Input Representation for Parsing Discourse Representation Structures Comparing English with Chinese

Chunliu Wang, Rik van Noord, Arianna Bisazza , Johan Bos
University of Groningen
chunliu.wang@rug.nl, r.i.k.Noord@rug.nl, a.bisazza@rug.nl, johan.bos@rug.nl

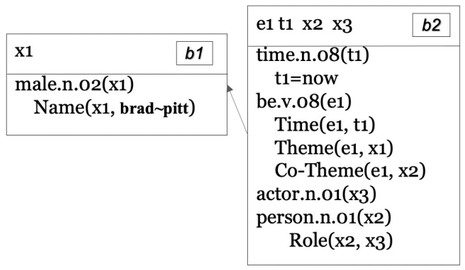
Questions

1. Can existing DRS parsing models achieve good results for Chinese?

2. Given the different writing systems used for English and Chinese, which input granularity is best for either language?

3. Is rule-based word segmentation beneficial for Chinese DRS parsing?

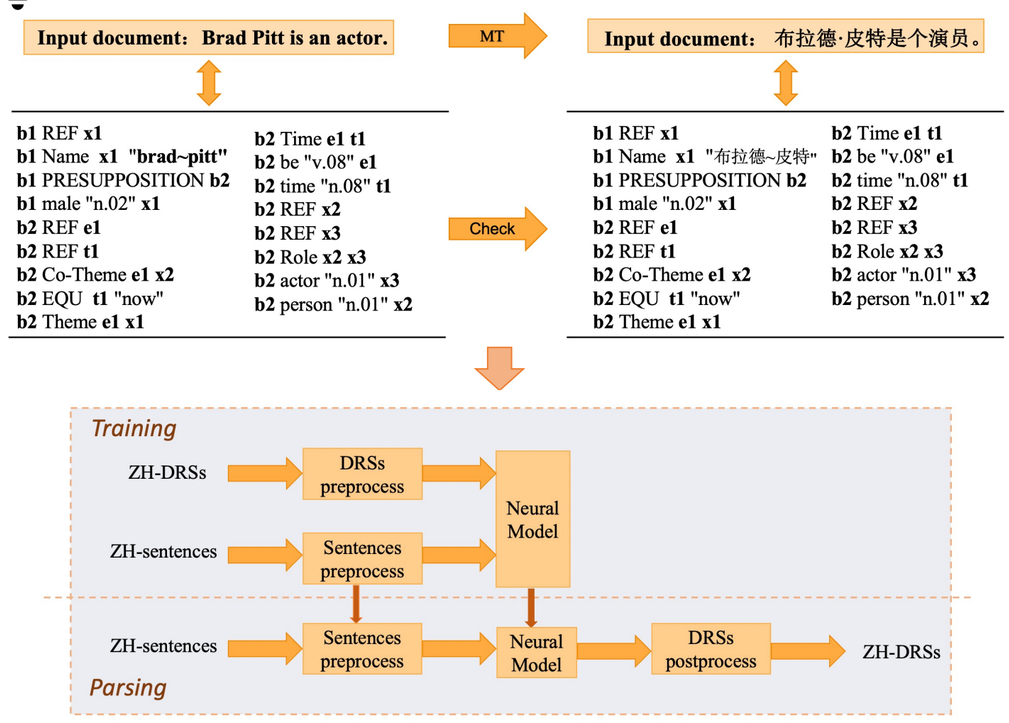
Example DRS: Brad Pitt is an actor.



b1 REF **x1**
b1 Name **x1** **"brad~pitt"**
b1 PRESUPPOSITION **b2**
b1 male "n.02" **x1**
b2 REF **e1**
b2 REF **t1**
b2 Co-Theme **e1 x2**
b2 EQU **t1** "now"
b2 Theme **e1 x1**
b2 Time **e1 t1**
b2 be "v.08" **e1**
b2 time "n.08" **t1**
b2 REF **x2**
b2 REF **x3**
b2 Role **x2 x3**
b2 actor "n.01" **x3**
b2 person "n.01" **x2**

Type	English input representation	Chinese input representation
Char (raw)	^ b r a d | ^ p i t t | i s | a n | a c t o r .	布 拉 德 · 皮 特 是 个 演 员 。
Char (continuous)	^ b r a d ^ p i t t i s a n a c t o r .	
Char (tokenized)	^ b r a d | ^ p i t t | i s | a n | a c t o r | .	布 拉 德 | · | 皮 特 | 是 | 个 | 演 员 | 。
Word	brad pitt is an actor .	布拉德 · 皮特 是 个 演员 。
BPE (5k)	^ b@ ra@ d ^ p@ it@ t is an ac@ tor@ .	布 拉 德 · 皮 特 是个 演 员。

Methodology



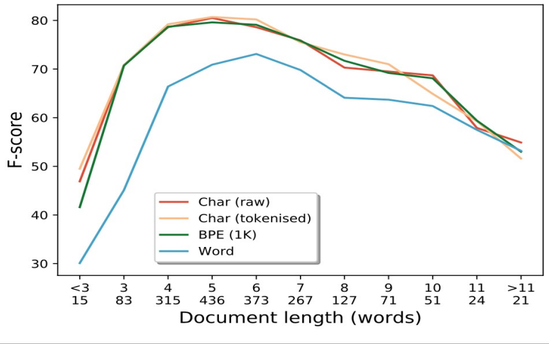
Evaluation

Gold DRS

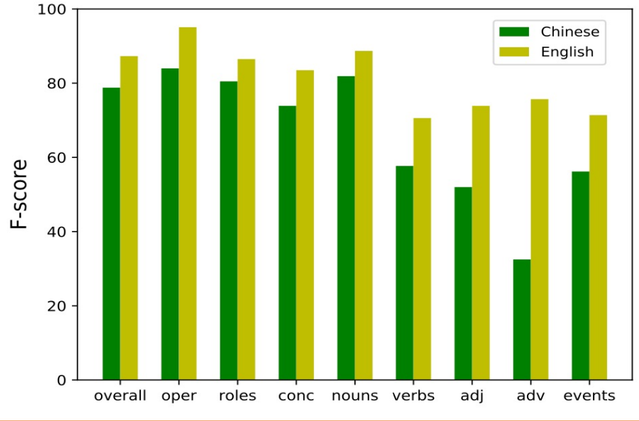
```
b1 REF t1
b1 EQU t1 "now"
b1 time "n.08" t1
b2 Time e1 t1
b1 NOT b2
b2 REF e1
b2 Agent e1 "speaker"
b2 work "v.02 e1
b2 Co-Agent e1 x1
b3 REF x1
b3 Name x1 "tom"
b3 male "n.02" x1
```

System output

```
b2 REF x1
b2 EQU x1 "now"
b2 time "n.08" x1
b1 Time x2 x1
b2 NOT b1
b1 REF x2
b1 Agent x2 "speaker"
b1 work "v.01 x2
b1 Goal x2 x3
b1 REF x3
b1 Name x3 "tom"
```



	English		Chinese	
Input type	**Dev**	**Test**	**Dev**	**Test**
Char (raw)	87.9	87.6	78.8	76.2
Char (continuous)	86.1	86.9		
Char (tokenised)	88.0	88.1	79.5	76.2
BPE (1k)	86.8	87.0	78.5	76.2
BPE (5k)	87.4	87.1	75.1	71.8
BPE (10k)	82.5	82.3	68.5	65.2
Word	84.5	83.2	74.7	71.6



Conclusions

1. Chinese based on projecting meaning representations from English translations gives remarkable performance

2. Characters are the preferred input representation for Chinese

3. Tokenisation (segmenting the text into words) of the input offers a small advantage for English, but not for Chinese

Source

1. https://github.com/wangchunliu/Chinese-DRS-parsing
2. https://github.com/wangchunliu/Chinese-DRS-data
3. https://pmb.let.rug.nl/