OntoNotes Data Plan

2009-02-24

Lance Ramshaw

Original Data Plan - English

Full ON Ann.	end Y1	end Y2	end Y3	end Y4	end Y5
English					
NW	300K WSJ		250K ECTB		
BN		200K			
BC			200K		
NG				200K	
WL				200K	
CTS					100K, 100K*

(* Wordsense annotation only)

Original Data Plan - Chinese

Full ON Ann.	end Y1	end Y2	end Y3	end Y4	end Y5
Chinese					
NW	250K ECTB			150K	
BN		300K			
BC			150K		
NG				150K*	
WL					100K, 50K*

(* Wordsense annotation only)

Original Data Plan - Arabic

Full ON Ann.	end Y1	end Y2	end Y3	end Y4	end Y5
Arabic					
NW		100K	100K	100K	
BN				100K	100K

New Priorities

- **Parallel Data**
  - **Beginning with BC data, parallel data has been maximized**
- **Genre Focus Shifts**
  - **Conversational Telephone Speech dropped as a goal**
- **Targeting Data Selection for Wordsense Coverage**
  - **Random data selection provides limited coverage of less frequent words**
    - **Also of less frequent senses even for frequent words**
  - **Selection of documents to maximize target words can improve things somewhat**
  - **More effective targeting using sentence selection, though that data is not useful for coreference annotation**
  - **Selected data to be translated into Chinese to provide additional parallel text**

Additional Data and Tasks

- **P2.0 and P2.5 Datasets**
  - **~15K per genre per language of parallel data**
  - **140K total of new English data**
- **Treebank Normalization**
  - **NMLs and Hyphenization**
  - **TB/PB Merge changes**
  - **Including updating the remaining 400K of non-ON, non-financial WSJ**
- **Updating to new Arabic Trees**
- **PropBank revisions**

Available Data from Years 1-3

English 1.1M *(373K par)*

- **NW 664K**
  - **347K WSJ**
  - ***317K ECTB-Eng****
- **BN 225K**
  - **225K TDT4 Eng Data***
- **BC 187K**
  - **131K Eng**
    - ***50K of that Eng-to-Chi***
  - ***56K Eng-from-Chi***

* Coref annotation only partially completed

Chinese 692K *(309K par)*

- **NW 254K**
  - ***254K ECTB-Chi***
- **BN 269K**
  - **269K TDT-4 Chi**
- **BC 169K**
  - **114K Chi**
    - ***50K of that Chi-to-Eng***
  - ***55K Chi-from-Eng***

Arabic 204K

- **NW 204K**
  - **204K ATB-Ara**

Y4 Data to be Annotated

English 375K *(115K par)*

- **WEB**
  - *45K LDC Eng-from-Chi*
  - 60K LDC Eng-from-Ara
  - *70K Wordsense Selected Eng Documents*
  - 200K Wordsense Selected Eng Sentences*
    - To be annotated for syntax, PropBank, and wordsense of target words and other words for which we don't already have strong models

Chinese 160K *(160K par)*

- **WEB**
  - *90K LDC Chi*
  - *~70K Wordsense Selected Documents Chi from Eng*
    - 70K English words, no exact count yet of Chinese words

Arabic 100K

- **NW**
  - 100K ATB-Ara*

P2.5 Data for Additional Y4 Annotation?

English

- NW, BN, BC, & Web
- ~30K/genre (141K total)

Chinese

- NW, BN, BC, & Web
- ~15K/genre

Arabic

- NW, BN, BC, & Web
- ~15K/genre

- Already Treebanked
- Could also be PropBanked
- Not contiguous segments, so perhaps not ideal for coreference annotation