University of Zurich[UZH]

Institute of Computational Linguistics

Coreference Resolution Evaluation for Higher Level Applications

Don Tuggener
tuggener@cl.uzh.ch

25. April 2014

Problem Introduction I

"Coreference resolution is an important preprocessing step for higher level applications."

- Commonly used **coreference resolution metrics** (MUC,BCUB,CEAF,BLANC) are **well-suited for comparing systems** under a "closed-world" assumption, i.e. **detached from other tasks**.
- Arguably **cannot assess the potential benefit** of using a coreference system in preprocessing in a **transparent**, **direct**, and **detailed** fashion.

Problem Introduction II

"Coreference resolution is an important preprocessing step for higher level applications."

- What does e.g. 56 % BCUB Recall mean for a Sentiment Analysis system?
- Performance on e.g. 3rd person pronouns and the (potentially pleonastic) *it* pronoun?

Our approach: Recall and Precision

Devise **Recall** and **Precision** measures which are **intuitive**, **straight-forward**, and they should provide **detailed analysis** of system output (any mention feature should be measurable).

- tp = correctly resolved true mentions
- fp = non-anaphoric mentions that have been resolved, but should not have been (e.g. "*it* is raining.")
- fn = unresolved true mentions

→ What about true mentions that have been linked to an incorrect antecedent?

- **wl** = true mentions linked to incorrect antecedents

Our approach: Recall and Precision

Adaption to Coreference Resolution: Introduce a new error class

- Recall = $\frac{tp}{tp+\mathbf{wl}+fn}$
  - Denominator extends over **all mentions in the gold standard** (key)
  - *"How many of the relevant mentions does the system resolve correctly?"*
- Precision = $\frac{tp}{tp+\mathbf{wl}+fp}$
  - Denominator extends over **all mentions in the system output** (response)
  - *"How many of the mentions resolved by the system are correct?"*

The crucial question: What is a correct antecedent?

Well, it depends → **Higher level applications**

- How do higher level applications use coreference systems?
- Their **requirements** towards a coreference system define what a **correct antecedent** is.
- We outline three application scenarios.

Scenario 1: Text/discourse structure based models

Track subsequent entity occurrence and (shifts in) their appearance contexts.

- Coherence models: Track sequential patterns of grammatical roles of re-occurring entities (Centering Theory, Entity Grid...)
- Event models: Track argument roles of re-occurring entities, derive event schemes

Rely on **correct and gap-less identification of entity occurrence** in discourse. → The direct antecedent of a mention in a response chain has to be the direct antecedent in the key chain.

Scenario 2: Inferred local nominal antecedents

Need for a nominal antecedent to help text understanding

- Summarization: Replace pronouns in target sentences with the closest nominal antecedent to ensure coherence of the summary.
- Machine translation:
  DE *"Der Mond ist sichtbar, er scheint"*
  FR *"La lune est visible, il* brille"*

Rely on **correct identification of a nominal antecedent** → The closest nominal antecedent of a mention in a response must be an antecedent of the mention in the key.

Scenario 3: Query based applications

Finding occurrence contexts of specific entities

- Sentiment Analysis: Find occurrences of entity X and derive polarity information from its context.
- Relation Mining: What kind of relations has entity X with what other entities?

Based on string query. Mentions need to be linked to an **antecedent** which is **accessible by string query**.
→ mentions need to link to the **anchor mention** of a coreference set (approximation: first nominal mention in the chain).

Analysis of three state-of-the-art coreference systems

Compare three systems

- CoNLL 2012 English test set
- MELA ranking $\frac{MUC+BCUB+CEAF_e}{3}$:

Berkley (Durret & Klein, 2013)	**61.62**
IMS (Björkelund & Farkas, 2012)	**57.42**
Stanford (Lee et al., 2013)	**55.69**

Analysis of three state-of-the-art coreference systems

Scenario 1: Immediate Antecedent

BERKLEY	R	P	F
NOUN	45.06	47.06	46.04
PRP	67.66	64.87	66.24
PRP$	74.49	74.32	74.41
TOTAL	56.60	56.91	56.76
IMS			
NOUN	38.01	43.09	40.39
PRP	**69.06**	**68.64**	**68.85**
PRP$	72.57	72.11	72.34
TOTAL	53.55	57.55	55.48
STANFORD			
NOUN	38.51	42.92	40.60
PRP	65.55	61.09	63.25
PRP$	66.12	65.70	65.91
TOTAL	51.70	52.69	52.19

Scenario 2: Inferred Antecedent

BERKLEY	R	P	F
NOUN	55.54	60.37	57.85
PRP	48.92	53.62	51.16
PRP$	61.95	66.80	64.28
TOTAL	52.94	58.04	55.37
IMS			
NOUN	46.90	54.96	50.61
PRP	43.04	**57.42**	49.20
PRP$	51.51	63.54	56.90
TOTAL	45.27	56.47	50.25
STANFORD			
NOUN	**50.03**	**57.62**	**53.56**
PRP	36.67	45.97	40.80
PRP$	40.64	52.38	45.77
TOTAL	43.01	51.73	46.97

Scenario 2: Comparison of 3rd person pronouns

Inferred Antecedent								
		Scores			Counts			
	Sys.	R	P	F	tp	wl	fn	fp
He (874)	Berk.	**75.89**	77.66	76.76	664	184	27	7
	IMS	63.73	73.77	68.39	557	193	***124***	5
	Stan.	61.90	69.63	65.54	541	***230***	103	6
She (309)	Berk.	48.25	47.95	48.10	152	***160***	3	5
	IMS	**51.59**	57.86	**54.55**	**162**	113	39	5
	Stan.	44.16	**62.10**	51.61	136	**78**	94	5

remember: key [NOUN - PRONOUN1 - PRONOUN2]
response [PRONOUN1 - PRONOUN2] → 2 fn (no nominal antecedent)

Scenario 3: Anchor mention based evaluation

Named Entity anchors

PERSON (18.69%)	
	Fϕ
BERK	67.11
IMS	***52.74***
STAN	61.61

ORG (9.63%)	
	Fϕ
BERK	67.62
IMS	56.85
STAN	**59.41**

GPE (13.28%)	
BERK	75.71
IMS	**76.21**
STAN	75.24

TOTAL (100%)	
BERK	63.41
IMS	55.24
STAN	55.27

Conclusions

- Scores are transparent, intuitive, explained easily
  - “Save and sound” definitions of Recall and Precision
- Counting based on tp, wl, fp, fn allows for explanation of the scores and detailed system comparison
  - Pronouns: Overall scores not that different, but *very* different scores on different pronoun lemmata
- Detailed analysis can also help developers to analyze and improve components of their systems

Thank you

Scorers are available for download and can easily be modified for other languages and measurement of specific mention features : http://www.cl.uzh.ch/research/coreferenceresolution.html

Chen, C. & Ng, V. (2013). *Linguistically Aware Coreference Evaluation Metrics*. IJCNLP.

Durrett, G. & Klein, D. (2013). *Easy Victories and Uphill Battles in Coreference Resolution*. EMNLP.

Björkelund, A. & Farkas, R. (2012). *Data-driven Multilingual Coreference Resolution Using Resolver Stacking*. Joint Conference on EMNLP and CoNLL, Shared Task.

Lee, H. et al.(2013). *Deterministic Coreference Resolution based on Entity-centric, Precision-ranked Rules*. Computational Linguistics, 39.

Other scores

Accuracy score: Only operate on resolved gold mentions.

Neglect fp and fn: Acc=$\frac{tp}{tp+wl}$

- Blend out the anaphoricity detection problem
- How accurate is the resolution strategy when it resolves gold mentions?

Gold mentions only simulation: $\frac{tp}{tp+wl+fn}$, i.e. no fp

Our approach: Recall and Precision

Sensitive to the “anaphoricity detection problem”:

- Anaphoric *it*-pronoun linked to an incorrect antecedent → wrong linkage. Affects Recall and Precision.
- Pleonastic *it*-pronoun is resolved → false positive. Only Precision is affected.

This is especially relevant for nouns, as they constitute the largest portion of the coreferent mentions. Determining their anaphoricity status is a major issue in coreference resolution.

Cataphora and chains not containing nominal mentions

total chains: 4532
key chains without nouns: 476 (10.5 %)
key chains starting with cataphora: 241 (5.32 %)
Chain without nouns: 1st and 2nd person pronouns (*I - me - ...*)
→ Relevant for inferred nominal antecedent evaluation only.

Our approach: Recall and Precision

Two special cases:

- A mention starts a coreference chain in the key but not in the response → Falsely resolved non-anaphoric mention → fp
  - Punishes incorrectly merging gold coreference sets:
    key [A-B],[C-D] response [A-B-C-D] → fp
- A mention starts a coreference chain in the response but not in the key → unresolved anaphoric mention → fn
  - Punishes splitting gold coreference sets:
    key [A-B-C-D] response [A-B],[C-D] → fn

Scenario 3: Query based applications

Finding occurrence contexts of specific entities

- **Entity Detection (ED)**: How many **anchor mentions** can be **aligned** in the key and the response (tp, fn → Recall), how many spurious anchors are returned (fp → Precision)?
- **Entity Mentions (EM)**: How many **references to aligned anchors** are found by a system (tp, fn → Recall), how many spurious mentions are returned (fp → Precision)?
- Harmonic mean of ED and EM to compare systems

Scenario 3: Anchor mention based evaluation

Named Entity anchors

		R	P	F	Fϕ
		PERSON (18.69%)			
BERK	ED	64.02	75.88	69.45	67.11
	EM	63.60	66.29	64.92	
IMS	ED	45.66	51.69	48.48	**52.74**
	EM	47.67	73.45	57.82	
STAN	ED	56.33	59.74	57.98	61.61
	EM	53.84	84.37	65.73	
		GPE (13.28%)			
BERK	ED	73.21	77.36	75.23	75.71
	EM	69.89	83.73	76.19	
IMS	ED	73.51	74.17	73.84	**76.21**
	EM	69.94	90.04	78.73	
STAN	ED	70.24	76.62	73.29	75.24
	EM	68.44	88.81	77.30	

		R	P	F	Fϕ
		ORG (9.63%)			
BERK	ED	62.78	67.13	64.88	67.62
	EM	66.87	74.78	70.60	
IMS	ED	44.98	54.30	49.20	56.85
	EM	57.26	81.66	67.32	
STAN	ED	49.68	58.56	53.75	**59.41**
	EM	57.25	79.05	66.41	
		TOTAL (100%)			
BERK	ED	58.65	53.19	55.79	63.41
	EM	72.65	74.28	73.45	
IMS	ED	47.16	42.66	44.80	55.24
	EM	65.88	79.40	72.01	
STAN	ED	48.62	41.40	44.72	55.27
	EM	65.66	80.48	72.32	

Scenario 1: Text/discourse structure based models

→ The direct antecedent of a mention in a response chain has to be the direct antecedent in the key chain.

key: [A - B - C]
response 1: [A - B] [C] → 1 tp (B), 1 fn (C), R: 0.5 P: 1 F: 0.75
response 2: [A] [B - C] → 1 tp (C), 1 fn (B), R: 0.5 P: 1 F: 0.75
response 3: [A - C] [B] → 1 fn (B), 1 wl (C), R: 0 P: 0, F: 0
mention B missing in 3, forms a gap/occurrence sequence is broken, C counted as wrong linkage

Scenario 2: Inferred local nominal antecedents

→ The closest nominal antecedent of a mention in a response must be an antecedent of the mention in the key.

key [NOUN - PRONOUN1 - PRONOUN2]
response 1: [NOUN- PRONOUN1] [PRONOUN2]
→ 1 tp, 1 fn, R: 0.5 P: 1.0 F: 0.75
response 2: [NOUN - PRONOUN2] [PRONOUN1]
→ 1 tp, 1 fn, R: 0.5 P: 1.0 F: 0.75
response 3: [NOUN] [PRONOUN1 - PRONOUN2]
→ 2 fn, R: 0 P: 0 F: 0
3 does not infer a nominal antecedent
Only resolving PRONOUN2 is still helpful, the gap by PRONOUN1 is not as crucial as in scenario 1

Scenario 3: Query based applications

→ Mentions need to link to the most representative **anchor mention** of a coreference set (Approximation: first nominal mention in the chain).

key [NAMED_ENTITY - NOUN - PRONOUN]
response 1: [NAMED_ENTITY - NOUN] [PRONOUN]
→ 1 tp, 1 fn, R: 0.5 P: 1 F: 0.75
response 2: [NAMED_ENTITY - PRONOUN] [NOUN]
→ 1 tp, 1 fn, R: 0.5 P: 1 F: 0.75
response 3: [NAMED_ENTITY] [NOUN - PRONOUN]
→ 2 fn, R: 0 P: 0 F: 0
Anchor is missing in 3, underlying entity cannot be inferred