CMU LTI @ KBP 2016 Event Track

Zhengzhong Liu

Jun Araki, Teruko Mitamura, Eduard Hovy

Language Technologies Institute
Carnegie Mellon University

And why the Chinese track is hard， and what can we do?

A Brief Introduction of the Models

Event Nugget Detection

1. We first use similar CRF model from last year.
    a. Participates in English and Chinese
2. We try a Neural Network model
    a. Participates in English

Guess how many tokens in this sentence actually annotated?

Mention Detection Feature Types

Freeman and his now ex-wife, Myrna Colley-Lee, had separated in December 2007 after 26 years of marriage.

	Lexical	Automatic Clusters	Hand-made Clusters
Trigger Head	“separate”	Brown Cluster ID Word Embedding POS tag	WordNet Hypernym
Trigger Context	Syntactic child head word	Entity Type in Context	WordNet Hypernym of context
Trigger Argument	SRL role head word	Entity Type of the argument head. Brown Cluster of the argument head.	Frame Net Role Name

Mention Detection Features

1. Main criticism: hand-crafted features
   a. Time consuming
   b. Need domain knowledge -> The exact reason that we don’t have a Spanish version.
2. Other criticism:
   a. May cause overfit.
3. Pros?
   a. Easy to work
   b. Easy to understood
   c. Resources for certain languages are sufficient
   d. Time consumption is reasonable

Resources Used

English:

1. Brown Cluster on TDT5
2. Frame Net (Parsed by Semafor)
3. PropBank (Parsed by Fanse)
4. Word Net

Chinese:

1. Brown Clusters on Gigaword
2. Synonym Dictionary *
3. SRL *

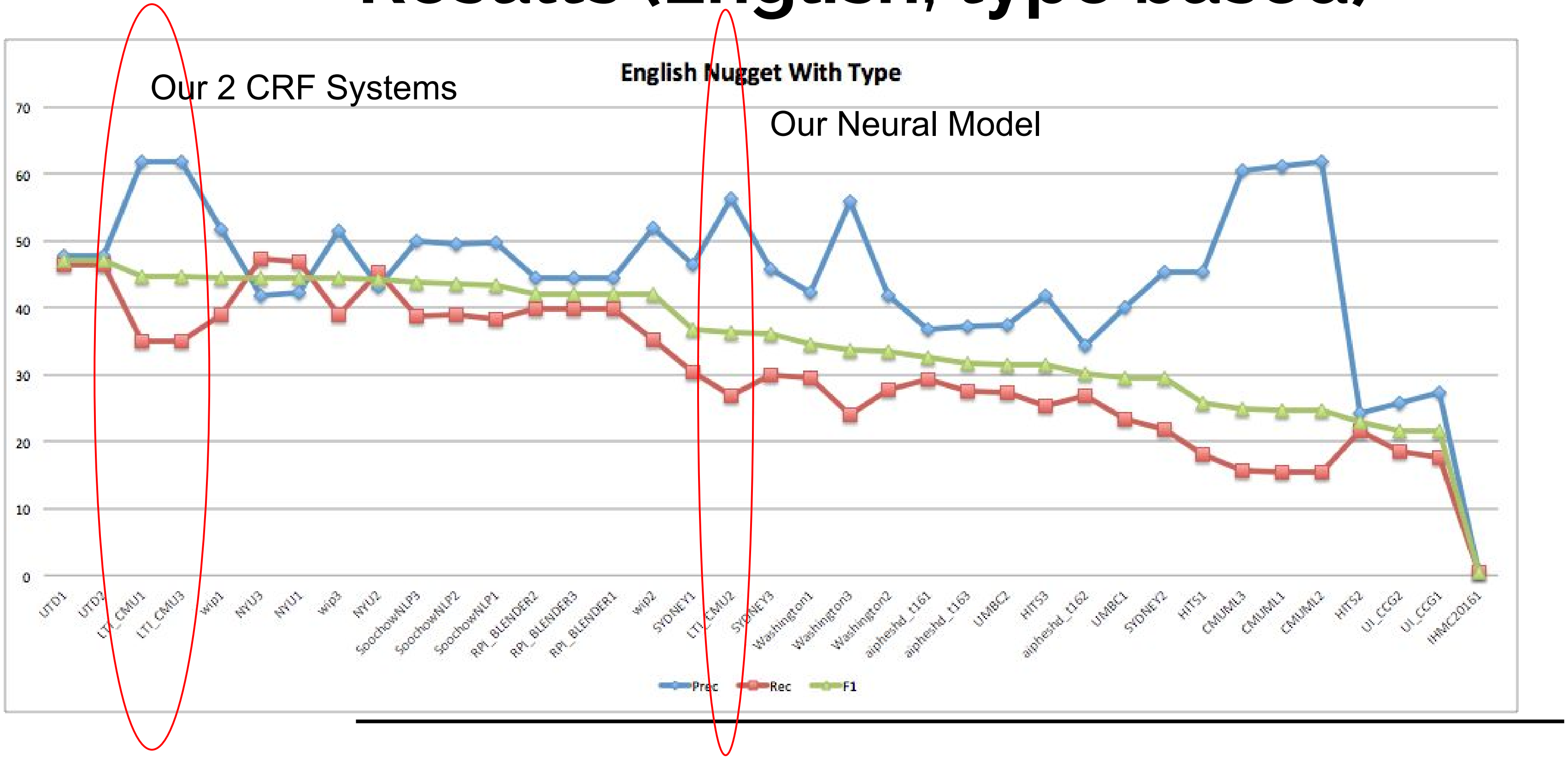
* From the LTP project by HIT

Neural Network Models

Argument structure is very important in nugget detection, will that help here? We haven’t tested that yet.

1. We adopt a bidirectional GRU
2. Trained on ACE corpus with Adam
3. Use and update pre-trained word embeddings (GloVe)
4. Pros?
   a. Relatively less resources needed : only pre-trained word vectors
   b. Less domain knowledge required
5. Cons?
   a. Cannot interpret weights: why it did well?
   b. Can a RNN model actually capture all kinds of information we needed?

Results (English, type based)



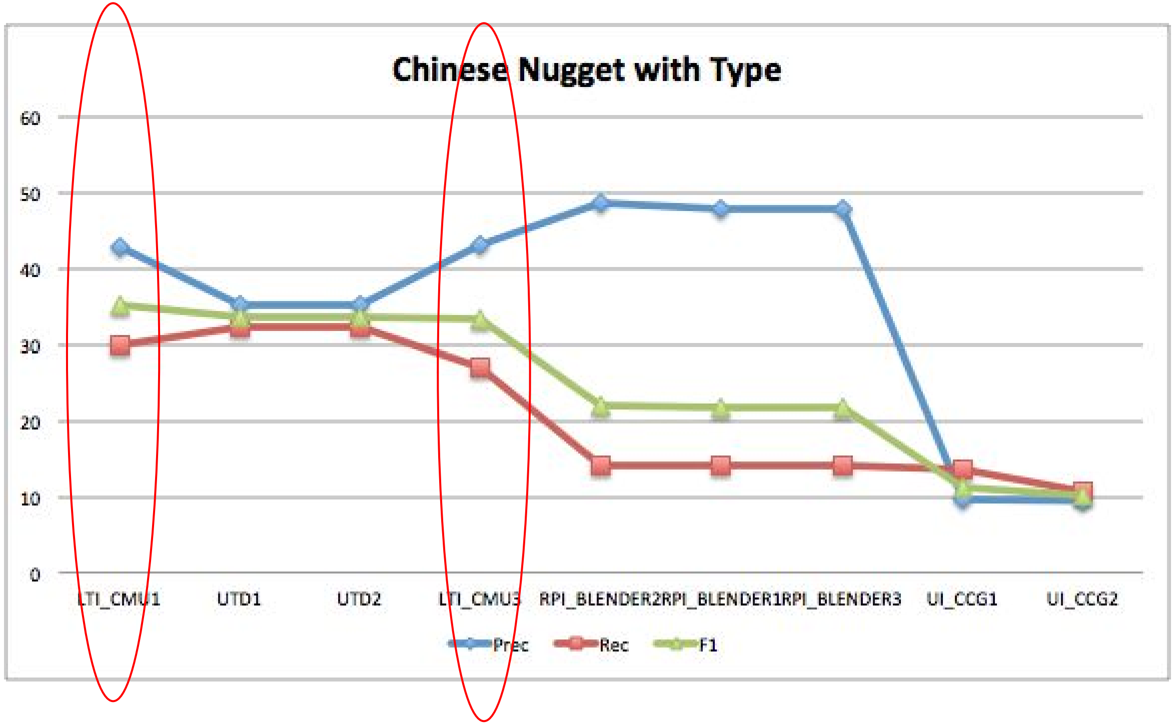
English Nugget With Type

Our 2 CRF Systems

Our Neural Model

Results (Chinese, type based)

Our 2 CRF Systems



Specific Features for Chinese Nugget

1. Chinese words can be easily combined with additional tokens to create new word, which may not be taggable:
   a. **侵略** 者 (invade + ~er = invader)
   b. **选举** 权 (election + ~right = election right)
2. We add features to see if the token modify anything.

Specific Features for Chinese Nugget

1. Chinese Character can have some important semantics
2. We use the a character level parsing to find out the Head Character for a verb
   a. 报告（报and告 are both base verb）
   b. 解雇（雇is base)

A note on Chinese Nuggets

1. We have suffered from a low recall problem in Chinese for quite a long time.
    a. We first simply add in features
2. We realize that it is the inconsistency in annotation cause the problem.
3. Also, the ambiguous single character mentions make the problem more serious

Some Examples

- 支持香港同胞争取[Personnel.Elect 选举]与 被[Personnel.Elect 选举]权!
- 司务长都是骑着二八去[TransferOwnership 买]菜 去。
- 海豹行动是绝密，塔利班竟然可以预先得 知?用个火箭就可以[Conflict.Attack打]下来，这个难度也实在是太高了吧。

TOP ERE Nugget Surface

1. Single token nuggets are very popular
2. These nuggets are very ambiguous
3. You can also see that most of them do not have an annotated rate of more than 50%.
4. In ACE 2005, top mentions are mostly 2-character mentions.

Event	Count	Actual	%				
打	170	593	28.67%	买	34	92	36.96%
说	148	949	15.60%	到	34	826	4.12%
死	131	410	31.95%	送	30	121	24.79%
杀	118	451	26.16%	击	28	329	8.51%
战争	96	223	43.05%	战	27	642	4.21%
占	55	189	29.10%	卖	24	94	25.53%
去	39	455	8.57%	死亡	24	33	72.73%

Our Solution (Or just hacks)

For the noisy annotation:

1. Probably the best thing to do is data clean up.
2. We use a heuristic that remove all Chinese sentences without nugget annotated
    a. Annotators are less likely to make mistakes when looking at one sentence
3. This improve the performance by 3 to 5 F1.

For single character nugget:

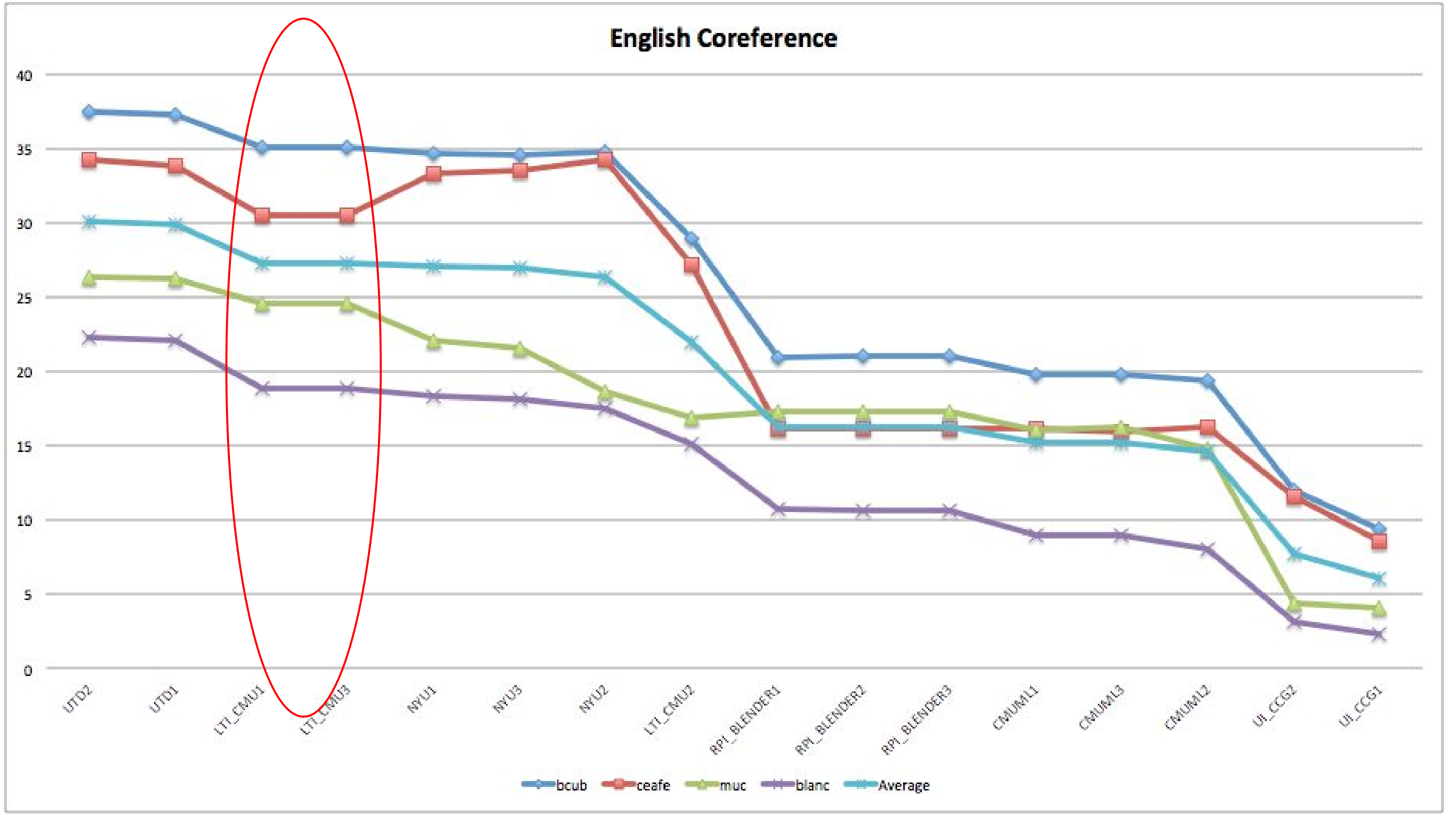
1. Argument is normally the main point for distinguishing.
2. Design features focusing on the argument.
3. We haven't assessed the impact of these features yet, but from development set, we see a couple F1 score improvement.

Event Coreference Model

Similarly, we need to migrate our English features to Chinese like what we did for event detection.

1. We continue use the Latent Antecedent Tree model
   a. A simple incremental antecedent selection model
   b. The key is that the update is done by comparing the predicted tree against one of the gold tree.
2. With regular matching features
   a. Trigger Match
   b. Argument Match
3. And some discourse clues
   a. Distance
   b. Structure of the forum (such as quotes)

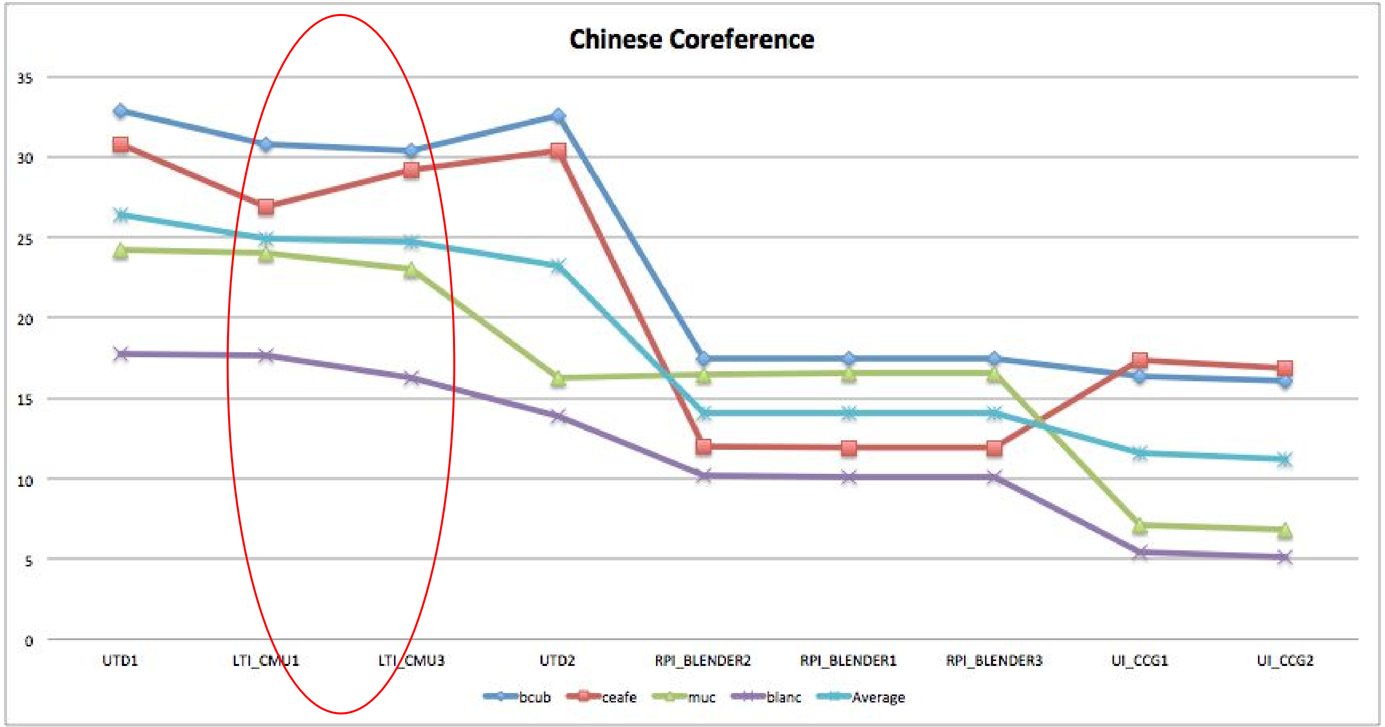
English Coreference



Chinese Coreference

Coreference performance is largely bottlenecked by Nugget Detection.

By manually inspecting the output, often the mentions in the coreference clusters are not event found in the first place.



Joint Decoding Not Helping?

1. We jointly decode the nugget detection CRF system with the latent tree coreference system.
2. We use Dual Decomposition to add constraints:
   a. When coreference, the mention type must be the same.
   b. Using binary variable **y(i,t)** to denote index **i** is of type **t** (=1) or not (=0).
   c. Using binary variable **z(i,j)** to denote index **i** and **j** are coreferent (=1) or not (=0)
   d. $y(i,t) - y(j,t) + z(i,j) - 1 <= 0$
3. We observe little performance gain because coreference links seems to rely too much on mention type.

We instead consider Joint Learning that consider the interaction of mention detection and coreference to be more fruitful.

We currently work on a model similar to Daumé & Marcu (2009) on joint NER and Entity Coreference, with a new approach to promote diversity.

The Chinese Challenge?
The Event Challenge.

More Data Problems

1. English and Spanish may suffer from the same annotation problem.
2. More importantly, the annotated data is always small and restricted.
3. Root causes:
    a. Event structures are complex and difficult to annotate.
    b. Deeper semantic understand may be required.

Current Paradigm

1. Annotate small set -> Train on small set -> Test
2. Annotation is difficult, and the training data is also not sufficient
3. For example, the nugget/coreference performance of this year has little improvement over last year:
    a. We are still doing surface level matching
4. However, there are interesting and difficult problems to think about:
    a. E.g. Why does two event mention coref when the arguments are not coreferent?

We need new paradigm

前苏联自1959年至1976年，先后十余次无人探测器“月球号”登临月球，据说1970年9月12日**发射**的月球16号，9月20日在月面丰富海软着陆，第一次使用钻头**采集**了120克月岩样口，装入回收舱的密封容器里，于24日带回地球。

Some missing annotations from the test set.

1. People have make progress on predicting event nuggets with small amount of supervision:
   a. Lifu Huang, Taylor Cassidy, Xiaocheng Feng, Heng Ji, Clare R Voss, Jiawei Han, and Avirup Sil. 2016. Liberal Event Extraction and Event Schema Induction. In *ACL 2016*.
   b. Haoruo Peng, Yangqi Song, and Dan Roth. 2016. Event Detection and Co-reference with Minimal Supervision. In *EMNLP 2016*.
2. However, the evaluation scheme do not favor these methods
   a. If annotators have biases over certain event nugget surface.
   b. Other nuggets may not get their credits.