The TAI System for Trilingual Entity Discovery and Linking Track in TAC KBP 2017

Tao Yang, Dong Du and Feng Zhang
Tencent AI Platform Department

Outline

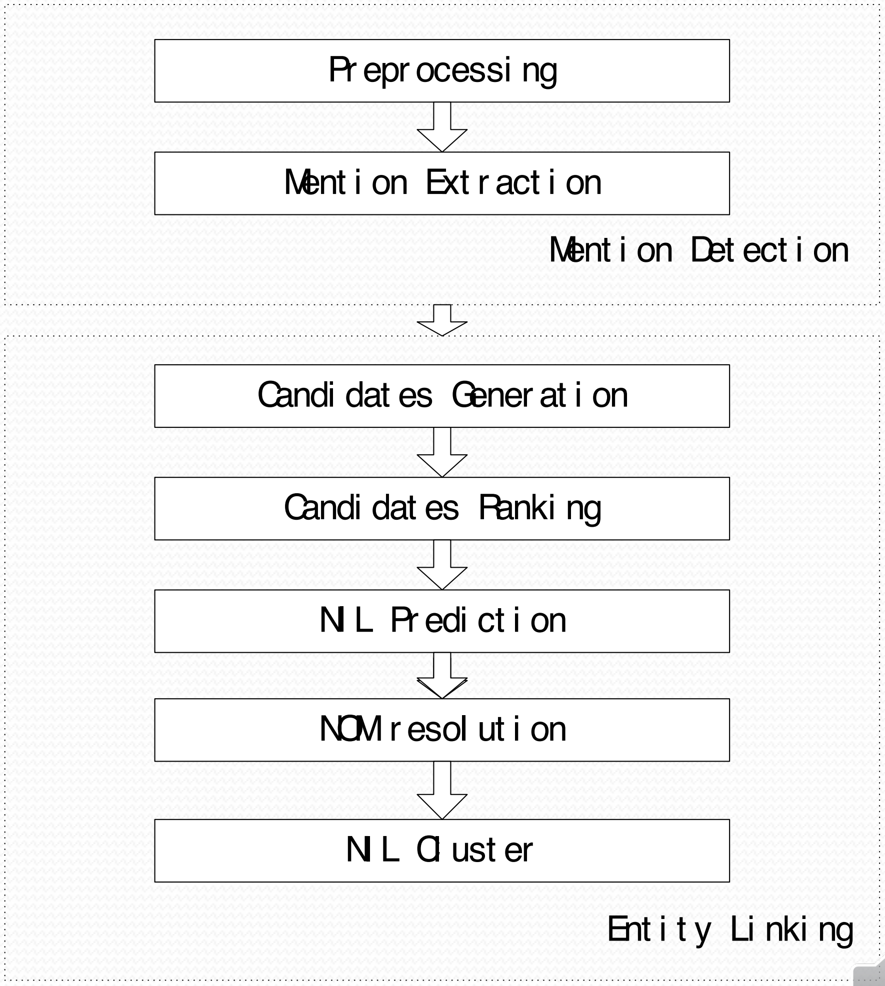
- Task Description
- The TAI System
  - Mention Detection
  - Entity Linking
- Results

Task Description

- Mention extraction and entity linking in three languages: Chinese, English and Spanish.
  - BaseKB as the target knowledge base
  - Two types of documents: newswire and discussion forum
  - Five entity types: PER, LOC, ORG, GPE, FAC
  - Two mention types: named (NAM) and nominal (NOM)
  - Cluster NIL mentions

The framwork of TAI System

- Two sub-systems
  - Mention Detection
    - Pre-processing
    - Mention extraction
  - Entity Linking
    - Candidates generation
    - Candidates ranking
    - NIL prediction
    - NOM Resolution
    - NIL Cluster

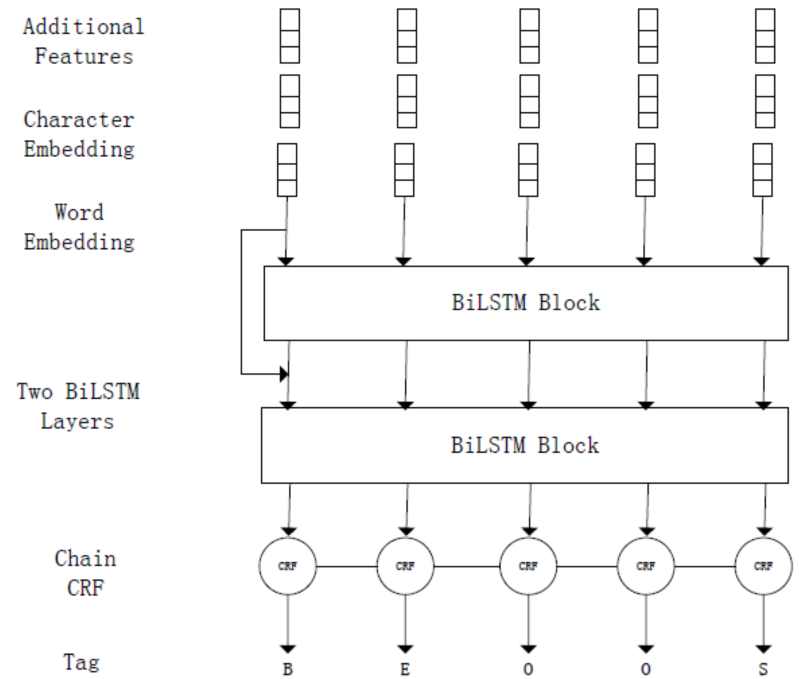


Mention Detection

- **Preprocessing**
  - Remove XML tags
  - Remove URLs and quote texts from the discussion forum
  - Convert traditional characters to simplified characters for Chinese
  - Extract the authors from newswire and discussion forum
  - Tokenize English and Spanish texts using CoreNLP tool
  - Character sequence instead of word sequence for Chinese

Mention Detection

- **Architecture**
  - Sequence labeling problem
  - Two-layers stacked BiLSTM + CRF model
  - Skip connections
  - Ensemble of two models
  - Multiple types of features
    - word embedding
    - character embedding
    - additional Features



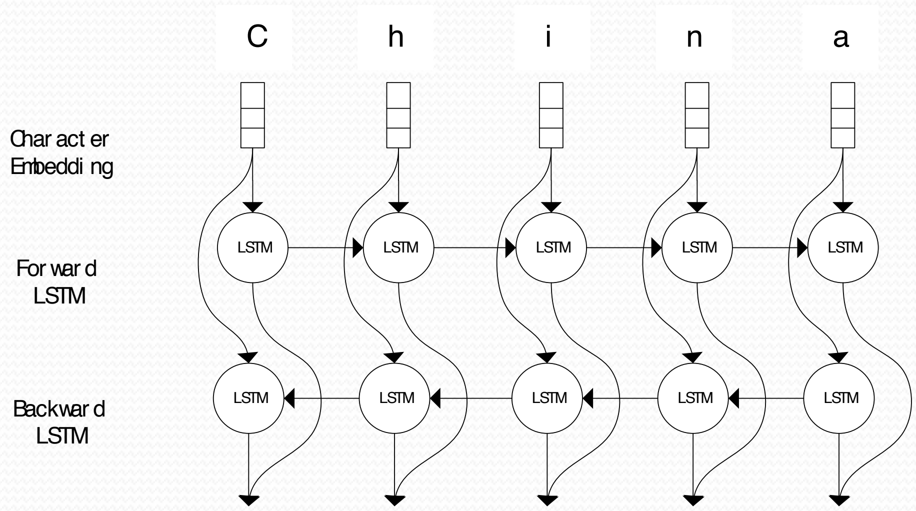
Mention Detection

- **Word Embedding Feature**
  - Pre-training from the Gigawords data
  - Training tool is wang2vec[1]
  - For Chinese, the character embeddings are enhanced by the positional character embeddings[2]

[1] Wang Ling etc. 2015. Two/too simple adaptations of word2vec for syntax problems.
[2] Xinxiong Chen etic. 2015. Joint learning of character and word embeddings

Mention Detection

- **Character Embedding**
  - Another BiLSTM to generate the character embeddings
    - Solve the out of vocabulary (OOV) problem
    - Model the word's prefix and suffix feature



Mention Detection

- **Additional Features**
  - Dictionary feature: collected entities from Wikipedia and Baike.
  - POS and NER feature: the POS and NER results produced by CoreNLP and QQseg.
  - Word boundary feature: indicates whether current Chinese character is at the word's boundary or inside the word.
  - NOM's feature: NOM mention's previous word

Entity Linking

- **Candidates generation**
  - Generate entities' aliases
    - BaseKB entities' name
    - Wikipedia's page title
    - Wikipedia's anchors
    - Wikipedia's disambiguate pages
    - Google translation service
    - Split the person's name
    - Baike aliases resource
  - Generate mention's candidate
    - Search the alias-to-entities dictionary, exact and fuzzy matching
    - Whole document searching for substring matching: such as "Bush" and "George Bush"

Entity Linking

- **Candidates Ranking**
  - Model: Pair-wise learning to rank model, called LambdaMART
    - The target entity should be ranked higher than any other entities.
  - Features:
    - Popular features
    - Type features
    - Matching features between context and entity
    - Semantic relatedness features

Entity Linking

- **Candidates Ranking - Popular Features**
  - Page rank score based on the Wikipedia's anchors
  - Page rank score based on the BaseKB
  - Wikipedia pages' language number

  - Mention linking probability

$$link\_prob(m,c) = \frac{count(m,c)}{\sum_{c'} count(m,c')}$$

Entity Linking

- **Candidates Ranking - Types Features**
  - Document types: NW or DF
  - Mention's entity types: PER, LOC, ORG, FAC and GPE
  - BaseKB's entity types

organization.organization
location.location
geography
location.country
location.administrative
division
location.statistical_region
people.person
architecture.structure
government.governmental_body
base.newsevents.news_reporting_organisation
government.government
government.legislative_committee
aviation.airport
education.educational_institution
base.prison.prison
government.governmental_jurisdiction

Table 1: The selected entity type in BaseKB as EL ranking features.

Entity Linking

- **Candidates Ranking - Matching features**
  - Word similarity between the entity and the context based on bag of words
  - Semantic similarity between the entity and the context based on DSSM model[1]
    - The framework of DSSM model is shown in figure 1.
    - Pre-training using the Wikipedia's anchors, and fine-tune using the training data
    - Pair-wise loss function:

$$L = max\{0, M - (cos(e_t, c) - (cos(e_i, c)))\}$$

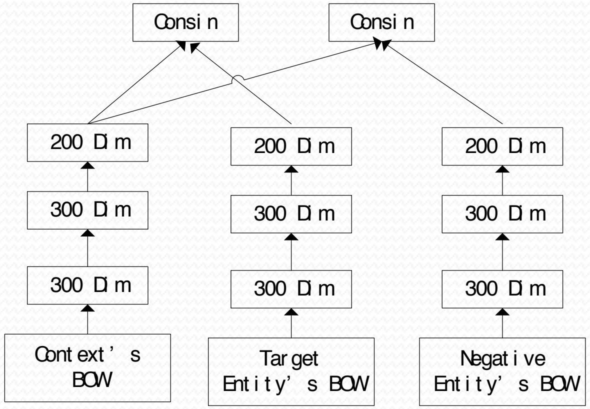


figure 1 framework of DSSM

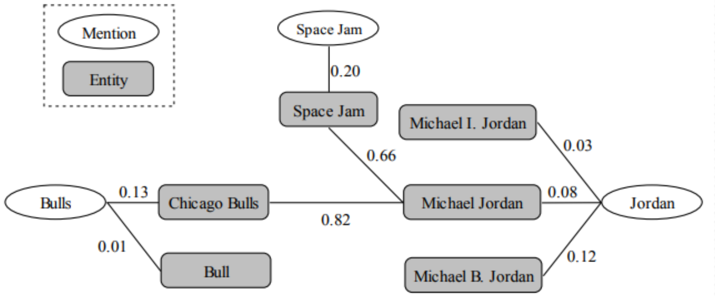
[1] Po-Sen Huang etc. 2013. Learning deep structured semantic models for web search using clickthrough data.

Entity Linking

- **Candidates Ranking - Semantic Relatedness Features**
  - Max WLM score between current entity and the other mentions' candidate entities

$$WLM(e_1, e_2) = 1 - \frac{log(max(|S(e_1)|, |S(e_2)|) - max(|S(e_1) \bigcap S(e_2)|))}{log(|W|) - log(min(|S(e_1)|, |S(e_2)|))}$$

  - Global coherent score[1]
    - Graph-based method
    - Mention-to-entity and entity-to-entity edges
    - Bag of words cosine and WLM score
    - Personalized page rank to resovle



[1] Xianpei Han etc. 2011. Collective entity linking in web text: a graph-based method.

Entity Linking

- **NIL Prediction:**
  - Motivation:
    - The top ranked entity may be not right
  - Model:
    - A binary classification is trained to make the decision
  - Features:
    - All the ranking model's features
    - Ranking score
    - Differential between 1^{st} and 2^{nd} score
    - Differential between the 1^{st} and mean score
    - Standard deviation of all the scores

Entity Linking

- **NOM resolution**
  - Link the mentions in the pre-compiled dictionary directly, such as “中方(Chinese Government)”
  - Link to the named mention with most occurring times in the document, such as “Country”
  - Link to the neatest named mention with the same type
  - For each pair <m_{nom}, m_{nam}>, a simple binary classification model is trained to classify whether m_{nom} can link to target m_{nam}, where m_{nam} is a named mention in m_{nom}’ context.

Entity Linking

- **NIL Cluster**
  - Authors and Body's mentions are clustered altogether
  - Clustering mentions in the same document, if mention span is the same
  - Clustering partial match mentions, if they are PER types
  - Special rules, such as "楼主" in Chinese discussion forum texts, always cluster it with the first author

Results

- **The trilingual results of our best run(according to the typed_mention_ceaf):**

strong_typed_mention_ceaf			strong_typed_all_match			typed_mention_ceaf		
Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
85.0	68.6	75.9	76.0	61.3	67.8	79.0	63.7	70.5

- **Conclusion**
  - Our system achieved competitive results
  - Nominal mentions' detection and linking is much harder than named mentions', need to try more complicated models or incorporate more features
  - NIL clustering is mainly based on rules, further exploration is needed

Thank you!

Q&A

rigorosyyang@tencent.com

Tencent AI Platform Department