The More Antecedents, the Merrier: Resolving Multi-Antecedent Anaphors

Hardik Vala[1], **Andrew Piper**[2], **Derek Ruths**[1]
[1]School of Computer Science, [2]Dept. of Languages, Literartures, & Cultures
McGill University
Montreal, Canada
hardik.vala@mail.mcgill.ca,
{andrew.piper, derek.ruths}@mcgill.ca

Abstract

Anaphor resolution is an important task in NLP with many applications. Despite much research effort, it remains an open problem. The difficulty of the problem varies substantially across different sub-problems. One sub-problem, in particular, has been largely untouched by prior work despite occurring frequently throughout corpora: the anaphor that has multiple antecedents, which here we call *multi-antecedent anaphors* or m-anaphors. Current coreference resolvers restrict anaphors to at most a single antecedent. As we show in this paper, relaxing this constraint poses serious problems in coreference chain-building, where each chain is intended to refer to a single entity. This work provides a formalization of the new task with preliminary insights into multi-antecedent noun-phrase anaphors, and offers a method for resolving such cases that outperforms a number of baseline methods by a significant margin. Our system uses local agglomerative clustering on candidate antecedents and an existing coreference system to score clusters to determine which cluster of mentions is antecedent for a given anaphor. When we augment an existing coreference system with our proposed method, we observe a substantial increase in performance (0.6 absolute CoNLL F1) on an annotated corpus.

1 Introduction

Anaphor resolution is a very difficult task in Natural Language Understanding, involving the complex interaction of discourse cues, syntactic rules, and semantic phenomena. It is closely related to the task of coreference resolution (Van Deemter and Kibble, 2000), for which a myriad of solutions have been proposed (Clark and Manning, 2015; Peng et al., 2015; Wiseman et al., 2015; Björkelund and Farkas, 2012; Lee et al., 2011; Stoyanov et al., 2010; Ng, 2008; Bergsma and Lin, 2006; Soon et al., 2001). However, given the complexity of the problem, a comprehensive approach remains elusive. The difficulty varies drastically across different cases (proper nouns, pronouns, gerunds, etc.), each of which involves different assumptions about and models of various linguistic phenomena (e.g., vocabulary, syntax, and semantics). As a result, state-of-the-art systems yield varying performance across sub-problems (Mitkov, 2014; Kummerfeld and Klein, 2013; Björkelund and Nugues, 2011; Recasens and Hovy, 2009; Stoyanov et al., 2009; Bengtson and Roth, 2008; Van Deemter and Kibble, 2000; Ng and Cardie, 2002b; Kameyama, 1997).

To avoid the complexity of the overarching resolution task, many current systems — whether learning-based (Clark and Manning, 2015; Peng et al., 2015; Wiseman et al., 2015; Durrett and Klein, 2013; Björkelund and Farkas, 2012) or rule-based (Lee et al., 2011) — focus on a restricted version of the problem, where candidate anaphors are linked to at most one antecedent, from which coreference chains are built by propagating the induced equivalence relation, with each chain corresponding to an entity (Van Deemter and Kibble, 2000).

While this single-antecedent inference task does resolve a very large number of anaphors in any given text, it leaves one quite common sub-problem virtually untouched: anaphors that link to multiple antecedents. These have sometimes been called split-antecedent anaphors; here we use the term *multi-antecedent anaphors* or m-anaphors in

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 2287–2296,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

order to emphasize the existence of more than one (possibly more than two) antecedents for a given anaphor. Consider the following examples:

(1) *[Elizabeth]$_1$ met [Mary]$_2$ at the park and [they]$_{1,2}$ began their stroll to the river.*

(2) *Mrs. Dashwood, having moved to another country, saw her [mother]$_1$ and [sister-in-law]$_2$ demoted to occasional visitors. As such, however, her old [kin]$_{1,2}$ were treated by her new family with quiet civility.*

Such cases present a challenge to state-of-the-art methods: certain features well-suited for the single-antecedent case do not apply (e.g. gender and pluarity) (Recasens and Hovy, 2009; Stoyanov et al., 2009; Bergsma and Lin, 2006), and strong long-distance effects cannot be ignored (Ingria and Stallard, 1989). Moreover, the presence of multiple antecedents for a single anaphor violates the separation between coreference chains.

In this paper, we address the multi-antecedent case of noun-phrase (NP) anaphor resolution in English, the most widely understood and studied form of coreference resolution (Ng, 2010; Ng, 2008). While we frame the general question of multi-antecedent inference, we restrict our analyses to one particular sub-problem: resolving the antecedents of the pronouns *they* and *them*. These pronouns best isolate the characteristics of m-anaphors (see Section 2 for more on the motivation of this choice). We propose a system for resolving *they* and *them* that models grouping compatibility of mentions through a maximum entropy pairwise model, independently from coreference of groupings, which is handled through an existing coreference resolution system leveraging corpus knowledge.

This paper makes four core contributions. First, it provides a generalization of the anaphor resolution problem to permit linking to multiple antecedents. Second, we characterize core properties of m-anaphors and their linguistic environments in a large, annotated corpus. Third, we provide a entity-centric system for specifically resolving multi-antecedent cases that outperforms a number of baselines. And, finally, we show how to pair our system with an existing coreference system and show a gain of 0.6 points (CoNLL F1) on the complete coreference resolution task (resolving all anaphors, single- and multi-antecedent).

The rest of the paper is organized as follows: We introduce the terminology and problem statement for split-antecedent resolution in Section 2. A summary of the data is given in Section 3 and the behaviour of split-antecedent anaphors is analyzed in Section 4. Our approach to antecedent prediction is presented in Section 5 and the results and analysis are reported in Section 6. Finally, we review related work in Section 7 and conclude and discuss future work in Section 8.

2 Problem

This section establishes the terminology used throughout the paper and reformulates the anaphor resolution problem to incorporate linking to multiple antecedents.

2.1 Terminology

We introduce the term m-anaphor for convenience as a special case of anaphor that has to multiple antecedents. For example, *they* and *kin* in Examples (1) and (2), respectively, from the Introduction are m-anaphors. By extension, 1-anaphors are anaphors that have only one antecedent.

Similarly, we define an m-antecedent as one of multiple antecedents of an m-anaphor and we refer to m-antecedents with the same m-anaphor as *siblings*. In Example (1) from the Introduction, *Elizabeth* and *Mary* are sibling m-antecedents of *they*, and in Example (2), *mother* and *sister-in-law* are sibling m-antecedents of *kin*.

Finally, we refer to anaphors with two, three, and four m-antecedents as *2-anaphors*, *3-anaphors*, and *4-anaphors*, respectively. We provide two more examples:

(3) *[Mr. Holmes]$_1$ stared off into the distance. [Watson]$_2$ simply walked off. [Both]$_{1,2}$ were troubled by the news.*

(4) *Virginia found herself alone with her [brother]$_1$, and then the thought of her [sister]$_2$ came to mind. [She]$_3$ remembered the camping trip [they]$_{1,2,3}$ embarked on a few summers ago.*

The anaphor in Example (3) is a 2-anaphor and the anaphor in Example (4) is a 3-anaphor.

2.2 Definition

We define the NP anaphor resolution problem similar to Wiseman et al. (2015), Durrett and Klein

Pronoun	# m-anaphors
they	278
them	165
we	140
you	43
everybody	12

Table 1: Counts of the most frequent m-anaphoric pronouns in P&P.

	they		*them*		Total	
	#	%	#	%	#	%
P&P	278	32.10	165	19.05	443	51.15
Scribner	243	12.96	79	4.21	322	17.17
Total	521	19.01	244	8.90	765	27.91

Table 2: Number of m-anaphoric *they* and *them* mentions and % of all *they* and *them* mentions that are m-anaphors.

(2013), and Hirschman (1997): Let $\mathcal{M}$ denote the set of all identified mentions in a document and let $M(x) \subseteq \mathcal{M}$ denote all mentions preceding a mention $x \in \mathcal{M}$. The objective of the task is, for each $x \in \mathcal{M}$, to find $C \subseteq M(x)$ such that all mentions in C are antecedent to x. If $C = \emptyset$, then x is non-anaphoric and if $|C| \geq 1$, then x is 1-anaphoric, and if $|C| > 1$, then x is m-anaphoric. Hence, this formulation generalizes the problem to account for multi-antecedent anaphors.

To constrain the scope of the study, we perform all our analyses on gold mentions, leaving the effect of imperfect mention detection as a problem for future work (this has been studied for the single-antecedent case in Stoyanov et al. (2009)). Moreover, we only consider mentions of *they* and *them* that are known to be m-anaphoric for three reasons. First, non-pronomial m-anaphors, i.e. proper and common nouns, are much more susceptible to long-distance effects and may require external knowledge to resolve. Second, by focusing on this case, we circumvent a host of very involved aspects of the complete m-anaphor resolution problem, i.e. determining whether a mention is m-anaphoric, 1-anaphoric, or not anaphoric at all. For example, *you* may refer to one person or multiple, *who* can be used as an interrogative (non-anaphoric) or reflexive pronoun (anaphoric)), pronouns such as *anyone* and *everyone* introduce many scoping difficulties, and pleonastic pronouns must be removed from the inference task entirely. Third, *they* and *them* are the most prevalent of all pronouns in our dataset (refer to Table 1).

3 Data

Our dataset comprises of the *Pride and Prejudice* novel (P&P) (121440 words) and 36 short stories from the *Scribner Anthology of Contemporary Short Fiction* (Martone et al., 1999) (Scribner) (total of 216901 words), representing an eclectic collection of stories from the modern era. For P&P, all mentions of character have been fully resolved to their antecedents, including mentions referencing multiple characters. For Scribner, all mentions of *they* and *them* are resolved (m-anaphoric, 1-anaphoric, and singleton), including those of non-person entities.

These stories were annotated by three annotators according to a slightly modified version of the ACE coreference resolution task formulation (Doddington et al., 2004) to allow multiple antecedents. Annotations were conducted through the *brat*[1] annotation tool (Stenetorp et al., 2012)) and the inter-annotator agreement on the shared texts (3 stories from Scribner + 7 chapters from P&P) was 86.5%.

Overall, in P&P, 1289 m-anaphors were discovered, of which 34 (2.6%) were proper nouns, 536 (41.6%) were common nouns, and 719 (55.8%) were pronouns. Table 2 shows the number of gold m-anaphoric *they* and *them* mentions and the percentage of all *they* and *them* mentions that are m-anaphoric.

Literary works were chosen over other textual modalities, e.g. news articles, because they showed a higher density of m-anaphors (a preliminary annotation exercise showed that literary works contained 37% more m-anaphors per word).

The dataset is partitioned according to a roughly, 60/20/20 split into training, validation, and testing sets, where the split is applied to the text of P&P (e.g. the first 60% of story text is used for training), and the collection of Scribner stories (e.g. 60% of the stories were used for training).

4 Behaviour of m-anaphors

m-anaphors present a novel class of anaphor for which very little knowledge exists. To better understand the linguistic behaviour of m-anaphors, we perform the following analyses. First, we examine first and second order statistics of our

[1] `http://brat.nlplab.org`

	First	Second
Avg. distance (# words)	17.08	33.50
Std. distance (# words)	23.80	40.66
Avg. distance (# sent.)	1.19	2.28
Std. distance (# sent.)	3.18	5.10
Avg. # intermediates	1.44	4.21
Std. # intermediates	2.33	4.44

Table 3: Average and standard deviations of the word distance, sentence distance, and number of intermediate mentions between the first and second most recent mentions to an m-anaphor.

dataset to gain insight into the distribution of m-anaphors across a number of dimensions. Second, we fit a maximum entropy model over common coreference features for distinguishing m-anaphoric and anaphoric mentions to evaluate the importance of various features in determining m-anaphoricity versus anaphoricity of mentions.

4.1 m-anaphor Statistics

The distribution of m-anaphors according to the number of referenced $m-$antecedents is as follows: 79.3% are 2-anaphors, 13.2% are 3-anaphors, 3.7% are 4-anaphors, and the remaining 3.8% refer to larger numbers of antecedents. Despite the bias towards 2-anaphors, the simple approach to m-anaphor resolution of taking the previous two mentions as m-antecedent siblings will fail according to Table 3. The usual presence of intermediate mentions between m-anaphors and their m-antecedents makes the resolution task non-trivial. Moreover, the large distances between m-anaphors and their antecedents attenuates any signal for coreference, introducing greater noise to the problem.

4.2 m-anaphoricity Features

The statistics discussed above shed light on the complexity of this problem. Here, we examine whether certain surface-level features of anaphoric phenomena from prior work exhibit any differences for m-anaphoric mentions over anaphoric ones. We construct a maximum entropy model from the training data over the combination of syntactic and semantic features in Table 4, inspired by Wiseman et al. (2015), Durrett and Klein (2013), and Recasens et al. (2013b). The binary classification decision is between m-anaphoric and 1-anaphoric mentions, coded as '1' and '0', respectively. Therefore, the estimated coefficients that are positive favor m-anaphoricity and those that are negative favor 1-anaphoricity.

Feature	Coefficient	p-value
Sentence position = first	0.16	0.13
Sentence position = last	-0.18	0.006
Dependency = subject	0.27	0.05
Dependency = object	0.08	0.24
Dependency = preposition	-0.22	0.07
Coordinated = true	0.29	0.08
Presence of negation	0.06	0.31
Presence of modality	0.04	0.21

Table 4: Features for m-anaphoricity versus 1-anaphoricity with coefficients estimated from a maximum entropy model, and associated p-values.

Except for the feature testing on the last sentence position, none of the results in Table 4 were able to reach statistical significance, suggesting at a surface level, m-anaphoricity and 1-anaphoricity behave very similarly and operate in similar linguistic environments. One possibility is that a deeper set of features is required for distinguishing m-anaphors from 1-anaphors. We identify this as an important topic for future work in this area.

5 m-anaphor Resolution

Our approach to m-anaphor resolution draws inspiration from mention pair models for coreference that make independent binary classification decisions (Ng, 2010). In our method, we employ a maximum entropy model that makes binary decisions on mention pairs as well, but the decision corresponds to "group compatibility" of mentions, i.e. to what degree can a given set of mentions be the sibling m-antecedents to the same m-anaphor. This model is embedded in an agglomerative clustering process, after which a coreference decision is made between clusters and the given m-anaphor. Thus, our model treats the grouping of candidate mentions into sibling sets independently from antecedent-anaphor linking.

5.1 Architecture

Given an m-anaphor g in document $\mathcal{D}$, the steps of our approach are as follows:

1. Mentions preceding g within a k-sentence window are extracted as candidate m-antecedents to g.

2. Perform an agglomerative clustering of the candidate mentions using similarity metric

SIM_1 and average-linkage criteria. Let $\mathcal{C}$ represent the clustering.

3. Each non-singleton cluster $C \in \mathcal{C}$ is scored according to the probability of coreference of the m-anaphor to the cluster. This is done by appealing to an external corpus comprising of sentences containing either *they* or *them*. The grouping of sentences in the document containing all of the mentions in C (and sentences in-between) are compared to each *they* or *them* sentence in the external corpus (depending on the identity of g) using similarity metric SIM_2. The sentence yielding the maximum similarity is selected. The probability of coreference is then calculated by replacing the sentence grouping with the extracted sentence and applying an existing coreference system $COREF$ between g and its counterpart (*they* or *them*) in the extracted sentence.

4. The cluster C_{max} producing the highest probability of coreference is predicted as the group of m-antecedents for g.

Again, inspired by mention-pair models for coreference resolution (Clark and Manning, 2015; Björkelund and Farkas, 2012; Ng and Cardie, 2002a), the SIM_1 similarity metric is defined as $\sigma(\mathbf{w}^\top \mathbf{x})$, where $\mathbf{w}$ is a weight vector and $\mathbf{x}$ is a feature vector defined for a pair of mentions. The parameter vector $\mathbf{w}$ is learned using the standard cross-entropy loss function in a maximum entropy model, where the target variable is a decision on whether the mentions pairs are siblings or not. The learning is conducted over the training set with L2-regularization.

For SIM_2, which is responsible for selecting replacement sentences, we experiment with two different similarity metrics: (1) longest common subsequence normalized by sentence length (LCS) and (2) a subset tree kernel (Collins and Duffy, 2002) with a bag-of-words extension as described in Moschitti (2006), which also describes a simple adaptation to forests (for multiple sentences). The named entity (NE) mentions in sentences are replaced by corresponding NE type placeholders (PERSON, LOCATION, etc. as described in Finkel et al. (2005)) before comparison.

In the experiments to follow, we adopt the classification mention-pair model, a component of the statistical coreference resolution system available in the Stanford CoreNLP suite[2] system, described in Clark and Manning (2015), as $COREF$ for scoring coreference. The external corpus was built from texts comparable to our dataset. 651,108 sentences containing one of *they* or *them* were mined from a larger corpus of 798 literary texts spanning the nineteenth and twentieth centuries (including novels such as *To The Lighthouse*, by Virginia Woolf). Lastly, the candidate m-antecedents are extracted from a 5-sentence pre-window of the given m-anaphor ($k = 5$) and the regularization parameter in learning is set to 0.20.

5.2 Clustering Features

Table 5 depicts the features we chose to use in the pairwise similarity metric (SIM_1) for agglomerative clustering of candidate m-antecedents. All are common to many coreference resolver systems (Durrett and Klein, 2013; Recasens et al., 2013b; Stoyanov et al., 2010). We distinguish between mention features (Columns 1 & 2), which are defined for each candidate m-antecedent in a pair, and pairwise features (Columns 3-5), which are defined over a pair of candidate m-antecedents.

Three features, in particular, deserve further discussion. Under morphosyntax (Column 3), *[Type Conjunctions]* is a placeholder for a number of conjunctive boolean features derived from the noun type (pronoun/proper/common) of each antecedent in a pairing: e.g., pronoun-pronoun, pronoun-proper, proper-pronoun. Similarly, *[Dependency Conjunctions]* is a placeholder for conjunctive boolean features derived from the grammatical dependency of each antecedent in a pairing: e.g., subject-subject, subject-object, object-subject. The *[# Dependency Pairings]* is an ordinal version of the *Dependency Conjunctions* feature set - a count of the number of occurrences rather than an indicator variable.

The ‘Governor = except’ feature triggers if one of the mentions in the mention pair is governed by *except* or *exclude*. It represents a form of negation of group membership (e.g. *Everyone except for Mary visited Castlebary*).

Features were extracted using the Stanford CoreNLP system (Manning et al., 2014) and animacy information was specifically obtained through the Stanford deterministic coreference resolution module (Lee et al., 2011).

[2] `http://stanfordnlp.github.io/CoreNLP/coref.html`

Morphosyntax (Mention)	Grammatical (Mention)	Morphosyntax (Pairwise)	Grammatical (Pairwise)	Semantic (Pairwise)

Type = pronoun
Type = proper noun
Animacy = animate
Animacy = unknown
Person = first
Person = third
Singular = true
Quantified = true
# Modifiers 	
	Sentence position = first
	Sentence position = last
	Dependency = subject
	Dependency = object
	Dependency = preposition
		Head match
		[Type Conjunctions]
			Word distance (max. 30)
			Sentence distance
			Coordination = and
			[Dependency Conjunctions]
				Governor = except
				# Conjunctive pairings
				[# Dependency Pairings]

Table 5: Features used in the clustering similarity metric, separated by category. The features *[Type Conjunctions]*, *[Dependency Conjunctions]*, and *[# Dependency Pairings]* are all placeholders for feature sets. See the text for details.

6 Experiments

In order to assess the performance of our method, we conduct two experiments. In the first, we assess performance of our system on the specific *they-them* m-anaphor resolution sub-task. Our system, and its variants, are compared against a number of baseline methods based on performance on the test set.

In the second experiment, we consider how our system improves the performance of a coreference resolution system when *all* anaphors (both 1-anaphors and m-anaphors) are considered.

6.1 Evaluation

Accuracy is measured in terms of the number of mention pairs correctly grouped as m-antecedents for a given m-anaphor — similar to previous works in anaphor resolution (Peng et al., 2015). We use the standard classification metrics for precision, recall, and F1-score. If $n_1, n_2, \ldots, n_N$ represent the number of gold m-antecedents for m-anaphors $g_1, g_2, \ldots, g_N$ in a document, and $m_1, m_2, \ldots, m_N$ are predicted, of which $k_1, k_2, \ldots, k_N$ are correct, then precision is defined as $\sum_i k_i / \sum_i m_i$ and recall as $\sum_i k_i / \sum_i n_i$, where i ranges from 1 to N.

In order to align ourselves with the gold labels, we adjust the predicted mention corresponding to an entity to the closest one preceding the given m-anaphor. Because a given entity may appear multiple times in a candidate mention window, the most recent one, relative to the m-anaphor, is not always the one carrying the strongest signal and hence is not always predicted as an antecedent. For the purposes of evaluation, such cases are considered correct. Automatic handling would involve a separate, single-antecedent coreference resolver, but given the thesis of this work is the multi-antecedent case, this choice is justified.

6.2 System Comparison

We first describe the various baselines and variants of our method we assess and then analyze the performance results.

Systems

- The "most-recent-k" baselines (denoted RECENT-k), which predict the most recent k mentions, relative to the m-anaphor, as the m-antecedents for k = 2, 3, 4.

- The random selection baseline (denoted RANDOM), which randomly predicts mentions in a 5-sentence pre-window as the antecedents according to a binomial with probability 0.5 (imposing the constraint that at least two must be predicted).

- A simple rule-based method (denoted RULE) which proceeds as follows:
  - If the m-anaphor occupies a subject or prepositional position, then predict the most recent mentions in subject positions if they are coordinated, otherwise take them from previous, distinct sentences. If no such mentions can be found take the most recent mentions in subject and object positions governed by the same verb.
  - If the m-anaphor occupies in object position, take the previous mentions in object or prepositional positions if they are coordinated, otherwise take them from previous, distinct sentences. If no such mentions can be found, take the most recent mentions in subject and object positions governed by the same verb.
  - Otherwise, take the two most recent mentions (usually arrive here if there is an error in the dependency parsing).

	Precision	Recall	F1
RECENT-2	21.46	17.68	19.39
RECENT-3	23.73	30.10	26.54
RECENT-4	21.43	38.82	27.62
RANDOM	30.02	29.11	29.56
RULE	39.23	17.45	24.16
LEE	**46.78**	9.91	16.36
M-LCS	41.35	37.81	39.50
M-TREE	41.94	**44.88**	**43.36**

Table 6: Test set performance of each system on the m-anaphor resolution task.

m-anaphor class	Precison	Recall	F1
2-anaphor	48.14	52.90	50.41
3-anaphor	35.92	34.77	35.34
4-anaphor	36.74	12.87	19.06

Table 7: Performance results of the M-TREE system on the different classes of m-anaphors.

- The system described in Lee et al. (2011) (denoted LEE), which performs some light m-anaphor resolution (solely for conjunctive cases).

- The two variants of the developed method, one using the LCS similarity metric (denoted M-LCS) and the other using the subset tree kernel (M-TREE).

Results and Discussion

Accuracy results on the test set for each of the systems are given in Table 6. Both the proposed systems, M-LCS and M-TREE, outperform all other methods by a substantial margin. The Stanford system achieves the highest precision, which is not surprising because it targets conjunctive mentions, which often serve as m-antecedents. Based on the analysis of Section 4, the poor performance of RECENT-2, RECENT-3, and RECENT-4 is expected.

The results for the best-performing system, M-TREE, on the different classes of m-anaphors is given in Table 7. M-TREE outperforms all other systems but exhibits a bias towards 2-anaphors, recent mentions, and mentions coordinated by conjunction. This is not surprising given such cases are the easiest to resolve.

6.3 Full Coreference Resolution

For the complete coreference resolution task, the M-TREE system can be integrated with an existing coreference system. For this experiment, we pair the full coreference resolution system of Clark and Manning (2015) with M-TREE, and we raise the prediction threshold of our model to 0.89, at which point precision on the validation set is 78.9. Moreover, we restrict ourselves to the P&P portion of the test set, given the Scribner stories only have gold labels for instances of *they* and *them*.

	MUC	B^3	CEAF$_e$	Avg.
CLARK	42.3	39.5	32.4	38.1
CLARK+M-TREE	**43.4**	**40.0**	**31.9**	**38.7**

Table 8: CoNLL metric scores for coreference resolution on the test portion of P&P for the Clark and Manning (2015) system, with (CLARK+M-TREE) and without (CLARK) the pairing with M-TREE.

The Clark and Manning (2015) system is first run over the test set, producing coreference chains which are then filtered for character entities using the approach of Vala et al. (2015). Our adjusted M-TREE system is then applied over all *they* and *them* mentions. Each such mention predicted as m-anaphoric is added to the coreference chains of the entities corresponding to the m-antecedent mentions.

To evaluate the accuracy against the gold mention clusters, each m-anaphoric *they* and *them* is added to each cluster containing a gold m-antecedent. The CoNLL metric scores (Bagga and Baldwin, 1998) of the coreference predictions are shown in Table 8, with the integrated system outperforming the Clark and Manning (2015) system by 0.6 average score (pairing the Clark and Manning (2015) system instead with an oracle m-anaphor resolver yields an average score of 44.8, an increase of 6.7 points).

7 Related Work

The formal problem statement for the noun phrase anaphor resolution we propose is an extension of the standard ACE (Doddington et al., 2004), MUC (Hirschman, 1997), and Ontonotes (Hovy et al., 2006) formulations, as well as the problem settings outlined in Wiseman et al. (2015) and Durrett and Klein (2013), to allow anaphors to link to multiple antecedents. Most previous works impose the constraint that anaphors can be assigned at most one antecedent. Some works cast the coreference resolution problem in an Integer Linear Programming framework, with an explicit constraint for assigning at most one antecedent to an anaphor (Peng et al., 2015; Denis et al., 2007).

The early work of Ingria and Stallard (1989) proposes the resolution of pronouns without the restriction they be linked to at most one antecedent. The method uses an indexing scheme for parse trees, similar to Hobb's algorithm (Hobbs, 1978), that eliminates candidates antecedents as more information is acquired. Those pronouns with multiple candidates remaining after tree-traversal are predicted as m-anaphors. The method considers each parse tree in isolation, and hence does not permit inter-sentential linking, a severe limitation in corpora such as the one offered in this work.

Other researchers have evaluated noun phrase coreference resolvers along a number of dimensions, including different classes of anaphors (Mitkov, 2014; Kummerfeld and Klein, 2013; Björkelund and Nugues, 2011; Recasens and Hovy, 2009; Stoyanov et al., 2009; Bengtson and Roth, 2008; Van Deemter and Kibble, 2000; Ng and Cardie, 2002b; Kameyama, 1997). This work explores a new class of anaphor, previously unstudied, and evaluates its impact on the coreference resolution problem.

Many state-of-the-art systems for coreference resolution, especially supervised, are constrained to the single-antecedent case (Clark and Manning, 2015; Peng et al., 2015; Wiseman et al., 2015; Björkelund and Farkas, 2012; Ng, 2010; Stoyanov et al., 2010; Ng, 2008; Soon et al., 2001). The most well-known, benchmark datasets for coreference resolution (e.g. Ontonotes and ACE-2005), do not offer gold annotations for multi-antecendet anaphors. Our work presents the first dataset for tackling this problem.

The Lee et al. (2011) is a deterministic system that attempts to resolve the "easy" multi-antecedent cases, namely those in which mentions are joined by some conjunction. Our system goes beyond and attempts to predict more difficult cases as well.

Many of the individual features we employ in our model appear in a variety of other coreference systems, especially those involving mention-pair models (Durrett and Klein, 2013; Recasens et al., 2013b; Stoyanov et al., 2010). Recasens et al. (2013a) attempts to perform coreference resolution under conditions where many standard features for coreference are not suited. Peng et al. (2015) resort to corpus counts of predicates as features, much in the same way we obtain counts of mention pairings according to simple predicates on dependency structures.

The system of Clark and Manning (2015) also makes uses of agglomerative clustering, although it's employed in merging coreference chains, rather than candidate antecedent groupings.

Last, resorting to an external corpus for sentence structures is common practice in the Natural Language Generation literature for producing phrases that are coherent and consistent(Krishnamoorthy et al., 2013; Bangalore and Rambow, 2000; Langkilde and Knight, 1998).

8 Conclusion

We introduced a new class of anaphors to the anaphor resolution problem, m-anaphors, and extended the problem formulation to incorporate them. We offered insights into the linguistic behaviour of m-anaphors, finding that surface-level syntactic and semantic features do not carry enough discriminative power in distinguishing them from 1-anaphors. Furthermore, we developed a system combining a mention-pair model, an existing coreference resolver, and corpus knowledge to resolve m-anaphors that scores higher than a number of baseline methods. Finally, we paired this system with a coreference resolver to solve the general coreference resolution task, showing that m-anaphor prediction can help boost performance.

An important component of the m-anaphor resolution problem that falls outside the scope of this study, but is important for practical application, is the detection of m-anaphoric mentions. Section 4 gives some insight into the problem but a much deeper investigation is necessary to devise a detection method.

Moreover, for simplicity, this study focused solely on m-anaphoric *they* and *them* mentions, but as explained earlier, m-anaphoric mentions can take many forms, each introducing their own particular complexities that warrant special attention.

Regarding the system developed for m-anaphor resolution, resorting to an external corpus to obtain well-formed sentences proved to be very computationally expensive. In future work, we look to incorporate methods that incur less cost, possibly tolerating some error in the formation of sentences without significantly degrading performance. Also, negation of group membership is a complex linguistic phenomenon that was handled in a crude manner in our system. We look to devote future work to handling such cases.

To promote further research into m-anaphors, we make all our data and software freely available at `http://www.github.com/networkdynamics/manaphor-acl2016`.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Citeseer.

Srinivas Bangalore and Owen Rambow. 2000. Exploiting a probabilistic hierarchical model for generation. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 42–48. Association for Computational Linguistics.

Eric Bengtson and Dan Roth. 2008. Understanding the value of features for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 294–303. Association for Computational Linguistics.

Shane Bergsma and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 33–40. Association for Computational Linguistics.

Anders Björkelund and Richárd Farkas. 2012. Data-driven multilingual coreference resolution using resolver stacking. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 49–55. Association for Computational Linguistics.

Anders Björkelund and Pierre Nugues. 2011. Exploring lexicalized features for coreference resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 45–50. Association for Computational Linguistics.

Kevin Clark and Christopher D Manning. 2015. Entity-centric coreference resolution with model stacking. In *Association of Computational Linguistics (ACL)*.

Michael Collins and Nigel Duffy. 2002. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 263–270. Association for Computational Linguistics.

Pascal Denis, Jason Baldridge, et al. 2007. Joint determination of anaphoricity and coreference resolution using integer programming. In *HLT-NAACL*, pages 236–243. Citeseer.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *LREC*, volume 2, page 1.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *EMNLP*, pages 1971–1982.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

Lynette Hirschman. 1997. {MUC-7 Coreference Task Definition}.

Jerry R Hobbs. 1978. Resolving pronoun references. *Lingua*, 44(4):311–338.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics.

Robert JP Ingria and David Stallard. 1989. A computational mechanism for pronominal reference. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, pages 262–271. Association for Computational Linguistics.

Megumi Kameyama. 1997. Recognizing referential links: An information extraction perspective. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 46–53. Association for Computational Linguistics.

Niveda Krishnamoorthy, Girish Malkarnenkar, Raymond J Mooney, Kate Saenko, and Sergio Guadarrama. 2013. Generating natural-language video descriptions using text-mined knowledge. In *AAAI*, volume 1, page 2.

Jonathan K Kummerfeld and Dan Klein. 2013. Error-driven analysis of challenges in coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 265–277.

Irene Langkilde and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Compu-*

tational Linguistics-Volume 1, pages 704–710. Association for Computational Linguistics.

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. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34. Association for Computational Linguistics.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Michael Martone, Lex Williford, and Rosellen Brown. 1999. *The Scribner Anthology of Contemporary Short Fiction: Fifty North American Stories Since 1970*. Touchstone.

Ruslan Mitkov. 2014. *Anaphora resolution*. Routledge.

Alessandro Moschitti. 2006. Making tree kernels practical for natural language learning. In *EACL*, volume 113, page 24.

Vincent Ng and Claire Cardie. 2002a. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Vincent Ng and Claire Cardie. 2002b. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 104–111. Association for Computational Linguistics.

Vincent Ng. 2008. Unsupervised models for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 640–649. Association for Computational Linguistics.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1396–1411. Association for Computational Linguistics.

Haoruo Peng, Daniel Khashabi, and Dan Roth. 2015. Solving hard coreference problems. *Urbana*, 51:61801.

Marta Recasens and Eduard Hovy. 2009. A deeper look into features for coreference resolution. In *Anaphora Processing and Applications*, pages 29–42. Springer.

Marta Recasens, Matthew Can, and Daniel Jurafsky. 2013a. Same referent, different words: Unsupervised mining of opaque coreferent mentions. In *HLT-NAACL*, pages 897–906.

Marta Recasens, Marie-Catherine de Marneffe, and Christopher Potts. 2013b. The life and death of discourse entities: Identifying singleton mentions. In *HLT-NAACL*, pages 627–633.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France, April. Association for Computational Linguistics.

Veselin Stoyanov, Nathan Gilbert, Claire Cardie, and Ellen Riloff. 2009. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 656–664. Association for Computational Linguistics.

Veselin Stoyanov, Claire Cardie, Nathan Gilbert, Ellen Riloff, David Buttler, and David Hysom. 2010. Reconcile: A coreference resolution research platform.

Hardik Vala, David Jurgens, Andrew Piper, and Derek Ruths. 2015. Mr. bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 769–774.

Kees Van Deemter and Rodger Kibble. 2000. On coreferring: Coreference in muc and related annotation schemes. *Computational linguistics*, 26(4):629–637.

Sam Wiseman, Alexander M Rush, Stuart M Shieber, Jason Weston, Heather Pon-Barry, Stuart M Shieber, Nicholas Longenbaugh, Sam Wiseman, Stuart M Shieber, Elif Yamangil, et al. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 92–100. Association for Computational Linguistics.