Coreference Resolution

Reading The Web, class project report

Brendan O'Connor

December 11, 2009

1 Introduction

Text contains information about various entities. To extract information and perform inference from the meaning of a text, it is often necessary to identify when various expressions corefer; that is, refer to the same underlying entity. For example, consider the following two sentences.

> While establishing a [refuge]$_1$ for [Catholics]$_2$, who faced increasing [persecution]$_3$ in [Anglican England]$_4$, the [Calverts]$_5$ were also interested in creating profitable [estates]$_6$. To this end, and to avoid [trouble]$_7$ with the [British government]$_8$, [they]$_9$ also encouraged [Protestant]$_{10}$ [immigration]$_{11}$.

To answer the question *Who or what encouraged immigration?*, it is necessary to resolve mention 9, "they," to the entity it refers to. This is an example of pronoun resolution, in which case there is usually an *antecedent*, a mention that occurred previously in the text, that author is intending the pronoun to stand for. In this example, we have bracketed all noun phrases, and there are eight candidates to be possible antecedents. How should an algorithm select the correct one?

Pronoun resolution is only one subcase of coreference resolution. Entities can also be referred to by multiple different names and common nouns. For example, in the following text, *USS Cole* is coreferent to *ship* but not *boat:*

> The US navy now says the [USS Cole] was being refueled when an explosion ripped through it in Yemen last week, killing 17. The revised accounting of the incident was given in a navy statement Friday raising new questions about how the small [boat] carrying the explosives was able to get near the [ship] and set off the blast.

The coreference resolution task is heavily bound up in various syntactic and semantic phenomena. In this project my goal was to develop a simple coreference resolution system, evaluate it with a standard annotated dataset, then focus on enhancing it with improved lexical resources. Unfortunately, the first two steps — constructing the base system and its evaluation — took most of the semester. I describe the current system, give error analysis, and show the results from an experiment to use web context distributional similarity to enhance antecedent selection.

2 Description of system

We follow the general approach of Haghighi and Klein [2009], a system based on rich syntactic and semantic features. It makes nearly deterministic decisions based on strong indicators of coreference. This makes it a useful approach for two reasons. First, it is a convenient baseline system with modular components, so it is amenable to adding new semantic strategies. Second, another contribution of this work is to confirm their findings — which is possible given their system is fairly straightforward to implement if you're comfortable with linguistic engineering — and make this basic implementation available for use in future work.[1]

To analyze a document, it performs the following steps.

1. Parse and recognize named entities in the text.
2. Find all mentions.
3. For every mention, find its antecedent, if any.
   (a) Make immediate decisions for certain specific syntactic patterns.
   (b) For a pronominal mention, filter previous mentions by matching syntactic type.
   (c) For nominal and proper mentions, filter previous mentions based on matching surface features and semantic compatibility.
   (d) Among remaining filtered antecedent candidates, choose the candidate with the smallest syntactic distance. If there are no candidates, resolve to NULL.
4. To partition mentions into entity clusters, take the transitive closure of these antecedent selection decisions.

The final step is necessary to fit into the MUC/ACE view of coreference resolution, in which the task is to approximate the gold clustering of mentions as well as possible. This is probably a reasonable view of the task for the support of information extraction and information retrieval applications.

This approach depends completely on getting individual antecedent selection decisions correct; it misses opportunities to use joint information and constraints across the document, and it also can allow a single bad decision to merge many non-coreferent mentions into the same cluster.

However, the process is very transparent, since it easily viewed as a series of individual decisions. Since it relies on syntactic and semantic resources, it also provides a motivating use case and testbed for improvements to these subsystems.

Perhaps surprisingly, Haghighi and Klein [2009] find their system has state-of-the-art performance, outperforming systems based on unsupervised learning, and approaching systems based on supervised learning; our system has similar characteristics. Most of our syntactic analysis system closely follows HK.

[1] This system was built jointly with Mike Heilman.

2.1 Subsystems

We use the Stanford Parser and Stanford Named Entity Recognizer. Little is dependent on these particular choices. The parser must be a Treebank-style constituent parser. The NER was used with a MUC model; i.e., it uses the standard tag system of PER, ORG, LOC, and DATE.

2.2 Mention identification

For an unlabeled piece of text, we mark most NPs as mentions. Specifically, the system takes all NPs that are the largest possible for their head word. The head word is defined by the Collins head rules Collins [1999]. For example, if the children are a sequence of noun tokens, the head will be the rightmost token; but if the subtree has a prepositional attachment like (NP NP (PP IN NP)), then the head is the head of the left NP. This prevents repetition of redundant noun phrases that are embedded inside each other; for example,

$_{NP}$[$_{NP}$[the revised accounting] of $_{NP}$[the incident]]

In this case, *accounting* is the head of *the revised accounting*, and it is also the head of *the revised accounting of the incident*. Both noun phrases are considered as belonging to the same mention; the highest-level NP is used for syntactic pattern matching. The internal noun phrase *the incident* remains its own mention, since it is the only and largest noun phrase whose head word is *incident*.

This mention identification strategy finds pronouns, common nouns, and named mentions. It was run on several reference texts from Wikipedia and other sources, and seemed to perform reasonably well.

For evaluating on annotated ACE data, we follow previous work and use the ACE data's definitions of mentions. This can and does cause conflicts when trying to reconcile annotators' definitions of phrases with the Treebank-style parses and Collins head rules.

2.3 Immediate match patterns

Appositives are fairly easy to identify from the parse tree, and are resolved immediately; for example, in the following cases, we start from the right NP and find the left side is the immediate sibling of an intervening comma token.

- [Lawrence Tribe], the Harvard Law School [Professor] ...
- [David Boies], Gore 's chief trial [lawyer] ...

We also implemented a recognizer for role appositives, e.g. *[Republican candidate] [George Bush]*. Unlike HK, we did not find this very helpful.

The other useful pattern is the predicate-nominative construction, in which the subject and object of the sentence is mediated by a form of the verb "to be." For example,

- [Lameu] was the first NHL [player] to become a team owner.
- The [Gridiron Club] is an [organization] of 60 Washington journalists.

2.4 Pronoun resolution

Pronominal mentions are identified through the parser's part-of-speech analysis; specifically, PRP and PRP$ nodes.

First, several syntactic patterns are checked for to reject (but never immediately accept) certain candidates:

- The "I-within-I" constraint: a pronoun cannot refer to a node that dominates it. Example from HK:
  - e.g. *Walmart says Gitano, its top-selling brand, is underselling.* $\Rightarrow$ *it* $\neq$ *Gitano*
- Reflexive required for a verb's object to corefer with the subject.
  - e.g. *The bank ruined it.* $\Rightarrow$ *it* $\neq$ *bank*
  - e.g. *The bank ruined itself.* $\Rightarrow$ *itself* $=$ *bank*
- Subjects cannot refer to NPs in an adjunct phrase.
  - e.g. *To call John, he picked up the phone* $\Rightarrow$ *he* $\neq$ *John*
  - e.g. *Because John likes cars, he bought a Ferrari.* $\Rightarrow$ *he = John*

Next, syntactic type compatibility plays a major role in filtering to allowable pronoun matches. The system identifies the following types from the pronoun.

- Gender: Male, Female, Unknown (e.g. he/his vs. she/her vs. they/it)
- Personhood: Pers, NotPers, Unknown (e.g. he/she vs. it/that)
- Number: Singular, Plural (e.g. he/she/it vs. they/them/those)

Type information is inferred for antecedent candidates. For nominal and proper mentions,

- Gender: for names, match against lists of common male and female names from the U.S. Census Bureau.[2] For common nouns, check against a WordNet-derived list.
- Personhood: if the NER system gave a tag, use PERSON vs. ORG, LOC, DATE. Otherwise, check against a WordNet-derived list.
- Number: use the parser's part-of-speech analysis: NN/NNP vs NNS/NNPS.

We derived lexicons from WordNet as follows. To create a list of words that are human beings, we started with the WordNet synset for *person* (as well as a few others like *man, woman*, and *child*), took the set of all synsets that are their hyponym descendants. Every WordNet synset is associated with a number of words (surface forms); we took

[2] http://www.census.gov/genealogy/names/

the single most frequent word from each of these synsets as our list of person words. It is important to use the most frequent sense — this is information that comes with WordNet, calculated from the SemCor Brown corpus annotations — because WordNet associates very obscure senses to many words. Since we do deterministic matching of the mention word against the word lists, including words that only infrequently have the sense of person causes a many bad word sense errors. (Doing this filtering when creating the word list is basically the same as the most frequent sense baseline used in word sense disambiguation systems.)

A similar procedure is used to assemble word lists for locations, organizations, groups, and times; these lists help give syntactic type information for common nouns, for which the NER system often gives no useful tags. The final word lists have several thousand items each, look pretty clean on inspection, and have been found useful in a separate application (question generation).

This component of the syntactic analysis system is the main deviation from HK.

2.4.1 Use of syntactic type filtering

Given that reliability of the identification of these various types differs — for example, number identification is quite reliable, but personhood is harder and we sometimes give up flagging as Unknown — we experimented with different rules for the strictness of matching. For example, plural vs. singular is less definite for certain types of entities like human organizations, which can be referred to as both "they" and "it."

Gender information made little impact on the ACE development data, which is newswire text, in which its is rare for pronouns of both genders to be used in the same document. For example, the word "she" appears in only 7 of 68 documents. Gender information actually slightly hurts performance, even when used as a very lax constraint. However, personhood and number matching was very useful.

We only used syntactic type information for matching pronouns to other mentions. We did not attempt to is not used for matching for nominal mentions, since they seemed harder to reliably identify, despite our usage of various lexical resources.

Enhancing syntactic type identification should be an avenue of future work, given that the current system uses a hodepodge of lexical resources, NER, and POS analysis, but is quite useful for coresolution performance. The WordNet-derived lists could be directly used as seeds in semi-supervised learning approaches.

The word lists are useful for the final system; removing them causes performance to decrease (Table 1).

2.5 Nominal and proper resolution

Common nouns and names (a.k.a. nominal and proper mentions) are also resolved by looking for an antecedent. Unlike pronouns, it is allowable for these to have a NULL reference; for example, the first few mentions in a document usually have no antecedent. It is arguable that the antecedent selection approach, while reasonable for pronouns, doesn't fit these cases as well.

In any case, we implement only one rule for this resolution, allowing a match of exact head words match. This does make precision errors (e.g. "Korean officials" and

“Iranian officials”).

Final selection is done through shortest path distance.

2.6 Shortest path distance

The above mechanisms yield a list of antecedent candidates. If there are zero candidates, we resolve to NULL. If there are multiple candidates, we choose the one that’s closest by the syntactic path distance through the parse tree. We allow crossing between sentences by linking all sentence parses in a right-branching structure. (This can be thought of as the simplest possible discourse structure.)

Path distance outperforms selection by simple surface distance. Consider the first example in this document. The mention *they* has two plural antecedent candidates, *Calverts* and *estates*; the latter is surface-closer, but since it is embedded in a predicate clause, the first sentence’s subject, *Calverts*, is actually syntactically closer. This is the right thing to do in this and other similar examples. Path distance is better at capturing saliency.

3 Evaluation

We integrated our system with annotated ACE Phase 2 newswire data, using the same development set of 68 articles defined by Bengtson and Roth [2008] and used by HK. Unfortunately, there were numerous issues integrating this data, both mundane (our copy had mangled filenames, and ACE’s byte offset numbers have subtle bugs), as well as more fundamental mismatches between how ACE defines mentions and how our system initially did.

We evaluate precision and recall of all pairwise resolutions. To calculate precision, take all coreferent mention pairs from all predicted clusters, and the percentage of these links that are correct — i.e., links that don’t cross gold cluster boundaries — is the precision. To calculate recall, the same procedure is used on gold clusters.

Pairwise precision, recall, and F1 have the problem that they unfairly penalize errors in large clusters. Other metrics are used to more fairly compensate for this, such as the b^3 metric, which averages the per-mention accuracies of being linked to coreferent mentions. Arbitrarily, we stick with pairwise F1 for the following analysis.

The system described above performs at 64.1% precision, 48.1% recall, and 55% F1 on this dataset, which is comparable to what HK09 report for their most similar system, albeit with lower precision and higher recall. See Table 1.

3.1 Error analysis

We perform error analysis by inspecting the accuracy rates of individual antecedent selection decisions; i.e., whether the chosen antecedent from the candidate list is indeed coreferent with the mention. Note that this accuracy rate has a non-trivial relationship with cluster-aware metrics like pairwise F1. For example, if the a bad antecedent is selected but the final cluster size is only those two mentions, that hurts precision by

Table 1: Pairwise F1 performance on Bengston and Roth's ACE dev set

	P	R	F1	
Our main system	64.1	48.1	55.0	
Remove word lists	63.6	47.9	54.6	i.e. WordNet, U.S. Census
Laxer pronoun resolution				
Remove gender typecheck	64.7	48.3	55.3	slight improvement (!)
Remove person typecheck	63.0	47.4	54.1	
Remove number typecheck	56.1	46.1	50.6	
Stricter pronoun resolution				
Never resolve pronouns	75.1	26.8	39.5	
Never resolve 2nd person	66.5	46.6	54.8	
Stricter Pro.-Pro.				
Never match pro-pro	67.3	41.8	51.5	
Strict typechecking	66.2	43.4	52.4	
Check gram. number	66.5	43.7	52.7	
Distrib. sim. exper.				
Match on cos > 0.5	46.6	48.8	47.7	Matches ~2% of pairs
Match on cos > 0.3	32.6	51.2	39.9	Matches ~5% of pairs
Post-syn semantics oracle	*65.4*	*72.0*	*68.6*	
Other systems				
HK09 "SynConstr"	71.3	45.4	55.5	
HK09 full	68.2	51.2	58.5	
BR08	55.4	63.7	59.2	

Table 2: Breakdown of antecedent selection decisions

Decision Type		Corr.	Incorr.	Acc.	Notes
Imm. Rules	Appositives	105	23	82%	
	Role Appos.	5	0	100%	
	Pred.-Nom.	28	5	85%	
Standard resolution pathways					
Pronoun resolutions		460	235	66%	
Non-pronoun resolutions		1133	407	74%	
NULL		964	509	*	Errors can be recovered later

only a single false positive. But if these two mentions end up merging two different gold clusters, false positives occur for every pair between the two gold clusters.

However, since accuracy rates should be monotonic in final cluster metrics, and we're not sure that pairwise F1 is the right thing to optimize anyway, we feel this exercise is still useful and also helps gain insight into the problem.

Table 2 breaks down the types of antecedent selection decisions the system makes. The first thing to note is that the immediate syntactic pattern matches are uncommon but relatively high accuracy. Inspecting individual examples reveals a few changes could further improve precision. Appositive errors include institutional affiliation and location specification constructs. Perhaps typechecks could solve errors like the following pairs, which currently get coreferenced:

- “[David Coler], [VOA News]” “[NPR news], [Washington]“
- “[Orange County], [Calif.]” “[Washington], [D.C.]”

(The last example is arguably an error in the annotations; the correct reading under most circumstances is as a single mention.)

It is surprising that role appositives are so rare. It is worth investigating if there exist examples in the data that the current system is missing.

Predicate-nominative errors are interesting. A number of errors are due to modal verbs being picked up by the syntactic rule. These should be eliminated by forcing a stricter, smaller set of allowed verbs, and perhaps handling negations. For example, the current system resolves the following mention pairs as coreferent:

- “[I]'ll be that [president],” he added...
- [Koetter] may not have been Arizona State's top [choice].

Though a few examples seem genuinely harder: “The Taliban are predominantly Sunni Muslim...”

However, the meat of possible improvements to the system is still in pronoun and non-pronoun resolution. Given our extensive work for pronouns and syntactic types, it is disappointing to see so many remaining errors. We performed simple ablation

Table 3: Antecedent selection breakdown for pronouns occuring at least 10 times

Corr.	Incorr.	Acc.	Pronoun
4	18	18%	your
17	23	42%	you
9	8	53%	our
22	19	54%	their
39	31	56%	they
11	8	58%	them
39	22	64%	i
9	4	69%	him
7	3	70%	my
23	10	70%	we
106	34	76%	he
30	9	77%	it
36	10	78%	its
67	15	82%	his
13	1	93%	she

experiments to ensure the type checking is helping, and indeed it is; see Table 1. But many pronoun resolution cases are still difficult.

There do not seem to be any especially easy types of pronouns. Even first- and second-person pronouns, which at first glance seem odd in a newswire corpus, often get resolved correctly. A subset of pronoun resolution accuracies is shown in Table 3.1. After considering this breakdown, we found that small precision/recall tradeoffs can be made by refusing to resolve second person pronouns; this seems a bit like overfitting, though.

Another odd case is pronoun-to-pronoun matches, which we didn't even consider when building the system. It turns out many of these work OK, even when matching between seemingly type-mismatches like "I" resolving to "he" — e.g. in dialogue or quotations. We experimented with adding more typechecking, and also adding grammatical number typechecking (first vs. second vs. third person pronouns), but they only gave precision gains at cost to recall. Reported in Table 1.

As an example how dialogue and speaker shifts can be difficult, in the following our system resolves "Ray Bourque" to "he":

- "We've always stuck together and we'll stick by Patrick," defenseman [Ray Bourque] said. "We know [he] is a quality person and a great family man."

Finally, for nominal matches, there is still low-hanging fruit with surface similarity matching. Our surface matching is still very primitive, only looking to see if head words match. The data has many adjective-to-proper matching cases like ("Israeli," "Israel") that could be achieved through a string similarity metric that roughly captures English morphology. (e.g. Jaro-Winkler, or perhaps simply Levenshtein with a constraint that strings must share a prefix.)

[A note on the "NULL" row of the table: a correct "NULL" decision means the mention is actually a singleton in the gold annotations. An incorrect "NULL" decision is trickier to analyze. It specifically means that among the previous mentions, there was a gold-coreferent mention, but instead NULL was chosen as the antecedent. This doesn't mean this mention will have a pair error with this should-have-been antecedent; they could later be connected through a completely different path if later mentions select them as antecedents. It's still important to be aware of these errors, though, since some of them represent recall errors for nominal mentions.]

4 Distributional similarity experiment

A big potential use case to learn and add semantic resources is for non-surface-matching matches between non-pronouns (names and common nouns). For example, "USS Cole" should be allowed to be an antecedent of "the ship." Finding possible matches in this manner — what HK call semantic compatibility filtering — has to rely on prior semantic information about the noun phrases.

Unfortunately, all the previous tasks took a long time to accomplish, and we had little time to build an in-depth lexical resources. However, we did try one experiment with distributional similarity from the category contexts dataset.

We looked at examples of two non-pronoun mentions being analyzed as a pair. The idea was to declare them semantically compatible if they had similar context vectors in web data. We harvested 625 cases where mentions were being compared, and the gold annotations knew they were coreferent, but our system didn't know to resolve them (because neither was a pronoun, and they didn't share head words). We also harvested 30,000 cases of pairs in the same situation, but were actually not coreferent in the gold. This can be viewed as a binary classification problem: is it possible to leverage semantic data to discriminate between the two cases?

If this seems too hard, remember that this experiment only deals with mentions that the syntactic filtering system didn't know how to deal with; this effectively imports some intelligence about the context.

In any case, this is a very difficult task. First, words can reasonably be coreferent in some situations but the same words will not be coreferent in others. In fact, many pairs of noun phrases harvested by the above procedure appear as both coreferent and not coreferent in the gold; we removed all positive pairs from the negative set, since their sizes were so skewed already.

Second, distributional similarity is a poor proxy for possibility of coreference in the case of entities that are different but of the the same or similar types. Distributional similarity knows that "US" and "China" are similar things — they share many contexts and behaviors — but these two words will almost never corefer.

We took the noun phrase pairs and ranked them by the simple cosine similarity of the noun phrases' context vectors. (We tried a few other similarity functions, including Jaccard, Spearman's rho, and cosine on log-counts, but got broadly similar results.) Among the 26,000 pairs for which we had context information for both NPs, there 523 coreferent pairs. Ranking pairs by cosine similarity gives an ROC AUC of 64.3%; and note that the ROC knows how to ignore class imbalances. Precision and recall, by

Table 4: Example high-similarity pairs; coreferent examples are over-represented

NP pair	Coref?	Notes
America, US	*	
fighters, men	*	*The freed men, all said to be fighters belonging...*
area, region	*	*flood waters throughout the region... the stricken area*
California, Texas		all impossible
China, Japan		
California, Florida		
Florida, Texas		
China, United States		
China, US		
Asia, China		
companies, executives		plausible but wrong
army, military		
friends, people		
people, politicians		

contrast, tend to be abysmally low; the best achievable F1 is less than 2%.

Table 4 shows some examples, drawn from the most similar NP pairs. Coreferent examples are in this table overrepresented; among the top 500 pairs, only 5% are coreferent. (This is higher percentage than the dataset's overall rate, which is why the AUC is better than even.)

We also evaluated in the full system by using a very strict cosine similarity threshold for nominal-nominal matching; if the NP pair's cosine similarity passed the threshold, it was counted as a potential antecedent match. This dramatically hurt precision, with some gains to recall. Table 1 reports the results for two thresholds.

It would be good to perform follow-up experiment that only used similarity comparisons for common nouns, or between a common noun and a name, since often, proper mentions with different surface forms actually refer to different entities, as is the case for "Florida" and "Texas" (though not "US" and "America").

Furthermore, it would be most helpful to use a knowledge resource that knows about mutual exclusion among entities; distributional similarity is the most extreme and low-level option, and it should not be too surprising it did not work.

5 Conclusion

We constructed a coreference system featuring rich syntactic features, most following Haghighi and Klein [2009] with a few interesting modifications. It performs near the state-of-the-art, is amenable to modification to support various text interpretation tasks, and it presents many possibilities for building its semantic compatibility filtering system. We are working a very near-term internal release of our current codebase for interested research groups at CMU.

References

Eric Bengtson and Dan Roth. Understanding the value of features for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Honolulu, Hawaii, October 2008. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D08-1031.

Michael Collins. Head-driven statistical models for natural language processing, 1999.

Aria Haghighi and Dan Klein. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1152–1161, Singapore, August 2009. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D/D09/D09-1120.