Building linguistically informed models for low-resource settings

Nanyun (Violet) Peng

Collaborations with Hoifung Poon, Chris Quirk, Kristina Toutanova, Wen-tau Yih, Max Ma, Eduard Hovy, Wasi Uddin Ahmad, Zhisong Zhang, Kai-Wei Chang, Ralph Weischedel, Kevin Knight, Lili Yao, Rui Yan
Low Resource Languages

• ایتھوپیا کی

25 اکتوبر، سبھالے ورك زیوڈے صدر منتخب بوئین- سبھالے ورك بر اعظم افریقا کی کسی بهی ریاست

• অ্যাঞ্জেলিনা জোলিএকজন জনপ্রিয় মার্কনি চলচ্চিত্র অভিনেত্রী। যুক্তরাষ্ট্রেরক্‌ যালফিওরিনয়ির লস অ্যাঞ্জেলেসেরে একটা সংস্কৃতমনা পরিবারা এই অস্কারজয়ী
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.
Low Resource Tasks

• Creative Composition
  – Poetry
    Two roads diverged in a yellow wood,
    And sorry I could not travel both
    And be one traveler, long I stood ……
  – Pun
    The magician got so mad he pulled his hare out.
  – Story
    The last person on Earth was alone in a room. There was a knock on the door….
Challenges in Low-Resource Settings

• HUGE gap on social media (low-resource) v.s news (high-resource) text:
  • informal language and insufficient annotations.
Challenges of Obtaining Training Data

• Constructing data sets is labor intensive
• Many different
  – Languages
  – Domains
  – Tasks
  – …
Building Robust Models For Low-Resource Settings

• Cross-Sentence N-ary Relation Extraction for Biomedical Domain (low resource domain)

• On Difficulties of Cross-lingual Transfer (low resource languages)

• Plan-and-Write Story Generation (low resource task)
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.

Peng et. al. (TACL2017)
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

Contextual Entity Representation

concatenation

Relation Classifier

Representation Learner

Word Input Text

\( w^{(1)} \ldots w^{(n-1)} w^{(n)} \)
$w_1, \ldots, w_{n-1}, w_n$
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.

Picture credit: colah's blog, 2015
This mutation has been shown to prevent the activation of BIM in response to getinib.
Directed Cyclic Graph

\[ \begin{align*}
  h_0 & \rightarrow A & h_1 & \rightarrow A & h_2 & \rightarrow A & h_3 & \rightarrow A & h_4 & \rightarrow A & h_5 & \rightarrow A & h_6 & \rightarrow A & h_t & \rightarrow A \\
  x_0 & \rightarrow h_0 & x_1 & \rightarrow h_1 & x_2 & \rightarrow h_2 & x_3 & \rightarrow h_3 & x_4 & \rightarrow h_4 & x_5 & \rightarrow h_5 & x_6 & \rightarrow h_6 & & & x_t & \rightarrow h_t
\end{align*} \]
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?
Training Graph LSTMs

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

\[ T790M \text{ is present as a minor clone in NSCLC} \]
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)**

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

Pairwise
Get Entity Representation
concatenation
Logistic Regression

Pairwise
Get Entity Representation
concatenation
Logistic Regression

N-ary
Get Entity Representation
Logistic Regression

Shared Representation Learner

GraphLSTM

Input Embeddings
\( w^{(1)} \) \( \cdots \) \( w^{(n-1)} \) \( w^{(n)} \)

Word Input Text
\( w^{(1)} \) \( \cdots \) \( w^{(n-1)} \) \( w^{(n)} \)
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 is Hard

• 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)

• We can approximate precision
  – Draw a random sample of relations from output, check precision manually

• 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 (Automatic Eval)

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>

More results please see Peng et. al. (TACL2017)
Building Robust Models For Low-Resource Settings

• Cross-Sentence N-ary Relation Extraction for Biomedical Domain (low resource domain)

• On Difficulties of Cross-lingual Transfer (low resource languages)

• Plan-and-Write Story Generation (low resource task)
Standard Neural Architectures for NLP

\[ s = \{s_1, \ldots, s_n\} \]  An encoder to produce contextualized representations

\[ x = \{w_1, \ldots, w_n\} \]  Embeddings for the input sentence

\[ h = \{h_1, \ldots, h_n\} \]  A decoder that makes (structured) predictions

\[ y = \{p_1, \ldots, p_n\} \]

Popular encoder and decoder: RNNs

Ahmad et. al. 2018
Cross-Lingual Transfer Learning

\[ s = \{s_1, \ldots, s_n\} \]

An encoder to produce contextualized representations

\[ x = \{w_1, \ldots, w_n\} \]

Multi-lingual embeddings for the input sentence

\[ h = \{h_1, \ldots, h_n\} \]

A decoder that makes (structured) predictions

\[ y = \{p_1, \ldots, p_n\} \]

Popular encoder and decoder: RNNs
Cross-Lingual Transfer Learning

\( s = \{s_1, \ldots, s_n\} \) An encoder to produce contextualized representations

\( h = \{h_1, \ldots, h_n\} \) A decoder that makes (structured) predictions

\( x = \{w_1, \ldots, w_n\} \) Multi-lingual embeddings for the input sentence

\( y = \{p_1, \ldots, p_n\} \)

Popular encoder and decoder: RNNs
Are RNNs Good Encoders/Decoders for Cross-lingual Transfer?

- Overfitting to language-specific order information (our hypothesis)
- Verify an examine our hypothesis on cross-lingual dependency parsing
  - We have UD annotation for over 70 languages
  - Parser is a bottom-level task, directly reflect the problems
But there were no buyers

$\text{But there were no buyers}$
Background: Deep Biaffine Parser

- **Graph-based** parser
- **Encoder**: Order-sensitive; **Decoder**: Order-free

Dozat and Manning (ICLR2017)
Background: Stack-pointer Networks (StackPtr) Dependency Parsing

- Transition-based
- Order: Top-down, depth-first
- Actions: "Point" to the next word to choose as a child
- Encoder: Order-sensitive; Decoder: Order-dependent

Ma et. al. (ACL2018)
Background: Multi-Head Self-Attention

- In the original paper:

\[
PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right)
\]

\[
PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right)
\]

Vaswani et. al. (NIPS 2017)

- Relative positional embeddings:

\[
\begin{align*}
a^V_{2,1} &= w^{-1}^V & a^V_{2,4} &= w^V_2 \\
a^K_{2,1} &= w^{-1}^K & a^K_{2,4} &= w^K_2 \\
a^V_{4,n} &= w^V_k & a^K_{4,n} &= w^K_k
\end{align*}
\]

Shaw et. al. (NAACL2018)

Flexible positional encoding (order-free)
Architectures for Cross-lingual Parser

• Embedding
  - Facebook MUSE
  - MUSE

• Encoders
  - BiLSTMs (order-sensitive) v.s.
  - Multi-Head Self-Attention (order-free)

• Decoders
  - Pointer Network (order-sensitive) v.s.
  - BiAffine Attention (order-free)

*Conneau et. al. ICLR2018*
Experiments

• Datasets:
  – Universal Dependency Treebanks (V2.2)
  – 31 languages, 12 families

• Setting:
  – Train and develop on English
  – Directly predict on the rest 30 languages (zero-shot)
## Datasets Details

<table>
<thead>
<tr>
<th>Language Families</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afro-Asiatic</td>
<td>Arabic (ar), Hebrew (he)</td>
</tr>
<tr>
<td>Austronesian</td>
<td>Indonesian (id)</td>
</tr>
<tr>
<td>IE.Baltic</td>
<td>Latvian (lv)</td>
</tr>
<tr>
<td>IE.Germanic</td>
<td>Danish (da), Dutch (nl), English (en), German (de), Norwegian (no), Swedish (sv)</td>
</tr>
<tr>
<td>IE.Indic</td>
<td>Hindi (hi)</td>
</tr>
<tr>
<td>IE.Latin</td>
<td>Latin (la)</td>
</tr>
<tr>
<td>IE.Romance</td>
<td>Catalan (ca), French (fr), Italian (it), Portuguese (pt), Romanian (ro), Spanish (es)</td>
</tr>
<tr>
<td>IE.Slavic</td>
<td>Bulgarian (bg), Croatian (hr), Czech (cs), Polish (pl), Russian (ru), Slovak (sk), Slovenian (sl), Ukrainian (uk)</td>
</tr>
<tr>
<td>Japanese</td>
<td>Japanese (ja)</td>
</tr>
<tr>
<td>Korean</td>
<td>Korean (ko)</td>
</tr>
<tr>
<td>Sino-Tibetan</td>
<td>Chinese (zh)</td>
</tr>
<tr>
<td>Uralic</td>
<td>Estonian (et), Finnish (fi)</td>
</tr>
</tbody>
</table>
Augmented dependency label features:
- Triple (ModifierPOS, HeadPOS, DependencyLabel), e.g. (PRON, VERB, obj)
- Feature selection: exists in > 24 languages and with > 0.1% relative frequency
- Feature value: left (modifier before head) frequency and right (modifier after head) frequency
- 52 feature types (104 features) total.
Word Order Characterizes Language Distances
Word Order Characterizes Language Distances
Selected Transfer Results of Different Architectures

Zero-shot Transfer UAS Results (Except for English)

Distances to English increase, Transfer performances decrease.

Selected Transfer Results of Different Architectures

RNNStack
TransformerStack
RNNGraph
TransformerGraph

order-dependent
order-free
Overall Comparisons of Order-Free v.s. Order-Dependent Encoders/Decoders
Case Study -- (ADP, NOUN, case)
Case Study -- (ADJ, NOUN, amod)
Case Study -- (AUX, VERB, aux)
Overall Performances

Average UAS (Over 31 languages)

- **RNNStack**
- **TransformerStack**
- **RNNGraph**
- **TransformerGraph**
Building Robust Models For Low-Resource Settings

• Cross-Sentence N-ary Relation Extraction for Biomedical Domain (low resource domain)

• On Difficulties of Cross-lingual Transfer (low resource languages)

• Plan-and-Write Story Generation (low resource task)
Story Generation

• What are in a story?
  – Characters, key events, morals, conflicts, sentiment…

• We want to incorporate all the aspects
  – Unfortunately, even human do not have clear understanding about what’s in a story. There are few annotations.

• Analyzing stories to generate stories with minimal or no supervision.

Yao & Peng et. al. (AAAI2019)
Problem of (Neural) Story Generation

- **Title:** bicycle path accident
  - **Generated Story:** sam bought a new bicycle. his bicycle was in an accident. his bicycle was in an accident. his bicycle was in an accident. his bicycle was totaled.
- **Title:** darth vader on earth
  - **Generated Story:** it was a very windy day. i 've never been to it before. i do not know what to do. i do not know what to do. i think it is a good idea.
Plan-and-Write Hierarchical Generation

• Can computer generate storylines automatically (given titles)?
  – Equip our system with the ability to model “what happens next”.
  – Mimic human writers’ common practice of writing sketches: have a big picture.
  – Computer and human can interactively modify the storylines, more fun interactions.
Interactive Generation Task

Title

Storyline

The Whole Story
Extracting Storylines

• The ROCStories dataset: 98,161 turker-written five-line stories with titles.

• Extract one word or phrase from one sentence using RAKE algorithm proposed in the IR community (2010).

• Use the word/phrase sequence as an approximation of the storyline.
Examples

Title: christmas shopping
Story: frankie had christmas shopping to do.
she went to the store.
inside, she walked around looking for gifts.
soon her cart was full.
she paid and took her things home.

Storyline (unsupervised extraction): frankie store gifts cart paid

Title: farm
Story: bogart lived on a farm.
he loved bacon.
he decided to buy a pig.
shortly after, he grew fond of the pig.
bogart stopped eating bacon.

Storyline (unsupervised extraction): farm bacon decided pig bogart
The **planning** component generates storylines from titles. The **writing** component generates stories from storylines and titles.

Tina made spaghetti for her boyfriend.

……

……

……
Dynamic and Static Schemas

Dynamic Schema

\[ \text{ctx} = [t; s_{1:i-1}] \]

At the plan step, we model: \( P(li|\text{ctx}, l_{1:i-1}) \)

At the write step, we model: \( P(s_i|\text{ctx}, l_{1:i}) \)

The probabilities are computed by some specifically designed fusion-RNN cells.

Static Schema

At the plan step, we model: \( P(li|t, l_{i-1}) \)

At the write step, we model: \( P(s_i|\text{ctx}, l_{1:i}) \)

The probabilities are computed by standard language models and sequence to sequence with attention models.
Some Observations

- Plan-and-Write strategies generate more interesting, less repetitive stories.
- Plan-and-Write strategies generate more on-topic stories.
- Static strategy works better than dynamic strategy.
**Generation Results**

**Without Storyline Planning**

**Title:** gymnastics  
**Story (generated):**  
i wanted to learn how to draw.  
so, i decided to go to the gym.  
i went to the local gym.  
i got a lot of good grades.  
i was very happy.

**With Storyline Planning**

**Title:** gymnastics  
**Storyline (generated):** wanted  
decided class practiced well  
**Story (generated):**  
i **wanted** to be a gymnast.  
i **decided** to learn how to do gymnastics.  
i decided to take a **class**.  
i **practiced** every day.  
i was able to do **well** on the **class**.
Without Storyline Planning

**Title:** rock jumping

**Story (generated):**
i was at the park with my friends.
i was playing with my friends.
i tripped over a rock.
i fell on the ground.

With Storyline Planning

**Title:** rock jumping

**Storyline (generated):** day
decided jumped fell broke

**Story (generated):**
one day , i decided to go rock jumping.
i jumped and fell.
i fell and broke my ankle.
i had to go to the hospital.
i learned to be more careful next time .
Quantitative Results on Repetition

Inter- and intra-story tri-grams repetition rates by sentences (curves) and for the whole stories (bars), the lower the better. We also conduct the same computation for four and five-grams and observed the same trends. As reference points, the whole story repetition rates on the human-written training data are 34% and 0.3% for the inter- and intra-story measurements respectively.
## User Preferences

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Dynamic v.s. Inc-S2S</th>
<th>Static v.s. Cond-LM</th>
<th>Static v.s. Dynamic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fidelity</td>
<td>35.8%</td>
<td>12.9%</td>
<td>0.42</td>
</tr>
<tr>
<td>Coherence</td>
<td>37.2%</td>
<td>28.6%</td>
<td>0.30</td>
</tr>
<tr>
<td>Interesting</td>
<td>43.5%</td>
<td>26.7%</td>
<td>0.31</td>
</tr>
<tr>
<td>Overall</td>
<td>42.9%</td>
<td>27.0%</td>
<td>0.34</td>
</tr>
</tbody>
</table>

The human study is conducted on Amazon Mechanical Turk (AMT). 233 users were participated in the study.
Analysis

Figure: The regression coefficient that shows which aspect is more important in human evaluation of stories.

Table: The storyline BLEU score (only BLEU-1 and BLEU-2) and the correlation between storylines and stories.

<table>
<thead>
<tr>
<th>Method</th>
<th>B1</th>
<th>B2</th>
<th>L-S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic</td>
<td>6.46</td>
<td>0.79</td>
<td>0.88</td>
</tr>
<tr>
<td>Static</td>
<td>9.53</td>
<td>1.59</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Yao & Peng et. al. (AAAI2019)
## Lingering Problems

Yao & Peng et. al. (AAAI/2019)

<table>
<thead>
<tr>
<th>Title (problem)</th>
<th>Generated Storyline</th>
<th>Generated Story</th>
</tr>
</thead>
<tbody>
<tr>
<td>pulled over (off-topic and strange language)</td>
<td>driving road car hit time</td>
<td>tom was driving to work. he was driving down the road. a car hit him in the car. tom hit the car. he had to get a new car.</td>
</tr>
<tr>
<td>saving money (inconsistent)</td>
<td>money wanted decided job pay</td>
<td>i had a lot of money. i wanted to save money. i decided to get a job. i got a job at a local bank. i was able to pay my bills.</td>
</tr>
<tr>
<td>my nephew’s vacation (off-topic and inconsistent)</td>
<td>family vacation trip decided time</td>
<td>i went on vacation with my family. we went on a vacation. we had a great time. we decided to go on a trip. we had a great time.</td>
</tr>
</tbody>
</table>
Summary

• Low-resource settings are everywhere in NLP
  – Low-resource languages
  – Low-resource domains
  – Low-resource tasks

• Need more robust/generalizable and intelligent NLP and machine learning algorithms

• Creative composition is a fun and challenging task!

• Alexa storytelling demo
Thanks!

Questions?

Demo: http://cwc-story.isi.edu/