Adaptive Clustering for Coreference Resolution with Deterministic Rules and Web-Based Language Models

Razvan C. Bunescu
School of EECS
Ohio University
Athens, OH 45701, USA
bunescu@ohio.edu

Abstract

We present a novel adaptive clustering model for coreference resolution in which the expert rules of a state of the art deterministic system are used as features over pairs of clusters. A significant advantage of the new approach is that the expert rules can be easily augmented with new semantic features. We demonstrate this advantage by incorporating semantic compatibility features for neutral pronouns computed from web n-gram statistics. Experimental results show that the combination of the new features with the expert rules in the adaptive clustering approach results in an overall performance improvement, and over 5% improvement in F_1 measure for the target pronouns when evaluated on the ACE 2004 newswire corpus.

1 Introduction

Coreference resolution is the task of clustering a sequence of textual entity mentions into a set of maximal non-overlapping clusters, such that mentions in a cluster refer to the same discourse entity. Coreference resolution is an important subtask in a wide array of natural language processing problems, among them information extraction, question answering, and machine translation. The availability of corpora annotated with coreference relations has led to the development of a diverse set of supervised learning approaches for coreference. While learning models enjoy a largely undisputed role in many NLP applications, deterministic models based on rich sets of expert rules for coreference have been shown recently to achieve performance rivaling, if not exceeding, the performance of state of the art machine learning approaches (Haghighi and Klein, 2009; Raghunathan et al., 2010). In particular, the top performing system in the CoNLL 2011 shared task (Pradhan et al., 2011) is a multi-pass system that applies tiers of deterministic coreference sieves from highest to lowest precision (Lee et al., 2011). The PRECISECONSTRUCTS sieve, for example, creates coreference links between mentions that are found to match patterns of apposition, predicate nominatives, acronyms, demonyms, or relative pronouns. This is a high precision sieve, correspondingly it is among the first sieves to be applied. The PRONOUNMATCH sieve links an anaphoric pronoun with the first antecedent mention that agrees in number and gender with the pronoun, based on an ordering of the antecedents that uses syntactic rules to model discourse salience. This is the last sieve to be applied, due to its lower overall precision, as estimated on development data. While very successful, this deterministic multi-pass sieve approach to coreference can nevertheless be quite unwieldy when one seeks to integrate new sources of knowledge in order to improve the resolution performance. Pronoun resolution, for example, was shown by Yang et al. (2005) to benefit from semantic compatibility information extracted from search engine statistics. The semantic compatibility between candidate antecedents and the pronoun context induces a new ordering between the antecedents. One possibility for using compatibility scores in the deterministic system is to ignore the salience-based ordering and replace it with the new compatibility-based ordering. The draw-

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 11–19,
Montréal, Canada, June 7-8, 2012.

back of this simple approach is that now discourse salience, an important signal in pronoun resolution, is completely ignored. Ideally, we would want to use both discourse salience and semantic compatibility when ranking the candidate antecedents of the pronoun, something that can be achieved naturally in a discriminative learning approach that uses the two rankings as different, but overlapping, features. Consequently, we propose an adaptive clustering model for coreference in which the expert rules are successfully supplemented by semantic compatibility features obtained from limited history web n-gram statistics.

2 A Coreference Resolution Algorithm

From a machine learning perspective, the deterministic system of Lee et al. (2011) represents a trove of coreference resolution features. Since the deterministic sieves use not only information about a pair of mentions, but also the clusters to which they have been assigned so far, a learning model that utilized the sieves as features would need to be able to work with features defined on pairs of clusters. We therefore chose to model coreference resolution as the greedy clustering process shown in Algorithm 1. The algorithm starts by initializing the clustering C with a set of singleton clusters. Then, as long as the clustering contains more than one cluster, it repeatedly finds the highest scoring pair of clusters $\langle C_i, C_j \rangle$. If the score passes the threshold $\tau = f(\emptyset, \emptyset)$, the clusters C_i, C_j are joined into one cluster and the process continues with another highest scoring pair of clusters.

Algorithm 1 CLUSTER(X,f)

Input: A set of mentions $X = \{x_1, x_2, ..., x_n\}$;
A measure $f(C_i, C_j) = \mathbf{w}^T \Phi(C_i, C_j)$.
Output: A greedy agglomerative clustering of X.

for $i = 1$ **to** n **do**
  $C_i \leftarrow \{x_i\}$
$C \leftarrow \{C_i\}_{1 \leq i \leq n}$
$\langle C_i, C_j \rangle \leftarrow \underset{p \in \mathcal{P}(C)}{\operatorname{argmax}} f(p)$
while $|C| > 1$ **and** $f(C_i, C_j) > \tau$ **do**
  replace C_i, C_j in C with $C_i \cup C_j$
  $\langle C_i, C_j \rangle \leftarrow \underset{p \in \mathcal{P}(C)}{\operatorname{argmax}} f(p)$
return C

The scoring function $f(C_i, C_j)$ is a linearly weighted combination of features $\Phi(C_i, C_j)$ extracted from the cluster pair, parametrized by a weight vector $\mathbf{w}$. The function $\mathcal{P}$ takes a clustering C as argument and returns a set of cluster pairs $\langle C_i, C_j \rangle$ as follows:

$$\mathcal{P}(C) = \{\langle C_i, C_j \rangle \mid C_i, C_j \in C,\ C_i \neq C_j\} \cup \{\langle \emptyset, \emptyset \rangle\}$$

$\mathcal{P}(C)$ contains a special cluster pair $\langle \emptyset, \emptyset \rangle$, where $\Phi(\emptyset, \emptyset)$ is defined to contain a binary feature uniquely associated with this empty pair. Its corresponding weight is learned together with all other weights and will effectively function as a clustering threshold $\tau = f(\emptyset, \emptyset)$.

Algorithm 2 TRAIN($\mathcal{C}$,T)

Input: A dataset of training clusterings $\mathcal{C}$;
The number of training epochs T.
Output: The averaged parameters $\overline{\mathbf{w}}$.

$\mathbf{w} \leftarrow \mathbf{0}$
for $t = 1$ **to** T **do**
  for all $C \in \mathcal{C}$ **do**
    $\mathbf{w} \leftarrow$ UPDATE(C,$\mathbf{w}$)
return $\overline{\mathbf{w}}$

Algorithm 3 UPDATE(C,$\mathbf{w}$)

Input: A gold clustering $C = \{C_1, C_2, ..., C_m\}$;
The current parameters $\mathbf{w}$.
Output: The updated parameters $\mathbf{w}$.

$X \leftarrow C_1 \cup C_2 \cup ... \cup C_m = \{x_1, x_2, ..., x_n\}$
for $i = 1$ **to** n **do**
  $\hat{C}_i \leftarrow \{x_i\}$
$\hat{C} \leftarrow \{\hat{C}_i\}_{1 \leq i \leq n}$
while $|\hat{C}| > 1$ **do**
  $\langle \hat{C}_i, \hat{C}_j \rangle = \underset{p \in P(\hat{C})}{\operatorname{argmax}} \mathbf{w}^T \Phi(p)$
  $\mathcal{B} \leftarrow \{\langle \hat{C}_k, \hat{C}_l \rangle \in \mathcal{P}(\hat{C}) \mid g(\hat{C}_k, \hat{C}_l | C) > g(\hat{C}_i, \hat{C}_j | C)\}$
  if $\mathcal{B} \neq \emptyset$ **then**
    $\langle \hat{C}_k, \hat{C}_l \rangle = \underset{p \in \mathcal{B}}{\operatorname{argmax}} \mathbf{w}^T \Phi(p)$
    $\mathbf{w} \leftarrow \mathbf{w} + \Phi(\hat{C}_k, \hat{C}_l) - \Phi(C_i, C_j)$
  if $\langle \hat{C}_i, \hat{C}_j \rangle = \langle \emptyset, \emptyset \rangle$ **then**
    return $\mathbf{w}$
  replace $\hat{C}_i$, $\hat{C}_j$ in $\hat{C}$ with $\hat{C}_i \cup \hat{C}_j$
return $\mathbf{w}$

Algorithms 2 and 3 show an incremental learning model for the weight vector $\mathbf{w}$ that is parametrized with the number of training epochs T and a set of training clusterings C in which each clustering contains the true coreference clusters from one document. Algorithm 2 repeatedly uses all true clusterings to update the current weight vector and instead of the last computed weights it returns an averaged weight vector to control for overfitting, as originally proposed by Freund and Schapire (1999). The core of the learning model is in the update procedure shown in Algorithm 3. Like the greedy clustering of Algorithm 1, it starts with an initial system clustering $\hat{C}$ that contains all singleton clusters. At every step in the iteration (lines 5–13), it joins the highest scoring pair of clusters $\langle \hat{C}_i, \hat{C}_j \rangle$, computed according to the current parameters. The iteration ends when either the empty pair obtains the highest score or everything has been joined into only one cluster. The weight update logic is implemented in lines 7–10: if a more accurate pair $\langle \hat{C}_k, \hat{C}_l \rangle$ can be found, the highest scoring such pair is used in the perceptron update in line 10. If multiple cluster pairs obtain the maximum score in lines 6 and 9, the algorithm selects one of them at random. This is useful especially in the beginning, when the weight vector is zero and consequently all cluster pairs have the same score of 0. We define the goodness $g(\hat{C}_k, \hat{C}_l | C)$ of a proposed pair $\langle \hat{C}_k, \hat{C}_l \rangle$ with respect to the true clustering C as the accuracy of the coreference pairs that would be created if $\hat{C}_k$ and $\hat{C}_l$ were joined:

$$g(\cdot) = \frac{\left| \{(x,y) \in \hat{C}_k \times \hat{C}_l \mid \exists C_i \in C : x, y \in C_i\} \right|}{|\hat{C}_k| \cdot |\hat{C}_l|} \quad (1)$$

It can be shown that this definition of the goodness function selects a cluster pair (lines 7–9) that, when joined, results in a clustering with a better pairwise accuracy. Therefore, the algorithm can be seen as trying to fit the training data by searching for parameters that greedily maximize the clustering accuracy, while overfitting is kept under control by computing an averaged version of the parameters. We have chosen to use a perceptron update for simplicity, but the algorithm can be easily instantiated to accommodate other types of incremental updates, e.g. MIRA (Crammer and Singer, 2003).

3 Expert Rules as Features

With the exception of mention detection which is run separately, all the remaining 12 sieves mentioned in (Lee et al., 2011) are used as Boolean features defined on cluster pairs, i.e. if any of the mention pairs in the cluster pair $\langle \hat{C}_i, \hat{C}_j \rangle$ were linked by sieve k, then the corresponding sieve feature $\Phi_k(\hat{C}_i, \hat{C}_j) = 1$. We used the implementation from the Stanford CoreNLP package[1] for all sieves, with a modification for the PRONOUNMATCH sieve which was split into 3 different sieves as follows:

- ITPRONOUNMATCH: this sieve finds antecedents only for neutral pronouns *it*.
- ITSPRONOUNMATCH: this sieve finds antecedents only for neutral possessive pronouns *its*.
- OTHERPRONOUNMATCH: this is a catch-all sieve for the remaining pronouns.

This 3-way split was performed in order to enable the combination of the discourse salience features captured by the pronoun sieves with the semantic compatibility features for neutral pronouns that will be introduced in the next section. The OTHERPRONOUNMATCH sieve works exactly as the original PRONOUNMATCH: for a given non-neutral pronoun, it searches in the current sentence and the previous 3 sentences for the first mention that agrees in gender and number with the pronoun. The candidate antecedents for the pronoun are ordered based on a notion of discourse salience that favors syntactic salience and document proximity (Raghunathan et al., 2010).

4 Discourse Salience Features

The IT/SPRONOUNMATCH sieves use the same implementation for finding the first matching candidate antecedent as the original PRONOUNMATCH. However, unlike OTHERPRONOUNMATCH and the other sieves that generate Boolean features, the neutral pronoun sieves are used to generate real valued features. If the neutral pronoun is the leftmost mention in the cluster $\hat{C}_j$ from a cluster pair $\langle \hat{C}_i, \hat{C}_j \rangle$, the corresponding normalized feature is computed as follows:

[1] http://nlp.stanford.edu/software/corenlp.shtml

1. Let $S_j = \langle S_j^1, S_j^2, ..., S_j^n \rangle$ be the sequence of candidate mentions that precede the neutral pronoun and agree in gender and number with it, ordered from most salient to least salient.

2. Let $A_i \subseteq \hat{C}_i$ be the set of mentions in the cluster $\hat{C}_i$ that appear before the pronoun and agree with it.

3. For each mention $m \in A_i$, find its rank in the sequence S_j:

$$rank(m, S_j) = k \Leftrightarrow m = S_j^k \quad (2)$$

4. Find the minimum rank across all the mentions in A_i and compute the feature as follows:

$$\Phi_{it/s}(\hat{C}_i, \hat{C}_j) = \left(\min_{m \in A_i} rank(m, S_j) \right)^{-1} \quad (3)$$

If A_i is empty, set $\Phi_{it/s}(\hat{C}_i, \hat{C}_j) = 0$.

The discourse salience feature described above is by definition normalized in the interval $[0, 1]$. It takes the maximum value of 1 when the most salient mention in the discourse at the current position agrees with the pronoun and also belongs to the candidate cluster. The feature is 0 when the candidate cluster does not contain any mention that agrees in gender and number with the pronoun.

5 Semantic Compatibility Features

Each of the two types of neutral pronouns is associated with a new feature that computes the semantic compatibility between the syntactic head of a candidate antecedent and the context of the neutral pronoun. If the neutral pronoun is the leftmost mention in the cluster $\hat{C}_j$ from a cluster pair $\langle \hat{C}_i, \hat{C}_j \rangle$ and c_j is the pronoun context, then the new normalized features $\Psi_{it/s}(\hat{C}_i, \hat{C}_j)$ are computed as follows:

1. Compute the maximum semantic similarity between the pronoun context and any mention in $\hat{C}_i$ that precedes the pronoun and is in agreement with it:

$$M_j = \max_{m \in A_i} comp(m, c_j)$$

2. Compute the maximum and minimum semantic similarity between the pronoun context and any mention that precedes the pronoun and is in agreement with it:

$$M_{all} = \max_{m \in S_j} comp(m, c_j)$$
$$m_{all} = \min_{m \in S_j} comp(m, c_j)$$

3. Compute the semantic compatibility feature as follows:

$$\Psi_{it/s}(\hat{C}_i, \hat{C}_j) = \frac{M_j - m_{all}}{M_{all} - m_{all}} \quad (4)$$

To avoid numerical instability, if the overall maximum and minimum similarities are very close ($M_{all} - m_{all} < 1e-4$) we set $\Psi_{it/s}(\hat{C}_i, \hat{C}_j) = 1$.

Like the salience feature $\Phi_{it/s}$, the semantic compatibility feature $\Psi_{it/s}$ is normalized in the interval $[0, 1]$. Its definition assumes that we can compute $comp(m, c_j)$, the semantic compatibility between a candidate antecedent mention m and the pronoun context c_j. For the possessive pronoun its, we extract the syntactic head h of the mention m and replace the pronoun with the mention head h in the possessive context. We use the resulting possessive pronoun context $pc_j(h)$ to define the semantic compatibility as the following conditional probability:

$$\begin{aligned} comp(m, c_j) &= \log P(pc_j(h)|h) \quad (5) \\ &= \log P(pc_j(h)) - \log P(h) \end{aligned}$$

To compute the n-gram probabilities $P(pc_j(h))$ and $P(h)$ in Equation 6, we use the language models provided by the Microsoft Web N-Gram Corpus (Wang et al., 2010), as described in the next section.

Figure 1 shows an example of a possessive neutral pronoun context, together with the set of candidate antecedents that agree in number and gender with the pronoun, from the current and previous 3 sentences. Each candidate antecedent is given an index that reflects its ranking in the discourse salience based ordering. We see that discourse salience does not help here, as the most salient mention is not the correct antecedent. The figure also shows the

In 1946, the nine justices dismissed a *case*$_{[7]}$ involving the *apportionment*$_{[8]}$ of congressional districts. That *view*$_{[6]}$ would slowly change. In 1962, the *court*$_{[3]}$ abandoned *its*$_{[5]}$ *caution*$_{[4]}$. Finding remedies to the unequal *distribution*$_{[1]}$ of political *power*$_{[2]}$ was indeed within ***its*** constitutional authority.

[3] P(*court's constitutional authority* | *court*)
$\approx exp(-5.91)$
[5] P(*court's constitutional authority* | *court*) (*)
$\approx exp(-5.91)$
[7] P(*case's constitutional authority* | *case*)
$\approx exp(-8.32)$
[2] P(*power's constitutional authority* | *power*)
$\approx exp(-9.30)$
[8] P(*app-nt's constitutional authority* | *app-nt*)
$\approx exp(-9.32)$
[4] P(*caution's constitutional authority* | *caution*)
$\approx exp(-9.39)$
[1] P(*dist-ion's constitutional authority* | *dist-ion*)
$\approx exp(-9.40)$
[6] P(*view's constitutional authority* | *view*)
$\approx exp(-9.69)$

Figure 1: Possessive neutral pronoun example.

The *letter*$_{[5]}$ appears to be an *attempt*$_{[6]}$ to calm the concerns of the current American *administration*$_{[7]}$. "I confirm my *commitment*$_{[1]}$ to the points made therein," Aristide said in the *letter*$_{[2]}$, "confident that they will help strengthen the ties between our two nations where *democracy*$_{[3]}$ and *peace*$_{[4]}$ will flourish." Since 1994, when ***it*** sent 20,000 troops to restore Aristide to power, the administration ...

[7] P(*administration sent troops* | *administration*)
$\approx exp(-6.00)$
[2] P(*letter sent troops* | *letter*)
$\approx exp(-6.57)$
[5] P(*letter sent troops* | *letter*)
$\approx exp(-6.57)$
[4] P(*peace sent troops* | *peace*)
$\approx exp(-7.92)$
[6] P(*attempt sent troops* | *attempt*)
$\approx exp(-8.26)$
[3] P(*democracy sent troops* | *democracy*)
$\approx exp(-8.30)$
[1] P(*commitment sent troops* | *commitment*)
$\approx exp(-8.62)$

Figure 2: Neutral pronoun example.

compatibility score computed for each candidate antecedent, using the formula described above. In this example, when ranking the candidate antecedents based on their compatibility scores, the top ranked mention is the correct antecedent, whereas the most salient mention is down in the list.

When the set of candidate mentions contains pronouns, we require that they are resolved to a nominal or named mention, and use the head of this mention to instantiate the possessive context. This is the case of the pronominal mention [5] in Figure 1, which we assumed was already resolved to the noun *court* (even if the pronoun [5] were resolved to an incorrect mention, the noun *court* would still be ranked first due to mention [3]). This partial ordering between coreference decisions is satisfied automatically by setting the semantic compatibility feature $\Psi_{it/s}(\hat{C}_i, \hat{C}_j) = 0$ whenever the antecedent cluster $\hat{C}_i$ contains only pronouns.

A similar feature is introduced for all neutral pronouns *it* appearing in subject-verb-object triples. The new pronoun context $pc_j(h)$ is obtained by replacing the pronoun *it* in the subject-verb-object context c_j with the head h of the candidate antecedent mention. Figure 2 shows a neutral pronoun context, together with the set of candidate antecedents that agree in number and gender with the pronoun, from an abridged version of the original current and previous 3 sentences. Each candidate antecedent is given an index that reflects its ranking in the discourse salience based ordering. Discourse salience does not help here, as the most salient mention is not the correct antecedent. The figure shows the compatibility score computed for each candidate antecedent, using Equation 6. In this example, the top ranked mention in the compatibility based ordering is the correct antecedent, whereas the most most salient mention is at the bottom of the list.

To summarize, in the last two sections we described two special features for neutral pronouns: the discourse salience feature $\Phi_{it/s}$ and the semantic compatibility feature $\Psi_{it/s}$. The two real-valued

Candidate mentions	Original context	N-gram context
capital, store, GE, side, offer	with ***its*** corporate tentacles reaching	GE's corporate tentacles
AOL, Microsoft, Yahoo, product	***its*** substantial customer base	AOL's customer base
regime, Serbia, state, EU, embargo	meets ***its*** international obligations	Serbia's international obligations
company, secret, internet, FBI	***it*** was investigating the incident	FBI was investigating the incident
goal, team, realm, NHL, victory	something ***it*** has not experienced since	NHL has experienced
Onvia, line, Nasdaq, rating	said Tuesday ***it*** will cut jobs	Onvia will cut jobs
coalition, government, Italy	but ***it*** has had more direct exposure	Italy has had direct exposure
Pinochet, arrest, Chile, court	while ***it*** studied a judge 's explanation	court studied the explanation

Table 1: N-gram generation examples.

features are computed at the level of cluster pairs as described in Equations 3 and 4. Their computation relies on the mention level rank (Equation 2) and semantic compatibility (Equation 6) respectively.

6 Web-based Language Models

We used the Microsoft Web N-Gram Corpus[2] to compute the pronoun context probability $P(pc_j(h))$ and the candidate head probability $P(h)$. This corpus provides smoothed back-off language models that are computed dynamically from N-gram statistics using the CALM algorithm (Wang and Li, 2009). The N-grams are collected from the tokenized versions of the billions of web pages indexed by the Bing search engine. Separate models have been created for the document body, the document title and the anchor text. In our experiments, we used the April 2010 version of the document body language models. The number of words in the pronoun context and the antecedent head determine the order of the language models used for estimating the conditional probabilities. For example, to estimate $P(\textit{administration sent troops} \mid \textit{administration})$, we used a trigram model for the context probability $P(\textit{administration sent troops})$ and a unigram model for the head probability $P(\textit{administration})$. Since the maximum order of the N-grams available in the Microsoft corpus is 5, we designed the context and head extraction rules to return N-grams with size at most 5. Table 1 shows a number of examples of N-grams generated from the original contexts, in which the pronoun was replaced with the correct antecedent. To get a sense of the utility of each context in matching the right antecedent, the table also shows a sample of candidate antecedents.

For possessive contexts, the N-gram extraction rules use the head of the NP context and its closest premodifier whenever available. Using the premodifier was meant to increase the discriminative power of the context. For the subject-verb-object N-grams, we used the verb at the same tense as in the original context, which made it necessary to also include the auxiliary verbs, as shown in lines 4–7 in the table. Furthermore, in order to keep the generated N-grams within the maximum size of 5, we did not include modifiers for the subject or object nouns, as illustrated in the last line of the table. Some of the examples in the table also illustrate the limits of the context-based semantic compatibility feature. In the second example, all three company names are equally good matches for the possessive context. In these situations, we expect the discourse salience feature to provide the additional information necessary for extracting the correct antecedent. This combination of discourse salience with semantic compatibility features is done in the adaptive clustering algorithm introduced in Section 2.

7 Experimental Results

We compare our adaptive clustering (AC) approach with the state of the art deterministic sieves (DT) system of Lee et al. (2011) on the newswire portion of the ACE-2004 dataset. The newswire section of the corpus contains 128 documents annotated with gold mentions and coreference information, where coreference is marked only between mentions that belong to one of seven semantic classes: person, organization, location, geo-political entity, facility, vehicle, and weapon. This set of documents has been used before to evaluate coreference resolution sys-

[2]http://web-ngram.research.microsoft.com

System	Mentions	P	R	F_1
DT	Gold, all	88.1	73.3	80.0
AC	Gold, all	**88.7**	**73.5**	**80.4**
DT	Gold, neutral	82.5	51.5	63.4
AC	Gold, neutral	**83.0**	**52.1**	**64.0**
DT	Auto, neutral	84.4	34.9	49.3
AC	Auto, neutral	**86.1**	**40.0**	**54.6**

Table 2: B^3 comparative results on ACE 2004.

tems in (Poon and Domingos, 2008; Haghighi and Klein, 2009; Raghunathan et al., 2010), with the best results so far obtained by the deterministic sieve system of Lee at al. (2011). There are 11,398 annotated gold mentions, out of which 135 are possessive neutral pronouns *its* and 88 are neutral pronouns *it* in a subject-verb-object triple. Given the very small number of neutral pronouns, in order to obtain reliable estimates for the model parameters we tested the adaptive clustering algorithm in a 16 fold cross-validation scenario. Thus, the set of 128 documents was split into 16 folds, where each fold contains 120 documents for training and 8 documents for testing. The final results were pooled together from the 16 disjoint test sets. During training, the AC's update procedure was run for 10 epochs. Since the AC algorithm does not need to tune any hyper parameters, there was no need for development data.

Table 2 shows the results obtained by the two systems on the newswire corpus under three evaluation scenarios. We use the B^3 version of the precision (P), recall (R), and F_1 measure, computed either on all mention pairs (all) or only on links that contain at least one neutral pronoun (neutral) marked as a mention in ACE. Furthermore, we report results on gold mentions (Gold) as well as on mentions extracted automatically (Auto). Since the number of neutral pronouns marked as gold mentions is small compared to the total number of mentions, the impact on the overall performance shown in the first two rows is small. However, when looking at coreference links that contain at least one neutral pronoun, the improvement becomes substantial. AC increases F_1 with 5.3% when the mentions are extracted automatically during testing, a setting that reflects a more realistic use of the system. We have also evaluated the AC approach in the Gold setting using only the original DT sieves as features, obtaining an F_1 of 80.3% for all mentions and 63.4% – same as DT – for neutral pronouns.

By matching the performance of the DT system in the first two rows of the table, the AC system proves that it can successfully learn the relative importance of the deterministic sieves, which in (Raghunathan et al., 2010) and (Lee et al., 2011) have been manually ordered using a separate development dataset. Furthermore, in the DT system the sieves are applied on mentions in their textual order, whereas the adaptive clustering algorithm AC does not assume a predefined ordering among coreference resolution decisions. Thus, the algorithm has the capability to make the first clustering decisions in any section of the document in which the coreference decisions are potentially easier to make. We have run experiments in which the AC system was augmented with a feature that computed the normalized distance between a cluster and the beginning of the document, but this did not lead to an improvement in the results, lending further credence to the hypothesis that a strictly left to right ordering of the coreference decisions is not necessary, at least with the current features.

The same behavior, albeit with smaller increases in performance, was observed when the DT and AC approaches were compared on the newswire section of the development dataset used in the CoNLL 2011 shared task (Pradhan et al., 2011). For these experiments, the AC system was trained on all 128 documents from the newswire portion of ACE 2004. On gold mentions, the DT and AC systems obtained a very similar performance. When evaluated only on links that contain at least one neutral pronoun, in a setting where the mentions were automatically detected, the AC approach improved the F_1 measure over the DT system from 58.6% to 59.1%. One reason for the smaller increase in performance in the CoNLL experiments could be given by the different annotation schemes used in the two datasets. Compared to ACE, the CoNLL dataset does not include coreference links for appositives, predicate nominals or relative pronouns. The different annotation schemes may have led to mismatches in the training and test data for the AC system, which was trained on ACE and tested on CoNLL. While we tried to control for these conditions during the evaluation of the AC system, it is conceivable that the differ-

System	Mentions	P	R	F_1
DT	Auto, *its*	86.0	46.9	60.7
AC	Auto, *its*	**91.7**	**47.5**	**62.6**

Table 3: B^3 comparative results on CoNLL 2011.

ences in annotation still had some effect on the performance of the AC approach. Another cause for the smaller increase in performance was that the pronominal contexts were less discriminative in the CoNLL data, especially for the neutral pronoun *it*. When evaluated only on links that contained at least one possessive neutral pronoun *its*, the improvement in F_1 increased at 1.9%, as shown in Table 3.

8 Related Work

Closest to our clustering approach from Section 2 is the error-driven first-order probabilistic model of Culotta et al. (2007). Among significant differences we mention that our model is non-probabilistic, simpler and easier to understand and implement. Furthermore, the update step does not stop after the first clustering error, instead the algorithm learns and uses a clustering threshold τ to determine when to stop during training and testing. This required the design of a method to order cluster pairs in which the clusters may not be consistent with the true coreference chains, which led to the introduction of the goodness function in Equation 1 as a new scoring measure for cluster pairs. The strategy of continuing the clustering during training as long as a an adaptive threshold is met better matches the training with the testing, and was observed to lead to better performance. The cluster ranking model of Rahman and Ng (2009) proceeds in a left-to-right fashion and adds the current discourse old mention to the highest scoring preceding cluster. Compared to it, our adaptive clustering approach is less constrained: it uses only a weak, partial ordering between coreference decisions, and does not require a singleton cluster at every clustering step. This allows clustering to start in any section of the document where coreference decisions are easier to make, and thus create accurate clusters earlier in the process.

The use of semantic knowledge for coreference resolution has been studied before in a number of works, among them (Ponzetto and Strube, 2006), (Bengtson and Roth, 2008), (Lee et al., 2011), and (Rahman and Ng, 2011). The focus in these studies has been on the semantic similarity between a mention and a candidate antecedent, or the parallelism between the semantic role structures in which the two appear. One of the earliest methods for using predicate-argument frequencies in pronoun resolution is that of Dagan and Itai (1990). Closer to our use of semantic compatibility features for pronouns are the approaches of Kehler et al. (2004) and Yang et al. (2005). The last work showed that pronoun resolution can be improved by incorporating semantic compatibility features derived from search engine statistics in the twin-candidate model. In our approach, we use web-based language models to compute semantic compatibility features for neutral pronouns and show that they can improve performance over a state-of-the-art coreference resolution system. The use of language models instead of search engine statistics is more practical, as they eliminate the latency involved in using search engine queries. Web-based language models can be built on readily available web N-gram corpora, such as Google's Web 1T 5-gram Corpus (Brants and Franz, 2006).

9 Conclusion

We described a novel adaptive clustering method for coreference resolution and showed that it can not only learn the relative importance of the original expert rules of Lee et al. (2011), but also extend them effectively with new semantic compatibility features. Experimental results show that the new method improves the performance of the state of the art deterministic system and obtains a substantial improvement for neutral pronouns when the mentions are extracted automatically.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful suggestions. This work was supported by grant IIS-1018590 from the NSF. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF.

References

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, Honolulu, Hawaii, October. Association for Computational Linguistics.

Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram version 1.

Koby Crammer and Yoram Singer. 2003. Ultraconservative online algorithms for multiclass problems. *J. Mach. Learn. Res.*, 3:951–991.

Aron Culotta, Michael Wick, and Andrew McCallum. 2007. First-order probabilistic models for coreference resolution. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 81–88, Rochester, New York, April. Association for Computational Linguistics.

Ido Dagan and Alon Itai. 1990. Automatic processing of large corpora for the resolution of anaphora references. In *Proceedings of the 13th conference on Computational linguistics - Volume 3*, COLING'90, pages 330–332.

Yoav Freund and Robert E. Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine Learning*, 37:277–296.

Aria Haghighi and Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1152–1161, Singapore, August.

Andrew Kehler, Douglas Appelt, Lara Taylor, and Aleksandr Simma. 2004. The (non)utility of predicate-argument frequencies for pronoun interpretation. In *HLT-NAACL 2004: Main Proceedings*, pages 289–296, Boston, Massachusetts, USA. 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.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 192–199.

Hoifung Poon and Pedro Domingos. 2008. Joint unsupervised coreference resolution with markov logic. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii, October.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. Conll-2011 shared task: modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nate Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher D. Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP'10)*, pages 492–501.

Altaf Rahman and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 968–977, Singapore, August. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2011. Coreference resolution with world knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 814–824, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kuansan Wang and Xiaolong Li. 2009. Efficacy of a constantly adaptive language modeling technique for web-scale applications. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '09, pages 4733–4736, Washington, DC, USA. IEEE Computer Society.

Kuansan Wang, Christopher Thrasher, Evelyne Viegas, Xiaolong Li, and Bo-june (Paul) Hsu. 2010. An overview of microsoft web n-gram corpus and applications. In *Proceedings of the NAACL HLT 2010 Demonstration Session*, HLT-DEMO '10, pages 45–48, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiaofeng Yang, Jian Su, and Chew Lim Tan. 2005. Improving pronoun resolution using statistics-based semantic compatibility information. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 165–172.