Improving Named Entity Recognition for Chinese Social Media with Word Segmentation Representation Learning

Nanyun Peng, Mark Dredze
Human Language Technology Center of Excellence, Center for Language and Speech Processing
Johns Hopkins University, Baltimore, Maryland USA
Bloomberg LP, New York, NY

Named Entity Recognition

- **Detecting** boundaries and **classifying** types of text chunks that correspond to entities:
  - *persons, organizations, locations*

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) fxxxer our time is not time, waited for you for two days what a speed..

Special Challenges of *Chinese Social Media*: *no word boundaries;* no in-domain training data for word segmentation.

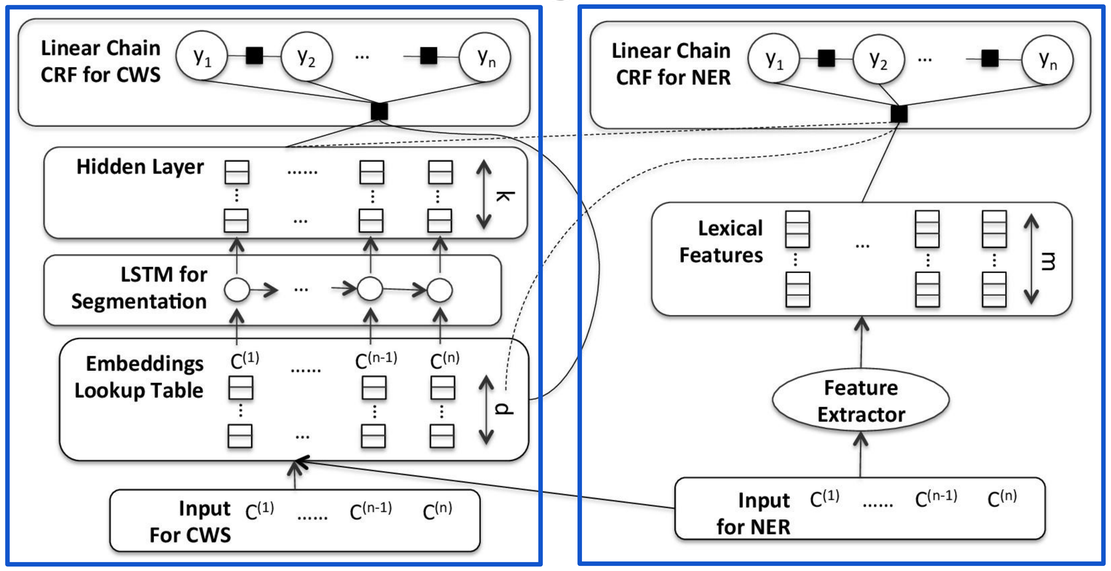
Chinese Word Segmentation

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 / 。

Detecting word boundaries. A *pre-processing step* for Chinese language processing tasks.
Helps NER: NE boundaries are *consistent* with word boundaries.
Domain mis-match: No annotations for Chinese word segmentation in *social media*, but annotations in *news domain* are abundant. ⇒ Instead of using pre-trained CWS models, we do *multi-task joint training*.

Code and data available at:
https://github.com/hltcoe/goldenhorse

A Multi-task Learning Framework



A joint model for Chinese word segmentation (left) and NER (right). The CRF for NER (upper left) has access to both the NER feature extractor and the representations produced by the LSTM for word segmentation. Learning propagates gradients back to both the LSTM and feature weights to adjust parameters.

Multi-task Joint Training

Segmentation Module:

$$\mathcal{L}_s(\boldsymbol{y}_s; \boldsymbol{x}_s, \Theta) = \frac{1}{K}\sum_k \Big[\log \frac{1}{Z(\boldsymbol{x}_s)^k} + \sum_i \Big(T_s(y^k_{i-1}, y^k_i) + s(y^k_i; \boldsymbol{x}^k_s, \boldsymbol{\Lambda}_s)\Big)\Big]$$

NER Module:

$$\mathcal{L}_n(\boldsymbol{y}_n; \boldsymbol{x}_n, \Theta) = \frac{1}{K}\sum_k \Big[\log \frac{1}{Z(\boldsymbol{x}_n)^k} + \sum_j \Lambda_j F_j(\boldsymbol{y}^k_n, \boldsymbol{x}^k_n, \boldsymbol{e}_w, \boldsymbol{h}_w)\Big],$$

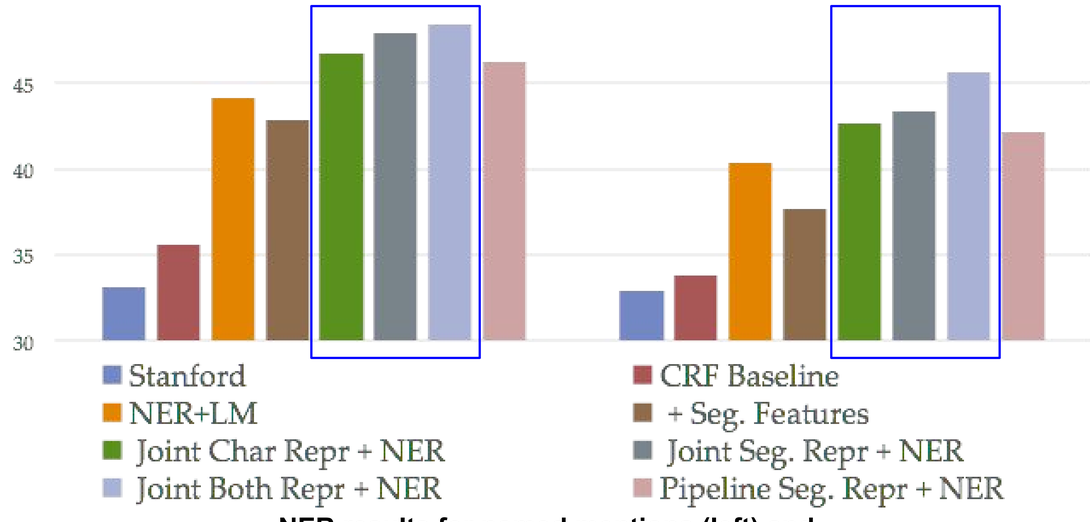
Multi-task Joint Learning:

$$\mathcal{L}_{joint}(\Theta) = \lambda \mathcal{L}_s(\boldsymbol{y}_s; \boldsymbol{x}_s, \Theta) + \mathcal{L}_n(\boldsymbol{y}_n; \boldsymbol{x}_n, \Theta)$$

Datasets

Segmentation: SIGHAN 2005 shared task, PKU portion, 43,963 sentences for training and 4,278 sentences as dev.
Weibo NER: 1890 Weibo messages annotated with named entities and nominal mentions.

Entity Type	Mentions		
	Name	Nominal	Total
Geo-political	243	0	243
Location	88	38	126
Organization	224	31	255
Person	721	636	1,357



NER results for named mentions (left) and name + nominal mentions (right) on weibo data.