Low Resource and Cross-Lingual Information Extraction

CSCI 699: Machine Learning for Knowledge Extraction & Reasoning

Nanyun (Violet) Peng
T790M is present as a minor clone in NSCLC, and may be selected for during therapy.

This mutation has been shown to prevent the activation of BIM in response to gefitinib, but can be overcome by an irreversible inhibitor of EGFR.
T790M is present as a minor clone in NSCLC, and may be selected for during therapy.

This mutation has been shown to prevent the activation of BIM in response to gefitinib, but can be overcome by an irreversible inhibitor of EGFR.
NER on Chinese Social Media

• Contains informal use of language: e.g. dialects, curse words, typos, etc.

成都(GPE.NAM)电信(ORG.NAM)到底有没的时间观念
哦，一托再托，日妈(PER.NOM)我们时间就不是时间哇，等了你两天啥子速度。

Chengdu(GPE.NAM) Telecom(ORG.NAM) do you have no
countcept of time, delay again and again, mother(PER.NOM)
(curse word) our time is not time, waited for you for two
days what a speed.
Challenges of Obtaining Training Data

• Constructing data sets is labor intensive
• Many different
  – Languages
  – Domains
  – Modalities
  – ...

Outline

• Low resource NER by multi-task multi-domain learning.
• Low resource relation extraction by structured modeling.
• Cross-lingual NER by learning multi-lingual Embeddings.
Named Entity Recognition (NER)

- Identifying entities (in social media domain, usually person, organization, location and GPE) boundaries and their type from the plain text.

成都(GPE.NAM)电信(ORG.NAM)到底有没的时间观念哦，一托再托，日妈(PER.NOM)我们时间就不是时间哇，等了你两天啥子速度。

Chengdu(GPE.NAM) Telecom(ORG.NAM) do you have no concept of time, delay again and again, mother(PER.NOM) (curse word) our time is not time, waited for you for two days what a speed.
Neural Sequence Tagging Models

Linear Chain CRF for NER

RNNs for Feature Learning

Embeddings

Input Texts

End-to-end training
Challenges for Social NER

- HUGE gap on social media (noisy) v.s news text:
  - informal language and insufficient annotations.

![Bar chart showing F1 scores for English and Chinese social NER](chart.png)

Ma and Hovy, 2016
Cherry and Guo, 2015
Yu et al, 2008
Challenges for Social Media NER

• A *typical setting* of low-resource IE.
• A hard domain
  – Social media: *Informal language, many new words*
  – Chinese: *No word boundaries.*
• Lack labeled data
  – We initiated this task and collected annotations from *Amazon Mechanical Turk.*
  – 1350 sentences for training the models.
  – 540 sentences for development and evaluation.
Ideas

• Leverage existing resources to learn representations that generalize across multiple types of data.
  – Multi-task Learning.
  – Domain Adaptation
Multi-task Learning of Word Embeddings and Named Entity Recognition

Model for Learning Word Representations

Model for Named Entity Recognition
Multi-task Learning of Word Embeddings and Named Entity Recognition

Sharing parameters
Multi-task Learning of Word Embeddings and Named Entity Recognition

Skip-gram model to learn word representations

\[ L_u(X; e_x) = \frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log p(x_{t+j} | x_t) \]

\[ p(x_i | x_j) = \frac{\exp(e_{x_i}^T e_{x_j})}{\sum_{i'} \exp(e_{x_i'}^T e_{x_j})} \]

Log-bilinear CRF model for named entity recognition

\[ L_s(Y; X; \theta; e_x) = \frac{1}{T} \sum_{t=1}^{T} \log p(y_t | x_t) \]

\[ P(y | x) = \frac{1}{Z(x)} \prod_{t=1}^{T} \exp\{\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t, e_x)\} \]

2 millions of unannotated text for training

1350 NER annotated data for training

Peng and Dredze, EMNLP 2015
成都(GPE.NAM) 电信(ORG.NAM)到底有没的时间观念
哦，一托再托，日妈(PER.NOM)我们时间就不是时间哇，
等了你两天啥子速度。

Chengdu(GPE.NAM) Telecom(ORG.NAM) do you have no
concept of time, delay again and again, mother(PER.NOM)
(curse word) our time is not time, waited for you for two
days what a speed.
Multi-task Learning of Word Segmentation and Named Entity Recognition

Model for Chinese Word Segmentation

Model for Named Entity Recognition
Multi-task Learning of Word Segmentation and Named Entity Recognition

Neural-CRF for Chinese word segmentation

\[
L_s(Y \mid X; \lambda^s, \Psi) = \frac{1}{T} \sum_{t=1}^{T} \log p(y_t \mid x_t)
\]

\[
P(y \mid x) = \frac{\prod_{t=1}^{T} \exp\{ \sum_{k=1}^{K} \lambda^s_k f_k(y_t, y_{t-1}, \Psi(x)) \}}{Z(x)}
\]

40k word segmentation annotated data for training

Neural-CRF for named entity recognition

\[
L_n(Y \mid X; \lambda^n, \Psi) = \frac{1}{T} \sum_{t=1}^{T} \log p(y_t \mid x_t)
\]

\[
P(y \mid x) = \frac{\prod_{t=1}^{T} \exp\{ \sum_{k=1}^{K} \lambda^n_k f_k(y_t, y_{t-1}, \Psi(x)) \}}{Z(x)}
\]

1350 NER annotated data for training

Peng and Dredze, ACL 2016
Domains for Languages

McDonald’s Seeks Its Fast-Food Soul
- NYTimes 3/7/2015

Nivre and McDonald (2008) used the output of one dependency parser to provide features for another.
- Stacking Dependency Parsers, Martins+ (EMNLP 2008)
Multi-task Multi-domain Learning

Task Specific Models

Domain Projections

Shared Representation Learner

Input data $w^{(1)}$ ...... $w^{(n-1)}$ $w^{(n)}$
Multi-task Multi-domain learning for sequence tagging

- Domains: news and social media
- Tasks: Chinese word segmentation and NER
- Datasets:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Train</th>
<th>#Dev</th>
<th>#Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>News CWS</td>
<td>39,567</td>
<td>4,396</td>
<td>4,278</td>
</tr>
<tr>
<td>News NER</td>
<td>16,814</td>
<td>1,868</td>
<td>4,636</td>
</tr>
<tr>
<td>Social CWS</td>
<td>1,600</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Social NER</td>
<td>1,350</td>
<td>270</td>
<td>270</td>
</tr>
</tbody>
</table>

2 millions of unannotated weibo message for training
NER on Chinese Social Media

Named Entity Recognition on Chinese social media
Closing The Gap

Ours new Chinese Social NER system leveraging heterogeneous sources for representation learning

Peng and Dredze, ACL Dep4NLP 2017
Outline

• Low resource NER by transfer learning.
• Low resource relation extraction by structured modeling.
• Cross-lingual NER by learning multi-lingual Embeddings.
T790M is present as a minor clone in NSCLC, and may be selected for during therapy.

This mutation has been shown to prevent the activation of BIM in response to gefitinib but can be overcome by an irreversible inhibitor of EGFR.
Knowledge Bases for Drug-Gene-Mutation Interaction

• People manually curate drug-gene-mutation interaction databases for precision medicine:
  – Gene Drug Knowledge Database (GDKD) (Dienstmann et al., 2015)
  – Clinical Interpretations of Variants in Cancer (CiViC) (Washington University School of Medicine)
Special Challenges

• N-ary relations:
  – Traditional feature-based classification method usually use features defined on the shortest syntactic dependency paths between two entities.
  – Such features are hard to define in the N-ary case.

• Cross sentence relations:
  – Traditional features become sparser and learning becomes harder.
A Representation Learning Framework

\[
\begin{align*}
&\text{Contextual Entity Representation} \\
&\text{concatenation} \\
&\text{Relation Classifier} \\
&R_1 \quad \ldots \quad R_k
\end{align*}
\]

\[
\text{Graph LSTM} \quad \text{Representation Learner}
\]

\[
\text{Word Input Text} \quad w^{(1)} \quad \ldots \quad w^{(n-1)} \quad w^{(n)}
\]
Contextual Entity Representation

\[ w^{(1)}, \ldots, w^{(n-1)}, w^{(n)} \]

concatenation

Relation Classifier

\[ R_1, \ldots, R_k \]

Graph LSTM

Representation Learner

Word Input Text
Long-Short Term Memory Networks (LSTMs)

Capture *long-term dependencies* of the input.

*However*, it still only explicitly models the dependencies between the adjacent inputs.
Linguistics Structures for Input Texts

T790M is present as a minor clone in NSCLC,

This mutation has been shown to prevent the activation of BIM in response to getinib.
Directed Cyclic Graph
Graph Long Short-Term Memory Networks (Graph LSTMs)

• Goals:
  – *different types* of dependencies: adjacency, *syntactic* dependencies, *coreferences*, and *discourse* relations.
  – *long-distance* dependencies: entities span sentences.

• Challenges: how to define a neural architecture over a cyclic graph?
Work beyond Linear-Chain

- NLP: Tree LSTM (Tai et. al. ACL 2015)
- Programming verification: Gated Graph Neural Network (Li et. al. ICLR 2016)
- Graph Convolutional Networks (Kipf and Welling, ICLR 2017)
Challenge in Training

• Existing approach
  – Unroll recurrence for a number of steps
  – Analogous to loopy belief propagation (LBP)

• Problems
  – Expensive: Many steps per epoch
  – Information does not propagate from distant nodes
Training Graph LSTMs

- *Provably*, all directed cyclic graph without self-loop can be decomposed into two DAGs.

T790M is present as a minor clone in NSCLC
Training Graph LSTMs

• Approximate a cyclic graph by two directed acyclic graphs (DAGs), and stack the DAGs.

Topological order is well-defined, back propagation training
Chain LSTMs v.s. Graph LSTMs

Linear-chain LSTM

\[
\begin{align*}
    i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\
    o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\
    \tilde{c}_t &= \tanh(W_c x_t + U_c h_{t-1} + b_c) \\
    f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\
    c_t &= i_t \odot \tilde{c}_t + f_t \odot c_{t-1} \\
    h_t &= o_t \odot \tanh(c_t)
\end{align*}
\]

Graph LSTM (one DAG)

\[
\begin{align*}
    i_t &= \sigma(W_i x_t + \sum_{j \in P(t)} U^m_{i(t,j)} h_j + b_i) \\
    o_t &= \sigma(W_o x_t + \sum_{j \in P(t)} U^m_{o(t,j)} h_j + b_o) \\
    \tilde{c}_t &= \tanh(W_c x_t + \sum_{j \in P(t)} U^m_{c(t,j)} h_j + b_c) \\
    f_{tj} &= \sigma(W_f x_t + U^m_{f(t,j)} h_j + b_f) \\
    c_t &= i_t \odot \tilde{c}_t + \sum_{j \in P(t)} f_{tj} \odot c_j \\
    h_t &= o_t \odot \tanh(c_t)
\end{align*}
\]
Multi-task Learning

Pairwise

Get Entity Representation

Logistic Regression

concatenation

Pairwise

Get Entity Representation

Logistic Regression

concatenation

N-ary

Get Entity Representation

Logistic Regression

Shared Representation Learner

GraphLSTM

Input Embeddings

Word Input Text

\[ w^{(1)} \ldots w^{(n-1)} w^{(n)} \]
Domain: Molecular Tumor Board

- Ternary interaction: (drug, gene, mutation)
- Distant supervision
  - Knowledge bases: GDKD + CIVIC
  - Text: PubMed Central articles (~1 million full-text articles)
- We got 3,462 paragraphs about drug-gene-mutation relations from distant supervision.
Evaluation of Distant Supervision Relation Extraction

• Since it extracts totally new relations from the web
  – There is no gold set of correct instances of relations!
    • Can’t compute precision (don’t know which ones are correct)
    • Can’t compute recall (don’t know which ones were missed)

• Instead, we can approximate precision (only)
  – Draw a random sample of relations from output, check precision manually

\[
\hat{P} = \frac{\text{# of correctly extracted relations in the sample}}{\text{Total # of extracted relations in the sample}}
\]

• Can also compute precision at different levels of recall.
  – Precision for top 1000 new relations, top 10,000 new relations, top 100,000
  – In each case taking a random sample of that set

• But no way to evaluate recall. Instead, we do Absolute Recall
Absolute Recall

<table>
<thead>
<tr>
<th></th>
<th>Drug</th>
<th>Gene</th>
<th>Mutation</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>DGKD + CiViC</td>
<td>16</td>
<td>12</td>
<td>41</td>
<td>59</td>
</tr>
<tr>
<td>Single-Sent</td>
<td>68</td>
<td>228</td>
<td>221</td>
<td>530</td>
</tr>
<tr>
<td>Cross-Sent</td>
<td>103</td>
<td>512</td>
<td>445</td>
<td>1461</td>
</tr>
</tbody>
</table>

Numbers of *distinct* drugs, genes and mutations and their interactions in the knowledge bases vs. PubMed scale automatic extraction.

- Machine reading extracted orders of magnitudes more knowledge
- Cross-sentence extraction triples the yield
Sample Precision

Precision

Random
BiLSTM
Graph LSTM
Automatic Evaluation

- Logistic Regression
- CNN
- Linear LSTM
- Graph LSTM
Multi-Task Learning

Code and data available at: http://hanover.azurewebsites.net/

<table>
<thead>
<tr>
<th></th>
<th>Drug-Gene-Mutation</th>
<th>Drug-Mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph LSTM</td>
<td>80.7</td>
<td>76.7</td>
</tr>
<tr>
<td>+ Multi-task</td>
<td>82.0</td>
<td>78.5</td>
</tr>
</tbody>
</table>

Peng et. al. (TACL 2017)
Outline

• Low resource NER by transfer learning.
• Low resource relation extraction by structured modeling.
• Cross-lingual NER by learning multi-lingual Embeddings.
For America’s allies in Asia, the outcome of President Trump’s summit meeting with Kim Jong-un of North Korea has been decidedly mixed.

ORGANIZATION LOCATION PERSON
Low-Resource Language

What?
- Endangered
- Critical
  - Important for language understanding
- Low density
  - few computational data exist
  - few online data exist
  - low-affluence
- Less commonly taught

Why?
- Disaster
- Outbreak
- Surveillance

Challenges
- Low or No data in the language
- No native speakers (experts)

How?
- Rich in English;
- Language sources share partial common knowledge;

Transfer Learning

Where? (data sources)
- LORELEI
- REFLEX
Transfer Learning

Store knowledge gained while solving one problem and applying it to a different but related problem.

Learn from **Source** data  
Apply on **Target** data
## Data & Stats

<table>
<thead>
<tr>
<th>Task</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoNLL 2002 &amp; 03</td>
<td>English, German, Dutch, Spanish</td>
</tr>
<tr>
<td>LORELEI</td>
<td>Hindi, Tamil, Uzbek, Turkish, Russian, etc</td>
</tr>
<tr>
<td>WIKI</td>
<td>Chinese</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lang</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>13095</td>
<td>3035</td>
<td>3223</td>
</tr>
<tr>
<td>de</td>
<td>11599</td>
<td>2666</td>
<td>2850</td>
</tr>
<tr>
<td>es</td>
<td>8322</td>
<td>1914</td>
<td>1517</td>
</tr>
<tr>
<td>nl</td>
<td>15232</td>
<td>2747</td>
<td>4957</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lang</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>20040</td>
<td>5020</td>
<td>4941</td>
</tr>
<tr>
<td>de</td>
<td>9561</td>
<td>3822</td>
<td>3001</td>
</tr>
<tr>
<td>es</td>
<td>16609</td>
<td>3884</td>
<td>3214</td>
</tr>
<tr>
<td>nl</td>
<td>10002</td>
<td>1868</td>
<td>2751</td>
</tr>
</tbody>
</table>

# of Sentences  # of Entity
How to build NER for a new language using
(1) Comparable Corpora
(2) English NER tagger

Wang, Peng, and Duh (IJCNLP 2017)
Idea

Cross-lingual NER Tagger

Bilingual Word Embeddings

约翰·霍普金斯大学
约翰·霍普金斯大学 是一所主校区位于美国马里兰州巴尔的摩市的研究型私立大学。截止至2012年，共有36名校友获诺贝尔奖[3]。

Johns Hopkins University
Johns Hopkins University is an American private research university in Baltimore, Maryland. Founded in 1876, the university was named for its benefactor, the entrepreneur Johns Hopkins.[5]

English NER Tagger
Training Bilingual Word Embeddings

Word2Vec

Mixed-Language Pseudo-Document

Johns Hopkins University 大学 is an American private research university in Baltimore.

Johns 約翰·霍普金斯 Hopkins University 大學 is a private research university in Baltimore, Maryland. Founded in 1876, the university was named for its benefactor, the entrepreneur Johns Hopkins.

Vulic and Moens (ACL2015)
Training Cross-lingual NER Tagger

1. Fixed Embeddings
2. Multi-task training

Word2Vec Objective

Bilingual Word Embeddings

Cross-lingual NER Tagger

English NER Tagger

Johns Hopkins University is an American private research university in Baltimore, Maryland. Founded in 1876, the university was named for its benefactor, the entrepreneur Johns Hopkins.[5]
Training Cross-lingual NER Tagger

1. Fixed Embeddings
2. Multi-task training

Johns Hopkins University is an American private research university in Baltimore, Maryland. Founded in 1876, the university was named for its benefactor, the entrepreneur Johns Hopkins. [5]

English NER Tagger
Training Cross-lingual NER Tagger

1. Fixed Embeddings
2. Multi-task training

Word2Vec Objective

Bilingual Word Embeddings

Cross-lingual NER Tagger

Johns Hopkins University

Johns Hopkins University is an American private research university in Baltimore, Maryland. Founded in 1876, the university was named for its benefactor, the entrepreneur Johns Hopkins.\[5\]

English NER Tagger
Training Cross-lingual NER Tagger

1. Fixed Embeddings
2. Multi-task training

Word2Vec Objective
Bilingual Word Embeddings
Cross-lingual NER Tagger

Johns Hopkins University
American private research university in Baltimore, Maryland. Founded in 1876, the university was named for its benefactor, the entrepreneur Johns Hopkins. [5]
Training Cross-lingual NER Tagger

1. Fixed Embeddings
2. Multi-task training

Word2Vec Objective

Bilingual Word Embeddings

Cross-lingual NER Tagger

English NER Tagger

Johns Hopkins University

American private research university in Baltimore, Maryland. Founded in 1876, the university was named for its benefactor, the entrepreneur Johns Hopkins.[5]
2 Types of Multi-task Training

Word2Vec Objective

- Chinese
- English

Cross-lingual NER Tagger

1. Fixed Embeddings
2. Multi-task training
   (a) Update all embeddings
   (b) Update Chinese, but fix English embeddings

Johns Hopkins University

Johns Hopkins University is an American private research university in Baltimore, Maryland. Founded in 1876, the university was named for its benefactor, the entrepreneur Johns Hopkins. [5]
Experiment Setup

1 million words

19k document pairs (comparable corpora)

Test set:
1K Chinese sentence

Cross-lingual NER Tagger

45k sentences by StanfordNER

Johns Hopkins University

Johns Hopkins University is an American private research university in Baltimore, Maryland. Founded in 1876, the university was named for its benefactor, the entrepreneur Johns Hopkins. [5]
Multi-task training (update Chinese, fix English) is effective!

Embedding dimension

<table>
<thead>
<tr>
<th>Embedding dimension</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>128</td>
<td>128</td>
</tr>
<tr>
<td>256</td>
<td>256</td>
</tr>
<tr>
<td>512</td>
<td>512</td>
</tr>
</tbody>
</table>