Coreference Resolution: A Survey

Pradheep Elango

Computer Sciences Department
University of Wisconsin, Madison

Abstract

Coreference resolution is the task of resolving noun phrases to the entities that they refer to. Much work has been done in the past in this area and the related area of anaphora resolution. In this paper, we present a literature survey that is divided into two broad categories. Discussed first are papers that are linguistically motivated - based on syntax, focus and Centering theory. We then discuss machine learning techniques that are applied to coreference resolution, which include decision trees, conditional random fields, clustering, cotraining, and others. Further, we discuss evaluation methods, coreference classes and examine simple genre-dependent noun phrase characteristics of the Brown corpus. Finally, a preliminary proposal is presented, along with future directions.

1 Introduction

Coreference is, in a sense, the hyperlink of natural language. On the one hand, it lends an element of style and cohesion to the human writer, while on the other hand, it adds another dimension of obscurity to the mechanical understanding of language.

Coreference resolution has been a core research topic in NLP. It has important applications in areas such as question answering, machine translation, automatic summarization and named entity extraction.

The subject has a large, growing body of literature. Algorithms for the problem of pronoun resolution have been developed since the seventies [16, 29]. While the early approaches incorporated a lot of domain and linguistic knowldege, the newer approaches have showed an inclination towards knowledge-lean methods which were helped by the emergence of more powerful automatic parsers and taggers. In this paper we present a survey of some of the more significant papers in coreference resolution. Before describing the outline of the paper, we will discuss the scope of the problem.

1.1 Coreference vs. Anaphora

Coreference resolution has sometimes been confused with anaphora resolution, and often the strict definition of the coreference resolution has not been follwed. Defining the coreference problem for practical purposes is not a trivial task. The MUC-6 document [21] details as to what to annotate for coreference.

However, these guidelines do not strictly follow the definitions of coreference. According to [34], the task of coreference is often confused with the task of anaphora resolution. Two noun phrases are said to be co-referring to each other if both of them resolve to a unique referent (unambiguously). However, a noun phrase A is said to be the anaphoric antecedent of a noun phrase B, if and only if A is required for the interpretation of B. Thus, coreference is an equivalence relation, whereas anaphora is neither reflexive nor symmetric (nor transitive). A number of coreferential links could be anaphoric relations but some anaphora relations such as bound anaphora are not coreference links.

For example, sentences like the following contain bound anaphora:

a. *Every dog* has *its* day.

b. *The man* who gave *his* paycheck to *his* wife was wiser than *the man* who gave *it* to his mistress.

where the anaphor and the antecedent are not coreferent.

Further, there could be other anaphoric relations like the following:

c. The boy entered *the room*. *The door* closed automatically.

where the relation between *the room* and *the door* is that of meronomy/holonomy. In these cases, the two noun phrases do not refer to a single entity.

Other contentious cases exist such as concepts that vary with time such as the President of a nation, the price of a stock, etc. The problem with defining coreference resolution this way is the inappropriate use of the term "coreference" to cover semantic relations such as those involving temperature and price.

1.2 Outline of the paper

In the following sections, some important approaches that have been taken to solve coreference resolution are discussed, which are broadly classified into two categories: linguistics-based (methods which rely heavily on linguistic and domain knowledge), and machine learning-based (methods which rely on data driven approaches). We present a discussion of Linguistic methods mainly related to pronoun anaphora and other closely related anaphora in Section 2. We then discuss the machine learning themes in Section 3. Corpus characteristics and evaluation metrics are discussed in Section 4. We conclude with future directions in Section 5.

2 Linguistic approaches

In this section, we present and discuss the approaches that were primarily linguistics-based, which include those based on Hobbs' algorithm, Centering theory and other ideas based on Centering.

2.1 Hobbs' Algorithm

Hobbs' algorithm [16] was one of the earliest approaches to pronoun resolution. The algorithm is mainly based on the syntactic parse tree of the sentences. It makes use of syntactic constraints when resolving pronouns. [1]

First, intra-sentential antecedents are proposed - the syntactic tree of the current sentence is searched in a breadth-first left-to-right fashion to find antecedents. The *contra-indexing* constraint is taken care inside the algorithm, by making sure that the path from the NP to the S node of the syntactic tree has at least one another NP on the way. If there are higher-level nodes in the current sentence, then antecedents resulting from a breadth-first left-to-right search of each subtree, are proposed. Then, parse trees of previous sentences in reverse chronological order are searched in the same fashion to propose antecedents. In essence, Hobbs' algorithm prefers entities that are within the same sentence, and entities that are closer pronoun in the same sentence. Depending on the position of the pronoun in the sentence, different entities in a sentence may become more relevant. When looking for antecedents in previous sentences, the antecedents that occur (or are realized) in the subject position are more salient, since a breadth-first left-to-tree search is performed starting at the root S node of the sentence. Depth of a node in the syntactic tree is thus a very important factor to determine discourse prominence.

2.2 Centering

Centering theory [14] was proposed in order to model the "relationships among focus of attention, choice of referring expression and perceived coherence of utterances within a discourse segment". The model draws inspiration from previous discourse processing papers, and its principles have been widely used directly or indirectly in much later work. A discourse can be any source of textual information, like newswire text, literary works, dialogues or samples from speech.

One of the main goals of Centering is to track the entities in focus in a given sentence. The most detailed earliest work in this line is Sidner's thesis [29], which provides an extensive analysis on immediate focus, including algorithms and rules to apply immediate focus to resolve the referent of a pronoun or a demonstrative noun phrase ("this/that x"). The centering framework is a generalization, in part, of Sidner's idea.

Sentences can be written in different ways. Any entity need not be refered to by the same expression and in the same grammatical role (such as subject/object) always. The use of pronouns helps to keep the main subject of the discourse in focus, which can change over different portions of the discourse. If the subject, object pair refers to the same entities in a given discourse for each sentence resolving the references will be a simple task. In other words, the *inference load on the hearer* will be low in such cases. In discourses, where the subject and object entities keep flipping back and forth, inference load will be high. Let's consider the following example:

```
a. Terry really goofs sometimes.
b. Yesterday was a beautiful day and
   he was excited about trying out his
   new sailboat.
c. He wanted Tony to join him on a
   sailing expedition.
d. He called him at 6 AM.
e. He was sick and furious at being woken
   up so early.
```

In the above example, from (a) to (d), the subject has been Terry and the object has been Tony (in (c) and (d)). The sentence (e) is however a deviation from these sentences, and uses Tony in the subject position and Terry in the object position. This change is rather weird, and calls for a high *inference load* on the hearer. In other words, more domain knowledge might be required to solve this problem, than simple grammatical rules.

Let us describe two terms which are used in the centering literature. An *entity* means an object, that could be the targets of a refering expression. An *utterance* is used to describe the basic unit, which could be a sentence or a clause or a phrase. Assume that utterances are numbered as $U_1, U_2, ...U_N$ for a given discourse.

Each utterance is assigned a set of *forward-looking centers*, $C_f(U)$, and a single *backward-looking center*, $C_b(U)$. The centers refer to the entities in focus. The backward-looking center indicates the entity which is in focus at the point of the current utterance. The forward-looking centers for an utterance depend on the entities in that utterance. There is no dependence on the C_f's of previous utterances. The centers in $C_f(U_n)$ are ranked in order to reflect its prominence in U_n. The most highly ranked element in $C_f(U_n)$ that is *realized* in U_{n+1} is $C_b(U_{n+1})$. The forward list thus gives a probable list of entities that could be the focus in the next utterance.

Pure centering theory does not describe sophisticated methods to rank the C_f's or algorithms for "realization". This makes it more general, and the details could depend on the specific language and domain for a particular application.

The theory is inherently local. To form the forward-looking centers, it looks only at the expressions in the current utterance. Further, to form each backward-looking center, it looks only at $C_f(U_{n-1})$. In a way, this resembles Markovian assumptions.

[1] Two main types of syntactic constraints are binding constraints and sortal constraints. Binding constraints are one of several constraints that govern resolution of pronouns. For example, conder the sentence
"Mark was talking to him".
"Mark was talking to himself."
In the first case, the pronoun *him* cannot have Mark as an antecedent, whereas in the second case the pronoun *himself* can be matched only with Mark. The above is an example of *contra-indexing*. Though easy to describe, such constraints are hard to be incorporated because it involves semantics (of the verb involved). So, in many cases simple checks like ensuring the two arguments of a verb are not identical are implemented.

Sortal constraints are those that are governed by semantics. For example, the sentence
"Mark sleeps. She smiled."
should not have Mark as the antecedent for *she*. This requires gender information, in this case. Generally speaking, sorts can be hierarchies that incorporate domain-specific constraints.

Three types of transitions are defined across subsequent pairs of utterances.

- Center Continuation: $C_b(U_{n+1}) = C_b(U_n)$ and this is the most highly ranked element of $C_f(U_{n+1})$.
- Center Retaining: $C_b(U_{n+1}) = C_b(U_n)$ but this is not the most highly ranked element of $C_f(U_{n+1})$.
- Center Shifting:The two backward centers are different.

One of the main tenets of the centering theory is that for higher perceived coherence, sequences of center continuation are preferred over sequences of retaining; and sequences of retaining are preferred over sequences of shifting.

Further, if any element of $C_f(U_n)$ is realized by a pronoun in U_{n+1}, then the $C_b(U_{n+1})$ must be realized by a pronoun too.

These basic rules define the original centering theory. It should be noted that the goal of centering theory is not to resolve pronouns. It mainly provides a model which can predict the focus of the next sentence. Further, pronouns refer to objects in focus; thus, centering theory has been applied to develop pronoun resolution algorithms. However, extensions are needed to handle plural, quantified noun phrases and indefinites.

2.3 Centering applied to Pronoun resolution

Brennan, Friedman and Pollard [7] proposed an algorithm (henceforth abbreviated as BFP) to resolve pronouns adhering to the Centering Theory principles. Using pronouns helps the reader/hearer to focus attention; according to the authors, not using pronouns leads to a less-fluent communication. Centering theory provides a framework to model what a sentence is speaking about. This idea can be used to find which entities are referred to by pronouns in a given sentence. BFP makes one extension to the original centering model, by introducing an additional transition called smooth shift and rough shift. Suppose Cp(U_n) is the highest ranked entity in the Cf-list of U_n. Cb(U_n) is compared to two entities - Cb(U_{n-1}) and Cp(U_n) - in order to define the nature of the transtion. BFP makes the transition more fine-grained in the event of a shift. Specifically, when Cb (U_n) != Cb(U_{n-1}), the case when Cb(U_n) = Cp(U_n) is called a smooth shift, and the case when they are not equal is called a rough shift. The transition preference then becomes *continue* $>$ *retain* $>$ *smooth shift* $>$ *rough shift*.

The algorithm consists of three main steps:

- Construct all possible $< Cb, Cf >$ pairs, by taking the cross-product of Cb, Cf lists.
- Filter these pairs by applying certain constraints
- Classify each pair based on the transition type, and rank the pairs. Choose the best ranked pair.

One practical problem that we should consider is the treatment of non-anaphoric noun phrases. In the example, *yesterday* in statement (b) should not be considered for coreference resolution, and the same holds for the noun phrase *beautiful day*. This therefore will call for knowledge either in the form of extensive grammatical information, or domain knowledge.

Left-Right Centering

Left-Right Centering(LRC) [33] is an algorithm based on BFP and Centering, motivated by the lack of support for the incremental resolution of pronouns and the computational overheads of generating pairs. The main differing step is when they process an utterance: For each pronoun, they search for an antecedent in the same sentence by looking at the partial Cf(U_n) to identify an entity that meets feature and binding constraints. If nothing is found then they look for entities in the previous sentence. The Cf list of U_n is formed by ranking entities in U_n. Though ranking can be any complex mix of syntax and semantic constraints, it is usually done using grammatical role (which can be approximately computed using a left-to-write breadth-first walk of the parse tree).

2.4 Pronoun resolution without centering

Strube's approach [31] models pronoun resolution avoiding the idea of backward centers. According to the algorithm, a list of entities called the S-List is maintained, and the best matching entity searched from that list in the order of the ranks is used to resolve pronouns (after certain constraints such as the binding contraints, agreement etc. have been applied). This allows one to incrementally resolve pronouns, which closer to how humans interpret pronouns. The approach does not involve transitions in the centering sense. However, it uses a hierarchy of entities, based on when they appeared in the discourse. Roughly, the hierarchy can be summarized as hearer-old discourse entities in U_i, hearer-old discourse entities in U_{i-1}, mediated discourse entities in U_i, mediated discourse entities in U_{i-1}, hearer-new discourse entities in U_i, and hearer-new discourse entities in U_{i-1}. Thus, this gives preference to intra-sentential entities.

2.5 Centering in Practice

Though there have been so many pronoun/anaphora resolution methods based on centering and many variants of the centering theory, the original claims of Centering theory was empirically validated only recently. In [27], Poesio et al. present a parametric view of the Centering theory, along with empirical evaluations. In particular, they note that while the preference for pronomalizing the backward-looking center is observed, the constraint on the uniqueness of the Cb is more dependent of the parameters. Furthermore, the "parameter space" in this case is large since the original centering theory leaves many details unspecified - such as the ranking of entities, what an utterance should be, how we compute a previous utterance - with the view that these factors should be language dependent.

Experiments with different choices for defining an utterance to determine the best utterance unit. An interesting observation is that just 1.1% of the utterances have more than once CB. Surprisingly, a large number of transitions are ZERO or NULL transitions (nearly 64% combined), which are those transitions that involve at least one utterance with no CB. And for about 55% of the cases, a CB is realized as a pronoun. The conflict between the local preference for a locally salient entity and the global preference for the main entity in the discourse is observed. The corpus included a text from the museum domain, and one from the pharmaceutical domain; the proportion

of transitions varied across these domains. Further, one more interesting observation is the fact that it is generally not common to use a single referring expression to an entity through the course of a discourse.

2.6 Bridging References

Bridging references,(or indirect anaphora or associative anaphora), arise when a reference to an object that is not directly mentioned, is made. For example, consider the following example:
When the detective got back to **the garage**, *the door* was unlocked.
Resolving bridging references require background knowledge. Few systems use the web for background knowledge [26, 8, 18]. These systems use search engine results by issuing queries that contain the referring expresion and a candidate antecedent to estimate the strength of a candidate link. Alternatively, WordNet [26] has been used to provide background knowledge. WordNet is used to determine whether there exists any direct relation (such as synonyms, hypernyms) or an indirect relation [10] between a candidate antecedent and the refering expression. However, WordNet is not a complete resource as Vieira et al. [35] point out 62% of the meronymy relations needed for bridging resolution in their corpus were not encoded in WordNet. It should be noted that there are other kinds of anaphora that we have not discussed such as verb phrase anaphora.

3 Machine Learning Approaches

In this section, a few approaches that are based on machine learning are presented. We consider a method to be machine-learning based, if it acquires knowledge using a learning algorithm and training data. Interest in the machine learning commmunity on the coreference resolution problem has risen since the mid-to-late 90's. Starting from simple statistical naive bayes-based model, we describe methods using decision trees and conditional random fields.

3.1 Naive Bayes'

A statistical approach to anaphora resolution was introduced by Ge, Hale and Charniak in '98 [13]. The probabilistic model includes several syntactic and semantic features which affect pronoun resolution. In this model, the random variable is the candidate antecedent for a given pronoun. Following is the description of the features:

- Distance between the pronoun and the candidate antecedent (closer ones are preferred).
- Syntactic structure. This helps to resolve binding constraints like the contra-indexing constraints in the Hobbs' algorithm.
- Gender, number and animacy. These constraints (that have been called agreement constraints and sortal constraints in different circles) can be implemented based on the actual words that occur.
- Mention count. Noun phrases that occur repeatedly get more preference. Probability that a proposed antecedent is correct given that it occurs a certain number of times, is computed.

The motivation for the mention count can be traced to Centering theory, according to which a continued topic is the highest-ranked candidate for a pronoun. However, locality and the preference among different transitions may not be directly modeled here. A modified version of Hobbs algorithm is used to compute distance between pronoun and a proposed antecedent, thus taking into account both the syntactic structure and the distance. The Hobbs algorithm also provides the antecedents for which the probability of the antecedent being the correct antecedent for the pronoun is computed.

3.2 Decision Trees

Coreference resolution can be cast as a pairwise classification task. Soon et al. [30] adopt a decision-tree approach for coreference resolution. The coreference resolution problem is cast as a classification problem, the question being whether two *markables* corefer or not. A markable could be a noun phrase or a pronoun, thus generalizing the coreference resolution problem beyond pronouns. All possible markables are identified during preprocessing steps. For learning, they employ a decision-tree approach to learn rules based on different features computed on pairs of markables. Some of their features, which do not use too much of syntactic information, are described briefly below:

- Distance feature: Distance between the two markables in terms of number of sentences is used.
- Agreement features: A few features to take care of gender and number agreement
- Type of markable: A few features which indicate the type of the markable, namely, demonstrative noun phrase, definite noun phrase, pronoun, reflexive pronoun, and proper noun.
- Semantic class agreement: Another interesting feature is the semantic class agreement feature which basically checks if the semantic class of the two markables, agree according to the WordNet hierarchy.
- Alias feature: Two markables have a positive alias feature if they share the same name, or if one markable has just the last name and the other has the complete name, or if one is the acronym of the other.

Training samples are automatically generated from the corpus - the positive samples are generated for immediately adjacent noun phrase pairs, and the negative samples are generated by using pairs that are not marked as coreferent.

Ng and Cardie [23] propose a slight modification to the above framework for generating the training data by treating non-pronominal NPs and pronominal NPs separately. Further, they extend the feature set by including a bunch of additional features. Notably, they include features that consider the grammatical role of the NPs (subject/object), and a lot of heuristics. However, the massive feature engineering effort does not nearly pay off as much, owing to data fragmentation problems.

Ng and Cardie [24] present another extension to the above framework by including a separate classifier for determining if a noun phrase is anaphoric or not. A maximum entropy model is used to train this classifier, and the overall coreference classifier uses the same decision-tree learning framework.

Yang et al. present a competition-based learning approach [39]; instead of using pairs of candidate antecedents and a anaphor for a given anaphor as training set, they use a pair of candidate antecedents such that one is positive and the other is negative, along with the anaphor during training. The motivation is that single candidate models are not sufficient to learn resolution; the twin candidate model helps in this regard by trying to learn the difference between a positive and a negative candidate for an anaphor. However, it is not particularly clear why they used the specific set up of positive and negative training examples set-up, and a decision tree for the purpose.

3.3 Conditional Random Fields

McCallum et al. apply CRFs [20] to attack the coreference problem. The authors propose three models for the purpose. The first model is a very general discriminative model where the dependency structure is unrestricted. The model considers the coreference decisions and the attributes of entities as random variables, conditioned on the entity mentions and the feature functions depend on the coreference decisions, **y**, the set of attributes, **a** as well as the mentions of the entities, **x**.

In the second model, the authors remove the dependence of the coreference variable, **y**, by replacing it with a binary valued random variable, Y_{ij} for every pair of mentions. Further the clique potentials are restricted to only pairs of mentions; and an additional term, in order to ensure there are no cyclic coreference errors, is added. Though the authors do not use the above two models for implementation, they point out various algorithms that can be used for inference and estimation.

The third model that they introduce does not include attributes as a random variable, and is otherwise similar to the second model. This model is used in their implementation. According to their results, their model performs a little better than the approach by Ng and Cardie. Still, the F1 results on NP coreference on the MUC-6 dataset is only about 73%.

The major advantages of using CRFs is that it can take care of transitive dependencies. For example, if a mention "Mr. Powell" and "Powell" are coreferent, then the chances of "Powell" and "she" corefering will be very low. To effect this, an additional term is included in the conditional for **y** that considers all possible triangle relations, with very high negative weights. The inference problem is analogous to graph partitioning, with an unknown number of partitions. The Correlational Clustering algorithm[4] is used to approximate the graph partitioning problem, which works by measuring the inconsistency incurred by including a node in a partition and minimizing the disagreements. In [32], a skip-chain CRF is introduced for the purpose of coreference on proper names, as a part of an information extraction task. An improved approach is presented in [11], which uses long-distance features and Gibbs' sampling for inference. According to [38], CRFs present a natural framework to integrate named entity extraction and coreference resolution of proper names. However, it is not clear if an integrated model will be useful for resolution of other types of noun phrases.

3.4 Coreference as clustering

In [9], Cardie uses a feature vector representation for each noun phrase, and then applies a clustering algorithm on these feature vectors. The clustering algorithm takes care to avoid triangle inconsistencies. The clustering algorithm resembles agglomerative clustering, checking at each merging step whether all the member of the two clusters merged are compatible with each other. This would avoid a noun phrase like "Mr. Powell" being clustered in a group that contains "she". However, this method is not completely unsupervised, as the distance metric used for comparison, uses fixed weights that are heuristically decided.

Wagstaff and Cardie [37] propose a modification of the clustering algorithm, called constrained clustering, to the noun phrase coreference problem. Specifically, their clustering algorithm accepts constraints in the form of "cannot-link" constraints and "must-link" constraints. "Cannot-link" constraints indicate noun phrases that cannot be grouped in the same cluster, whereas "must-link" constraints indicate NPs that should be grouped together. Most of the constraints used in their experiments were "cannot-link" constraints, where each of these constraints model a specific linguistic constraint, such as gender constraint, number constraint, semantic class compatability, article constraints etc. It will be useful to remark here that not all these constraints are perfect, for e.g., the number constraint uses very simple heuristics, such as looking for "s" or "es" endings. Further, there is no clear demarcation of which constraint rules over the other, so there is a sort of hierarchy among these constraints for assigning preference.

Finley and Joachims [12] describe an approach to supervised clustering. The algorithm learns a similarity measure to produce desired clusterings. This contrasts with using pairwise classification, where the target concept to be learned is "same cluster or not". The sparsity of such pairs in the training data (only around 1.6% of the pairs in MUC-6 training set are coreferent) make the data set imbalanced. However, Ng and Cardie use only the closest preceding non-pronominal noun phrase for a given phrase to make a positive pair, and all coreferent pairs between the two are paired to form negative samples. Further, like the CRF approach this algorithm too can take care of transitive dependencies. The main difference with the CRF approach is that the objective to be maximized is the margin instead of the likelihood as in the CRF method. The objective function is the same as the one used in [4]. One problem is that the number of constraints in this formulation can grow faster than exponential with the number of items. Suitably optimizing the objective function is an NP-complete problem, and therefore approximate inference methods are adopted. One of the advantages is that this method can handle transitive dependencies; however, when there is not much transitive dependency in the dataset, the performance of this method is only comparable to that of pairwise classification. Further, it is not clear whether this approach is better than just correlation clustering or the CRF approach discussed earlier.

3.5 Co-training for Coreference Resolution

Muller et al. present a co-training approach to cofererence resolution in [22]. Parallel to the original co-training framework, they divide the data into two views; however, the views here

are features and do not necessarily provide a natural feature split. Interestingly, they divide the data into three sets based on the noun phrase form and argue how the algorithm performs with respect to the different forms, with respect to their dataset (drawn from German short stories).

Ng and Cardie [25] approach the problem in a different way. Since a natural feature split does not exist in coreference resolution data sets (or is hard to find), they advocate the use of single-view bootstrapping algorithms. In their experiments with MUC-6 and MUC-7 datasets, they find that self-training algorithm performs slightly better than Blum and Mitchell's co-training algorithm. A greedy approach is used to find a good feature split for co-training. However, they argue that using multiple learners instead of multiple views will be useful in such scenarios, since different learners can have different biases and induce different hypotheses in a complementary manner. In their experiments, they find that their single-view, multiple learner bootstrapping algorithms perform much better than the multiple-view single learner cotraining algorithms.

3.6 Corpus-based approaches

Harabigiu et al. [15] extend data mining approaches to the problem of coreference resolution. They use the annotated coreference chains from MUC-6 and MUC-7 datasets to generate more coreference data. One interesting result from the paper is that the number of anaphor-to-proper noun links is around 29.1%, and the number of coreference links between two common noun phrases is around 10%. Further, nearly 83% of coreference links in the MUC-6 corpus is resolved with simple rules/features such as repetition, alias, common head, etc. Multiple knowledge rules are then combined using the entropy of the rule as a measure of confidence of the rule. The best partitioning of a given set of noun phrases is then computed by maximizing an objective function which resembles the one used in correlation clustering. In [19] corpus-based approaches for obtaining knowledge are used for resolving other-anaphora . For example, if we want to find instances of comparative anaphora (such as *another such facility, other repercussions*), we need knowledge to indicate that the referring expression and an antecedent are related. A corpus is used to mine for patterns that contain patterns such as "X and other Y", "X other than Y" to find related noun phrases. The rationale is that a pattern that occurs frequently indicates a relation between the two noun phrases (or the type of the noun phrases, if one of them is a proper noun). Apart from sparsity issues, one problem is coming up with useful patterns. We believe a co-training approach, similar to the lines of [17] could help generate more patterns.

Bean and Riloff [6] developed a system that learns relations between words and the different contexts in which they can appear, in an unsupervised manner. Identifying the contextual roles of noun phrases is important for coreference. Consider the following example:

a. The boys were kidnapped by masked men.
b. After **they** were released ....
c. After **they** blindfolded ...,

In case (b), "they" refers to the boys, and in (c), "they" refers to the masked men. Resolving such references requires background knowledge that kidnapped people are released, etc. Encoding such knowledge is a largely unsolved problem. However, a large corpus can facilitate mining of interesting patterns. Bean and Riloff adopt an unsupervised corpus-based approach for this purpose. Context is represented by a *case frame* that can be thought of a phrase with a filler, such as "murder of $< NP >$", "killed $< NP >$", "$< agent >$ added", "$< agent >$ stated".

The system is divided into modules. One module constructs a case frame network, by identifying related frames, with the underlying assumption that frames of words that co-occur are related (based on synonyms or events). Thus, given a case frame, one can list all other case frames that can be expected to occur with it. The hope is that the situation expressed in the example above will be captured here: "$< NP >$ were kidnapped" can be expected to co-occur (or be related) with "$< NP >$ were released".

Another module identifies related case frames for words. If two words co-occur, then they are related to each other's case frames also. So, given a case frame, one can list all the words that can be expected to occur with that frame.

Finally, another module learns relations between frames and the semantic type of the words. The semantic type is looked up from WordNet. The hope is that this will be an useful generalizing step. The authors point out that WordNet classes could be coarse and noisy due to polysemy. For boot-strapping the system, it is seeded with an initial set of easy-to-identify patterns for co-occurring noun phrases.

The other common features such as gender, number, etc. appear in this model as separate knowledge sources. The multiple knowledge sources are then combined using the Dempster-Shafer probabilistic model [28] to resolve noun phrases. The Dempster-Shafer theory can be used to combine multiple evidences with differing beliefs, in order to find the most likely hypothesis. However, in [28] it is argued that this theory is not very well understood, and one could perform similar analysis with a complete Bayesian model.

4 Evaluation and Corpus Characteristics

Evaluation Metrics

Most systems use the MUC-6 and MUC-7 datasets for evaluation. The MUC-6 evaluation metrics are based on the model-theoretic scoring scheme in Vilain et al. [36]. Precision and recall metrics based on the number of missing links and number of links for a coreference chain. The number of missing links are based on the number of partitions that are generated for a given coreference chain, and similarly the number of incorrect links that are present in a partition are used to compute precision. However, this technique weighs all link errors equally. A link error linking two large groups could do more damage than one linking two small groups. Bagga and Baldwin [3] present their B-CUBED evaluation algorithm to deal with this issue.

Genre	Total # NPs	%Proper nouns	%Pronouns
PRESS:REPORTAGE	27002	22.22	8.35
PRESS:EDITORIAL	15977	14.92	12.86
PRESS:REVIEWS	10584	18.13	9.56
RELIGION	10462	11.17	14.61
SKILL-AND-HOBBIES	21271	9.18	10.35
POPULAR-LORE	28810	12.00	12.42
BELLES-LETTRES	46356	13.85	14.92
MISCELLANEOUS	18039	14.74	6.15
LEARNED	46193	8.87	7.25
FICTION:GENERA	18428	12.46	25.97
FICTION:MYSTERY	15442	11.86	29.81
FICTION:SCIENCE	3761	12.79	25.02
FICTION:ADVENTURE	18795	10.72	28.26
FICTION:ROMANCE	18779	11.76	31.96
HUMOR	5649	10.94	23.40

Table 1: **Brown Corpus Statistics.** *Press reports contain a large percentage of proper nouns whereas fiction texts contain a large percentage of pronouns.*

Corpus characteristics

Bagga[2] presents a classification of coreference classes, which could help analyzing the strengths and weaknesses of a given coreference resolution system. The major classes defined are appositives, syntactic equatives, proper names, pronouns, quoted speech pronouns, demonstratives, and so on. Proper names and pronouns form 27.8% and 21.0% of the coreference classes in the WSJ corpus, whereas the class of coreference that requires external world knowledge amounts to only 5.9%.

Another useful way of analyzing a corpus will be by computing such a distribution for different genre of texts. We have performed a simple characterization of noun phrases using the Brown corpus. From Table 1, we see that the type of text also determines the percentage of proper nouns and pronouns. Fiction texts tend to have more pronouns, whereas the press reports contain more proper nouns. Many systems perform very well on proper name coreference; however, this forms a major portion only for press reports. Improving system performance for other noun phrase types is thus important.

Performance in a given class will depend on a specific ability of the system. For example, a good named entity recognizer is required for the appositives, syntactic equatives, and proper nouns. This classification can be very useful for performance analysis; on similar lines, a corpus-based evaluation for anaphoric and non-anaphoric noun phrases is provided in [5]. Definite noun phrases (such as *the door, the cat*) is another hard category to resolve. One difficulty is that while pronouns are mostly anaphoric, definite NPs are not so. So, determining the anaphoricity of a definite NP is another problem. Further, the distance of such definite anaphora from the antecedent could be considerably larger compared to the distance of a pronoun from its antecedent. Moreover, it is generally believed pronouns refer to entities in focus, while definite NPs refer to those not in focus. Allen [1] identifies an interesting class of

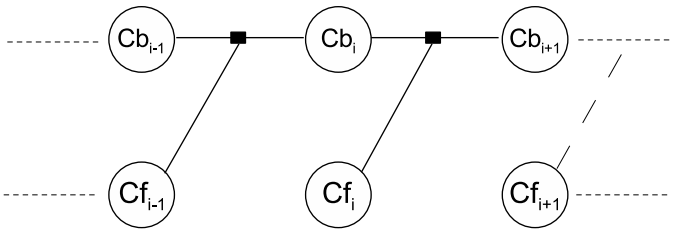

Figure 1: **A Centering-based CRF model.** *The figure shows a factor graph with factors involving forward centers, and backward centers.*

definite noun phrases, called *existential* noun phrases, which uniquely specify an object/concept and therefore do not need an antecedent. Bean and Riloff [5] have developed a system that identifies recognizes exisitential definite NPs. To resolve definite NP anaphors, systems use WordNet for a notions such as semantic compatibility between concepts, synonymy and hyponymy. In the past researchers have pointed out that [15] more than 30% of the missed coreference links is due to the lack of sufficent semantic consistency information.

5 Conclusions and Future Directions

In this paper, we have discussed several approaches to coreference resolution, surveying the classical approaches as well as the state-of-the-art systems. On the one hand, many linguistic approaches have been strongly based on syntax, linguistic and psycholinguistic principles. On the other hand, machine learning approaches focus on using more easily computable syntax information, automatically learning the parameters (which are heuristically decided in non-learning methods) based on training data, and of late, on using easily available unlabeled data.

The performance of an algorithm depends on the coreference class and the feature set, among other issues. For example, one can fine-tune the system for optimizing resolution proper nouns, if the underlying dataset consists of press reports.Analyzing the characteristics of the underlying data set provides an idea of the distribution of the different coreference classes in the corpus, which could help in deciding on a particular coreference resolution algorithm.

Evaluation, however, is conducted generally on the standard MUC-6 and MUC-7 datasets. With the availability of limited annotated data, there seems to be little choice on a standard evaluation data set. However, the narrow domain it represents motivates the need for wider variety of annotated data.

One line of future work that we identify is combining the strengths of the two themes, using more of the richer machine learning models with the linguistic ideas. As a starter, we present an initial proposal of using a CRF based on a centering model. Figure 1 shows the outline of the model, which is a product of factors, each modeling a set of relations (features) defined among the backward center of a given clause, the previous backward center, and the previous forward center list. These factors can thus model the preferences for the different kinds of center transitions. With the flexibility of the CRF model, context-dependent transition preferences can also be modeled. As discussed in section 2.5, using a clause as an utterance unit would be a reasonable assumption. Thus text is

represented as a set of utterances (clauses), which in turn is represented by the feature space representations of their forward center lists. The feature representation of the forward list could include factors such as grammatical role, gender, number, etc. Techniques for inference and estimation for similar linear chain CRFs are discussed in [32].

Further, another line of future work could be in characterizing the differences between two machine learning approaches for reference resolution empirically. While models have grown more complex, there has been no work that directly compares one algorithm to another with respect to hand-crafted datasets. For example, consider the decision tree approach and CRFs. Both these algorithms work very differently. The former is a rule-learning technique whereas the latter is a discriminative model. Yet, coreference resolution systems can use these as black boxes for pairwise classification. If transitive dependencies are not present, how much different will these methods perform? Another line of comparison would be among the supervised clustering model and CRFs. It will be interesting to carefully craft datasets that will expose the behavior of these algorithms empirically.

Coreference resolution, generally regarded as an "AI-complete" problem, is an area with interesting research problems.

References

[1] J. Allen. *Natural language understanding*. Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA, 1988.

[2] B. Amit. Evaluation of coreferences and coreference resolution systems, 1998.

[3] B. Amit and B. Baldwin. Algorithms for scoring coreference chains, 1998.

[4] N. Bansal, A. Blum, and S. Chawla. Correlation clustering. *Machine Learning*, 56(1-3):89–113, 2002.

[5] D. L. Bean and E. Riloff. Corpus-based identification of non-anaphoric noun phrases. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 373–380, Morristown, NJ, USA, 1999. Association for Computational Linguistics.

[6] D. L. Bean and E. Riloff. Unsupervised learning of contextual role knowledge for coreference resolution. In *HLT-NAACL*, pages 297–304, 2004.

[7] S. E. Brennan, M. W. Friedman, and C. J. Pollard. A centering approach to pronouns. In *Proceedings of the 25th annual meeting on Association for Computational Linguistics*, pages 155–162, Morristown, NJ, USA, 1987. Association for Computational Linguistics.

[8] R. Bunescu. Associative anaphora resolution: A web-based approach. In *Proceedings of the EACL-2003 Workshop on the Computational Treatment of Anaphora*, pages 47–52, Budapest, Hungary, 2003.

[9] C. Cardie and K. Wagstaff. Noun phrase coreference as clustering, 1999.

[10] J. Fan, K. Barker, and B. Porter. Indirect anaphora resolution as semantic path search. In *K-CAP '05: Proceedings of the 3rd international conference on Knowledge capture*, pages 153–160, New York, NY, USA, 2005. ACM Press.

[11] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[12] T. Finley and T. Joachims. Supervised clustering with support vector machines. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 217–224, New York, NY, USA, 2005. ACM Press.

[13] N. Ge, J. Hale, and E. Charniak. A statistical approach to anaphora resolution. *In Proceedings of the Sixth Workshop on Very Large Corpora*, 1998.

[14] B. J. Grosz, S. Weinstein, and A. K. Joshi. Centering: a framework for modeling the local coherence of discourse. *Comput. Linguist.*, 21(2):203–225, 1995.

[15] S. M. Harabagiu, R. C. Bunescu, and S. J. Maiorano. Text and knowledge mining for coreference resolution. In *NAACL '01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, pages 1–8, Morristown, NJ, USA, 1995. Association for Computational Linguistics.

[16] J. Hobbs. Resolving pronoun references. *Readings in natural language processing*, pages 339–352, 1986.

[17] R. Jones. Ph.d. dissertation, 2003.

[18] K. Markert, N. Modjeska, and M. Nissim. Using the web for nominal anaphora resolution, 2003.

[19] K. Markert and M. Nissim. Comparing knowledge sources for nominal anaphora resolution. *Computational Linguistics*, 31(3):367–402, 2005.

[20] A. McCallum and B. Wellner. Conditional models of identity uncertainty with application to proper noun coreference. In *In Neural Information Processing Systems (NIPS)*, 2004.

[21] MUC-6. Muc-6. message understanding conference.

[22] C. Muller, S. Rapp, and M. Strube. Applying co-training to reference resolution. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 352–359, Morristown, NJ, USA, 2001. Association for Computational Linguistics.

[23] V. Ng and C. Cardie. Improving machine learning approaches to coreference resolution. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 104–111, Morristown, NJ, USA, 2001. Association for Computational Linguistics.

[24] V. Ng and C. Cardie. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[25] V. Ng and C. Cardie. Bootstrapping coreference classifiers with multiple machine learning algorithms. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 113–120, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[26] M. Poesio, R. Mehta, A. Maroudas, and J. Hitzeman. Learning to resolve bridging references. In *ACL*, pages 143–150, 2004.

[27] M. Poesio, R. Stevenson, B. D. Eugenio, and J. Hitzeman. Centering: A parametric theory and its instantiations. *Comput. Linguist.*, 30(3):309–363, 2004.

[28] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 2003.

[29] C. L. Sidner. Focusing for interpretation of pronouns. *Computational Linguistics*, 7(4):217–231, 1981.

[30] W. M. Soon, H. T. Ng, and D. C. Y. Lim. A machine learning approach to coreference resolution of noun phrases. *Comput. Linguist.*, 27(4):521–544, 2001.

[31] M. Strube. Never look back: an alternative to centering. In *Proceedings of the 17th international conference on Computational linguistics*, pages 1251–1257, Morristown, NJ, USA, 1998. Association for Computational Linguistics.

[32] C. Sutton and A. McCallum. An introduction to conditional random fields for relational learning. in introduction to statistical relational learning. edited by lise getoor and ben taskar. mit press, 2006.

[33] J. R. Tetreault. Analysis of syntax-based pronoun resolution methods. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 602–605, Morristown, NJ, USA, 1999. Association for Computational Linguistics.

[34] K. van Deemter and R. Kibble. On coreferring: coreference in muc and related annotation schemes. *Computational Linguistics*, 26(4):629–637, 2000.

[35] R. Vieira and M. Poesio. An empirically based system for processing definite descriptions. *Comput. Linguist.*, 26(4):539–593, 2000.

[36] M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. A model-theoretic coreference scoring scheme. In *MUC6 '95: Proceedings of the 6th conference on Message understanding*, pages 45–52, Morristown, NJ, USA, 1995. Association for Computational Linguistics.

[37] K. Wagstaff. Intelligent Clustering with Instance-Level Constraints, 2002.

[38] B. Wellner, A. McCallum, F. Peng, and M. Hay. An integrated, conditional model of information extraction and coreference with application to citation matching. In *AUAI '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 593–601, Arlington, Virginia, United States, 2004. AUAI Press.

[39] X. Yang, G. Zhou, J. Su, and C. L. Tan. Coreference resolution using competition learning approach. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 176–183, Morristown, NJ, USA, 2003. Association for Computational Linguistics.