
<tr>
<td>+C</td>
<td>1.69</td>
<td>24.22</td>
<td>48.53</td>
<td>56.26</td>
<td>83.38</td>
</tr>
<tr>
<td>+CL</td>
<td>9.62</td>
<td>25.97</td>
<td>49.54</td>
<td>56.29</td>
<td>83.37</td>
</tr>
<tr>
<td>+CLS</td>
<td>3.21</td>
<td>25.43</td>
<td>50.67</td>
<td>56.34</td>
<td>84.02</td>
</tr>
<tr>
<td>+CLSH</td>
<td>7.70</td>
<td>30.48</td>
<td>53.73</td>
<td>58.09</td>
<td>84.68</td>
</tr>
<tr>
<td>+CLSHD</td>
<td>12.12</td>
<td>35.82</td>
<td>57.33</td>
<td>63.27</td>
<td>86.00</td>
</tr>
</tbody>
</table>

C: Character embedding; L: Shared LSTM; S: Language-specific
H: Highway Networks; D: Dropout

- Generally, all components improve the performance.
- Sharing the LSTM layer slightly hurts the performance in the “high-resource” setting.
- Language-specific Layer can impair the performance in extreme low-resource settings because this layer is trained only on the target task data.
EXPERIMENTS - EFFECT OF THE AMOUNT OF AUXILIARY TASK DATA

• Does our model heavily rely on the amount of auxiliary task data?
  • The performance goes up when we increase the sample rate from 0 to 0.2 for auxiliary task data.
  • However, we do not observe substantial improvement when we further increase the sample rate.
• Using only 1% auxiliary data, our model already obtains 3.7%-9.7% absolute F-score gains.
Does our model heavily rely on the amount of auxiliary task data?

- The performance goes up when we increase the sample rate from 0 to 0.2 for auxiliary task data.
- However, we do not observe substantial improvement when we further increase the sample rate.
- Using only 1% auxiliary data, our model already obtains 3.7%-9.7% absolute F-score gains.
REFERENCES

• Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. 2016. Multilingual language processing from bytes. In NAACL HLT