Corpus Linguistics
(L415/L615)
Annotation: Motivation & Limitations

Markus Dickinson
Department of Linguistics, Indiana University
Fall 2015
Motivation for Annotation

Searching

Example: Temporal NPs seem to behave differently from other NPs

(1) a. The economic and foreign ministers . . . will meet in Australia next week
     b. The exact amount of the refund will be determined next year based on actual collections made until December 31 of this year.
     c. The next morning, with a police escort, busloads of executives and their wives raced to the Indianapolis Speedway, . . .
     d. Earlier this year, Japanese investors snapped up a similar fund.
     e. One Egg King . . . can crack about 20,000 eggs an hour.
Form-Based Searching

Can we list out all possible forms & patterns for these NPs?

- all time units: day, week, ...
- all specific designations: today, tomorrow, ...
- other ways to refer to time: summer, once (in a blue moon), ...

How do the syntactic patterns differ for these cases?
Annotation-Based Searching

Some corpora have POS with temporal distinctions
- e.g., SUSANNE corpus has: NNT1, NNT2, NNT1c, NNT1h, NNT1m

Some corpora have syntactic distinctions:
- e.g., Penn Treebank has TMP functional annotation, e.g., NP–TMP
  - Distributions differ based on what they modify
    - VP modifier: 2,959 occurrences
    - S modifier: 1,069 occurrences
    - NP modifier: 636 occurrences
Approximations via Annotation

With only POS annotation, we can approximate the syntactic information

- Based on earlier examples: two consecutive NPs seems like a good place to start
  - Overgenerate: finds other phenomena, e.g., ditransitives
  - Undergenerate: miss cases not next to other NPs
Precision and Recall in Searching

- **Precision**: of the returned items, which are true hits?
  - Related to overgeneration
  - Returning *deed* for a search for past tense verbs diminishes the precision

- **Recall**: of the items to return, which are successfully returned?
  - Related to undergeneration
  - Failing to return *dove* for a past tense verb diminishes the recall

Annotation can ideally help both:
- e.g., a past tense verb POS tag can help both problems
Inconsistencies

It is important to note that both humans & automatic systems make mistakes in annotating . . . though of different kinds

▶ Humans: result of fatigue, often non-systematically inconsistent
  ▶ Though: systematic errors arise from confusing or difficult distinctions and/or guidelines

▶ Automatic systems are systematically inconsistent
  ▶ When faced with the same data, will make the same decision

Semi-automatic annotation processes thus tend to work best

▶ Likewise, error detection software, interannotator agreement (IAA) calculations, etc. ensure better quality

In general: what are the consequences for us? How bad is the impact of errors?
Inconsistencies

Examples

Some generally recognized issues in automatic POS annotation (for English):

1. Distinguishing proper names from beginnings of sentences (e.g., *Doctors tend to ... vs. I saw Doctor Robert*)
2. Distinguishing past participles from past tense verbs (e.g., *exploded*)
3. Recognizing unknown words correctly (e.g., *familiarly*)

And for humans, the issues might be more conceptual:

1. Distinguishing adjectives from nouns (e.g., *the red*)
2. Distinguishing participles from adverbs & prepositions (e.g., *ran up the hill vs. ran up the bill*)

For us: may have to formulate less-specific queries to accommodate errors
Inconsistencies

All About Me

If you’re really interested in annotation consistency, buy me a cup of tea & we can chat for a long while

- **DECCA (Detection of Errors and Correction in Corpus Annotation),** http://decca.osu.edu
  - with Detmar Meurers & Adriane Boyd
  - Core idea: same lexical context predicts same annotation; violations are potential errors

- **DAPS (Detecting Anomalous Parse Structures)**
  - with Amber Smith: see my publications page, http://cl.indiana.edu/~md7/papers/
  - Core idea: syntactic rules unlike other syntactic rules are potential errors
Limitations

We’ve discussed a lot about limitations already, but some quick points:

- Negative evidence is unable to be found, even with annotation
  - The picture with learner corpora is a bit murkier
- Rare occurrences are less likely to appear in (manually-)annotated corpora as they tend to be smaller
- Annotation has the potential to contain errors
- Annotation is prone to researcher bias
  - Always consult the guidelines!