Multilingual model using cross-task embedding projection

Jin Sakuma
The University of Tokyo

Naoki Yoshinaga
Institute of Industrial Science, the University of Tokyo
Needs for resources in deep learning

Improvements in various tasks by deep learning

- Task-specific representation learning needs more data

Translation performance (DE->EN) [Sennrich+ 2019]

- Large gain with rich resources
- Small gain with poor resources

Need massive data for every pair of task and language.
Gaps in available resources across langs

Among 7097 languages in the world [Simons+ 2018], massive resources are obtainable only in few

🖤 Universal dependencies project covers only 76 [Nivre+ 201X]

Large gap in model performances among languages

Can we exploit resources of resource-rich languages for training in resource-poor languages?
Problem settings of this study

Available resources:
- Labeled data for training in the source language
- Raw corpora in both languages

Applicable to various target languages and tasks
Problem settings of this study

Available resources:
- Labeled data for training
- Raw corpora in both languages
- Applicable to various target languages

No cross-lingual resources

Source
- Annotated data
- Raw corpus (e.g. Wikipedia)

Target
- No annotated data
- Raw corpus (e.g. Wikipedia)
Preliminary: Cross-lingual word embeddings (CLWE)

Language-independent representation of words [Mikolov+ 13]

- Words from two lang. are represented in a shared space
- Similar words from different languages are close
Existing multilingual models

Fix the emb. layer to general CLWE during training [Duong+ 17, Chen+ 18]

- Enables cross-lingual transfer
- The embedding layer is not optimized for the task
Related work:
Task-specific CLWE with specialized dict.

Utilize task-specific bilingual dict. to obtain CLWE [Gouws+ 15]

The embedding layer is optimized for the task
Additional cross-lingual resources are required
Related work:

Task-specific CLWE with specialized dict.

Utilize task-specific bilingual dict. to obtain CLWE [Gouws+ 15]

In this study:

**Obtain task-specific** CLWE without relying on any cross-lingual resources

- The embedding layer is optimized for the task
- Requires additional cross-lingual resources
Proposal: Multilingual model with task-spec emb.

Project general CLWE to the emb. layer optimized for the task by **cross-task embedding projection**.
Locally linear mapping: 
Idea: local topology of embeddings

Assumption:
- Words **adequately close** in the general CLWE are also close in task-specific space

![Diagram showing word embeddings in general CLWE and ideal task-specific CLWE](image-url)
Locally linear mapping:
Step 1: selecting nearest neighbors

For each **target word** (bien), select **k-nearest neighbors** in the general CLWE.
Locally linear mapping:
Step 2: local topology in general space

In general CLWE, learn linear combination of nearest neighbors that reconstructs the target word

\[ \hat{\alpha}_{w^*} = \arg\min_{\alpha_w} Y_w^{\text{gen}} - \sum_{i \in N_w} \alpha_{wi} X_i^{\text{gen}} \]
Locally linear mapping:
Step 3: task-specific word embeddings

Compute task-specific word emb. of the target word as the linear combination with the induced weights

General CLWE

\[
\hat{\alpha}_{w*} = \arg \min_{\alpha_{w*}} Y^\text{gen}_w - \sum_{i \in N_w} \alpha_{wi} X^\text{gen}_i \|^2
\]

Task-specific word emb.

\[
Y^\text{spec}_w = \sum_{i \in N_w} \hat{\alpha}_{wi} X^\text{spec}_w
\]
Proposal: Hyperparameter search

Dev. set in the target language is required to tune the hyperparameter $k$ (size of nearest neighbors)

Tuning to the task (no additional resources)

Assume the best $k$ is independent of language

Apply LLM to the embeddings of the source language and evaluate on the dev. set of the source language

(Tuning to the task/language)

Utilize small development set (100 examples) of the target language
Experimental setup (1/2)

Goal:
- Does our task-specific word embeddings improve the multilingual model?

Task:
- Topic classification task (and sent. analysis)

Languages:
- Source language: English (en)
- Target languages:
  - Danish (da), Italian (it), French (fr), Swedish (sv)

Datasets:
- RCV1/2 dataset (four topics)
Experimental setup (2/2):
Models to compare

Compare the following two models to evaluate the effect of task-specific CLWE

- Experiments on more models on the paper

CLWE fixed
- SoftMax
- 1-Layer FFNN

CLWE opt (proposal)
- SoftMax
- 1-Layer FFNN

General CLWE

Task-specific CLWE

LLM
### Results:

#### Topic classification task

Classification accuracies in four languages

<table>
<thead>
<tr>
<th>Method</th>
<th>$k$-tuning</th>
<th>en-da</th>
<th>en-it</th>
<th>en-fr</th>
<th>en-sv</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLWE fixed</td>
<td>-</td>
<td>0.621</td>
<td>0.535</td>
<td>0.772</td>
<td>0.816</td>
</tr>
<tr>
<td>CLWE opt (Proposed)</td>
<td>task</td>
<td>0.672</td>
<td><strong>0.623</strong></td>
<td>0.885</td>
<td><strong>0.831</strong></td>
</tr>
<tr>
<td>CLWE opt (Proposed)</td>
<td>task/lang</td>
<td><strong>0.687</strong></td>
<td>0.615</td>
<td>0.879</td>
<td>0.830</td>
</tr>
</tbody>
</table>

- CLWE opt outperforms the baseline
- Tuning $k$ for task and language is not necessary
Conclusion and future work

Conclusion

- Proposed a method to build a multilingual model with task-specific word embeddings
- Evaluated our method on real tasks and confirmed its effectiveness

Future work

- Evaluate this method on wider range of tasks, languages, and models
- Further improve the quality of locally linear mapping