LOT Winter School 2005

Discourse Anaphora: Theories, Data and Applications

Bonnie Webber

School of Informatics

University of Edinburgh

January 2005

Day 2: Heuristic Approaches to Anaphor Resolution

Outline

- Baseline algorithms for resolving coreferential pronouns: [Hobbs 78], [Lappin & Leass 95]
- Heuristic approaches to resolving coreferential pronouns: Centering
- Interlude: Performance evaluation [Stuckardt 2003; Byron 2001]
- Discussion
- Summary

Questions to Ask of an Algorithm

- What factors does the algorithm (try to) take account of?
- What is its coverage? Does it handle all/some instances of a particular form or all/somm instances serving a particular function?
- When are resolution decisions made?
- What is the effect of a resolution decision? Does it change the state of the system?
- What has the algorithm been evaluated on?
- How well does it do?

[Hobbs 78]

[Hobbs 78] established a baseline in terms of syntax-guided resolution of coreferential pronouns.

Input: Parse tree of each sentence in the text up to and including current S.

Overview: Parse trees are searched for the antecedent in order of *recency*. *Syntactic preferences* are approximated by the order in which search performed.

Grammar fragment defining SynStruc, used in search order:

S $\rightarrow$ NP VP	PP $\rightarrow$ preposition NP
NP $\rightarrow$ (Det) Nominal ({PP | Rel})*	Rel $\rightarrow$ wh-word S
NP $\rightarrow$ pronoun	VP $\rightarrow$ verb NP (PP)*
NP $\rightarrow$ npr	Nominal $\rightarrow$ (adj)* noun (PP)*
Det $\rightarrow$ determiner	Det $\rightarrow$ NP 's

$\Rightarrow$ Baseline Algorithms - [Hobbs 1978]

1. Begin at the NP node immediately dominating the pronoun.

2. Go up to the first NP or S above it. Call the node X and the path to it p.

3. Do a LR breadth-first traversal of all branches below X to the left of p . Propose as antecedent any NP node encountered that has an NP or S between it and X.

4. If X is the highest S in the sentence, consider the parse trees of previous Ss in *recency* order and traverse each in turn in LR breadth-first order. When an NP is encountered, propose it as an antecedent. If X is not the highest S, go to step 5.

5. From X, go up to the first NP or S above it. Call it X and the path to it p.

6. If X is an NP and p doesn't pass through the Nominal that X immediately dominates, propose X as an antecedent.

7. Do a LR breadth-first traversal of all branches below X to the left of p. Propose any NP encountered as the antecedent.

8. If X is an S, do a LR breadth-first traversal of all branches below X to the right of p, but don't go below any NP or S encountered. Propose any NP encountered as the antecedent.

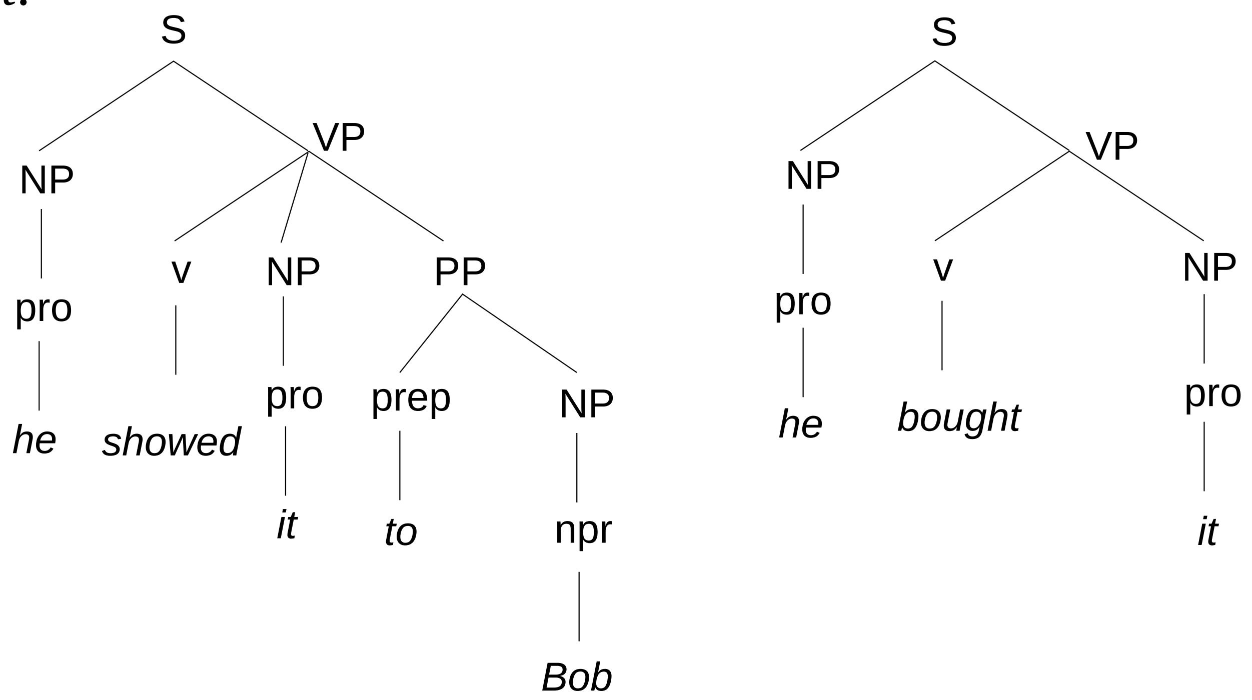
9. Go to step 4.

Note:

- When an NP is proposed as antecedent, gender/number agreement is checked.
- Algorithm covers *he*, *she*, *they*, *his (he+'s)*, *her (she)*, *them (they)*, *their (they+'s)*. Also *it*, when not referring to a clause or a time/weather construction.
- Performance of algorithm is improved by applying simple selectional restrictions – e.g. Dates, places, large fixed objects can't move.

Applying Hobbs' algorithm

(1) John saw a beautiful Integra at the dealership. *He* showed *it* to Bob. *He* bought *it*.



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

Performance of the Hobbs' Algorithm

Applied to one sample each of technical writing, fiction and news magazine.

	#	C_0	C_1-C_0	C_2-C_1	C_3-C_2	C_9-C_8	Alg correct	Alg correct after selection
he	139	126	10	2	0	1	130	130
she	7	7	0	0	0	0	7	7
it	71	64	4	1	2	0	55	59
they	83	74	9	0	0	0	73	79
Total	300	271	23	3	2	1	265	275

C_0 = entities in current & previous S, if pronoun **before** verb; entities in current S if **after** verb.

C_1 = entities in current & previous S.

C_n = entities in current & previous n Ss.

Hobbs' Algorithm: Instances of conflict

Important: Hobbs observed that in over half the cases (168/300 = 56%), there was only *one* nearby plausible antecedent against which to resolve the pronoun (although he say precisely what he means by "nearby").

	Conflicts before selection	Alg correct	Conflicts after selection	Alg correct
he	31	22	31	22
she	0	0	0	0
it	48	33	44	33
they	53	43	45	41
Total	132	98	120	96

Now [Kehler et al 2004a] are trying to exploit these "no conflict" cases in a "self-training" approach to pronoun resolution.

[Lappin & Leass 95]

[Lappin & Leass 95] was the first algorithm to be developed for, applied to and evaluated on a substantial text corpus: IBM computer training manuals.

Main features: It approximates a discourse model through *equivalence classes* – bundling together all references to the same entity.

Contrast:

- [Hobbs 1878] organises the search space, such that the first NP found that satisfies constraints is taken to be a pronoun's antecedent.
- [Lappin & Leass 95] uses an empirically developed weighting scheme on equivalence classes. The class with the highest weight is taken to be the pronoun's referent.

Overview

- L&L associates initial *salience value* with a new referential NP based on a set of *salience factors* ≈ *updating DM* with new entity.

- L&L resolves pronouns against entities in DM.
- L&L updates salience values of entities in DM before the next sentence is processes, halving their value and contributing to a *recency* preference.
- Since several NPs (with different salience values based on salience factors) may refer to the same referent, need a way of combining all their contributions. So all co-referring NPs are put into the same *equivalence class*. The weight that a salience factor assigns to an entity is the highest of the weights it assigns to the members of its equivalence class.
- Mention of an entity in successive sentences results in weights being added. Multiple mentions within the same sentence results in the maximum value of each saliency factor being assigned.

[Lappin & Leass 95] Salience Value

1. A referential NP is paired with its *salience value*, which is the sum of its salience factors:

sentence recency (added to current S)	100
subject emphasis (added if NP is S subject)	80
existential emphasis (added if “there is NP ...”)	70
direct object emphasis	50
indirect object or oblique complement emphasis	40
non-adverbial emphasis (added if NP isn’t in demarcated advP)	50
head noun emphasis (added if NP isn’t embedded in another NP)	80

L&L: Resolving a 3rd person pronoun (L$\rightarrow$R)

1. Identify potential equiv classes (from up to the previous four sentences) whose salience value exceeds some threshhold.

2. Remove any that don't agree in number or gender with the pronoun.

3. Remove any that fail binding (i.e., intra-sentential syntactic coreference) constraints.

4. Compute the total salience value of each remaining referent by adding salience factors corresponding to role parallelism (+35) and cataphora (-175), to model these preferences and dispreferences.

5. Select referent (equiv class) with highest salience value. Add pronoun to its equiv class. In case of tie, choose closest in terms of string position (direction independent). Update salience values.

Applying L&L algorithm

(3) Sue drives an Alfa Romeo.

ID	Equiv Class$_i$	Value$_i$	Equiv Class$_o$	Value$_o$
e_{Sue}	*Sue*	310	*Sue*	310
e_{Alfa}	*Alfa Romeo*	280	*Alfa Romeo*	280

(4) **She** drives too fast.

ID	Equiv Class$_i$	Value$_i$	Equiv Class$_o$	Value$_o$
e_{Sue}	*Sue*	155	*Sue, she*$_4$	310
e_{Alfa}	*Alfa Romeo*	140	*Alfa Romeo*	140

(5) Mary races **her** on week–ends

ID	Equiv Class$_i$	Value$_i$	Equiv Class$_o$	Value$_o$
e_{Mary}			*Mary*	310
e_{Sue}	*Sue, she*$_4$	155	*Sue, she*$_4$, *her*$_5$	280
e_{Alfa}	*Alfa Romeo*	70	*Alfa Romeo*	70

(6) **She** goes to Laguna Seca.

ID	Equiv Class$_i$	Value$_i$	Equiv Class$_o$	Value$_o$
e_{Mary}	*Mary*	155	*Mary, she*$_6$	310
e_{Sue}	*Sue, she*$_4$, *her*$_5$	140	*Sue, she*$_4$, *her*$_5$	140
e_{Alfa}	*Alfa Romeo*	35	*Alfa Romeo*	35
e_{LS}	*Laguna Seca*	270	*Laguna Seca*	270

(6’) **She** often beats **her**.

ID	Equiv Class$_i$	Value$_i$	Equiv Class$_o$	Value$_o$
e$_{Mary}$	*Mary*	155	*Mary*, $she_{6'}$	310
e$_{Sue}$	*Sue*, she_4, her_5	140	*Sue*, she_4, her_5, $her_{6'}$	280

Performance of the Lappin & Leass algorithm

RAP was tuned on a corpus consisting of 560 3^{rd} person pronouns (including reflexives and reciprocals) from five different computer manuals.

There was some lexical and syntactic substitution to improve parses, but not so as to change syntactic relations.

Results of training phase	Total	Intersentential cases	Intrasentential cases
Number of pronouns	560	89	471
Number of correct resolutions	475 (85%)	72 (81%)	403 (86%)

Results of testing phase	Total	Intersentential cases	Intrasentential cases
Number of pronouns	360	70	290
Number of correct resolutions	310 (86%)	52 (74%)	258 (89%)

Interesting observation: Applied to the same test cases, the Hobbs' algorithm and RAP agreed on 83% of cases, showing significant convergence between salience measured by RAP and the configurational prominence used in the Hobbs' algorithm on a language (English) with relatively fixed word order which is usually a good indicator of grammatical roles.

⇒ Centering Theory

Centering [Grosz, Joshi & Weinstein 83, 95] is a theory of

- *Entity Coherence*, in terms of how entities are introduced and discussed in a discourse, and
- *Entity Salience*, predicting which entities are most salient at a given time and hence require least effort to access.

The hearer's constantly changing *local attentional state* tracks changes in the entities introduced into the discourse and in the extent of their salience.

Main claims of Centering

- *Coherence*: All but the first utterance of a *segment* have a unique main link with the previous utterance, called the *backward looking center* (C_b) of the utterance.

  This unique link simplifies the complexity of procedures required to integrate an utterance into the discourse.

- *Salience*: The set of entities *realised* by an utterance can be rank-ordered by their *salience*.

  The is called the *CF-list* of an utterance, the list of its *forward-looking centers*.

Further hypotheses

- Identity of C_b is determined by this ranking. For utterance U_j, $C_b(U_j)$ is the highest ranked element of $C_f(U_{j-1})$ that is realised in U_j.

- If any entity is realised in an utterance by a pronoun, its C_b must be. (**Rule 1** in [GJW 95].

Coherence and Transitions

Centering takes (local) discourse coherence to depend on how entities are introduced and discussed.

Thus it needs means of characterising how certain ways of introducing and discussing entities are more or less coherent than other ways.

It does so by means of *transitions* between adjacent utterances.

Preferred center ($C_p(U_j)$): First element in CF-list of U_j.

Consider transitions between utterance U_{j-1} and U_j.

	$B_j = B_{j-1}$ or no B_{j-1}	$B_j \neq B_{j-1}$
$B_j = P_j$	continue (CON)	smooth–shift
$B_j \neq P_j$	retain (RET)	rough–shift

Transition Preferences

1. Some approaches to centering have preferences for *sequences* of transitions – e.g. CON-CON $\gg$ SHIFT-SHIFT

   [Di Eugenio 1998] shows that these larger patterns are predictive of the use of null/explicit pronouns in Italian.

2. Other approaches express preferences for individual transitions – e.g. CON $\gg$ RET $\gg$ smooth–shift $\gg$ rough–shift

3. [Strube & Hahn 1999] consider “cheap” transition pairs preferable to “expensive” ones, where

   - A transition is “cheap” if the CB of the current utterance is correctly predicted by the CP of the previous one – i.e., $B_j = P_{j-1}$
   - A transition is “expensive” if $B_j \neq P_{j-1}$

Summary of Centering Algorithm [BFP 87]

1. CREATE a list of ref exprs (REs) for U_n ordered by grammatical role.

2. IDENTIFY all possible Cf lists for U_n (i.e., a set of lists) by associating each RE with each discourse entity it can refer to.

3. IDENTIFY all possible Cb's for U_n (i.e. a set of elements). These are all discourse entities from U_{n-1} plus NIL (to allow for the absense of a Cb).

4. GENERATE all possible Cb_n/Cf_n combinations (called "anchors").

5. FILTER each anchor by *binding constraints* and adherence to centering rules.

   - Go through Cf_{n-1} keeping only those which appear in Cf of anchor. If anchor Cb does not equal its first elt, eliminate anchor. (Constraint on Cb realises highest ranking elt appearing in U_n.)
   - If no entity realised as a pronoun in anchor Cf equals anchor Cb, eliminate anchor. (Constraint: if anything pronominalised, C_b is.)

6. CLASSIFY by transition type and RANK by transition ordering.

7. SELECT highest ranking assignment.

(7) Sue drives an Alfa Romeo.

(8) **She** drives too fast.

Cf_8 = (E1: Sue)

Cb_8 = E1: Sue

Cp_8 = E1: Sue

continue

(9) Mary races **her** on week–ends.

Cf_9 = (R1:E2, R2:her$\rightarrow$E1)

Cb_9 = E1: Sue

Cp_9 = E2: Mary

retain: $Cb_9 = Cb_8$ *but* $Cb_9 \neq Cp_9$

(10) **She** goes to Laguna Seca.

i. *continue*: Cb_{10}=Cb_9 and Cb_{10}=Cp_{10}

Cb_{10} = E1: Sue

Cf_{10} = (R3: E1)

Cp_{10} = E1: Sue

ii. *retain*: $Cb_{10} \neq Cb_9$ but Cb_{10}=Cp_{10}

Cb_{10} = E2: Mary

Cf_{10} = (R3: E2)

Cp_{10} = E2: Mary

10’. **She** often beats **her**.

i. $Cb_{10'}$ = E2: Mary

$Cf_{10'}$ = (R3: E2, R4: E1)

$Cp_{10'}$ = E2: Mary

smooth-shift: $Cb_{10'} \neq Cb_9$ but $Cb_{10'} = Cp_{10'}$

ii. $Cb_{10'}$ = E2: Mary

$Cf_{10'}$ = (R3: E1, R4: E4)

$Cp_{10'}$ = E1: Sue

rough-shift: $Cb_{10'} \neq Cb_9$ and $Cb_{10'} \neq Cp_{10'}$

Strube's incremental model [Strube 98]

Centering theory operates at the sentential level, and so fails to account for intrasentential uses of pronouns.

Strube's model is incremental at the phrase level.

The model consists of:

- The **S–list**: a list of salient discourse entities which describes the local attentional state of the Hearer.
- An **insertion operation** on that list which orders ($\prec$) the items according to their information status (given,new,mediated).

Ranking constraints on the S–list

Ranking is based on Prince's familiarity scale:

- OLD: evoked and unused entities.
- NEW: brand–new entities.
- MEDIATED: inferrables, containing–inferrables, anchored brand–new entities.

- If X is old, Y mediated then X $\prec$ Y.
- If X is old, Y new then X $\prec$ Y.
- If X is mediated, Y new then X$\prec$ Y.
- Else % (i.e. they have the same status)
  - if S_Y precedes S_X then X $\prec$ Y.
  - if X and Y occur in the same sentence and X precedes Y then X$\prec$ Y.

Resolution algorithm [Strube 98]

Process a text from left to right and for each word W:

1. If W is a pronoun, (i) resolve pronoun to first compatible element in the S–list and (ii) update s–list.

2. If W is a sentence final dot, remove all entities from S–list which are not realised in the sentence.

Example [Strube 98]

a. Sue drives an Alfa Romeo.

$\text{sue}_O\text{:sue} \prec \text{alfa}_N\text{:alfa}$

b. **She** drives too fast.

$\text{sue}_O\text{:she}$

c. Mary

$\text{mary}_O\text{:mary} \prec \text{sue}_O\text{:she}$

races **her** on week–ends

$\text{mary}_O\text{:mary} \prec \text{sue}_O\text{:her}$

d. **She** goes to Laguna Seca.

$\text{mary}_O\text{:she} \prec \text{sue}_O\text{:her}$

$\text{mary}_O\text{:she} \prec \text{LagunaSeca}_O \prec \text{sue}_O\text{:her}$

d'. **She** often beats **her**.

$\text{mary}_O\text{:she} \prec \text{sue}_O\text{:her}$

$\text{mary}_O\text{:she} \prec \text{sue}_O\text{:her}$

LRC Algorithm [Tetreault 01]

Like Strube, Tetreault disliked the lack of incremental processing in BFP. So his *Left-Right Centering* (LRC) algorithm resolves pronouns incrementally.

1. Take Cb(U_{n-1}) and Cf(U_{n-1}) from previous utterance.
2. Parse and extract incrementally from U_n all references to discourse entities. For each pronoun,
   (a) Search for an antecedent *intrasententially* in Cf-partial(U_n) that meets feature and binding constraints. If found, continue reference extraction. Otherwise
   (b) Search for an antecedent *intersententially* in Cf(U_{n-1}) that meets feature and binding constraints.

3. Create Cf-list for U_n by ranking discourse entities mentioned in U_n according to grammatical function. (Sorting by grammatical function approximated through a LR breadth-first walk of the parse tree.)

	NY Times		Fiction	
algorithm	right	success rate (%)	right	success rate (%)
BFP	1004	59.4	241	46.4
S-list (Strube)	1211	71.7	337	66.1
LRC	1268	74.9	372	72.1
Hobbs	1298	76.8	414	80.1
LRC-F	1362	80.4	420	81.1

LRC-F varies LRC by moving entities in prepended (topicalised) phrases to the back of the Cf-list.

Detailed Analysis of Centering [Poesio et al 2004]

[Poesio et al 2004] examine the concepts that are *parameters* of centering, that are designed to be tuned, to better capture the entity coherence properties of different languages.

- *Utterance* and *previous utterance*: sentence? tensed clause? any clause? any clause other than embedded clause? VP coordination as one clause or two?
- *Realisation*: directly through an NP? via associative reference to an NP? via traces? via all pronouns or only 3^{rd} person pronouns
- *Ranking*: by grammaical function? by order of (direct) realisation? by position in Prince's *givenness hieraarchy*? by thematic role?
- *Pronominalisation of CB*: with a 3^{rd} person (singular) pronoun? with a trace? with a demonstrative pronoun? with a plural pronoun?
- *Span of transitions*: over successive utterances? over triplets of utterances?
- *Transition preferences*: strictly by type? "cheap" versus "expensive" transitions?

⇒ Interlude: Performance Evaluation

- Methods of evaluating coreference resolution & anaphor resolution
- Methods for reporting results: Comparability

Background

Consider a test that assigns a label A or $\bar{A}$ to items in a set, each of which is really one of A or $\bar{A}$. (Such a test would be called a *classifier* in machine learning, a *diagnostic* in medicine.)

The test can be characterised by its *Confusion Matrix*:

	Label		
Actual	A	$\bar{A}$	
A	true positive (TP)	false negative (FN)	Recall = TP/(TP+FN)
$\bar{A}$	false positive (FP)	true negative (TN)	
	Precision = TP/(TP+FP)		

Suppose one is interested in how good the test is on A decisions.

Precision: The proportion of items labelled A (i.e., row A) that are true As

Recall: The proportion of true As (i.e., column A) that are labelled A

If one is interested in both A and $\bar{A}$ decisions,

Accuracy: The proportion of correct decisions (true As labelled A and true $\bar{A}$s labelled $\bar{A}$)

Accuracy = (TP + TN)/(TP + FN + FP + TN)

Assessing Coreference and Anaphor Resolution

Coreference resolution has been assessed in terms of the classes of coreferring expressions it finds – i.e., assessing *sets*.

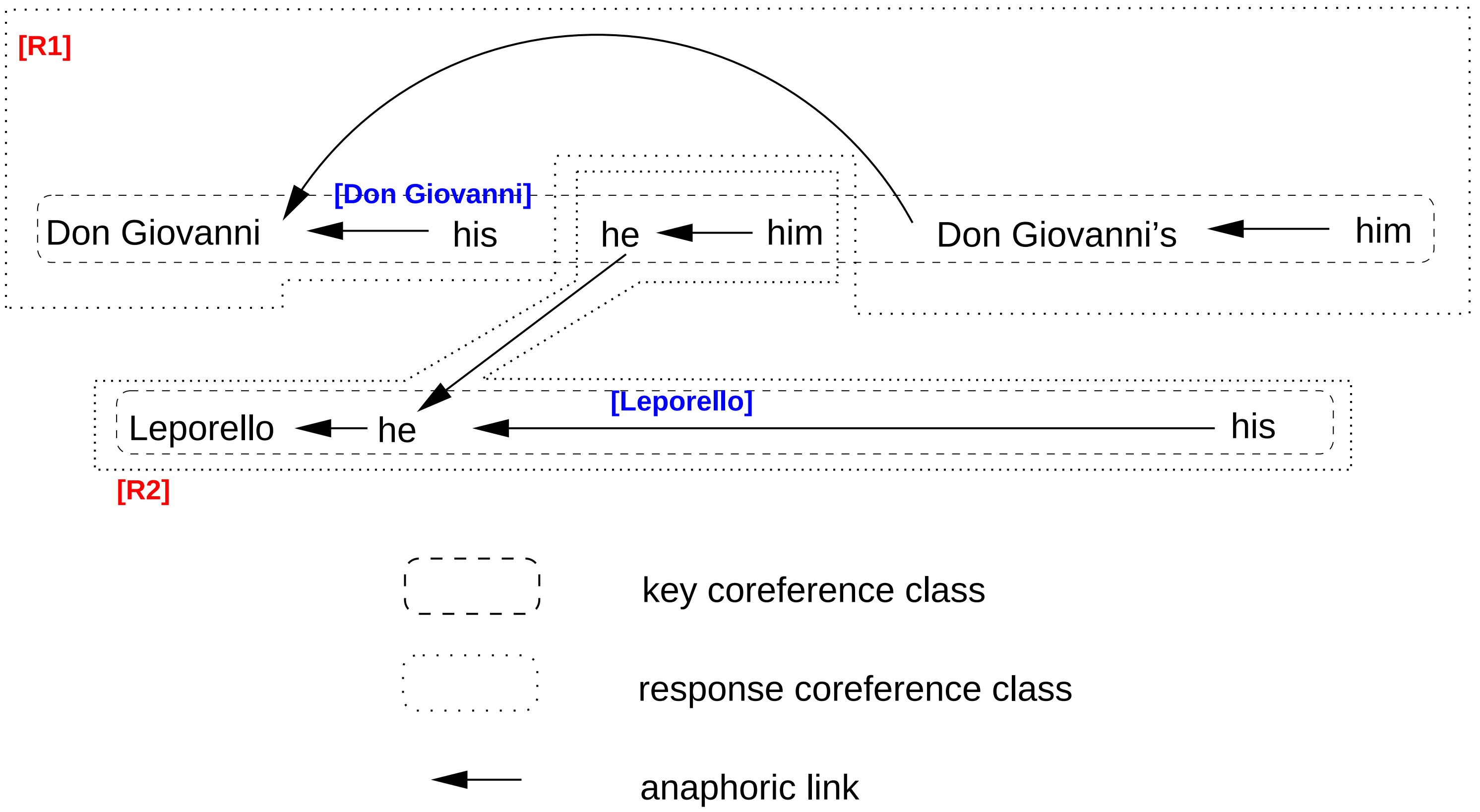
Anaphor resolution has been assessed in terms of the accuracy with each anaphoric expression is linked to its antecedent – i.e., assessing *links*.

Coreference resolution has used the output of *anaphor resolution* to compute its *coreference classes* through reflexive-transitive closure over anaphor/antecedent links.

First way of assessing coreference [Vilain et al 1996] defined **recall** and **precision** in terms of coreference sets.

Model-theoretic Scoring of Coreference [Vilain et al 1996]

Suppose one has a *key* identifying the true coreference classes (*key classes*). Also have response coreference classes (*response classes*) computed from closure over links.



- Two *key classes*: DG, L.
- Two *response classes*: R1, R2

Size of a class = $|C|$ = minimum number of links needed to create it

Recall

- $\mathrm{Recall}_{DG} \equiv \mathrm{TP}/|\text{true As}| = |DG \cap R1|/|DG| = 3/5 = 60\%$
- $\mathrm{Recall}_{L} \equiv \mathrm{TP}/|\text{true As}| = |L \cap R2|/|L| = 2/2 = 100\%$
- Overall recall = $(|DG \cap R1| + |L \cap R2|)/(|DG| + |L|) = 5/7 = 71.4\%$

Precision

- $\mathrm{Precision}_{R1} = |DG \cap R1|/|R1| = 3/3 = 100\%$
- $\mathrm{Precision}_{R2} = |L \cap R2|/|R2| = 2/4 = 50\%$
- Overall precision = $(|DG \cap R1| + |L \cap R2|)/(|R1| + |R2|) = 5/7 = 71.4\%$

Alternative way of Scoring Coreference

[Stuckhardt 2000] notes that this scoring method equates errors that have different consequences for applications:

(11) Leporello $\xleftarrow{-}$ he $\xleftarrow{+}$ him $\xleftarrow{+}$ his

(- = incorrectly linked to)

(12) Leporello $\xleftarrow{+}$ he $\xleftarrow{+}$ him $\xleftarrow{-}$ his

Stuckhardt's method is based on partitioning anaphor-antecedent pairs (α, γ) with respect to *key classes*.

- O_{++} α, γ belong to same *key class*
- O_{+-} α, γ belong to different *key classes*
- $O_{+\_}$ α corresponds to key class instance, but γ has not been assigned
- $O_{+?}$ α corresponds to key class instance, but γ doesn't.

Precision = $|O_{++}| / (|O_{++}| + |O_{+-}| + |O_{+?}|)$ (TP/resolution decisions)

Recall = $|O_{++}| / (|O_{++}| + |O_{+-}| + |O_{+?}| + |O_{+\_}|)$ (TP/total pronouns)

Applied to (11) and (12)

$\text{Precision}_1 1 = 0/(0+1+0) = 0\%$

$\text{Recall}_1 1 = 0/(0+1+2+0) = 0\%$

$\text{Precision}_1 2 = 2/(2+0+1) = 2/3 = 66.7\%$

$\text{Recall}_1 2 = 2/(2+0+0+1) = 1/3 = 66.7\%$

Methods for Reporting Results: Comparability

[Byron 2001] points out that **recall** scores for pronoun resolution are hard to compare due to differences in what pronouns in the test data the *true positives* are compared to. (**N.B.** The same thing can be said for any form with both anaphoric and non-anaphoric functions, or any form with different anaphoric functions.)

[Byron 2001] advocates a standard two-part form for reporting performance.

1. Description of test data
   - Corpus type: genre, size
   - Lexical coverage: list of each distinct pronoun *form* included in the study
   - Exclusions (in addition to excluded *forms*):
     - Are instances of pronouns function *non-referentially* excluded? e.g.

       "it" in *it-clefts*: *It* was in 1848 that revolution ...

       "it" in *extraposition*: *It* strikes me as odd that he doesn't drink beer.

       pronouns in *idioms*: "hit *it* off", "give *'em* hell"

- Are instances of pronouns functioning *cataphorically* excluded?
- In dialogue corpora, are instances of pronouns in abandoned fragments (*restarts*) excluded?
- Are plural pronouns with *split antecedents* excluded? e.g.

  (13) Pat hates Sam, but *their* sisters get along well.
- Is quoted speech excluded?
- Are pronouns with *clausal (AO)* antecedents excluded?
- Are non-corefential pronoun anaphors excluded? e.g. *type* reference?
- Are pronouns whose antecedent is outside some window of attention excluded?

2. Performance measurements

   - Precision, Recall for each type of pronoun within scope
   - **Resolution rate** = C/(T+E), where

     C = number of pronouns resolved correctly

     T = number of pronouns in the evaluation set

     E = number of excluded referential pronouns

⇒ Summary

- Considering only referential pronouns with NP antecedents, a “just in time” approach, using syntax plus number/gender agreement is surprisingly effective. (Hobbs’ approach)
- While “anticipatory” approaches, based on maintaining a partially or fully ordered set of entities made salident by various cognitively plausible factors seem more realistic, they work only equally well. (Lappin & Leass’ approach; Centering)
- It is important to specify what data types and tokens an algorithm applies to and what it does with what it’s not meant to apply to – *manual exclusion? accepted errors? algorithmic exclusion?*
- We’ll see much more attention to this aspect of algorithms when looking at recognizing and resolving anaphoric definite NPs.