Introduction to Coreference Resolution

Abbas Ghaddar

9 April 2015

Course: IFT 6010

Université de Montréal

Definition

- Coreference resolution is the task of determining the right mention and which expressions in a text refer to.

- Example extracted by Stanford coreNLP :

(1) Quebec is a province in east-central Canada .

(4) It is bordered to the west by the province of Ontario ,.....

(7) Quebec is Canada 's second most populous province ,after Ontario.

(19) The province is sometimes referred to as `` La belle province '' **(22)** The Quebec Act of 1774 expanded the territory of the province....

History

History

- **Hobbs 1978**
  - *Hobbs, Jerry R., 1978, ``Resolving Pronoun References'', Lingua, Vol. 44, pp. 311-338.*
- **Lappin and Leass 1994**
  - Andrew Kehler, Douglas Appelt, Lara Taylor, and Aleksandr Simma. 2004. Competitive Self-Trained Pronoun
- **Centering Theory 1987**
  - Brennan, Susan E., Marilyn W. Friedman, and Carl J. Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics,* pages 155-162.

Hobbs 1978

- Hobbs (1978) proposes an algorithm that searches parse trees for antecedents of a pronoun.
  - Starting at the NP node immediately dominating the pronoun
  - Search previous trees, in order of recently, left-to-right, breadth-first.
  - Looking for the first match of the correct gender and number(male-female/singular-plural)

Example

The castle in Camelot remained the residence of the king until 536 when he moved it to London.

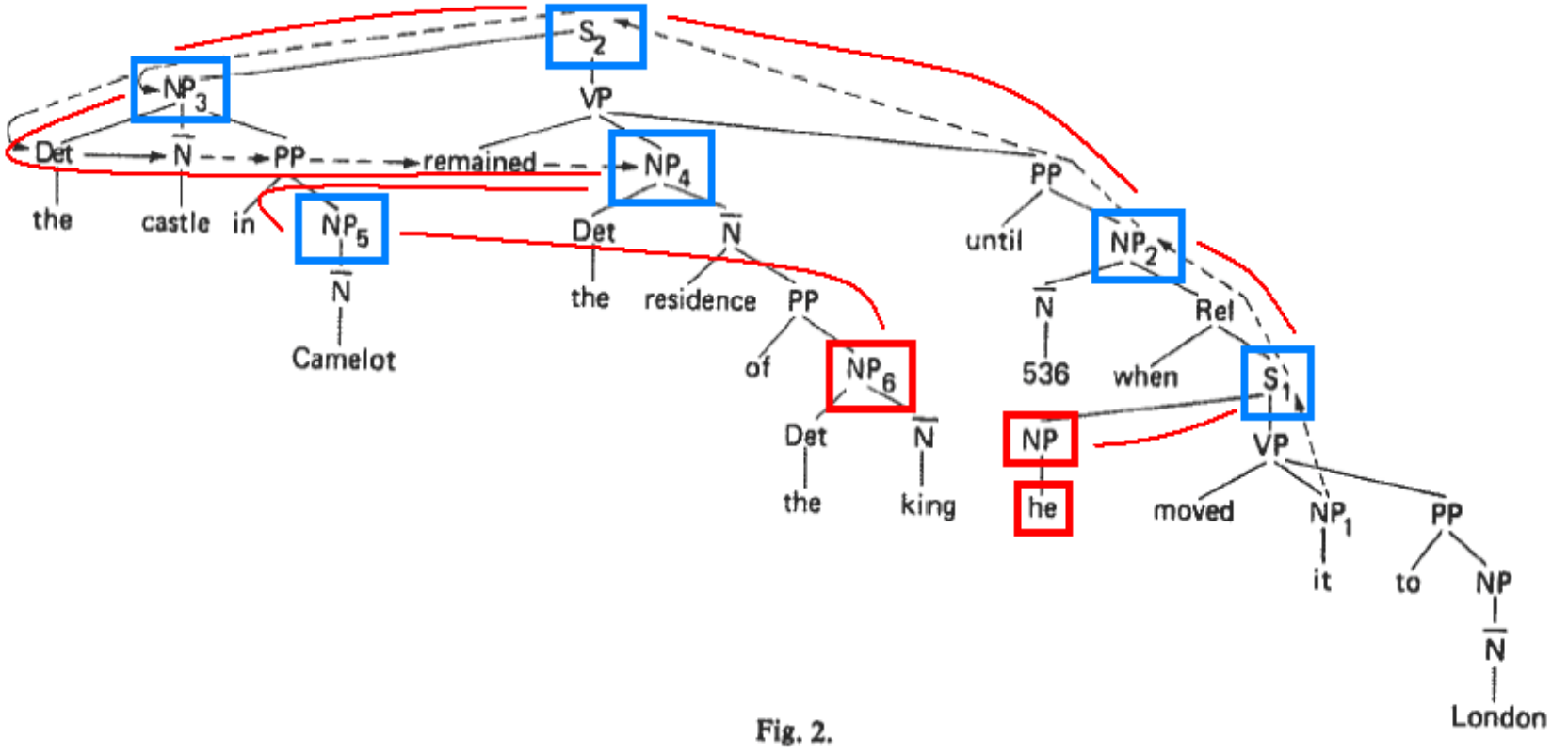


Fig. 2.

Hobbs results

- Hobbs analyzed, by hand, 100 consecutive examples from three “very different” texts.
- Assumed “the correct parse” was available.

- The algorithm was correct 72.7%.

- Hobbs concludes that the naïve approach provides a high baseline.
- Semantic algorithms will be necessary for much of the rest, but will not perform better for some time.

Lappin and Leass 1994

Lappin and Leass 1994 propose a discourse model in which potential referents have degrees of salience(weight).

They try to resolve (pronoun) references by finding highly salient referents compatible with pronoun agreement features.

Salience Factor	Salience Value
Sentence recency	100
Subject emphasis	80
Existential emphasis	70
Accusative emphasis	50
Indirect object emphasis	40
Non-adverbial emphasis	50
Head noun emphasis	80

Lappin and Leass 1994

The steps taken to resolve a pronoun are as follows:

- Collect potential referents (four sentences back);
- Remove potential referents that don't semantically agree;
- Remove potential referents that don't syntactically agree;
- Compute salience values for the rest potential referents;
- Select the referent with the highest salience value.

Hobbs' algorithm applied to same data is 82% accurate against 87% scored by Lappin and Leass Algorithm

Centering Theory

Basic ideas:

- A discourse has a focus, or center.
- The center typically remains the same for a few sentences, then shifts to a new object.
- The center of a sentence is typically pronominalized.
- Once a center is established, there is a strong tendency for subsequent pronouns to continue to refer to it.

Algorithm structure:

-First: Filtering based on hard constraints in order to detect the center

-Then: Ranking based on some soft constraints

Centering vs Hobbs

- Marilyn A. Walker. 1989 manually compared a version of centering to Hobbs on 281 examples from three genres of text.

- Reported 81.8% for Hobbs, 77.6% centering.

Stanford CoreNLP

Stanford CoreNLP

- . The toolkit provides most of the common core natural language processing (NLP) steps, from tokenization through to coreference resolution.

- NER Recognizes:

named (PERSON, LOCATION, ORGANIZATION, MISC)

numerical (MONEY, NUMBER, DATE, TIME, DURATION,SET)

- Truecase:

e.g. "NASA" should be all upper case otherwise, this information will be lost.

Annotator	Ara-bic	Chi-nese	Eng-lish	Fre-nch	Ger-man
Tokenize	✓	✓	✓	✓	✓
Sent. split	✓	✓	✓	✓	✓
Truecase			✓		
POS	✓	✓	✓	✓	✓
Lemma			✓		
Gender			✓		
NER		✓	✓		✓
RegexNER	✓	✓	✓	✓	✓
Parse	✓	✓	✓	✓	✓
Dep. Parse		✓	✓		
Sentiment			✓		
Coref.			✓		

Mentions Detection

- Candidate Mentions in each sentence are:
  - All noun phrase (NP)
  - Possessive pronoun
  - Named entity mentions

- **With Constraints:**

  1- Remove a mention if a larger mention with the same head word exists, e.g., remove *The five insurance companies* in *The five insurance companies approved to be established this time*.

  2- Discard numeric entities such as percent, money, cardinals, and quantities, e.g., 9%, $10, 000, Tens of thousands, 100 miles.

  3- Remove mentions with partitive or quantifier expressions, e.g., a total of 177 projects.

Mentions Detection-cont'd

4- Remove pleonastic it pronouns, detected using a set of known expressions, e.g., It is possible that

5- Discard adjectival forms of nations, e.g., American.

6- Remove stop words in a predetermined list of 8 words, e.g., there, ltd., hmm.

- When it's done, all singletons(not coreferenced) are filtered out before scoring(CoNLL Constraint).

Mention Processing

- Once mentions are extracted, we sort them by sentence number, and left-to-right breadth-first traversal order in syntactic trees in the same sentence (Hobbs, 1977).
- All sieves traverse the candidate list until they find a coreferent antecedent according to their criteria or decline to propose a solution (in the hope that one of the subsequent models will solve it).

Ordered sieves

1. **Mention Detection Sieve**
2. **Discourse Processing Sieve**
3. Exact String Match Sieve
4. **Relaxed String Match Sieve**
5. Precise Constructs Sieve (e.g., appositives)

6-8. Strict Head Matching Sieves A-C

9. **Proper Head Word Match Sieve**
10. **Alias Sieve**
11. Relaxed Head Matching Sieve
12. **Lexical Chain Sieve**
13. Pronouns Sieve

Mention Processing-cont'd

Passes	MUC			B^3			Pairwise		
	P	R	F1	P	R	F1	P	R	F1
{1}	95.9	31.8	47.8	99.1	53.4	69.4	96.9	15.4	26.6
{1,2}	95.4	43.7	59.9	98.5	58.4	73.3	95.7	20.6	33.8
{1,2,3}	92.1	51.3	65.9	96.7	62.9	76.3	91.5	26.8	41.5
{1,2,3,4}	91.7	51.9	66.3	96.5	63.5	76.6	91.4	27.8	42.7
{1,2,3,4,5}	91.1	52.6	66.7	96.1	63.9	76.7	90.3	28.4	43.2
{1,2,3,4,5,6}	89.5	53.6	67.1	95.3	64.5	76.9	88.8	29.2	43.9
{1,2,3,4,5,6,7}	83.7	74.1	78.6	88.1	74.2	80.5	80.1	51.0	62.3

The Modules of the Multi-Pass Sieve

- Exact Match:
  - This model links two mentions only if they contain exactly the same extent text, including modifiers and determiners, e.g.,

    the Shahab 3 ground-ground missile

- Relaxed String Match:
  - This sieve considers two nominal mentions as coreferent if the strings obtained by dropping the text following their head words are identical, e.g.,

    [Clinton] and [Clinton, whose term ends in January].

The Modules of the Multi-Pass Sieve

- Precise Constructs:
  - *Appositive* – the two nominal mentions are in an appositive construction, e.g.,

    [Israel's Deputy Defense Minister], [Ephraim Sneh] , said . . .

  - *Predicate nominative* – the two mentions (nominal or pronominal) are in a copulative subject-object relation, e.g., [The New York-based College Board] is [a nonprofit organization that administers the SATs and promotes higher education]

  - *Role appositive* – the candidate antecedent is headed by a noun and appears as a modifier in an NP whose head is the current mention, e.g., [[actress] Rebecca Schaeffer].

The Modules of the Multi-Pass Sieve

- Precise Constructs-cont'd :
  - *Acronym* – both mentions are tagged as NNP and one of them is an acronym of the other, e.g

    [Agence France Presse] . . . [AFP].

  - *Demonym* – one of the mentions is a demonym of the other, e.g., [Israel] . . . [Israeli]. For demonym detection a static list of countries and their gentilic forms from Wikipedia is used.

  - *Relative pronoun* – the mention is a relative pronoun that modifies the head of the antecedent NP, e.g.,

    [the finance street [which] has already formed in the Waitan district].

The Modules of the Multi-Pass Sieve

- Proper Head Word Match:
  - This sieve marks two mentions headed by proper nouns as coreferent if they have the same head word and satisfy the following constraints:
    - No location mismatches: For example, [Lebanon] and [southern Lebanon] are not coreferent.
    - No numeric mismatches: [people] and [around 200 people] are not coreferent.
    - Pronoun distance: Sentence distance between a pronoun and its antecedent cannot be larger than 3.

.

The Modules of the Multi-Pass Sieve

- Alias Sieve:
  - Two mentions headed by proper nouns are marked as aliases, if they appear in the same **Wikipedia infobox** or **Freebase record** or in the same **synset in WordNet**. For example, America Online and AOL are aliases in FreeBase record.

- Lexical Chain Sieve:
  - This sieve marks two nominal mentions as coreferent if they are linked by a WordNet lexical chain. This sieve correctly links Britain with country, and plane with aircraft

The Modules of the Multi-Pass Sieve

- Pronouns:
  - Pronouns are enforced to agree some constraints of the coreferent mentions. The following attributes are used for these constraints:
    - Number: Singular or plural based on some constraint.
    - Gender: Using Stanford Gender output.
    - Person: person attributes is assigned only to some pronouns.
    - Animacy: animacy attributes is set using NER labels, e.g.,

      PERSON is animate whereas LOCATION is not

More Details on Pronouns sieve and other sieves in reference 6 and 7

Evaluation

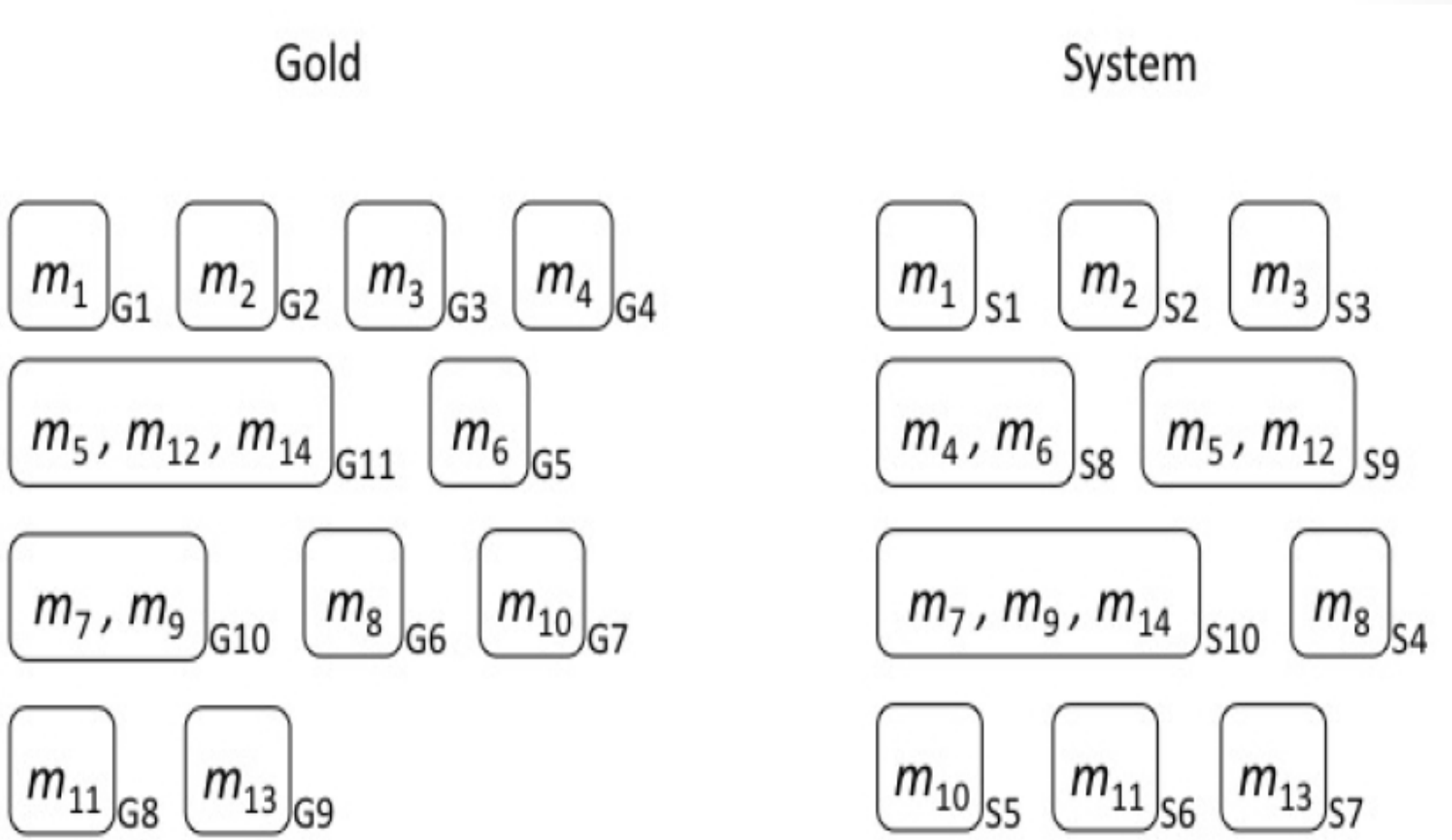
Evaluation

- In evaluating we need to compare the true set of entities (the gold partition, GOLD, produced by human expert) with the predicted set of entities (the system partition, SYS, produced by the system)

$[\text{Eyewitnesses}]_{m_1}$ reported that $[\text{Palestinians}]_{m_2}$ demonstrated today Sunday in $[\text{the West Bank}]_{m_3}$ against $[\text{the } [\text{Sharm el-Sheikh}]_{m_4} \text{ summit to be held in } [\text{Egypt}]_{m_6}]_{m_5}$. In $[\text{Ramallah}]_{m_7}$, $[\text{around 500 people}]_{m_8}$ took to $[[\text{the town}]_{m_9}\text{'s streets}]_{m_{10}}$ chanting $[\text{slogans}]_{m_{11}}$ denouncing $[\text{the summit}]_{m_{12}}$ and calling on $[\text{Palestinian leader Yasser Arafat}]_{m_{13}}$ not to take part in $[\text{it}]_{m_{14}}$.

Fig. 1. Example of coreference (from ACE-2004).

Evaluation


Gold
m_1 G1
m_2 G2
m_3 G3
m_4 G4
m_5, m_{12}, m_{14} G11
m_6 G5
m_7, m_9 G10
m_8 G6
m_{10} G7
m_{11} G8
m_{13} G9
System
m_1 S1
m_2 S2
m_3 S3
m_4, m_6 S8
m_5, m_{12} S9
m_7, m_9, m_{14} S10
m_8 S4
m_{10} S5
m_{11} S6
m_{13} S7

Evaluation Metric

- **MUC** (Vilain etal. 1995)
- **B3** (Bagga and Baldwin 1998)
- **CEAF** (Luo 2005)
- **BLANC** (Recasens & Hovy, 2010)

MUC

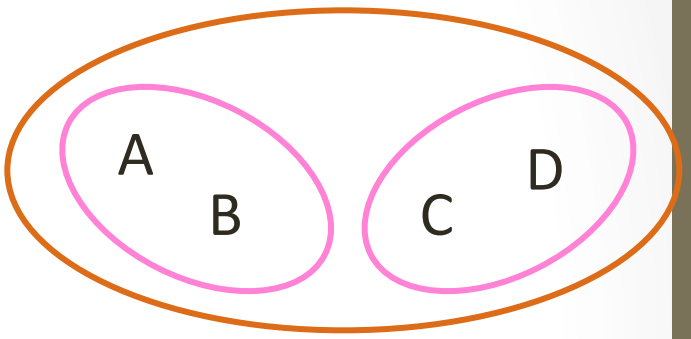
- Identify the **minimum number of link modifications** required to make the set of mentions identified by the system as coreferring perfectly align to the gold-standard set. That is, the total number of mentions minus the number of entities.
  - S_i is a coreference chain
  - $p(S_i)$ is a partition of Si relative to the system response.

$$Recall = \frac{\sum(|S_i| - |p(S_i)|)}{\sum(|S_i| - 1)} \qquad F = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$$

Precision, on the other hand, is defined in the opposite way by switching the role of Gold and System response.

MUC-Example

- Gold = [A, B, C, D]
- System Response = [A, B], [C, D]



Recall $\frac{4 - 2}{3} = 0.66$

Precision $\frac{(2 - 1) + (2 - 1)}{(2 - 1) + (2 - 1)} = 1.0$

F-measure $\frac{2 * 2/3 * 1}{2/3 + 1} = 0.79$

*B*3

- Problem with MUC:
  - Only gain points for links
  - All errors are equal
  - Cannot represent singleton entities
- Instead of looking at the links, B-CUBED metric measures the accuracy of coreference resolution based on individual mentions.
- R_{m_i} is the response chain and K_{m_i} is the key chain(gold)

$$Precision(m_i) = \frac{|R_{m_i} \cap K_{m_i}|}{|R_{m_i}|} \quad Recall(m_i) = \frac{|R_{m_i} \cap K_{m_i}|}{|K_{m_i}|}$$

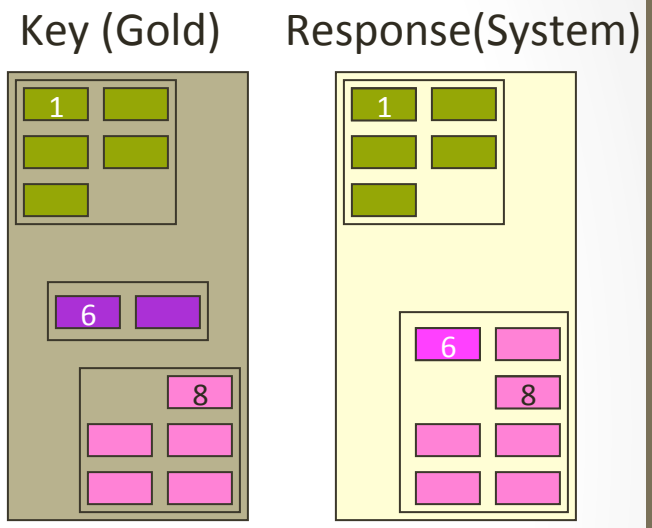
The overall precision and recall are computed by averaging them over all mentions.

B-Cubed Example

$$Pr(m_1) = \frac{|Common|}{|Response\ mentions|} = \frac{5}{5}$$

$$Pr(m_6) = \frac{|Common|}{|Response\ mentions|} = \frac{2}{7}$$

$$Pr(m_8) = \frac{|Common|}{|Response\ mentions|} = \frac{5}{7}$$



$$Precision = \frac{1}{12}\left(Pr(m_1) + Pr(m_2) + \ldots + Pr(m_{12})\right) = 0.76$$

Recall = 1.0

F-Measure = 0.865

Existing $B3$ variants

B_0^3 and B_{all}^3 Stoyanov et al. (2009)

Example: Key : {a b c}
Response: {a b d}

B_0^3 discards all twinless system mentions (i.e. mention d) and penalizes recall by setting recall(mi)= 0 for all twinless key mentions (i.e. mention c).

$$Pr_{B_0^3} = \frac{1}{2}\left(\frac{2}{2} + \frac{2}{2}\right) = 1.0$$

$$Rec_{B_0^3} = \frac{1}{3}\left(\frac{2}{3} + \frac{2}{3} + 0\right) \doteq 0.444$$

$$F_{B_0^3} = 2 \times \frac{1.0 \times 0.444}{1.0 + 0.444} \doteq 0.615$$

Existing $B3$ variants-cont'd

Example: Key : {a b c}
Response: {a b d}

B^3_{all} retains twinless system mentions. It assigns $1/|R_{m_i}|$ to a twinless system mention as its precision and similarly $1/|K_{m_i}|$ to a twinless key mention as its recall.

$$Pr_{B^3_{all}} = \frac{1}{3}(\frac{2}{3} + \frac{2}{3} + \frac{1}{3}) \doteq 0.556$$
$$Rec_{B^3_{all}} = \frac{1}{3}(\frac{2}{3} + \frac{2}{3} + \frac{1}{3}) \doteq 0.556$$
$$F_{B^3_{all}} = 2 \times \frac{0.556 \times 0.556}{0.556 + 0.444} \doteq 0.556$$

- Other B3 variants:
  - $B^3_{r\&n}$ Rahman & Ng (2009)
  - B^3_{sys} Cai and Strube (2010)

CEAF (Luo 2005)

- Luo (2005) criticizes the *B*3 algorithm for using entities more than one time, because *B*3 computes precision and recall of mentions by comparing entities containing that mention.

$$Precision = \frac{\Phi(g^*)}{\sum_i \phi(R_i, R_i)} \qquad Recall = \frac{\Phi(g^*)}{\sum_i \phi(K_i, K_i)}$$

$$\Phi(g^*) = \max \begin{cases} \phi_3(K_i, R_j) = |K_i \cap R_j| \\ \phi_4(K_i, R_j) = \dfrac{2|K_i \cap R_j|}{|K_i| + |R_j|} \end{cases}$$

CEAF Example

Example:

Key : {a b c}
Response: {a b d}

$$\phi_3(K_1, R_1) = 2\,(K_1 : \{abc\}; R_1 : \{abd\})$$
$$\phi_3(K_1, K_1) = 3$$
$$\phi_3(R_1, R_1) = 3$$

$$Pr_{CEAF_{orig}} = \frac{2}{3} = 0.667$$
$$Rec_{CEAF_{orig}} = \frac{2}{3} = 0.667$$
$$F_{CEAF_{orig}} = 2 \times \frac{0.667 \times 0.667}{0.667 + 0.667} = 0.667$$

- Other CEAF variants:
  - $CEAF_{r\&n}$ Rahman & Ng (2009)
  - $CEAF_{sys}$ Cai and Strube (2010)

BLANC (Recasens & Hovy, 2010)

- This measure implements the *Rand index* (Rand, 1971) which has been originally developed to evaluate clustering methods

		SYS		
		Coreference	Non-coreference	Sums
GOLD	Coreference	rc	wn	$rc + wn$
	Non-coreference	wc	rn	$wc + rn$
Sums		$rc + wc$	$wn + rn$	L

Where L = N*(N-1)/N

N is the total number of mentions in a document D

L is the total number of mention pairs (i.e., pairwise links) in D, thereby including both coreference and non-coreference links

BLANC-cont'd

Score	Coreference	Non-coreference	
P	$\mathrm{P}_c = \frac{rc}{rc+wc}$	$\mathrm{P}_n = \frac{rn}{rn+wn}$	BLANC-P $= \frac{\mathrm{P}_c+\mathrm{P}_n}{2}$
R	$\mathrm{R}_c = \frac{rc}{rc+wn}$	$\mathrm{R}_n = \frac{rn}{rn+wc}$	BLANC-R $= \frac{\mathrm{R}_c+\mathrm{R}_n}{2}$
F	$\mathrm{F}_c = \frac{2\mathrm{P}_c\mathrm{R}_c}{\mathrm{P}_c+\mathrm{R}_c}$	$\mathrm{F}_n = \frac{2\mathrm{P}_n\mathrm{R}_n}{\mathrm{P}_n+\mathrm{R}_n}$	BLANC $= \frac{\mathrm{F}_c+\mathrm{F}_n}{2}$

rc : the number of right coreference links.

wc: the number of wrong coreference links.

rn: the number of right non-coreference links.

wn: the number of wrong non-coreference links.

BLANC Example

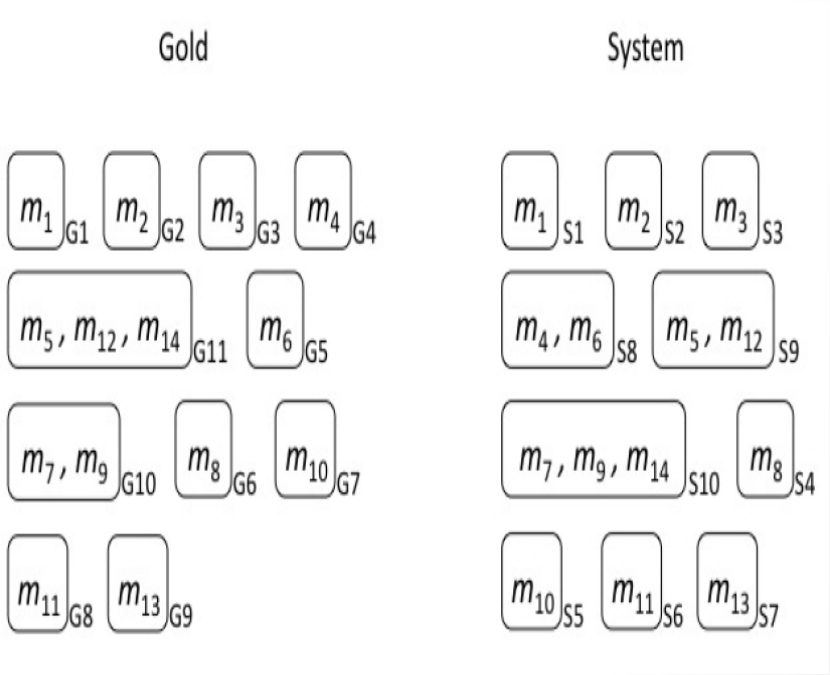
		SYS Coreference	SYS Non-coreference	Sums
GOLD	Coreference	2	2	4
	Non-coreference	3	84	87
Sums		5	86	91

rc {m5-m12, m7-m9}

wc {m4-m6, m7-m14, m9-m14}

wn {m5-m14,m12-m14}

rn {The rest are 84 right non-coreference links.}



Real World Evaluation

Stanford coreNLP detailed results:

MUC			B^3			CEAFE			BLANC			
R	P	F1	R	P	F1	R	P	F1	R	P	F1	avg F1
61.8	57.5	59.6	68.4	68.2	68.3	43.4	47.8	45.5	70.6	76.2	73.0	57.8

The evaluation of coreference has been a tricky issue and there does not exist a silver bullet, researchers often use only one or two measures when evaluating their systems.

CoNLL-2011 Shared Task Official results:

System	MD	MUC	B-CUBED	$CEAF_m$	$CEAF_e$	BLANC	Official
	F	F^1	F^2	F	F^3	F	$\frac{F^1+F^2+F^3}{3}$
lee	**70.70**	59.57	68.31	**56.37**	**45.48**	73.02	**57.79**
sapena	43.20	59.55	67.09	53.51	41.32	71.10	55.99
chang	64.28	57.15	**68.79**	54.40	41.94	**73.71**	55.96
nugues	68.96	58.61	65.46	51.45	39.52	71.11	54.53
santos	65.45	56.65	65.66	49.54	37.91	69.46	53.41
song	67.26	**59.95**	63.23	46.29	35.96	61.47	53.05
stoyanov	67.78	58.43	61.44	46.08	35.28	60.28	51.92

Evaluation Issues

- "A long-standing weakness in the area of anaphora resolution: the inability to fairly and consistently compare anaphora resolution algorithms due not only to the difference of evaluation data used, but also to the diversity of pre-processing tools employed by each system." (Barbu & Mitkov, 2001)

- "Haghighi and Klein (2010) compare four state-of-the-art systems on three different corpora and report B3 scores between 63 and 77 points. While the corpora used in (Haghighi and Klein, 2010) are different from the one in this shared task, our result of 68 B3 suggests that our system's performance is competitive."(Lee et al., 2011)

Error Samples - Quebec Wiki

the ministere des Transports du Quebec	REP ??
Quebec	☺
Canada 's second most populous province	☺
The name `` Quebec ''	😐
the highest in Quebec	☹
a religious tourism destination	☺☺
Opera de Quebec	☹
Quebec , such as the Festival d'ete de Quebec	☹☹

Sentence	Head	Text
4	14 (gov)	James Bay - The Gulf
38	41	James - The Gulf
195	19	James Murray -Quebec governor, Born 1721

Error Samples - Canada Wiki

Canada 's Atlantic coast	Rep???
Canada	☺
a North American country consisting of ten provinces and three territories	☺
The name Canada	😐
the Atlantic coast	☹
Canada and the United States	☹☹
Canada , Italy , the United Kingdom , Norway , and Russia	Come on!!!

Sentence	Head	Text
67	17 (gov)	Alberta and Saskatchewan
83	27	Quebec and Alberta

Sentence	Head	Text
68	28 (gov)	World War I. Volunteers sent to the Western Front
70	10	World War

Future Work

- Many errors in coreference resolution come from semantic mismatches due to inadequate world knowledge.
- Most researches focus on implementing external knowledge base(Wikipedia, FreeBase, WorldNet,…) in order to align KB concepts with text mentions.
- Stanford(2011):
  - Obama{Singular, Male, Person, Animated}
- By including Wikipedia category concept- Ratinov and Roth (2012):
  - Obama{Singular, Male, Person, Animated, American, President}

Error Example:

Obama tried to quit smoking …,Michelle Obama said that he had successfully quit smoking .

Reference

- [1] *Hobbs, Jerry R., 1978, ``Resolving Pronoun References'', Lingua, Vol. 44, pp. 311-338.*
- [2] *Natural Language Processing, B. Grosz, K. Sparck-Jones, and B. Webber, editors, pp. 339-352, Morgan Kaufmann Publishers, Los Altos, California.*
- [3] Andrew Kehler, Douglas Appelt, Lara Taylor, and Aleksandr Simma. 2004. Competitive Self-Trained Pronoun Interpretation. In *Proceedings of NAACL 2004*, 33-36.
- [4] Grosz, Barbara J., Aravind Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. Computational Linguistics, 21(2):203-225
- [5] Brennan, Susan E., Marilyn W. Friedman, and Carl J. Pollard. 1987. A centering approach to pronouns. In Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics, pages 155-162.

Reference

- [6] Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and C. Manning 2010. A Multi-Pass Sieve for Coreference Resolution. In EMNLP.

- [7] Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task.

- [8] Manning et al.,2014. The Stanford CoreNLP Natural Language Processing Toolkit

- [9] Bagga, Amit & Breck Baldwin (1998). Algorithms for scoring coreference chains. In *Proceedings of the 1st International Conference on Language Resources and Evaluation,* Granada, Spain, 28–30 May 1998, pp. 563–566.

- [10] Luo, Xiaoqiang (2005). On coreference resolution performance metrics. In *Proceedings of the Human Language Technology Conference and the 2005 Conference on EmpiricalMethods in Natural Language Processing,* Vancouver, B.C., Canada, 6–8 October 2005, pp. 25–32.

Reference

- [11] Rahman, Altaf & Vincent Ng (2009). Supervised models for coreference resolution. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 6-7 August 2009, pp. 968–977.

- [12] B. Amit and B. Baldwin. 1998. Algorithms for scoring coreference chains. In MUC-7.

- [13] Cai, J., and Strube, M. 2010. Evaluation metrics for end-to-end coreference resolution systems. In Proceedings of SIGDIAL, pp. 28{36. University of Tokyo, Japan.

- [14] Hovy and RECASENS 2010, BLANC: Implementing the Rand index for coreference evaluation

End