OntoNotes issues

Christopher Manning and Jenny Finkel
Stanford University

OntoNotes

- The concept is great
- We're excited by the possibilities for doing new forms of learning and inferences over these joint representations
- This is mainly just random complaining about how the data could be more consistent, which I should have assembled and passed on earlier, but at least this workshop was a forcing function…

OntoNotes v2.0

- These observations are based on working with OntoNotes v2.0, not the recently released OntoNotes v2.9
- But, our first impression is that most of these issues are unchanged in OntoNotes v2.9.
- (And we're really looking forward to the WSJ material being consistent with the rest of the data in treebank annotation style!)

Parsing and NER

- Reasonable desideratum (we think!)
  - A Named Entity will correspond to some contiguous subset of the children of some node in the tree.
    - Otherwise, the constituency claims contradict between the Named Entity and the treebank structure
      - "Crossing brackets"
  - There are 55,665 named entities
  - 656 fail to meet this criterion
    - Okay, only 1%. But, why? For good reasons or bad reasons?

Generally bad reasons

- Sentence final acronyms:
  - … left the U.S. Later, he returned …
    - You get the final period being both the period of the acronym and the sentence-ending period
    - NER data has "U.S." as the NE
    - Treebank gives priority to the end-sentence function, and has tokens of
      - U.S
      - .
    - (That Treebank treatment has always been kind of weird, maybe it should really be "U.S." and ".")
    - About 150 cases

Generally bad reasons

- Determiners, adjectives, and PPs inside NPs
  - (NP (NP (DT the) (NNP District)) (PP (IN of) (NP (NNP Columbia))))
  - NE is just "District of Columbia"
- Here the linguistically weird Penn Treebank structure of having determiners inside the base NP, whereas linguists would have (minimally)
  - (NP (DT the) (N' (NNP District) (PP (IN of) (NP (N' (NNP Columbia)))))

  becomes a problem
  - Common. About 285 cases.

Harder reasons

- Reduced conjunctive NEs

  (NP (NML (NNP North) (CC and) (NNP South)) (NNP Korea))

  <ENAMEX TYPE="PERSON">Kim</ENAMEX> received the award for improving ties between <ENAMEX TYPE="GPE">North</ENAMEX> and <ENAMEX TYPE="GPE">South Korea</ENAMEX> .

Harder reasons

- Similar issues like with areas:

  (NML (QP (CD 40) (SYM x) (CD 60)) (HYPH -) (NN foot))

  Workers on deck dwarfed by the <ENAMEX TYPE="CARDINAL">40</ENAMEX> x <ENAMEX TYPE="QUANTITY">60 - foot</ENAMEX> gaping wound .

Harder reasons

- Extra NP node required for NE structure

We have with us Captain <ENAMEX TYPE="PERSON">Kent Ringboard</ENAMEX> , on board the <ENAMEX TYPE="ORG">Royal Caribbean</ENAMEX> `` <ENAMEX TYPE="PRODUCT">Radiance of the Seas</ENAMEX> . ''

(NP (DT the)
(NP
(ADJP (JJ Royal) (JJ Caribbean))
(`` ``) (NNP Radiance))
(PP (IN of) (NP (DT the) (NNPS Seas))))

Coreference

- 67,500 mentions
- Again, about 553 (1%) aren't consistent with parse tree constituency
- Coreference is inconsistent about whether (POS 's) is part of entity mention
- Null trace elements inconsistently included in mentions
- Similar issues to NER with inclusion/exclusion of determiners in mentions

Coreference: harder cases

Stephen R. Barnett Professor of Law <COREF ID="116" TYPE="IDENT">University of California Berkeley</COREF> , Calif .

```
([TOP]
 ([NP]
   ([NP PERSON] ([NNP] [Stephen 0]) ([NNP] [R. 1]) ([NNP] [Barnett 2]))
   ([NP]
     ([NP] ([NNP] [Professor 3]))
     ([PP] ([IN] [of 4])
       ([NP] ([NNP] [Law 5])))))
   ([NP-LOC]
     ([NP] ([NNP] [University 6]))
     ([PP] ([IN] [of 7])
       ([NP] ([NNP] [California 8]))))
   ([NP-LOC]
     ([NP] ([NNP] [Berkeley 9]))
     ([,] [, 10])
     ([NP GPE] ([NNP] [Calif 11])))
   ([.] [. 12])))
```

Coreference vs. NER

- Cases where there are both a coref mention and an NER mention which clearly should refer to the same thing, but the spans are different
  - Lots of cases, many turning on similar issues of whether to include the POS 's in the NER vs. coref mention
  - The FBI 's
    - Whole thing is coref mention
    - Just FBI is an NER ORG
  - Coref and NER are inconsistent about whether salutations (Mr/Mrs/Dr/etc) are part of the entity.
  - Coref includes them, but NER does not.

Treatment of word indices

- Do traces (empty elements) count as words when determining indices in other components?
  - Coref and NER look similar in that both are sentences with XML markup, but coref includes traces and NER doesn't.
  - WSD and Propbank are annotated by indexing into the words (or tree leaves), and for Propbank the index includes traces, but for WSD it doesn't.

Too generous pattern-based TIME and DATE matching?

- (abc/00/0001.name):
- It 's a goal that seems possible when on <ENAMEX TYPE="TIME">homecoming night</ENAMEX> the starting kicker is <ENAMEX TYPE="DATE">14 - year - old Nikita Kargalskiy</ENAMEX> , who may be <ENAMEX TYPE="QUANTITY">5,000 miles</ENAMEX> from his hometown in <ENAMEX TYPE="GPE">Russia</ENAMEX> .

- Half/CARDINAL of/O the/O Palestinian/NORP population/O is/O under/O the/DATE age/DATE of/DATE 14/DATE

More time & date

Word	Named Entity	Count
today	DATE	461
today	Not an entity	40
today	TIME	14
tonight	Not an entity	22
tonight	TIME	126

NER in general has issues?

Word	Named Entity	Count
Mexicans	NORP	4
Mexicans	Not an entity	4

Works of art???

Stanley Cup USS Cole State of the Union Message

Headliners Gallup Poll Nobel Peace Prize

USA Today CNN Financial News Entertainer of the Year

the Journal of the American Medical Association

More boring NER stuff, but it would be good if they were fixed…

- For organizations, about a third of the time they are labeled ORG and about two thirds of the time they are labeled ORGANIZATION.
- Similarly for LOCATION/LOC
- "Coalition of the Left and Progress" should actually be one entity (a since-renamed Greek political party: http://en.wikipedia.org/wiki/Synaspismos
  - <ENAMEX TYPE="ORDINAL">First</ENAMEX> came his predictable fusillade : He charged the <ENAMEX TYPE="ORGANIZATION">Coalition of the Left</ENAMEX> and <ENAMEX TYPE="ORGANIZATION">Progress</ENAMEX> had sold out its leftist tenets by collaborating in a right-wing plot aimed at ousting <ENAMEX TYPE="ORGANIZATION">PASOK</ENAMEX> and thwarting the course of socialism in <ENAMEX TYPE="GPE">Greece</ENAMEX> .