Simple Coreference Resolution with Rich Syntactic and Semantic Features

Aria Haghighi and Dan Klein
Computer Science Division
UC Berkeley
{aria42, klein}@cs.berkeley.edu

Abstract

Coreference systems are driven by syntactic, semantic, and discourse constraints. We present a simple approach which completely modularizes these three aspects. In contrast to much current work, which focuses on learning and on the discourse component, our system is deterministic and is driven entirely by syntactic and semantic compatibility as learned from a large, unlabeled corpus. Despite its simplicity and discourse naivete, our system substantially outperforms all unsupervised systems and most supervised ones. Primary contributions include (1) the presentation of a simple-to-reproduce, high-performing baseline and (2) the demonstration that most remaining errors can be attributed to syntactic and semantic factors external to the coreference phenomenon (and perhaps best addressed by non-coreference systems).

1 Introduction

The resolution of entity reference is influenced by a variety of constraints. Syntactic constraints like the binding theory, the i-within-i filter, and appositive constructions restrict reference by configuration. Semantic constraints like selectional compatibility (e.g. a *spokesperson* can *announce* things) and subsumption (e.g. *Microsoft* is a *company*) rule out many possible referents. Finally, discourse phenomena such as salience and centering theory are assumed to heavily influence reference preferences. As these varied factors have given rise to a multitude of weak features, recent work has focused on how best to learn to combine them using models over reference structures (Culotta et al., 2007; Denis and Baldridge, 2007; Klenner and Ailloud, 2007).

In this work, we break from the standard view. Instead, we consider a vastly more modular system in which coreference is predicted from a deterministic function of a few rich features. In particular, we assume a three-step process. First, a self-contained syntactic module carefully represents syntactic structures using an augmented parser and extracts syntactic paths from mentions to potential antecedents. Some of these paths can be ruled in or out by deterministic but conservative syntactic constraints. Importantly, the bulk of the work in the syntactic module is in making sure the parses are correctly constructed and used, and this module's most important training data is a treebank. Second, a self-contained semantic module evaluates the semantic compatibility of headwords and individual names. These decisions are made from compatibility lists extracted from unlabeled data sources such as newswire and web data. Finally, of the antecedents which remain after rich syntactic and semantic filtering, reference is chosen to minimize tree distance.

This procedure is trivial where most systems are rich, and so does not need any supervised coreference data. However, it is rich in important ways which we argue are marginalized in recent coreference work. Interestingly, error analysis from our final system shows that its failures are far more often due to syntactic failures (e.g. parsing mistakes) and semantic failures (e.g. missing knowledge) than failure to model discourse phenomena or appropriately weigh conflicting evidence.

One contribution of this paper is the exploration of strong modularity, including the result that our system beats all unsupervised systems and approaches the state of the art in supervised ones. Another contribution is the error analysis result that, even with substantial syntactic and semantic richness, the path to greatest improvement appears to be to further improve the syntactic and semantic modules. Finally, we offer our approach as a very strong, yet easy to implement, baseline. We make no claim that learning to reconcile disparate features in a joint model offers no benefit, only that it must not be pursued to the exclusion of rich, non-reference analysis.

2 Coreference Resolution

In coreference resolution, we are given a document which consists of a set of mentions; each

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1152–1161,
Singapore, 6-7 August 2009. ©2009 ACL and AFNLP

mention is a phrase in the document (typically an NP) and we are asked to cluster mentions according to the underlying referent entity. There are three basic mention types: proper (*Barack Obama*), nominal (*president*), and pronominal (*he*).[1] For comparison to previous work, we evaluate in the setting where mention boundaries are given at test time; however our system can easily annotate reference on all noun phrase nodes in a parse tree (see Section 3.1.1).

2.1 Data Sets

In this work we use the following data sets:

Development: (see Section 3)

- **ACE2004-ROTH-DEV**: Dev set split of the ACE 2004 training set utilized in Bengston and Roth (2008). The ACE data also annotates pre-nominal mentions which we map onto nominals. 68 documents and 4,536 mentions.

Testing: (see Section 4)

- **ACE2004-CULOTTA-TEST**: Test set split of the ACE 2004 training set utilized in Culotta et al. (2007) and Bengston and Roth (2008). Consists of 107 documents.[2]

- **ACE2004-NWIRE**: ACE 2004 Newswire set to compare against Poon and Domingos (2008). Consists of 128 documents and 11,413 mentions; intersects with the other ACE data sets.

- **MUC-6-TEST**: MUC6 formal evaluation set consisting of 30 documents and 2,068 mentions.

Unlabeled: (see Section 3.2)

- **BLIPP**: 1.8 million sentences of newswire parsed with the Charniak (2000) parser. No labeled coreference data; used for mining semantic information.

- **WIKI**: 25k articles of English Wikipedia abstracts parsed by the Klein and Manning (2003) parser.[3] No labeled coreference data; used for mining semantic information.

2.2 Evaluation

We will present evaluations on multiple coreference resolution metrics, as no single one is clearly superior:

- Pairwise F1: precision, recall, and F1 over all pairs of mentions in the same entity cluster. Note that this over-penalizes the merger or separation of clusters quadratically in the size of the cluster.

- b^3 (Amit and Baldwin, 1998): For each mention, form the intersection between the predicted cluster and the true cluster for that mention. The precision is the ratio of the intersection and the true cluster sizes and recall the ratio of the intersection to the predicted sizes; F1 is given by the harmonic mean over precision and recall from all mentions.

- MUC (Vilain et al., 1995): For each true cluster, compute the number of predicted clusters which need to be merged to cover the true cluster. Divide this quantity by true cluster size minus one. Recall is given by the same procedure with predicated and true clusters reversed.[4]

- CEAF (Luo, 2005): For a similarity function between predicted and true clusters, CEAF scores the best match between true and predicted clusters using this function. We use the ϕ_3 similarity function from Luo (2005).

3 System Description

In this section we develop our system and report developmental results on ACE2004-ROTH-DEV (see Section 2.1); we report pairwise F1 figures here, but report on many more evaluation metrics in Section 4. At a high level, our system resembles a pairwise coreference model (Soon et al., 1999; Ng and Cardie, 2002; Bengston and Roth, 2008); for each mention m_i, we select either a single-best antecedent amongst the previous mentions $m_1, \ldots, m_{i-1}$, or the NULL mention to indicate the underlying entity has not yet been evoked. Mentions are linearly ordered according to the position of the mention head with ties being broken by the larger node coming first.

[1] Other mention types exist and are annotated (such as prenominal), which are treated as nominals in this work.

[2] The evaluation set was not made available to non-participants.

[3] Wikipedia abstracts consist of roughly the first paragraph of the corresponding article

[4] The MUC measure is problematic when the system predicts many more clusters than actually exist (Luo, 2005; Finkel and Manning, 2008); also, singleton clusters do not contribute to evaluation.

While much research (Ng and Cardie, 2002; Culotta et al., 2007; Haghighi and Klein, 2007; Poon and Domingos, 2008; Finkel and Manning, 2008) has explored how to reconcile pairwise decisions to form coherent clusters, we simply take the transitive closure of our pairwise decision (as in Ng and Cardie (2002) and Bengston and Roth (2008)) which can and does cause system errors.

In contrast to most recent research, our pairwise decisions are not made with a learned model which outputs a probability or confidence, but instead for each mention m_i, we select an antecedent amongst $m_1, \ldots, m_{i-1}$ or the NULL mention as follows:

- **Syntactic Constraint:** Based on syntactic configurations, either force or disallow coreference between the mention and an antecedent. Propagate this constraint (see Figure 4).
- **Semantic/Syntactic Filter**: Filter the remaining possible antecedents based upon compatibility with the mention (see Figure 2).
- **Selection**: Select the 'closest' mention from the set of remaining possible antecedents (see Figure 1) or the NULL antecedent if empty.

Initially, there is no syntactic constraint (improved in Section 3.1.3), the antecedent compatibility filter allows proper and nominal mentions to corefer only with mentions that have the same head (improved in Section 3.2), and pronouns have no compatibility constraints (improved in Section 3.1.2). Mention heads are determined by parsing the given mention span with the Stanford parser (Klein and Manning, 2003) and using the Collins head rules (Collins, 1999); Poon and Domingos (2008) showed that using syntactic heads strongly outperformed a simple rightmost headword rule. The mention type is determined by the head POS tag: proper if the head tag is NNP or NNPS, pronoun if the head tag is PRP, PRP$, WP, or WP$, and nominal otherwise.

For the selection phase, we order mentions $m_1, \ldots, m_{i-1}$ according to the position of the head word and select the closest mention that remains after constraint and filtering are applied. This choice reflects the intuition of Grosz et al. (1995) that speakers only use pronominal mentions when there are not intervening compatible mentions. This system yields a rather low 48.9 pairwise F1 (see BASE-FLAT in Table 2). There are many, primarily recall, errors made choosing antecedents for all mention types which we will address by adding syntactic and semantic constraints.

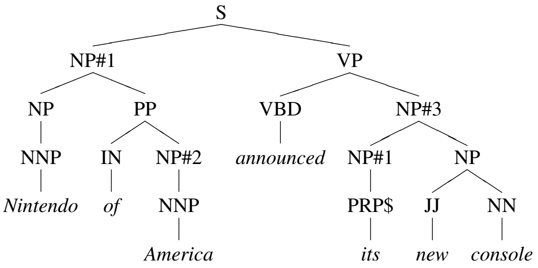


Figure 1: Example sentence where closest tree distance between mentions outperforms raw distance. For clarity, each mention NP is labeled with the underlying entity id.

3.1 Adding Syntactic Information

In this section, we enrich the syntactic representation and information in our system to improve results.

3.1.1 Syntactic Salience

We first focus on fixing the pronoun antecedent choices. A common error arose from the use of mention head distance as a poor proxy for discourse salience. For instance consider the example in Figure 1, the mention *America* is closest to *its* in flat mention distance, but syntactically *Nintendo of America* holds a more prominent syntactic position relative to the pronoun which, as Hobbs (1977) argues, is key to discourse salience.

Mapping Mentions to Parse Nodes: In order to use the syntactic position of mentions to determine anaphoricity, we must associate each mention in the document with a parse tree node. We parse all document sentences with the Stanford parser, and then for each evaluation mention, we find the largest-span NP which has the previously determined mention head as its head.[5] Often, this results in a different, typically larger, mention span than annotated in the data.

Now that each mention is situated in a parse tree, we utilize the length of the shortest tree path between mentions as our notion of distance. In

[5] If there is no NP headed by a given mention head, we add an NP over just that word.

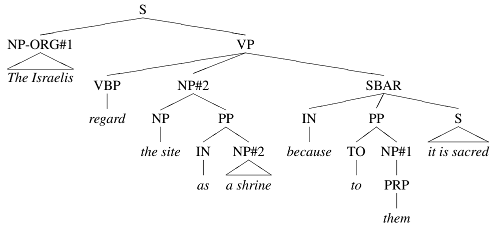


Figure 2: Example of a coreference decision fixed by agreement constraints (see Section 3.1.2). The pronoun *them* is closest to *the site* mention, but has an incompatible number feature with it. The closest (in tree distance, see Section 3.1.1) compatible mention is *The Israelis*, which is correct

particular, this fixes examples such as those in Figure 1 where the true antecedent has many embedded mentions between itself and the pronoun. This change by itself yields 51.7 pairwise F1 (see BASE-TREE in Table 2), which is small overall, but reduces pairwise pronoun antecedent selection error from 51.3% to 42.5%.

3.1.2 Agreement Constraints

We now refine our compatibility filtering to incorporate simple agreement constraints between coreferent mentions. Since we currently allow proper and nominal mentions to corefer only with matching head mentions, agreement is only a concern for pronouns. Traditional linguistic theory stipulates that coreferent mentions must agree in number, person, gender, and entity type (e.g. animacy). Here, we implement person, number and entity type agreement.[6]

A number feature is assigned to each mention deterministically based on the head and its POS tag. For entity type, we use NER labels. Ideally, we would like to have information about the entity type of each referential NP, however this information is not easily obtainable. Instead, we opt to utilize the Stanford NER tagger (Finkel et al., 2005) over the sentences in a document and annotate each NP with the NER label assigned to that mention head. For each mention, when its NP is assigned an NER label we allow it to only be compatible with that NER label.[7] For pronouns, we deterministically assign a set of compatible NER values (e.g. personal pronouns can only be a PERSON, but *its* can be an ORGANIZATION or LOCATION). Since the NER tagger typically does not label non-proper NP heads, we have no NER compatibility information for nominals.

gore	president	florida	state
bush	governor	lebanese	territory
nation	people	arafat	leader
inc.	company	aol	company
nation	country	assad	president

Table 1: Most common recall (missed-link) errors amongst non-pronoun mention heads on our development set. Detecting compatibility requires semantic knowledge which we obtain from a large corpus (see Section 3.2).

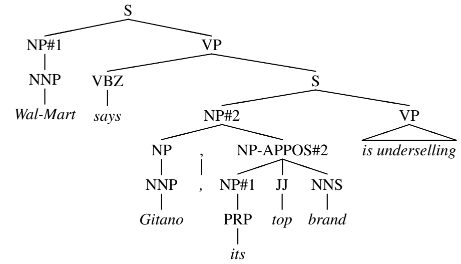


Figure 4: Example of interaction between the appositive and i-within-i constraint. The i-within-i constraint disallows coreference between parent and child NPs unless the child is an appositive. Hashed numbers indicate ground truth but are not in the actual trees.

We incorporate agreement constraints by filtering the set of possible antecedents to those which have compatible number and NER types with the target mention. This yields 53.4 pairwise F1, and reduces pronoun antecedent errors to 42.5% from 34.4%. An example of the type of error fixed by these agreement constraints is given by Figure 2.

3.1.3 Syntactic Configuration Constraints

Our system has so far focused only on improving pronoun anaphora resolution. However, a plurality of the errors made by our system are amongst non-pronominal mentions.[8] We take the approach that in order to align a non-pronominal mention to an antecedent without an identical head, we require evidence that the mentions are compatible.

Judging compatibility of mentions generally requires semantic knowledge, to which we return later. However, some syntactic configurations

[6] Gender agreement, while important for general coreference resolution, did not contribute to the errors in our largely newswire data sets.

[7] Or allow it to be compatible with all NER labels if the NER tagger doesn't predict a label.

[8] There are over twice as many nominal mentions in our development data as pronouns.

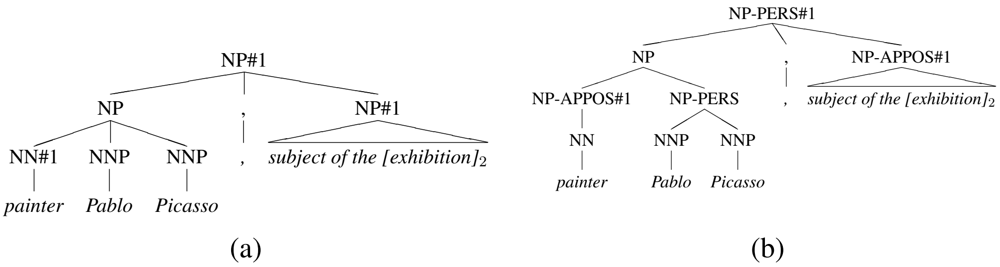


Figure 3: NP structure annotation: In (a) we have the raw parse from the Klein and Manning (2003) parser with the mentions annotated by entity. In (b), we demonstrate the annotation we have added. NER labels are added to all NP according to the NER label given to the head (see Section 3.1.1). Appositive NPs are also annotated. Hashes indicate forced coreferent nodes

guarantee coreference. The one exploited most in coreference work (Soon et al., 1999; Ng and Cardie, 2002; Luo et al., 2004; Culotta et al., 2007; Poon and Domingos, 2008; Bengston and Roth, 2008) is the appositive construction. Here, we represent apposition as a syntactic feature of an NP indicating that it is coreferent with its parent NP (e.g. it is an exception to the i-within-i constraint that parent and child NPs *cannot* be coreferent). We deterministically mark a node as NP-APPOS (see Figure 3) when it is the third child in of a parent NP whose expansion begins with (NP , NP), and there is not a conjunction in the expansion (to avoid marking elements in a list as appositive).

Role Appositives: During development, we discovered many errors which involved a variant of appositives which we call 'role appositives' (see *painter* in Figure 3), where an NP modifying the head NP describes the role of that entity (typically a person entity). There are several challenges to correctly labeling these role NPs as being appositives. First, the NPs produced by Treebank parsers are flat and do not have the required internal structure (see Figure 3(a)). While fully solving this problem is difficult, we can heuristically fix many instances of the problem by placing an NP around maximum length sequences of NNP tags or NN (and JJ) tags within an NP; note that this will fail for many constructions such as *U.S. President Barack Obama*, which is analyzed as a flat sequence of proper nouns. Once this internal NP structure has been added, whether the NP immediately to the left of the head NP is an appositive depends on the entity type. For instance, *Rabbi Ashi* is an apposition but *Iranian army* is not. Again, a full solution would require its own model, here we mark as appositions any NPs immediately to the left of a head child NP where the head child NP is identified as a person by the NER tagger.[9]

We incorporate NP appositive annotation as a constraint during filtering. Any mention which corresponds to an appositive node has its set of possible antecedents limited to its parent. Along with the appositive constraint, we implement the i-within-i constraint that any non-appositive NP cannot be be coreferent with its parent; this constraint is then propagated to any node its parent is forced to agree with. The order in which these constraints are applied is important, as illustrated by the example in Figure 4: First the list of possible antecedents for the appositive NP is constrained to only its parent. Now that all appositives have been constrained, we apply the i-within-i constraint, which prevents *its* from having the NP headed by *brand* in the set of possible antecedents, and by propagation, also removes the NP headed by *Gitano*. This leaves the NP *Wal-Mart* as the closest compatible mention.

Adding these syntactic constraints to our system yields 55.4 F1, a fairly substantial improvement, but many recall errors remain between mentions with differing heads. Resolving such cases will require external semantic information, which we will automatically acquire (see Section 3.2).

Predicate Nominatives: Another syntactic constraint exploited in Poon and Domingos (2008) is the predicate nominative construction, where the object of a copular verb (forms of the verb *be*) is constrained to corefer with its subject (e.g. *Microsoft is a company in Redmond*). While much less frequent than appositive configurations (there are only 17 predicate nominatives in our devel-

[9] Arguably, we could also consider right modifying NPs (e.g., *[Microsoft [Company]$_1$]$_1$*) to be role appositive, but we do not do so here.

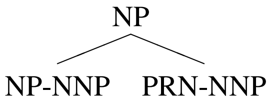
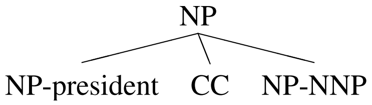
Path	Example
	*America Online Inc. (AOL)*
	*[President and C.E.O] Bill Gates*

Figure 5: Example paths extracted via semantic compatibility mining (see Section 3.2) along with example instantiations. In both examples the left child NP is coreferent with the rightmost NP. Each category in the interior of the tree path is annotated with the head word as well as its subcategorization. The examples given here collapse multiple instances of extracted paths.

opment set), predicate nominatives are another highly reliable coreference pattern which we will leverage in Section 3.2 to mine semantic knowledge. As with appositives, we annotate object predicate-nominative NPs and constrain coreference as before. This yields a minor improvement to 55.5 F1.

3.2 Semantic Knowledge

While appositives and related syntactic constructions can resolve some cases of non-pronominal reference, most cases require semantic knowledge about the various entities as well as the verbs used in conjunction with those entities to disambiguate references (Kehler et al., 2008).

However, given a semantically compatible mention head pair, say *AOL* and *company*, one might expect to observe a reliable appositive or predicative-nominative construction involving these mentions somewhere in a large corpus. In fact, the Wikipedia page for *AOL*[10] has a predicate-nominative construction which supports the compatibility of this head pair: *AOL LLC (formerly America Online) is an American global Internet services and media company operated by Time Warner.*

In order to harvest compatible head pairs, we utilize our BLIPP and WIKI data sets (see Section 2), and for each noun (proper or common) and pronoun, we assign a maximal NP mention node for each nominal head as in Section 3.1.1; we then annotate appositive and predicate-nominative NPs as in Section 3.1.3. For any NP which is annotated as an appositive or predicate-nominative, we extract the head pair of that node and its constrained antecedent.

The resulting set of compatible head words, while large, covers a little more than half of the examples given in Table 1. The problem is that these highly-reliable syntactic configurations are too sparse and cannot capture all the entity information present. For instance, the first sentence of Wikipedia abstract for *Al Gore* is:

> *Albert Arnold "Al" Gore, Jr. is an American environmental activist who served as the 45th Vice President of the United States from 1993 to 2001 under President Bill Clinton.*

The required lexical pattern *X who served as Y* is a general appositive-like pattern that almost surely indicates coreference. Rather than opt to manually create a set of these coreference patterns as in Hearst (1992), we instead opt to automatically extract these patterns from large corpora as in Snow et al. (2004) and Phillips and Riloff (2007). We take a simple bootstrapping technique: given a set of mention pairs extracted from appositives and predicate-nominative configurations, we extract counts over tree fragments between nodes which have occurred in this set of head pairs (see Figure 5); the tree fragments are formed by annotating the internal nodes in the tree path with the head word and POS along with the subcategorization. We limit the paths extracted in this way in several ways: paths are only allowed to go between adjacent sentences and have a length of at most 10. We then filter the set of paths to those which occur more than a hundred times and with at least 10 distinct seed head word pairs.

The vast majority of the extracted fragments are variants of traditional appositives and predicate-nominatives with some of the structure of the NPs

[10] http://en.wikipedia.org/wiki/AOL

	MUC			b^3			Pairwise			CEAF		
System	P	R	F1	P	R	F1	P	R	F1	P	R	F1
ACE2004-ROTH-DEV												
BASIC-FLAT	73.5	66.8	70.0	80.6	68.6	74.1	63.6	39.7	48.9	68.4	68.4	68.4
BASIC-TREE	75.8	68.9	72.2	81.9	69.9	75.4	65.6	42.7	51.7	69.8	69.8	69.8
+SYN-COMPAT	77.8	68.5	72.9	84.1	69.7	76.2	71.0	43.1	53.4	69.8	69.8	69.8
+SYN-CONSTR	78.3	70.5	74.2	84.0	71.0	76.9	71.3	45.4	55.5	70.8	70.8	70.8
+SEM-COMPAT	77.9	74.1	75.9	81.8	74.3	77.9	68.2	51.2	58.5	72.5	72.5	72.5
ACE2004-CULOTTA-TEST												
BASIC-FLAT	68.6	60.9	64.5	80.3	68.0	73.6	57.1	30.5	39.8	66.5	66.5	66.5
BASIC-TREE	71.2	63.2	67.0	81.6	69.3	75.0	60.1	34.5	43.9	67.9	67.9	67.9
+SYN-COMPAT	74.6	65.2	69.6	84.2	70.3	76.6	66.7	37.2	47.8	69.2	69.2	69.2
+SYN-CONSTR	74.3	66.4	70.2	83.6	71.0	76.8	66.4	38.0	48.3	69.6	69.6	69.6
+SEM-COMPAT	74.8	**77.7**	**79.6**	79.6	**78.5**	79.0	**57.5**	57.6	57.5	**73.3**	**73.3**	**73.3**
Supervised Results												
Culotta et al. (2007)	-	-	-	86.7	73.2	79.3	-	-	-	-	-	-
Bengston and Roth (2008)	**82.7**	69.9	75.8	**88.3**	74.5	**80.8**	55.4	**63.7**	**59.2**	-	-	-
MUC6-TEST												
+SEM-COMPAT	**87.2**	**77.3**	**81.9**	84.7	**67.3**	**75.0**	**80.5**	**57.8**	**67.3**	72.0	72.0	72.0
Unsupervised Results												
Poon and Domingos (2008)	83.0	75.8	79.2	-	-	-	63.0	57.0	60.0	-	-	-
Supervised Results												
Finkel and Manning (2008)	89.7	55.1	68.3	**90.9**	49.7	64.3	74.1	37.1	49.5	-	-	-
ACE2004-NWIRE												
+SEM-COMPAT	**77.0**	**75.9**	**76.5**	79.4	74.5	76.9	**66.9**	**49.2**	**56.7**	71.5	71.5	71.5
Unsupervised Results												
Poon and Domingos (2008)	71.3	70.5	70.9	-	-	-	62.6	38.9	48.0	-	-	-

Table 2: Experimental Results (See Section 4): When comparisons between systems are presented, the largest result is bolded. The CEAF measure has equal values for precision, recall, and F1.

specified. However there are some tree fragments which correspond to the novel coreference patterns (see Figure 5) of parenthetical alias as well as conjunctions of roles in NPs.

We apply our extracted tree fragments to our BLIPP and WIKI data sets and extract a set of compatible word pairs which match these fragments; these words pairs will be used to relax the semantic compatibility filter (see the start of the section); mentions are compatible with prior mentions with the same head or with a semantically compatible head word. This yields 58.5 pairwise F1 (see SEM-COMPAT in Table 2) as well as similar improvements across other metrics.

By and large the word pairs extracted in this way are correct (in particular we now have coverage for over two-thirds of the head pair recall errors from Table 1.) There are however word-pairs which introduce errors. In particular city-state constructions (e.g. *Los Angeles, California*) appears to be an appositive and incorrectly allows our system to have *angeles* as an antecedent for *california*. Another common error is that the % symbol is made compatible with a wide variety of common nouns in the financial domain.

4 Experimental Results

We present formal experimental results here (see Table 2). We first evaluate our model on the ACE2004-CULOTTA-TEST dataset used in the state-of-the-art systems from Culotta et al. (2007) and Bengston and Roth (2008). Both of these systems were supervised systems discriminatively trained to maximize b^3 and used features from many different structured resources including WordNet, as well as domain-specific features (Culotta et al., 2007). Our best b^3 result of 79.0 is broadly in the range of these results. We should note that in our work we use neither the gold mention types (we do not model pre-nominals separately) nor do we use the gold NER tags which Bengston and Roth (2008) does. Across metrics, the syntactic constraints and semantic compatibility components contribute most to the overall final result.

On the MUC6-TEST dataset, our system outper-

	PROPER	NOMINAL	PRONOUN	NULL	TOTAL
PROPER	21/451	8/20	-	72/288	101/759
NOMINAL	16/150	99/432	-	158/351	323/933
PRONOUN	29/149	60/128	15/97	1/2	105/376

Table 3: Errors for each type of antecedent decision made by the system. Each row is a mention type and the column the predicted mention type antecedent. The majority of errors are made in the NOMINAL category.

forms both Poon and Domingos (2008) (an unsupervised Markov Logic Network system which uses explicit constraints) and Finkel and Manning (2008) (a supervised system which uses ILP inference to reconcile the predictions of a pairwise classifier) on all comparable measures.[11] Similarly, on the ACE2004-NWIRE dataset, we also outperform the state-of-the-art unsupervised system of Poon and Domingos (2008).

Overall, we conclude that our system outperforms state-of-the-art unsupervised systems[12] and is in the range of the state-of-the art systems of Culotta et al. (2007) and Bengston and Roth (2008).

5 Error Analysis

There are several general trends to the errors made by our system. Table 3 shows the number of pairwise errors made on MUC6-TEST dataset by mention type; note these errors are not equally weighted in the final evaluations because of the transitive closure taken at the end. The most errors are made on nominal mentions with pronouns coming in a distant second. In particular, we most frequently say a nominal is NULL when it has an antecedent; this is typically due to not having the necessary semantic knowledge to link a nominal to a prior expression.

In order to get a more thorough view of the cause of pairwise errors, we examined 20 random errors made in aligning each mention type to an antecedent. We categorized the errors as follows:

- Sem. Compat: Missing information about the compatibility of two words e.g. *pay* and *wage*. For pronouns, this is used to mean that we incorrectly aligned a pronoun to a mention with which it is not semantically compatible (e.g. *he* aligned to *board*).
- Syn. Compat: Error in assigning linguistic features of nouns for compatibility with pronouns (e.g. disallowing *they* to refer to *team*).
- Head: Errors involving the assumption that mentions with the same head are always compatible. Includes modifier and specificity errors such as allowing *Lebanon* and *Southern Lebanon* to corefer. This also includes errors of definiteness in nominals (e.g. *the people in the room* and *Chinese people*). Typically, these errors involve a combination of missing syntactic and semantic information.
- Internal NP: Errors involving lack of internal NP structure to mark role appositives (see Section 3.1.3).
- Prag. / Disc.: Errors where discourse salience or pragmatics are needed to disambiguate mention antecedents.
- Process Error: Errors which involved a tokenization, parse, or NER error.

The result of this error analysis is given in Table 4; note that a single error may be attributed to more than one cause. Despite our efforts in Section 3 to add syntactic and semantic information to our system, the largest source of error is still a combination of missing semantic information or annotated syntactic structure rather than the lack of discourse or salience modeling.

Our error analysis suggests that in order to improve the state-of-the-art in coreference resolution, future research should consider richer syntactic and semantic information than typically used in current systems.

6 Conclusion

Our approach is not intended as an argument against the more complex, discourse-focused approaches that typify recent work. Instead, we note that rich syntactic and semantic processing vastly reduces the need to rely on discourse effects or evidence reconciliation for reference resolution. Indeed, we suspect that further improving the syntactic and semantic modules in our system may produce greater error reductions than any other

[11] Klenner and Ailloud (2007) took essentially the same approach but did so on non-comparable data.

[12] Poon and Domingos (2008) outperformed Haghighi and Klein (2007). Unfortunately, we cannot compare against Ng (2008) since we do not have access to the version of the ACE data used in their evaluation.

Mention Type	SEM. COMPAT	SYN. COMPAT	HEAD	INTENAL NP	PRAG / DISC.	PROCESS ERROR	OTHER	Comment
NOMINAL	7	-	5	6	2	2	1	2 general appos. patterns
PRONOUN	6	3	-	6	3	3	3	2 cataphora
PROPER	6	-	3	4	4	4	1	

Table 4: Error analysis on ACE2004-CULOTTA-TEST data by mention type. The dominant errors are in either semantic or syntactic compatibility of mentions rather than discourse phenomena. See Section 5.

route forward. Of course, a system which is rich in all axes will find some advantage over any simplified approach.

Nonetheless, our coreference system, despite being relatively simple and having no tunable parameters or complexity beyond the non-reference complexity of its component modules, manages to outperform state-of-the-art unsupervised coreference resolution and be broadly comparable to state-of-the-art supervised systems.

References

B. Amit and B. Baldwin. 1998. Algorithms for scoring coreference chains. In *MUC7*.

Eric Bengston and Dan Roth. 2008. Understanding the value of features for corefernce resolution. In *Empirical Methods in Natural Language Processing*.

E. Charniak. 2000. Maximum entropy inspired parser. In *North American Chapter of the Association of Computational Linguistics (NAACL)*.

Mike Collins. 1999. Head-driven statistical models for natural language parsing.

A Culotta, M Wick, R Hall, and A McCallum. 2007. First-order probabilistic models for coreference resolution. In *NAACL-HLT*.

Pascal Denis and Jason Baldridge. 2007. Global, Joint Determination of Anaphoricity and Coreference Resolution using Integer Programming. In *HLT-NAACL*.

Jenny Finkel and Christopher Manning. 2008. Enforcing transitivity in coreference resolution. In *Association of Computational Linguists (ACL)*.

Jenny Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse.

Aria Haghighi and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics.

Marti Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Conference on Natural Language Learning (COLING)*.

J. R. Hobbs. 1977. Resolving pronoun references. *Lingua*.

Andrew Kehler, Laura Kertz, Hannah Rohde, and Jeffrey Elman. 2008. Coherence and coreference revisited.

D. Klein and C. Manning. 2003. Accurate unlexicalized parsing. In *Association of Computational Linguists (ACL)*.

Manfred Klenner and Etienne Ailloud. 2007. Optimization in coreference resolution is not needed: A nearly-optimal algorithm with intensional constraints. In *Recent Advances in Natural Language Processing*.

Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. 2004. A Mention-Synchronous Coreference Resolution Algorithm Based on the Bell Tree. In *Association of Computational Linguists*.

X Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*.

Vincent Ng and Claire Cardie. 2002. Improving Machine Learning Approaches to Coreference Resolution. In *Association of Computational Linguists*.

Vincent Ng. 2008. Unsupervised models of coreference resolution. In *EMNLP*.

W. Phillips and E. Riloff. 2007. Exploiting role-identifying nouns and expressions for information extraction. In *Recent Advances in Natural Language Processing (RANLP)*.

Hoifung Poon and Pedro Domingos. 2008. Joint unsupervised coreference resolution with Markov Logic. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*.

R. Snow, D. Jurafsky, and A. Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. In *Neural Information Processing Systems (NIPS)*.

W.H. Soon, H. T. Ng, and D. C. Y. Lim. 1999. A machine learning approach to coreference resolution of noun phrases.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *MUC-6*.