"Simple Issues"

Rethinking Anaphora and Information Structure

Anders Bjorkelund Aoife Cahill Kevin Crooks Vincent Ng
Massimo Poesio Sameer S Pradhan Marta Recasens
Olga Uryupina Yannick Versley

Overview

- MIN-imal span
- Non-referring expressions
  - Expletives
  - Idioms
  - Quantifiers
  - Predication
- Singleton mentions
- Markable Spans
  - Disjoint
  - Split Antecedent
  - Sub-token annotation
- Pre-modifiers
- Zeroes
- (Events)

MIN-imal Spans

- Problem: rigid matching for mention boundaries, too restrictive for:
  - Training (material lost owing to parsing errors)
  - Evaluation (lower scores)
- Not so critical for OntoNotes (English), *but* may be different for
  - Other domains
  - Other languages

MIN-imal Spans (Solution 1)

- Annotate MAX, compute MIN automatically using headword heuristics
  - \+ No extra cost
  - – Requires gold trees *and* head rules *and* possibly name information
  - – Does not handle:
    - New York
    - high school
    - University of Trento
    - ...
- TuebaDZ corpus had some experience with this approach, but had to use more complicated rules.

MIN-imal Spans (Solution 2)

- Annotate MAX, compute MIN automatically, correct manually
  - \+ Better quality
  - – Extra cost (Massimo seems to differ)
- ARRAU corpus has some experience with this approach

MIN-imal Spans: Vote!

- Yes, Let's annotate MIN (5)
- No, Let's take the money (4)

Non-referring Expressions

- Semantically vacuous (expletives, idioms, ...)
- Predicatives, quantifiers, ...

Expletives

- Can we annotate them reliably?
- Anders: ML performance on 3 pronouns in OntoNotes:
  - It (68%); You (76%), We (51%)
  - But, the actual annotation sometimes looks arbitrary
- Some problemmatic cases
  - "You": generic vs. personal (esp. telephone conversation)
  - "It": in connection with congnition verbs

Idioms

- ".. kicked the bucket. The bucket is.."
  - is *the bucket* referential?
- Can we allow coreference between non-referential markables?

Quantifiers

- No standard practice (ACE vs OntoNotes vs ARRAU)
  - In ACE "Nobody ... Nobody" is considered coreferent!
- "Everybody loves his mom"
  - Referential?
  - Co-referential?
- "One" ?

Predication

- It is a mess here.
  - ACE considers this coreferential, unless the assertion is no longer true.
  - ARRAU annotates this as non-referential
  - OntoNotes ignores it as it can be recovered from Propbank (copula) and adds a special realation APPOS for apposition.

Predication (contd.)

- Examples
  - Our lion, Fluffy, should get more veal.
  - Fluffy, the lion, should get more veal.
  - As a lion, Fluffy should get more veal.
  - Fluffy should get more veal as a lion.
- Could some linguists kindly help us here?

Singleton Mentions

- Should we annotate mentions that do not participate in coreference chains?
- OntoNotes and TubaDZ did not annoate, but ARRAU and Prague folks do.

+ One can evaluate and optimize entity mention detection separately from coreference
+ Provide information on referentialilty
– Cost
– Can of worms
– Singleton can boost scores esp. B^3, although one could remove them for scoring purposes.

Singleton: Vote!

- Yes, given the time and money let's (9)
- No (0)

Disjoint Markables

- Jane and John Smith
  - How do we annotate Jane?
- More common in the medical domain
  - "upper and lower lobe lung cancer"
- NMLs might provide a step towards resolving this issue

Split Antecedents

- "Jane doesn't trust John. They ..."
  - We should probably annotate these for future generations, so far nobody even tries to resolve them automatically.

Subtoken Annotation

- "Anti-American", "New York-based", ...
  - Clitics
  - Compounds (German)
- OntoNotes Solution: annotate as tokens, but keep track of offsets.
  - \+ The corpus might get re-tokenized! In OntoNotes, almost all sub-token markables dissapeared after re-tokenizing.

Pre-modifiers

- French President
- US President
- FBI agent
  - OntoNotes did not annotate "adjectival" modifiers (or, acronyms)
  - For Chinese, the adjectival-ness decision is made using the NORP named entity

Pre-modifier: Solutions

- Ignore all pre-modifiers (we don't like it)
- Connect all (like MUC did. We don't like it)
- Possible Tests
  - If XY can be expressed as Y of X, then annotate X (English only)
  - If it can be referred by a pronoun later
  - So **stomach** cancer, but not ***coffee** table.

Zeroes

- OntoNotes (English)
  - small PROs are not linked, but can be recovered from Treebank/PropBank
- OntoNotes (Arabic/Chinese)
  - small PROs from Treebank are annotated
- Annotate Traces?
  - If syntax layer is available, then should not annotate traces
  - If syntax layer is not available, then only annotate when really-really necessary.
    - Jane came home. **[Jane]** Annotated 20 documents.
    - Jane came home and **[*Jane]** watched a movie.
- There are issues with genericity that come up during for languages with dropped subjects/objects.

(Events)

- This is not a simple issue.
- ACE: agreement was very low
- OntoNotes decided to side-step the bigger issue by creating event chains for only cases that have a nominal referent back.
- NAACL 2012 Events Worskshop
- Might make sense to annoate on trees rather than plain text.