Procesamiento del Lenguaje Natural, Revista nº 42, marzo de 2009, pp. 87-96 recibido 15-01-09, aceptado 02-03-09

Global joint models for coreference resolution and named entity classification

Modelos juntos globales para la resolución de la correferencia y de la clasificación de las entidades nombradas

Pascal Denis
Alpage Project-Team
INRIA and Université Paris 7
30, rue Château des Rentiers
75013 Paris, FRANCE
pascal.denis@inria.fr

Jason Baldridge
Department of Linguistics
University of Texas at Austin
1 University Station B5100
Austin, TX 78712-0198 USA
jbaldrid@mail.utexas.edu

Resumen: En este artículo, combinamos modelos de correferencia, anaforicidad y clasificación de las entidades nombradas, como un problema de inferencia junta global utilizando la Programación Lineal Entera (ILP). Nuestras restricciones garantizan: (i) la coherencia entre las decisiones finales de los tres modelos locales, y (ii) la transitividad de las decisiones de correferencia. Este enfoque proporciona mejoras significativas en el f-score sobre los corpora ACE con las tres métricas de evaluación principales para la correferencia: MUC, B^3, y CEAF. A través de ejemplos, modelos de oráculo y nuestros resultados, se muestra también que es fundamental utilizar estas tres métricas y, en particular, que no se puede confiar únicamente en la métrica MUC.
Palabras clave: Resolución de la correferencia, entidades nombradas, aprendizaje automático, Programación Lineal Entera (ILP)

Abstract: In this paper, we combine models for coreference, anaphoricity and named entity classification as a joint, global inference problem using Integer Linear Programming (ILP). Our constraints ensure: (i) coherence between the final decisions of the three local models, and (ii) transitivity of multiple coreference decisions. This approach provides significant f-score improvements on the ACE datasets for all three main coreference metrics: MUC, B^3, and CEAF. Through examples, oracle models, and our results, we also show that it is fundamental to use all three of these metrics, and in particular, to never rely solely on the MUC metric.
Keywords: Coreference Resolution, Named Entities, Machine Learning, Integer Linear Programming (ILP)

1 Introduction

Coreference resolution involves imposing a partition on a set of mentions in a text; each partition corresponds to some entity in a discourse model. Early machine learning approaches for the task which rely on local, discriminative pairwise classifiers (Soon, Ng, and Lim, 2001; Ng and Cardie, 2002b; Morton, 2000; Kehler et al., 2004) made considerable progress in creating robust coreference systems, but their performance still left much room for improvement. This stems from two main deficiencies:

- **Decision locality**. Decisions are made independently of others; a separate clustering step forms chains from pairwise classifications. But, coreference clearly should be conditioned on properties of an entity as a whole.
- **Knowledge bottlenecks**. Coreference involves many different factors, e.g., morphosyntax, discourse structure and reasoning. Yet most systems rely on small sets of shallow features. Accurately predicting such information and using it to constrain coreference is difficult, so its potential benefits often go unrealized due to error propagation.

More recent work has sought to address these limitations. For example, to address decision locality, McCallum and Wellner (2004) use conditional random fields with

ISSN 1135-5948

model structures in which pairwise decisions influence others. Denis (2007) and Klenner (2007) use integer linear programming (ILP) to perform global inference via transitivity constraints between different coreference decisions.[1] Haghighi and Klein (2007) provide a fully generative model that combines global properties of entities across documents with local attentional states. Denis and Baldridge (2008) use a ranker to compare antecedents for an anaphor simultaneously rather than in the standard pairwise manner. To address the knowledge bottleneck problem, Denis and Baldridge (2007) use ILP for joint inference using a pairwise coreference model and a model for determining the anaphoricity of mentions. Also, Denis and Baldridge (2008) and Bengston and Roth (2008) use models and features, respectively, that attend to particular types of mentions (e.g., full noun phrases versus pronouns). Furthermore, Bengston and Roth (2008) use a wider range of features than are normally considered, and in particular use predicted features for later classifiers, to considerably boost performance.

In this paper, we use ILP to extend the joint formulation of Denis and Baldridge (2007) using named entity classification and combine it with the transitivity constraints (Denis, 2007; Klenner, 2007). Intuitively, we only should identify antecedents for the mentions which are likely to have one (Ng and Cardie, 2002a), and we should only make a set of mentions coreferent if they are all instances of the *same* entity type (eg, PERSON or LOCATION). ILP enables such constraints to be declared between the outputs of independent classifiers to ensure coherent assignments are made. It also leads to global inference via both constraints on named entity types and transitivity constraints since both relate multiple pairwise decisions.

We show that this strategy leads to improvements across the three main metrics proposed for coreference: the MUC metric (Vilain et al., 1995), the B^3 metric (Bagga and Baldwin, 1998), and CEAF metric (Luo, 2005). In addition, we contextualize the performance of our system with respect to cascades of multiple models and oracle systems that assume perfect information (e.g. about entity types). We furthermore demonstrate the inadequacy of using only the MUC metric and argue that results should *always* be given for all three. We include a simple composite of the three metrics, called MELA, for Mention, Entity, and Link Average score.[2]

2 Data and evaluation

We use the ACE corpus (Phase 2) for training and testing. The corpus has three parts: NPAPER, NWIRE, and BNEWS, and each set is split into a `train` part and a `devtest` part. The corpus text was preprocessed with the OpenNLP Toolkit[3] (i.e., a sentence detector, a tokenizer, and a POS tagger). In our experiments, we consider only *true* ACE mentions instead of detecting them; our focus is on evaluating pairwise local approaches versus the global ILP approach rather than on building a full coreference resolution system.

Three primary metrics have been proposed for evaluating coreference performance: (i) the **link** based MUC metric (Vilain et al., 1995), (ii) the **mention** based B^3 metric (Bagga and Baldwin, 1998), and (iii) the **entity** based CEAF metric (Luo, 2005). All these metrics compare the set of chains $\mathcal{S}$ produced by a system against the true chains $\mathcal{T}$, and report performance in terms of *recall* and *precision*. They however differ in how they computes these scores, and each embeds a different bias.

The MUC metric is the oldest and still most commonly used. MUC operates by determining the number of *links* (i.e., pairs of mentions) that are common to $\mathcal{S}$ and $\mathcal{T}$. Recall is the number of common links divided by the total number of links in the $\mathcal{T}$; precision is the number of common links divided by the total number of links in $\mathcal{S}$. By focusing on the links, this metric has two main biases, which are now well-known (Bagga and Baldwin, 1998; Luo, 2005) but merit re-emphasis due its continued use as the sole evaluation measure. First, it favors systems that create large chains (hence, fewer entities). For instance, a system that produces a single chain achieves 100% recall without severe degradation in precision. Second, it ignores recall for single mention entities, since no link can be found in these; however, putting such mentions in the wrong chain does hurt precision.[4]

[1] These were independent, simultaneous developments.

[2] Interestingly, *mela* means "gathering" in Sanskrit, so this acronym seems appropriate.

[3] Available from `opennlp.sf.net`.

[4] It is worth noting that the MUC corpus for which

$\mathcal{T} = \{m_1, m_3, m_5\}, \{m_2\}, \{m_4, m_6, m_7\}$
$\mathcal{S}_1 = \{m_1, m_2, m_3, m_6\}, \{m_4, m_5, m_7\}$
$\mathcal{S}_2 = \{m_1, m_2, m_3, m_4, m_5, m_6, m_7\}$

Figure 1: Two competiting partitionings for mention set $\{m_1, m_2, m_3, m_4, m_5, m_6, m_7\}$.

The B^3 metric addresses the MUC metric's shortcomings, by computing recall and precision scores for each mention m. Let S be the system chain containing m, T be the true chain containing m. The set of correct elements in S is thus $|S \cap T|$. The recall score for a mention m is thus computed as $\frac{|S \cap T|}{|T|}$, while the precision score for m is $\frac{|S \cap T|}{|S|}$. Overall recall/precision is obtained by averaging over the individual mention scores. The fact that this metric is mention-based by definition solves the problem of single mention entities. It also does not favor larger chains, since they will be penalized in the precision score of *each* mention.

The Constrained Entity Aligned F-Measure[5] (CEAF) aligns each system chain S with *at most one* true chain T. It finds the best one-to-one mapping between the set of chains $\mathcal{S}$ and $\mathcal{T}$, which is equivalent to finding the optimal alignment in a bipartite graph. The best mapping is that which maximizes the similarity over pairs of chains (S_i, T_i), where the similarity of two chains is the number of common mentions between them. For CEAF, recall is the total similarity divided by the number of mentions in all the $\mathcal{T}$, while precision is the total similarity divided by the number of mentions in $\mathcal{S}$. Note that when true mentions are used, CEAF assigns the same recall and precision: this is because the two systems partition the same set of mentions.

A simple example illustrating how the metrics operate is presented in Figure 1 (see Luo (2005) for more examples). $\mathcal{T}$ is the set of true chains, $\mathcal{S}_1$ and $\mathcal{S}_2$ are the partitions produced by two hypothetical resolvers. Recall, precision, and f-score for these metrics are given in Table 1.

	MUC			B^3			CEAF
	R	P	F	R	P	F	F
$\mathcal{S}_1$	.50	.40	.44	.62	.45	.52	.57
$\mathcal{S}_2$	1.0	.66	.79	1.0	.39	.56	.43

Table 1: Recall (R), precision (P), and f-score (F) using MUC, B^3, and CEAF for partitionings of Figure 1

The bias of the MUC metric for large chains is shown by the fact that it gives better recall *and* precision scores for $\mathcal{S}_2$ even though this partition is completely uninformative. More intuitively, B^3 highly penalizes the precision of this partition: precision errors are here computed for each mention. CEAF is the harshest on $\mathcal{S}_2$, and in fact is the only metric that prefers $\mathcal{S}_1$ over $\mathcal{S}_2$.

MUC is known for being an applicable metric when one is only interested in precision on pairwise links (Bagga and Baldwin, 1998). Given that much recent work —including the present paper— seeks to move beyond simple pairwise coreference and produce good entities, it is crucial that they are scored on the other metrics as well as MUC. Most tellingly, our results show that both B^3 and CEAF scores can show degradation even when MUC appears to show an improvement.

3 Base models

Here we define the three base classifiers for pairwise coreference, anaphoricity, and named entity classification. They form the basis for several cascades and joint inference with ILP. Like Kehler *et al.* (2004) and Morton (2000), we estimate the parameters of all models using maximum entropy (Berger, Pietra, and Pietra, 1996); specifically, we use the limited memory variable metric algorithm (Malouf, 2002).[6] Gaussian priors for the models were optimized on development data.

3.1 The coreference classifier

Our coreference classifier is based on that of Soon, Ng, and Lim (2001), though the features have been extended and are similar (though not equivalent) to those used by Ng and Cardie (2002a). Features fall into 3 categories: (i) features of the anaphor, (ii) features of antecedent mention, and (iii) pairwise features (i.e., such as distance between

the metric was devised does not annotate single mention entities. However, the ACE corpus *does* include such entities.

[5]We use the *mention-based* CEAF measure (Luo, 2005). This is the same metric as ECM-F (Luo et al., 2004) used by Klenner (2007).

[6]This algorithm is implemented in Toolkit for Advanced Discriminative Modeling (`tadm.sf.net`).

the two mentions). We omit details here for brevity (details on the different feature sets can be found in Denis (2007)); the ILP approach could be equally well applied to models using other, extended feature sets such as those discussed in Denis and Baldridge (2008) and Bengston and Roth (2008).

Using the coreference classifier on its own involves: (i) estimating $P_C(\text{COREF}|\langle i,j\rangle)$, the probability of having a coreferential outcome given a pair of mentions $\langle i,j\rangle$, and (ii) applying a selection algorithm that picks one or more mentions out of the candidates for which $P_C(\text{COREF}|\langle i,j\rangle)$ surpasses a given threshold (here, .5).

$$P_C(\text{COREF}|\langle i,j\rangle) = \frac{exp(\sum_{k=1}^{n} \lambda_k f_k(\langle i,j\rangle, \text{COREF}))}{Z(\langle i,j\rangle)}$$

where $f_k(i,j)$ is the number of times feature k occurs for i and j, λ_k is the weight assigned to feature k during training, and $Z(\langle i,j\rangle)$ is a normalization factor over both outcomes (COREF and $\neg$COREF).

Training instances are constructed based on pairs of mentions of the form $\langle i,j\rangle$, where j and i describe an anaphor and an antecedent candidate, respectively. Each such pair is assigned a label, either COREF or $\neg$COREF, depending on whether or not the two mentions corefer. We followed the sampling method of Soon, Ng, and Lim (2001) for creating the training material for each anaphor: (i) a *positive instance* for the pair $\langle i,j\rangle$ where i is the closest antecedent for j, and (ii) a *negative instance* for each pair $\langle i,k\rangle$ where k intervenes between i and j.

Once trained, the classifier can be used to choose pairwise coreference links–and thus determine the partition of entities–in two ways. The first is to pick a unique antecedent with *closest-first* link-clustering (Soon, Ng, and Lim, 2001); this is the standard strategy, referred to as $\mathsf{COREF}_{closest}$. The second is to simply take all links with probability above .5, which we refer to as $\mathsf{COREF}_{above\_.5}$. The purpose of including this latter strategy is primarily to demonstrate an easy way to improve MUC scores that actually degrades B^3 and CEAF scores. This strategy indeed results in positing significantly larger chains, since each anaphor is allowed to link to several antecedents.

3.2 The anaphoricity classifier

Ng and Cardie (2002a) introduced the use of an anaphoricity classifier to act as a filter for coreference resolution to correct errors where non-anaphoric mentions are mistakenly resolved or where anaphoric mentions failed to be resolved. Their approach produces improvements in precision, but larger losses in recall. Ng (2004) improves recall by optimizing the anaphoricity threshold. By using joint inference for anaphoricity and coreference, Denis and Baldridge (2007) avoid cascade-induced errors without the need to separately optimize the threshold. They realize gains in both recall and precision; however, they report only MUC scores. As we will show, these improvements do not hold for B^3 and CEAF.

The task for the anaphoricity determination component is the following: one wants to decide for each mention i in a document whether i is anaphoric or not. This task can be performed using a simple classifier with two outcomes: ANAPH and $\neg$ANAPH. The classifier estimates the conditional probabilities $P(\text{ANAPH}|i)$ and predicts ANAPH for i when $P(\text{ANAPH}|i) > .5$. The anaphoricity model is as follows:

$$P_A(\text{ANAPH}|i) = \frac{exp(\sum_{k=1}^{n} \lambda_k f_k(i, \text{ANAPH}))}{Z(i)}$$

The features used for the anaphoricity classifier are quite simple. They include information regarding (i) the mention itself, such as the number of words and whether it is a pronoun, and (ii) properties of the potential antecedent set, such as whether there is a previous mention with a matching string. This classifier achieves 80.8% on the ENTIRE ACE corpus (BNEWS: 80.1, NPAPER: 82.2, NWIRE: 80.1).

3.3 The named entity classifier

Named entity classification involves predicting one of the five ACE class labels. The set of named entity types $\mathcal{T}$ are: FACility, GPE (geo-political entity), LOCation, ORGanization, PERson. The classifier estimates the conditional probabilities $P(t|i)$ for each $t \in \mathcal{T}$ and predicts the named entity type $\widehat{t}$ for mention i such that $\widehat{t} = argmax_{t\in\mathcal{T}} P(t|i)$.

$$P_E(t|i) = \frac{exp(\sum_{k=1}^{n} \lambda_k f_k(i,t))}{Z(i)}$$

The features for this model include: (i) the string of the mention, (ii) features defined over the string (e.g., capitalization, punctuations, head word), (iii) features describing the word and POS context around the mention. The classifier achieves 79.5% on the ENTIRE ACE corpus (BNEWS: 79.8, NPAPER: 73.0, NWIRE: 72.7).

4 Base model results

This section describes coreference performance when the pairwise coreference classifier is used alone with closest-first clustering ($\mathsf{COREF}_{closest}$) or with the liberal all-links-above-.5 clustering ($\mathsf{COREF}_{above\_.5}$), or when $\mathsf{COREF}_{closest}$ is constrained by the anaphoricity and named entity classifiers as filters in a cascade or by gold-standard information as filters in oracle systems. The cascades are:

- $\mathsf{CASCADE}_{a \to c}$: the anaphoricity classifier specifies which mentions to resolve
- $\mathsf{CASCADE}_{e \to c}$: the named entity classifier specifies which antecedents have the same type as the mention to be resolved; others are excluded from consideration
- $\mathsf{CASCADE}_{a,e \to c}$: the two classifiers acting as combined filters

We also provide results for the corresponding oracle systems which have perfect knowledge about anaphoricity and/or named entity types: $\mathsf{ORACLE}_{a,c}$, $\mathsf{ORACLE}_{e,c}$, and $\mathsf{ORACLE}_{a,e,c}$.

Table 2 summarizes the results in terms of recall (R), precision (P), and f-score (F) on the three coreference metrics: MUC, B^3, and CEAF. The first thing to note is the contrast between $\mathsf{COREF}_{closest}$ and $\mathsf{COREF}_{above\_.5}$. Recall that the only difference between the two clustering strategies is that the latter creates strictly larger entities than the former by adding all links above .5. By doing so, it gains about 10% in R for both MUC and B^3. However, whereas MUC does not register a drop in precision, B^3 P is 14% lower, which produces an overall 1% drop in F. CEAF punishes this strategy even more, with a 3.6% drop. Note that the resulting composite MELA scores are almost identical. Given the nature of the two strategies $\mathsf{COREF}_{closest}$ and $\mathsf{COREF}_{above\_.5}$, these differences across metrics strongly support arguments that MUC is too indiscriminate and can in fact be gamed (knowingly or not) by simply creating larger chains.

Table 2 also shows that cascades in general fail to produce significant F improvements over the pairwise model $\mathsf{COREF}_{closest}$. These systems are far behind the performance of their corresponding oracles. This tendency is even stronger when both classifiers filter possible assignments: $\mathsf{CASCADE}_{a,e \to c}$ does much worse than $\mathsf{COREF}_{closest}$ on all metrics. In fact, this system has the lowest F on the B^3 evaluation metric, suggesting that the errors of the two filters accumulate in this case. In contrast, the corresponding oracle, $\mathsf{ORACLE}_{a,e,c}$, achieves the best results across all measures. It does so by capitalizing on the improvements given by the separate oracles.

Furthermore, note that the use of the two auxiliary models have complementary effects on the MUC and B^3 metrics, in both the cascade and the oracle systems. Thus, the use of the anaphoricity classifier improves recall (suggesting that some true anaphors get "rescued" by this model), while the the use of the named entity model leads to precision improvements (suggesting that this model manages to filter out incorrect candidates that would have been chosen by the coreference model). In the case of the oracle systems, these gains translate in overall F improvements. But, as noted, this is generally not the case with the cascade systems. Only $\mathsf{CASCADE}_{a \to c}$ shows significant gains with MUC and CEAF (and not with B^3). $\mathsf{CASCADE}_{e \to c}$ underperforms in all three metrics. This latter system indeed shows a large drop in recall, suggesting that this model filter is overzealous in filtering true antecedents.

The oracle results suggest that joint modeling could deliver large performance gains by not falling prey to cascade errors. In the next section, we build on previous ILP formulations and show such improvements can indeed be realized.

5 Integer programming formulations

ILP is an optimization framework for global inference over the outputs of various base classifiers (Roth and Yih, 2004). Previous uses of ILP for NLP tasks include eg. Roth

System	MUC			B^3			CEAF	MELA
	R	P	F	R	P	F	R/P/F	F-avg
$\mathsf{COREF}_{closest}$	60.8	72.6	66.2	62.4	77.7	69.2	62.3	65.9
$\mathsf{COREF}_{above\_.5}$	70.3	72.7	71.5	73.2	63.7	68.1	58.7	66.1
$\mathsf{CASCADE}_{a\rightarrow c}$	64.9	72.3	68.4	65.6	74.1	69.6	63.4	67.1
$\mathsf{CASCADE}_{e\rightarrow c}$	56.3	75.2	64.4	59.6	82.4	69.2	61.6	65.1
$\mathsf{CASCADE}_{a,e\rightarrow c}$	61.3	68.8	64.8	62.5	73.8	67.7	61.9	64.8
$\mathsf{ORACLE}_{a,c}$	75.6	75.6	75.6	71.4	70.7	71.1	71.5	72.7
$\mathsf{ORACLE}_{e,c}$	62.5	81.3	70.7	62.9	85.5	72.4	65.2	69.4
$\mathsf{ORACLE}_{a,e,c}$	83.2	83.2	83.2	79.0	78.2	78.6	78.7	80.2

Table 2: Recall (R), precision (P), and f-score (F) using MUC, B^3, and CEAF on the entire ACE corpus for the basic coreference system, the cascade systems, and the corresponding oracle systems.

and Yih (2004), Barzilay and Lapata (2006), and Clarke and Lapata (2006). Here, we provide several ILP formulations for coreference. The first formulation $\mathsf{ILP}_{c,a}$ is based on Denis and Baldridge (2007) and performs joint inference over the coreference classifier and the anaphoricity classifier. A second formulation $\mathsf{ILP}_{c,e}$ combines the coreference classifier with the named entity classifier. A third formulation $\mathsf{ILP}_{c,a,e}$ combines all three models together. In each of these joint formulation, a set of *consistency* constraints mutually constrain the ultimate assignments of each model. Finally, a fourth formulation $\mathsf{ILP}_{c,a,e|trans}$ adds to $\mathsf{ILP}_{c,a,e}$ a set of transitivity constraints (similar to those of Klenner (2007)). These latter constraints ensure better *global* coherence between the various pairwise coreference decisions, hence making this fourth formulation both a joint and a global model.

For solving the ILP problem, we use CPLEX, a commercial LP solver.[7] In practice, each document is processed to define a distinct ILP problem that is then submitted to the solver.

5.1 $\mathsf{ILP}_{c,a}$: anaphoricity-coreference formulation

The $\mathsf{ILP}_{c,a}$ system of Denis and Baldridge (2007) brings the two decisions of coreference and anaphoricity together by including both in a single objective function and enforcing *consistency* constraints on the final outputs of both tasks. More technically, let first $\mathcal{M}$ denotes the set of mentions, and $\mathcal{P}$ the set of possible coreference links over $\mathcal{M}$: $\mathcal{P} = \{\langle i,j\rangle | \langle i,j\rangle \in \mathcal{M}\times\mathcal{M} \text{ and } i<j\}$. Each model introduces a set of indicator variables: (i) coreference variables $\langle i,j\rangle \in 0,1$ depending on whether i and j corefer or not, and (ii) anaphoricity variables $x_{\langle i,j\rangle} \in 0,1$ depending on whether j is anaphoric or not. These variables are associated with assignment costs that are derived from the model probabilities $p_C = P_C(\text{COREF}|i,j)$ and $p_A = P_A(\text{ANAPH}|j)$, respectively. The cost of commiting to a coreference link is $c^C_{\langle i,j\rangle} = -log(p_C)$ and the complement cost of choosing not to establish a link is $\overline{c}^C_{\langle i,j\rangle} = -log(1-p_C)$. Analogously, we define costs on anaphoricity decisions as $c^A_j = -log(p_A)$ and $\overline{c}^A_j = -log(1-p_A)$, the costs associated with making j anaphoric or not, respectively. The resulting objective function takes the following form:

$$min \sum_{\langle i,j\rangle\in\mathcal{P}} c^C_{\langle i,j\rangle}\cdot x_{\langle i,j\rangle} + \overline{c}^C_{\langle i,j\rangle}\cdot(1-x_{\langle i,j\rangle}) + \sum_{j\in\mathcal{M}} c^A_j\cdot y_j + \overline{c}^A_j\cdot(1-y_j)$$

subject to:

$$x_{\langle i,j\rangle}\in\{0,1\} \qquad \forall\langle i,j\rangle\in\mathcal{P}$$
$$y_j\in\{0,1\} \qquad \forall j\in\mathcal{M}$$

The final assignments of $x_{\langle i,j\rangle}$ and y_j variables are forced to respect the following two consistency constraints (where $\mathcal{M}_j$ is the set of all mentions preceding mention j in the document):

Resolve all anaphors: if a mention is anaphoric (y_j=1), it *must* have at least one antecedent.

$$y_j \leq \sum_{i\in\mathcal{M}_j} x_{\langle i,j\rangle} \qquad \forall j\in\mathcal{M}$$

[7] http://www.ilog.com/products/cplex/

Resolve only anaphors: if a pair of mentions $\langle i, j\rangle$ is coreferent ($x_{\langle i,j\rangle}$=1), then j is anaphoric (y_j=1).

$$x_{\langle i,j\rangle} \leq y_j \qquad \forall\langle i, j\rangle \in \mathcal{P}$$

These constraints make sure that the anaphoricity classifier are not taken on faith as they were with $\mathsf{CASCADE}_{a\rightarrow c}$. Instead, we optimize over consideration of both possibilities in the objective function (relative to the probability output by the classifier) while ensuring that the final assignments respect the signifance of what it is to be anaphoric or non-anaphoric.

5.2 $\mathsf{ILP}_{c,e}$: entity-coreference formulation

In this second joint formulation, we combine coreference decisions with named entity classification. New indicator variables for the assignments of this model are introduced, namely $z_{\langle i,j\rangle}$, where $\langle i, t\rangle \in \mathcal{M} \times \mathcal{T}$. Since entity classification is not a binary decision, each assigment variable encode a mention i and a named entity type t. Each of these variables have an associated cost $c^E_{\langle i,t\rangle}$, which is the probability that mention i has type t: $c^E_{\langle i,t\rangle} = -log(P_E(t|i))$. The objective function for this formulation is:

$$min \sum_{\langle i,j\rangle\in\mathcal{P}} c^C_{\langle i,j\rangle} \cdot x_{\langle i,j\rangle} + \overline{c}^C_{\langle i,j\rangle} \cdot (1-x_{\langle i,j\rangle}) + \sum_{\langle i,t\rangle\in\mathcal{M}\times\mathcal{T}} c^E_{\langle i,t\rangle} \cdot z_{\langle i,t\rangle}$$

subject to:

$$z_{\langle i,t\rangle} \in \{0, 1\} \qquad \forall\langle i, t\rangle \in \mathcal{M} \times \mathcal{T}$$
$$\sum_{i\in\mathcal{M}} z_{\langle i,t\rangle} = 1 \qquad \forall i \in \mathcal{M}$$

The last constraint ensures that each mention is only assigned a unique named entity type. Consistency between the two models is ensured with the constraint:

Coreferential mentions have the same entity type: if i and j are coreferential ($x_{\langle i,j\rangle}$=1), they must have the same type ($z_{\langle i,t\rangle} - z_{\langle j,t\rangle} = 0$):

$$1 - x_{\langle i,j\rangle} \geq z_{\langle i,t\rangle} - z_{\langle j,t\rangle} \quad \forall\langle i, j\rangle \in \mathcal{P}, \forall t \in \mathcal{T}$$
$$1 - x_{\langle i,j\rangle} \geq z_{\langle j,t\rangle} - z_{\langle i,t\rangle} \quad \forall\langle i, j\rangle \in \mathcal{P}, \forall t \in \mathcal{T}$$

These constraints above make sure that the coreference decisions (the x values) are informed by the named entity classifier and vice versa. Furthermore, because these constraints ensure like assignments to coreferent pairs of mentions, they have a "propagating" effect that makes the overall system global. Coreference assignments that have low cost (i.e., high confidence) can influence named entity assignments (e.g., from a ORG to a PER). This in turn influences other coreference assignments involving further mentions radiating out from one core, highly likely assignment.

5.3 $\mathsf{ILP}_{c,a,e}$: anaphoricity-entity-coreference formulation

For the third joint model, we combine all three base models with an objective function that is the composite of those of $\mathsf{ILP}_{c,a}$ and $\mathsf{ILP}_{c,e}$ and incorporate all the constraints that go with them. By creating a triple joint model, we get constraints between anaphoricity and named entity classification for free, as a result of the interaction of the consistency constraints between anaphoricity and coreference and of those between named entity and coreference. For example, if a mention of type t is anaphoric, then there must be at least one mention of type t preceding it.

5.4 Adding transitivity constraints

The previous formulations relate coreference decisions to the decisions made by two auxiliary models in a joint formulation. In addition one would also like to make coreference decisions dependent on one another, thus ensuring globally coherent entities. This is achieved through the use transitivity constraints that relate triples of mentions $\langle i, j, k\rangle \in \mathcal{M}\times\mathcal{M}\times\mathcal{M}$, where $i < j < k$ (Denis, 2007; Klenner, 2007). These constraints directly exploit the fact that coreference is an equivalence relation.

Transitivity: if $x_{\langle i,j\rangle}$ and $x_{\langle j,k\rangle}$ are coreferential pairs (i.e., $x_{\langle i,j\rangle} = x_{\langle j,k\rangle} = 1$), then so is $x_{\langle i,k\rangle}$:

$$x_{\langle i,k\rangle} \geq x_{\langle i,j\rangle} + x_{\langle j,k\rangle} - 1 \quad \forall\langle i, j, k\rangle \in M_{i,j,k}$$

Euclideanity: if $x_{\langle i,j\rangle}$ and $x_{\langle i,k\rangle}$ are coreferential pairs (i.e., $x_{\langle i,j\rangle} = x_{\langle i,k\rangle} = 1$), then so is $x_{\langle j,k\rangle}$.

$$x_{\langle j,k\rangle} \geq x_{\langle i,j\rangle} + x_{\langle i,k\rangle} - 1 \quad \forall \langle i,j,k\rangle \in M_{i,j,k}$$

Anti-Euclideanity: if $x_{\langle i,k\rangle}$ and $x_{\langle j,k\rangle}$ are coreferential pairs (i.e., $x_{\langle i,k\rangle} = x_{\langle j,k\rangle} = 1$), then so is $x_{\langle i,j\rangle}$:

$$x_{\langle i,j\rangle} \geq x_{\langle i,k\rangle} + x_{\langle j,k\rangle} - 1 \quad \forall \langle i,j,k\rangle \in M_{i,j,k}$$

Enforcing **Anti-Euclideanity** alone guarantees that the final assignment will not produce any "implicit" anaphors: that is, a configuration wherein $x_{\langle j,k\rangle} = 1$, $x_{\langle i,k\rangle} = 1$, and $y_j = 0$. The interaction of this constraint with **resolve only anaphors** indeed guarantees that such configuration cannot arise, since all three equalities cannot hold together. This means that mention j must be a good match for mention i as well as for mention k.

Note that one could have one unique transitivity constraint if we had symmetry in our model; concretely, capturing symmetry means: (i) adding a new indicator variable $x_{\langle j,i\rangle}$ for each variable $x_{\langle i,j\rangle}$, and (ii) making sure $x_{\langle j,i\rangle}$ agrees with $x_{\langle i,j\rangle}$.

Enforcing each of these constraints above means adding $\frac{1}{6} \times n \times (n-1) \times (n-2)$ constraints, for a document containing n mentions. This means close to $500,000$ of these constraints for a document containing just 100 mentions. The inclusion of such a large set of constraints turned out to be difficult, causing memory issues with large documents (some of the ACE documents have more than 250 mentions). Consequently, we investigated during development various simpler scenarios, such as enforcing these constraints for documents that had a relatively small number of mentions (e.g., 100) or just using one of these types of constraint (in particular **Anti-Euclideanity** given the way it interacts with the discourse status assignments). In the following, $\mathsf{ILP}_{c,a,e|trans}$ will refer to the $\mathsf{ILP}_{c,a,e}$ formulation augmented with the **Anti-Euclideanity** constraints.

6 ILP Results

Table 3 summarizes the scores for the different ILP systems, along with $\mathsf{COREF}_{closest}$. Like Denis and Baldridge (2007), we find that joint anaphoricity and coreference ($\mathsf{ILP}_{c,a}$) greatly improves MUC F. However, we also see that this model suffers from the same problem as $\mathsf{COREF}_{above\_.5}$: performance on the other metrics go down. This is in fact unsurprising: $\mathsf{COREF}_{above\_.5}$ can be viewed as an unconstrained ILP formulation; similarly, $\mathsf{ILP}_{c,a}$ takes all links above .5 subject to meeting the constraints on anaphoricity. The constraining effect of anaphoricity improves MUC R and P and B^3 R over $\mathsf{COREF}_{above\_.5}$, but not B^3 P nor CEAF. Despite the encouraging MUC scores, more is thus needed.

The next thing to note is that joint named entity classification and coreference ($\mathsf{ILP}_{c,e}$) nearly beats $\mathsf{COREF}_{closest}$ across the metrics, but fails for CEAF. As for $\mathsf{ILP}_{c,a}$, $\mathsf{ILP}_{c,e}$ can also be viewed as constraining $\mathsf{COREF}_{above\_.5}$: in this case, precision is improved (compare MUC: 72.7 to 75.0 and B^3: 63.7 to 71.2), while still retaining over half the gain in recall that $\mathsf{COREF}_{above\_.5}$ obtained over $\mathsf{COREF}_{closest}$. In doing so, the degradation in CEAF is just 1%, compared to $\mathsf{ILP}_{c,a}$'s 3.4%. In addition to improving coreference resolution performance, this joint formulation also yields a slight improvement on the named entity classification: specifically, accuracy for that task went from 79.5% to over 80.0% using the $\mathsf{ILP}_{c,e}$ model.

Joint inference over all three models ($\mathsf{ILP}_{c,a,e}$) delivers larger improvements for both MUC and B^3 without any CEAF degradation, thus mirroring the improvements found with the corresponding oracle. In particular, R is boosted nearly to the level of $\mathsf{COREF}_{above\_.5}$ without the dramatic loss in P (in fact P is better than $\mathsf{COREF}_{closest}$ for MUC). By adding the Anti-Euclideanity constraint to this formulation ($\mathsf{ILP}_{c,a,e|trans}$), we see the best across-the-metric scores of any system. For MUC and B^3, *both* P and R are boosted over $\mathsf{COREF}_{closest}$, and there is a jump of 4% for CEAF. Both the MUC and CEAF improvements for $\mathsf{ILP}_{c,a,e|trans}$ are in line with the improvements that Klenner (2007) found using transitivity, though it should be noted that he scored on all mentions, not just true mentions as we do here.

The composite MELA metric provides an interesting overall view, showing step-wise improvements through the addition of the various models and the global constraints.

These results are in sharp contrast with those obtained by the cascade model $\mathsf{CASCADE}_{a,e\rightarrow c}$: recall that this system, while also using the two auxiliary models as filters was worse than $\mathsf{COREF}_{closest}$. The joint ILP formulation is clearly better able to integrate the extra information provided by the anaphoric-

System	MUC			B^3			CEAF	MELA
	R	P	F	R	P	F	R/P/F	F
$\mathsf{COREF}_{closest}$	60.8	72.6	66.2	62.4	77.7	69.2	62.3	65.9
$\mathsf{COREF}_{above\_.5}$	70.3	72.7	71.5	73.2	63.7	68.1	58.7	66.1
$\mathsf{ILP}_{c,a}$	73.2	73.4	73.3	75.3	62.0	68.0	58.9	66.7
$\mathsf{ILP}_{c,e}$	66.2	75.0	70.4	69.6	71.2	70.4	61.2	67.3
$\mathsf{ILP}_{c,a,e}$	69.6	75.4	72.4	72.2	69.7	70.9	62.3	68.5
$\mathsf{ILP}_{c,a,e\|trans}$	63.7	77.8	70.1	65.6	81.4	72.7	66.2	69.7

Table 3: Recall (R), precision (P), and f-score (F) using the MUC, B^3, and CEAF evaluation metric on the entire ACE dataset for the ILP coreference systems.

ity and named entity classifiers. In doing so, it does not require fine-tuning thresholds, and it can further benefit from constraints, such as transitivity.

Further experiments reveal that bringing the other transitivity constraints into the ILP formulation results in additional *precision* gains, although not in overall F gains. The effect of these constraints is to withdraw incoherent links, rather than producing new links. At the global level, this results in the creation of smaller, more coherent clusters of mentions. In some cases, this will lead to a single entity being split across multiple chains. Switching on these constraints may therefore be useful for certain applications where precision is more important than recall.

Though in general CEAF appears to be the most discriminating metric, this point brings up the reason why using CEAF on its own is not ideal. When one entity is split across two or more chains, all the links between the mentions are indeed correct and will thus be useful for applications like information retrieval. MUC and B^3 give points to such assignments, whereas only the largest of such chains will be used for CEAF, leaving the others—and their correct links—out of the score. It is also interesting to consider MUC and B^3 as they can be useful for teasing apart the behavior of different models, for example, with $\mathsf{ILP}_{c,a,e}$ compared to $\mathsf{COREF}_{closest}$, where CEAF was the same but the others were different.

There is an interesting point of comparison with our results using rankers rather than classifiers and using models specialized to particular types of mentions (Denis and Baldridge, 2008). This work does not use ILP, but the best system there, with f-scores of 71.6, 72.7, and 67.0 for MUC, B^3, and CEAF, respectively, actually slightly beats $\mathsf{ILP}_{c,a,e|trans}$, our best ILP system. This underscores the importance of attending carefully to the base classifiers and features used (see also Bengston and Roth (2008) in this regard). The ILP approach in this paper could straightforwardly swap in these better base models. We expect this to lead to further performance improvements, which we intend to test in future work, as well as testing the performance of these models and methods when using predicted, rather than gold, mentions.

7 Conclusion

We have shown that joint inference over coreference, anaphoricity, and named entity classification using ILP leads to improvements for all three main coreference metrics: MUC, B^3, and CEAF. The fact that B^3 and CEAF scores were also improved is significant: the ILP formulations tend to construct larger coreference chains—these are rewarded by MUC without precision penalties, but B^3 and CEAF are not as lenient.

As importantly, we have provided a careful study of cascaded systems, oracle systems and the joint systems with respect to all of the metrics. We demonstrated that the MUC metric's bias for larger chains leads it to give much higher scores while performance according to the other metrics actually drops. Nonetheless, B^3 and CEAF also have weaknesses; it is thus important to report all of these scores. We also include the MELA score as a simple at-a-glance composite metric.

Acknowledgments

We would like to thank Nicholas Asher, David Beaver, Andrew Kehler, Ray Mooney, and the three anonymous reviewers for their comments, as well as the audience at the workshop for their questions. This work was supported by NSF grant IIS-0535154.

References

Bagga, A. and B. Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of LREC 1998*, pages 563–566.

Barzilay, Regina and Mirella Lapata. 2006. Aggregation via set partitioning for natural language generation. In *Proceedings of HLT-NAACL 2006*, pages 359–366, New York City, USA.

Bengston, Eric and Dan Roth. 2008. Understanding the value of features for coreference resolution. In *Proceedings of EMNLP 2008*, pages 294–303, Honolulu, Hawaii.

Berger, A., S. Della Pietra, and V. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Clarke, James and Mirella Lapata. 2006. Constraint-based sentence compression: An integer programming approach. In *Proceedings of COLING-ACL 2006*, pages 144–151.

Denis, P. 2007. *New Learning Models for Robust Reference Resolution.* Ph.D. thesis, University of Texas at Austin.

Denis, P. and J. Baldridge. 2007. Joint determination of anaphoricity and coreference resolution using integer programming. In *Proceedings of HLT-NAACL 2007*, Rochester, NY.

Denis, Pascal and Jason Baldridge. 2008. Specialized models and ranking for coreference resolution. In *Proceedings of EMNLP 2008*, pages 660–669, Honolulu, Hawaii.

Haghighi, A. and D. Klein. 2007. Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of ACL 2007*, pages 848–855, Prague, Czech Republic.

Kehler, A., D. Appelt, L. Taylor, and A. Simma. 2004. The (non)utility of predicate-argument frequencies for pronoun interpretation. In *Proceedings of HLT-NAACL 2004*.

Klenner, M. 2007. Enforcing coherence on coreference sets. In *Proceedings of RANLP 2007*.

Luo, X. 2005. On coreference resolution performance metrics. In *Proceedings of HLT-NAACL 2005*, pages 25–32.

Luo, Xiaoqiang, Abe Ittycheriah, Hogyan Jing, Nanda Kambhatla, and Salim Roukos. 2004. A mention-synchronous coreference resolution algorithm based on the bell tree. In *Proceedings of ACL 2004*, pages 135–142, Barcelona, Spain.

Malouf, R. 2002. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the Sixth Workshop on Natural Language Learning*, pages 49–55, Taipei, Taiwan.

McCallum, A. and B. Wellner. 2004. Conditional models of identity uncertainty with application to noun coreference. In *Proceedings of NIPS 2004*.

Morton, T. 2000. Coreference for NLP applications. In *Proceedings of ACL 2000*, Hong Kong.

Ng, V. 2004. Learning noun phrase anaphoricity to improve coreference resolution: Issues in representation and optimization. In *Proceedings of ACL 2004*.

Ng, V. and C. Cardie. 2002a. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of COLING 2002*.

Ng, V. and C. Cardie. 2002b. Improving machine learning approaches to coreference resolution. In *Proceedings of ACL 2002*, pages 104–111.

Roth, Dan and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of CoNLL*.

Soon, W. M., H. T. Ng, and D. Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Vilain, M., J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings fo the 6th Message Understanding Conference (MUC-6)*, pages 45–52, San Mateo, CA. Morgan Kaufmann.