Joint Anaphoricity Detection and Coreference Resolution with Constrained Latent Structures

Emmanuel Lassalle
Alpage Project-team
INRIA & Univ. Paris Diderot
Sorbonne Paris Cité, F-75205 Paris
emmanuel.lassalle@ens-lyon.org

Pascal Denis
Magnet Team
INRIA Lille - Nord Europe
Avenue Heloïse, F-59650 Villeneuve d'Ascq
pascal.denis@inria.fr

Abstract

This paper introduces a new structured model for learning anaphoricity detection and coreference resolution in a joint fashion. Specifically, we use a latent tree to represent the full coreference and anaphoric structure of a document at a global level, and we jointly learn the parameters of the two models using a version of the structured perceptron algorithm. Our joint structured model is further refined by the use of pairwise constraints which help the model to capture accurately certain patterns of coreference. Our experiments on the CoNLL-2012 English datasets show large improvements in both coreference resolution and anaphoricity detection, compared to various competing architectures. Our best coreference system obtains a CoNLL score of 81.97 on gold mentions, which is to date the best score reported on this setting.

Introduction

Resolving coreference in a text, that is, partitioning *mentions* (noun phrases, verbs, etc) into referential *entities*, is a challenging task in NLP leading to many different approaches (Ng 2010). Anaphoricity detection, on the other hand, consists in deciding whether a mention is *anaphoric* (aka *discourse-old*) or *non-anaphoric* (*discourse-new*).[1] This task is strongly related to coreference resolution and has been mainly addressed as a preliminary task to solve, leading to pipeline architectures (Ng and Cardie 2002a; Ng 2004; Denis and Baldridge 2008).

An important drawback of pipelined models is that errors tend to propagate from anaphoricity detection to coreference resolution, hence ultimately hurting the performance of the downstream system. In order to avoid error propagation, (Denis and Baldridge 2007) propose a joint inference scheme using Integrer Linear Programming (ILP) to maximize the scores of two models. In this case, inference is performed jointly but the two models are still trained independently. (Poon and Domingos 2008) perform joint learning using using Markov Logic Networks, but sampling techniques are needed to perform inference. (Rahman and Ng 2011) propose a ranking approach wherein, for each mention taken in the order of the text, the decision to link it to a previous mention and to classify it as discourse-new is taken jointly. In this approach, the decision is local to the mention and the previous context, but crucially does not take into account the next mentions in the document. Other approaches simply use the output of an anaphoricity classifier as feature for the coreference model (Bengtson and Roth 2008; Durrett, Hall, and Klein 2013).

In this paper, we employ latent trees to represent the full coreference and anaphoricity structure of a document. We extend latent tree models (Yu and Joachims 2009; Fernandes, dos Santos, and Milidiú 2012; Chang, Samdani, and Roth 2013; Björkelund and Kuhn 2014) by introducing two kinds of edges: the first ones encode coreference links, while the second ones represent *discourse-new* elements. Basically, a latent coreferent tree links together the mentions that make up the same entity. We restrict the shape of latent trees by allowing only one "backward link" per mention so as to be able to define a coherent structure when introducing discourse-new links. This also allows us to compute the structure easily from a weighted graph using a greedy "Best-First" algorithm. Our main contribution is to provide the first system that learns coreference resolution and anaphoricity detection both in a *joint* and *global* fashion. The model is joint in that parameters for the two models are estimated together, so that changes in the anaphoricity detection model directly affect the estimation of the coreference resolution parameter (and vice versa). The model is global in that parameters are learned in a way that minimizes a loss that is defined at the document level. We additionally define a set of must-link and cannot-link constraints on the structure, which helps the model on certain types of coreference links.

Our experiments on the English CoNLL-2012 datasets compare pipeline vs. joint models and local vs. global versions of them, always obtaining better coreference results with joint models. More precisely, the CoNLL score systematically improves as one goes from pipeline to joint models as well as from local to global models. The constrained version of our global joint model obtains the best results overall, and achieves performance that is well above the state-of-the-art. At the same time, we observe that anaphoricity detection also largely improves in the global joint model.

[1] In this paper, we slightly overload these terms by taking an non-anaphor to denote the first mention of an entity (in the order of the text), and an anaphor any mention that is not.

Joint Latent Representation of the Coreference Structure

This section first discusses the relationship between anaphoricity and coreference, and then define a tree structure to represent the coreference structure of a document.

Anaphoricity and Coreference

Once mentions are identified in a text, they have to be linked together to form coreference clusters. Determining whether a mention is *discourse-new* or *anaphoric* helps reducing the search space for coreference resolution. For example, in the text in Figure 1, m_1, m_2 and m_5 are discourse-new (i.e., they don't have backward coreference link to preceding mentions), while all other mentions are anaphora (i.e., they have outgoing backward coreference links).

> **[Peter]**$_{m_1}$ told **[John]**$_{m_2}$ **[he]**$_{m_3}$ did not manage to open the door of **[his]**$_{m_4}$ apartment because **[the key]**$_{m_5}$ broke off in the lock. **[[His]**$_{m_6}$ **friend]**$_{m_7}$ managed to get **[it]**$_{m_8}$ out and open, which made **[him]**$_{m_9}$ very happy.

Figure 1: A simple example: only mentions with coreference links (i.e., non-singleton) are annotated.

A common approach is to detect anaphoricity before coreference in a pipeline architecture: mentions are classified as anaphoric or not based on a locally trained model, and these classifications are used to constrain the decisions of the coreference model. An important drawback of such systems is that errors tend to propagate, which in turn requires a careful tuning of the confidence threshold used in anaphoricity classification (Ng 2004).

Joint Representation

We use latent tree representations of coreference clusters, which have proven efficient for globally learning to resolve coreference (Fernandes, dos Santos, and Milidiú 2012; Chang, Samdani, and Roth 2013; Yu and Joachims 2009; Björkelund and Kuhn 2014). We start from an undirected weighted graph of pair mentions and a collection of trees is computed, each tree representing a cluster. Two methods have been used for building such trees from weighted graphs: running a Maximum Spanning Tree (MST) algorithm on the graph or using a BESTFIRST decoding strategy (i.e., only the highest scoring backward link is selected for each mention, provided it exists). It is easy to see that the set of links selected by this latter method has no cycle, and then is also a spanning forest (Björkelund and Kuhn 2014). Both methods yield spanning forests but the structured generated by the latter method have a more restricted topology. This second method will be referred to as BESTFIRST MST.

There are at least two main motivations for using BESTFIRST MST over standard MST. First, the experiments carried out by (Chang, Samdani, and Roth 2013) suggest that BESTFIRST trees achieve better results than MST on coreference resolution. Second, BESTFIRST MST appears to be better suited from an algorithmic point of view. Thus, the BESTFIRST strategy can be easily extented by defining a single *rooted* tree for representing the partition of mentions. The root is a dummy mention added to the other mentions and placed before them in the order of the text. Root-mention links directly encode the fact that the mention is a discourse-new, while mention-mention links are coreference links (see Figure 2). This interpretation of root-mention links is guaranteed by BESTFIRST MST because no coreference path can be created between a mention linked to the root and a previous mention. We give a sketch of proof for this result: if such a path existed, the rightmost element (in the order of the text) of the set of mentions occurring along the path would necessarily have two backward links, which is not possible with the BESTFIRST strategy. By contrast, this kind of path is allowed in unrestricted MST: e.g., imagine that we have 2 mentions m_1 and m_2 and that the MST contains links $(root, m_2)$ and (m_1, m_2). We see that the semantics of "root-mention" links is not preserved in that case.

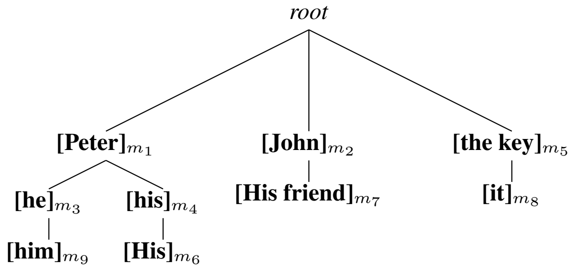

Figure 2: A latent tree representing the coreference structure of the text in Figure 1.

Formally, for a given document with mention $\{m_1, \ldots, m_n\}$, we consider a complete weighted undirected graph $G = (V, E, \omega)$, where V is the set of mentions plus an additional *root* (ordered as $root \preceq m_1 \preceq \cdots \preceq m_n$), E all the pairs formed using V, $\omega : E \rightarrow \mathbb{R}$ a weighting function decomposed as follows:

$$\begin{cases} \omega(m_i, m_j) = \omega_c(m_i, m_j) & 1 \leq i, j \leq n \\ \omega(root, m_j) = \omega_n(m_j) & 1 \leq j \leq n \end{cases}$$

where $\omega_c : E \rightarrow \mathbb{R}$ quantifies the confidence that a pair is coreferent, and $\omega_n : V \rightarrow \mathbb{R}$ the confidence that a mention is discourse-new. We define $\mathcal{S}^{best}(G)$ as the set of spanning trees on G such that every mention can only have at most one link to a previous mention (or to the *root*). We want to compute the following structure:

$$\hat{T} = \operatorname*{arg\,max}_{T \in \mathcal{S}^{best}(G)} \sum_{e \in E_T} \omega(e)$$

with E_T the set of links of the spanning tree T. It is easy to see why it is a global decision, that jointly incorporates coreference and anaphoricity, by decomposing the objective function as:

$$\sum_{(m,m') \in E_T^{coref}} \omega_c(m, m') + \sum_{m \in V_T^{new}} \omega_n(m)$$

where E_T^{coref} is the set of links (m_i, m_j) in tree T such that $m_i \neq root$ and V_T^{new} the set of mentions m_j such that there is a link $(root, m_j)$ in the tree.

Because we have restricted the shape of spanning trees, we can compute the *argmax* easily by using a BESTFIRST strategy: for each mention, the backward edge with the highest weight is selected, and links the mention either to a previous mention (i.e., it is anaphoric) or to the root (i.e., it is discourse-new). From a topological point of view, our tree is similar to the one used in (Fernandes, dos Santos, and Milidiú 2012). The difference is that they do not have weights on root-mention links (no global anaphoricity detection), and they compute the structure with Chu-Liu-Edmonds Algorithm (Chu and Liu 1965; Edmonds 1965). However, as pointed out by (Björkelund and Kuhn 2014), because MST is computed on a oriented graph with only backward edges, it is sufficient to use a BESTFIRST strategy to build it.

Constrained Structures

A way to integrate prior knowledge of the coreference structure is to use constraints on mention pairs: we can add *must-link* (knowledge of a coreference link) and *cannot-link* (impossibility of linking two mentions) constraints in the computation of the spanning tree. These constraints can be generated by finding patterns in the text using accurate rules. In this case, the BESTFIRST strategy only creates backward a backward link for mention that do not appear at the right position of a must-link, and backward links are selected among those which are not cannot-links.

Learning Latent Structures

This section explains how ω_c and ω_n are learned from data, and formulate the problem of learning coreference and anaphoricity as a joint structured classification problem.

Structured Learning

In this formulation, the learning algorithm observes a set of examples (i.e., annotated documents of the training set) $\mathcal{T} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{T}$, where $\boldsymbol{x}_i$ is an instance from a structured input space $\mathcal{X}$ (the space of documents) and $\boldsymbol{y}_i$ is a *structured* label from an output space $\mathcal{Y}$ whose size is exponential in the size of $\boldsymbol{x}_i$. Suppose we have at hand an *embedding function* $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space with inner product $\langle\cdot,\cdot\rangle$. We additionally assume that $\mathcal{H} = \mathcal{H}_c \oplus \mathcal{H}_n$ and that the subspaces are equipped with their own inner products $\langle\cdot,\cdot\rangle_c$ and $\langle\cdot,\cdot\rangle_n$ induced from $\langle\cdot,\cdot\rangle$. Now, given a document $\boldsymbol{x}$ with n mentions $m_1, \ldots, m_n$, we create a graph G with additional *root*. Let $\boldsymbol{y}$ be a BESTFIRST MST on G with coreference links E_y^{coref} and anaphoric mentions V_y^{new}. We restrict Φ to the following decomposed form ($\oplus$ is the concatenation operator to join the features):

$$\Phi(\boldsymbol{x},\boldsymbol{y}) = \sum_{(m,m')\in E_y^{coref}} \phi_c(\boldsymbol{x},m,m') \quad \oplus \sum_{m\in V_y^{new}} \phi_n(\boldsymbol{x},m)$$

where $\phi_c(\boldsymbol{x},m,m')$ (resp. $\phi_n(m)$) is an embedding of the mention pair (m,m') (resp. of the mentions m) in $\mathcal{H}_c$ (resp. $\mathcal{H}_n$), depending on information contained in document $\boldsymbol{x}$ (e.g. a feature representation). To score the pair $(\boldsymbol{x},\boldsymbol{y})$, we suppose we have at hand two weight vectors $\boldsymbol{w}_c \in \mathcal{H}_c$ and $\boldsymbol{w}_n \in \mathcal{H}_n$. The score associated to $(\boldsymbol{x},\boldsymbol{y})$ is:

$$\langle \boldsymbol{w}, \Phi(\boldsymbol{x},\boldsymbol{y})\rangle = \sum_{(m,m')\in E_y^{coref}} \langle \boldsymbol{w}_c, \phi_c(\boldsymbol{x},m,m')\rangle_c + \sum_{m\in V_y^{new}} \langle \boldsymbol{w}_n, \phi_n(\boldsymbol{x},m)\rangle_n$$

The relationship with the weighting function on G described before is the following: $\omega_c(m,m') = \langle \boldsymbol{w}_c, \phi_c(\boldsymbol{x},m,m')\rangle_c$ and $\omega_n(m) = \langle \boldsymbol{w}_n, \phi_n(\boldsymbol{x},m)\rangle_n$. Predicting $\boldsymbol{x}$'s BESTFIRST MST amounts to computing the following *argmax*:

$$T(\boldsymbol{x}) = \underset{T\in S^{best}(G)}{\arg\max}\ \langle \boldsymbol{w}, \Phi(\boldsymbol{x},\boldsymbol{y})\rangle$$

Now, we must address the problem of learning relevant weights vectors $\boldsymbol{w}_c$ and $\boldsymbol{w}_n$. The problem is that we cannot observe directly *gold* BESTFIRST MSTs since we are only given annotation describing the coreference partition of documents, and that there is no unique tree corresponding to a given partition.

Latent Structure Perceptron-based Learning

For our purpose, we consider that $\mathcal{H}_c = \mathbb{R}^m$ and $\mathcal{H}_n = \mathbb{R}^p$ and we use the canonical dot products. The goal of learning is to acquire the weight vectors $\boldsymbol{w}_c \in \mathbb{R}^m$ and $\boldsymbol{w}_n \in \mathbb{R}^p$. These are estimated with the online algorithm in Figure 3.

This algorithm is only given a sequence $\{\boldsymbol{x}_i, P_i\}_{i=1}^{T}$ of documents, where P_i is the gold coreference partition for $\boldsymbol{x}_i$. Typically, there are several possibilities for representing a partition by a tree, and we need to select one at each round. Starting from initial weight vectors $\boldsymbol{w}_c^{(0)}$ and $\boldsymbol{w}_n^{(0)}$, it iterates N times over the training examples, giving a total of $N \times T$ iterations. At each round i, a *true* tree is computed for $\boldsymbol{x}_i$, and it plays the role of the *gold* label in structured learning. We selecte the tree $y_i^{\boldsymbol{w}}$ with the best current weight which is compatible with the partition P_i (in the algorithm, the set of those trees is denoted by $\tilde{\mathcal{Y}}_i$). This *gold tree* is easily computed by removing non-coreferent edges from the complete graph according to the *gold* clustering, and by applying the BESTFIRST strategy.

Next, we compute a predicted structure in *max-loss* mode (Crammer et al. 2006).[2] This corresponds to a trade-off between maximizing the weight of the predicted tree and finding a tree "far from the *true* tree", according to a loss counting the number of edges the structures does not have in common (the approach is similar to (Fernandes, dos Santos, and Milidiú 2012)). The predicted tree is computed by modifying the weights of the graph (by just adding one to all links not in the *true* tree), and applying the extended BESTFIRST. The weight vectors are updated by the difference of *true* and *max-loss* predicted label in a structured perceptron manner (the difference is projected in $\mathcal{H}_c$ and $\mathcal{H}_n$ to update $\boldsymbol{w}_c$ and $\boldsymbol{w}_n$ respectively). The final weight vectors is obtained by averaging over the weight vectors compiled after each round.

[2] The more direct *prediction-based* (i.e., without loss) update always gave lower results in our development experiments, so we do not detail this learning mode here.

```
**Require:** Training data: $\mathcal{T} = \{(\boldsymbol{x}_i, P_i\}_{i=1}^{T}$
**Ensure:** Weight vectors $\boldsymbol{w}_c$ and $\boldsymbol{w}_n$
$\boldsymbol{w}_c^{(0)} = 0; \boldsymbol{w}_n^{(0)} = 0; \boldsymbol{v}_c = 0; \boldsymbol{v}_n = 0; i = 0$
for $n : 1..N$ do
    for $t : 1..T$ do
        Compute *true* label $y_i^{\boldsymbol{w}}$ from $P_i$ and $(\boldsymbol{w}_c^{(i)}, \boldsymbol{w}_n^{(i)})$:
        $y_i^{\boldsymbol{w}} = \arg\max_{y \in \tilde{\mathcal{Y}}_i} \Big\{ \sum_{(m,m') \in E_y^{coref}} \langle \boldsymbol{w}_c, \phi_c(\boldsymbol{x}, m, m') \rangle_c + \sum_{m \in V_y^{new}} \langle \boldsymbol{w}_n, \phi_n(\boldsymbol{x}, m) \rangle_n \Big\}$
        Compute *max-loss* prediction $\tilde{y}$:
        $\tilde{y} = \arg\max_{y \in \mathcal{Y}} \Big\{ \sum_{(m,m') \in E_y^{coref}} \langle \boldsymbol{w}_c, \phi_c(\boldsymbol{x}, m, m') \rangle_c + \sum_{m \in V_y^{new}} \langle \boldsymbol{w}_n, \phi_n(\boldsymbol{x}, m) \rangle_n + l(y, y_i^{\boldsymbol{w}}) \Big\}$
        $\boldsymbol{w}_c^{(i+1)} = \boldsymbol{w}_c^{(i)} + \sum_{(m,m') \in E_{y_i^{\boldsymbol{w}}}^{coref}} \phi_c(\boldsymbol{x}, m, m') - \sum_{(m,m') \in E_{\tilde{y}}^{coref}} \phi_c(\boldsymbol{x}, m, m')$
        $\boldsymbol{w}_n^{(i+1)} = \boldsymbol{w}_n^{(i)} + \sum_{m \in V_{y_i^{\boldsymbol{w}}}^{new}} \phi_n(\boldsymbol{x}, m) - \sum_{m \in V_{\tilde{y}}^{new}} \phi_n(\boldsymbol{x}, m)$
        $v_c = v_c + \boldsymbol{w}_c^{(i+1)}$
        $v_n = v_n + \boldsymbol{w}_n^{(i+1)}$
        $i = i + 1$
    end for
end for
$\boldsymbol{w}_c = \frac{\boldsymbol{v}_c}{(N \times T)}; \boldsymbol{w}_n = \frac{\boldsymbol{v}_n}{(N \times T)}$
```

Figure 3: Structured perceptron learning with averaging, in *max-loss* mode, for joint coreference-anaphoricity learning.

Weight averaging is common in online learning for it helps reducing overfitting (Freund and Schapire 1998).

Coreference resolution and anaphoricity detection are learned both *jointly* and *globally*: jointly because backward coreference links are in balance with "anaphoricity links", and globally because the update is achieved on a tree representing the complete coreference structure of the document.

Constrained Learning

Simple *must-links* and *cannot-links* rules can be applied before using the learning model. Specifically, we remove cannot-links from the graph and add obligatory edges to the tree and complete the tree by applying the BestFirst rule on mentions that are not at the right hand position of a must-link and by avoiding removed links. This is done both during training and inference.

Systems Description

Local vs. Structured models

Our different coreference systems are based on pairwise representation of mention pairs, meaning we re-employ standard pairwise features. We define a *baseline* system, referred to as "local model" which is a simple pairwise model trained using an averaged perceptron, and using a BestFirst decoder (Ng and Cardie 2002b; Bengtson and Roth 2008). This model uses anaphoricity in the form of a feature corresponding to the output of an anaphoricity classifier (Bengtson and Roth 2008; Durrett, Hall, and Klein 2013).[3]

We also define a joint local model JointBestFirst whose behavior is the same as BestFirst with the difference that the root can be an antecedent (using two separate weight vectors to represent coreference and anaphoricity). It is joint in that the decision of creating a backward link competes that of classifying the mention as discourse-new. But it is still local because the model is updated for each mention.

As opposed to these local models, we set up a global joint model, JointBestFirst^{struct}. We compare this model to its version without anaphoricity BestFirst^{struct} (global coreference, but not joint). For the two global models, we also define additional constrained versions, JointBestFirst^{constr} and BestFirst^{constr}.

Pipeline vs. Joint models

We compare joint models to their pipeline equivalents by using a classifier of anaphoricity upstream. PipeBestFirst is the pipelined version of BestFirst: that is, backward links are forbidden for mentions detected as discourse-new. Conversely, a mention that is classified as anaphoric by the anaphoricity model must have a backward link (taking the one with the highest score, even if negative). Similarly, we define pipeline versions of the structured models, PipeBestFirst^{struct} and PipeBestFirst^{constr}.

Feature sets

Our system uses a classical set of features used for mention pair classification (for more details see (Bengtson and Roth 2008; Rahman and Ng 2011). These include: grammatical types and subtypes, string and substring match, apposition and copula, distance (number of separating men-

[3] This model is also trained using the averaged perceptron.

tions/sentences/words), gender and number match, synonymy/hypernymy and animacy, family name (based on lists), named entity types, syntactic features (gold parse), a morphological feature indicating if a verb is derived from a noun and anaphoricity detection. In addition, we use products of the above features with grammatical types, which we found to improve the results of all our models.

For the anaphoricity classifier, we also use the features of (Ng and Cardie 2002a; Ng 2004; Denis and Baldridge 2008). These include: number of words in the mention; binary features indicating if it is pronoun, speech pronoun, reflexive pronoun, proper name, definite description, quantified description, possessive description or bare noun; the position in text; if the mention is embedded in another mention; if the string/the head matches that of a preceding mention; if the mention is an apposition or acronym of a preceding mention.

Constraints

We defined a small set of constraints. Our *must-links* are given, first, by our own implementation of sieve 1 of (Lee et al. 2011), which accurately matches patterns involving the speaker of sentences (e.g. *He* said: "*I* believe you"), and second, by exact string matches of proper nouns. We also use several sets of *cannot-links*, coming from number, gender, and (un)animated mismatches, as well as i-within-i constraints. In addition, in all our models (constrained or not), we set cannot-links between pronouns to disallow pronoun antecedent for pronouns, as in (Ng and Cardie 2002b).

Experiments

Our objective is to compare our joint structured model with their pipelined and local counterparts, for both coreference resolution and anaphoricity detection.

Experimental setup

Our systems are evaluated on the English part of CoNLL-2012 Corpus (Pradhan et al. 2012). We use the official Train/Dev/Test split sets and we test our models in the *closed mode* in which features are built only from provided data (with the exception of two additional sources: WordNet and (Bergsma and Lin 2006)'s gender and number data).

Our evaluation is restricted to *gold mentions*, to avoid introducing noise from detected mention and focus more on anaphoricity detection and coreference resolution. The score of a full end-to-end coreference depends strongly on the quality on the mention detection (heuristic filtering of mention may be required to build a robust coreference system). Here we focus only on evaluating the clustering power of the models. The task may be easier, but the results should only be compared to those of systems tested on gold mentions.

We use three metrics: MUC (Vilain et al. 1995), B^3 (Bagga and Baldwin 1998), and Entity-based CEAF (or $CEAF_e$) (Luo 2005). Following (Pradhan et al. 2012), we also report a global F1-score, referred to as the CoNLL score, which is an unweighted average of the MUC, B^3 and $CEAF_e$ F1 scores. Micro-averaging is used when reporting our scores for entire CoNLL-2012 dataset.

All our models are (local or structured) linear models, learned with the average perceptron algorithm with 30 iterations on the corpus (sufficient to obtain stable scores on the Dev set). In structured learning, we used the max-loss learning mode, associated with the tree loss. Our baseline for anaphoricity detection is a simple averaged perceptron.

Results and Discussion

Coreference resolution Looking at Table 1, one first observes that, whether pipeline or joint, structured models perform better than local models and constrained models better than unconstrained models.

Notice that the local pipelined model PIPEBESTFIRST slightly improves over the local BESTFIRST (from 72.96 to 73.46), but its structured version $\text{PIPEBESTFIRST}^{struct}$ performs a little worse than $\text{BESTFIRST}^{struct}$ (from 75.07 down to 74.4), mostly due to precision losses on MUC and B^3 and a corresponding loss in recall on $CEAF_e$. It is unclear whether these differences are truly significant, but this might mean that deciding whether to use the pipeline as a hard constraint or to only propagate anaphoricity values through a feature depends on the chosen coreference model.

Turning to the joint models, let us first observe that JOINTBESTFIRST outperforms BESTFIRST by a little over one and a half CoNLL point (from 72.96 up to 74.5). The performance gains come from large improvements in precision on MUC and B^3 (and a corresponding improvement in recall on $CEAF_e$). But the gains are much more significant with the structured version $\text{JOINTBESTFIRST}^{struct}$ and the constrained version $\text{JOINTBESTFIRST}^{constr}$: there, the CoNLL score increases by more than 5 points, due to improvements of 3.5 in B^3, and of close to 11.8 in $CEAF_e$. On all three metrics, gains are found in both recall and precision. Finally, our best model, $\text{JOINTBESTFIRST}^{constr}$, obtains a CoNLL score of 81.97, which is, up to our knowledge, the best score achieved on gold mentions. By comparison, (Chang, Samdani, and Roth 2013) obtained a maximum score of 77.43 as previous highest score.

Anaphoricity detection Anaphoricity detection is evaluated in the different models as follows: after resolving anaphoricity and coreference, we label the first mention of each cluster as discourse new and all the rest as anaphoric and compare this with the gold partition. Accuracy results are reported in the last column of Table 1. Anaphoricity scores in the pipeline models are the same and equal to the score of the local anaphoricity model. Before discussing the results in detail, note that gold mentions contain 22.7% of discourse-new mentions and 77.3% of discourse-old mentions. This distribution is biased towards anaphoric mentions, presumably more so than if singleton entities had been annotated in the CoNLL-2012 dataset.

Looking at the results, we first observe that the two local models BESTFIRST and JOINTBESTFIRST have worse performance than the baseline anaphoricity model, and this even though JOINTBESTFIRST obtains better coreference results than the pipeline version. This suggests that anaphoricity should not be addressed locally without tak-

	MUC			B^3			CEAF$_e$			CoNLL	anaphoricity accuracy
	P	R	F1	P	R	F1	P	R	F1		
Local Models											
BestFirst	89.75	77.03	82.9	84.23	65.95	73.98	52.03	76.68	61.99	72.96	84.39
PipeBestFirst	83.63	84.21	83.92	66.98	74.93	70.73	66.5	64.97	65.73	73.46	90.09
JointBestFirst	91.95	76.89	83.75	88.16	65.7	75.29	53.07	82.12	64.47	74.5	87.16
Strutured Models											
BestFirst^{struct}	85.36	84.97	85.16	69.89	75.16	72.43	67.11	68.12	67.61	75.07	90.34
PipeBestFirst^{struct}	84.61	82.19	84.9	67.94	75.56	71.55	97.54	65.99	66.76	74.4	90.09
JointBestFirst^{struct}	87.45	86.33	86.89	75.89	76.32	76.1	77.79	81.13	79.42	80.8	**98.52**
Constrained Models											
BestFirst^{constr}	86.44	85.42	85.93	73.87	76.33	75.08	67.28	69.93	68.58	76.53	90.31
PipeBestFirst^{constr}	85.35	85.81	85.58	71	76.82	73.8	68.24	67.03	67.63	75.67	90.09
JointBestFirst^{constr}	88.4	87.04	**87.71**	78.49	78.14	**78.31**	77.93	81.92	**79.88**	**81.97**	98.01

Table 1: Coreference resolution on CoNLL-2012 Test Set English (gold mentions).

ing the whole context of the document. Structured and constrained pipeline models BestFirst^{struct} and BestFirst^{constr} do not worsen anaphoricity detection quality after coreference resolution, but do not improve it either.

On the contrary, joint models JointBestFirst^{struct} and JointBestFirst^{constr} show a very significant improvement over the local anaphoricity model, especially on discourse-new mentions. The accuracy achieved by the global joint model is very high compared to the other configuration (i.e., 98.52), and we saw that it also resulted in strong improvements on the coreference side. Overall, the large improvement in anaphoricity detection confirms that coreference entities are much better segmented in our joint model. We finally notice that, because of their deterministic aspect, the constraints of BestFirst^{constr} and JointBestFirst^{constr} slightly hinder the quality of anaphoricity detection compared to BestFirst^{struct} and JointBestFirst^{struct}.

Related Work

Our joint structured approach presented directly extends recent work on latent tree structured models (Fernandes, dos Santos, and Milidiú 2012; Chang, Samdani, and Roth 2013; Yu and Joachims 2009; Björkelund and Kuhn 2014). These models are similar to ours, but do not include anaphoricity information nor are they used to jointly learn anaphoricity dectection and coreference resolution.

This type of approaches breaks away from the standard mention-pair models (Soon, Ng, and Lim 2001; Ng and Cardie 2002b; Bengtson and Roth 2008; Stoyanov et al. 2010; Björkelund and Farkas 2012; Lassalle and Denis 2013) and ranking models (Denis and Baldridge 2008; Rahman and Ng 2011). Other structured output models to coreference include correlation clustering (Finley and Joachims 2005) and probabilistic graphical model-based approaches (McCallum and Wellner 2004; Culotta et al. 2007). These learning models are more complex in that they also attempt to enforce transitivity. Other transitivity enforcing models use Integer Programming-based (Klenner 2007; Denis and Baldridge 2009). Due to their much higher complexity, these global decoding schemes are used in combination with locally-trained models. Coreference resolution has also been framed as a (hyper)graph-cut problem (Nicolae and Nicolae 2006; Cai and Strube 2010). Several other models have attempted to depart from the mention pair representation altogether, trying to model cluster-mention or cluster-cluster relations (Luo et al. 2004; Haghighi and Klein 2010; Rahman and Ng 2011; Stoyanov and Eisner 2012).

A number of previous work has attempted to model anaphoricity detection, and to combine it with coreference resolution. (Ng and Cardie 2002a) show empirically that the pipeline setting typically lead in drops in coreference performance. (Ng 2004) shows that one can get coreference improvement, but this requires careful tuning of the anaphoricity classification threshold. (Denis and Baldridge 2008) uses an anaphoricity classifier combined with a mention ranking model. Previous joint approaches using ILP (Denis and Baldridge 2007) or Markov Logic Network (Poon and Domingos 2008) (or more recently (Bögel and Frank 2013)) have the drawback of formulating a problem which is NP-complete, and may be very time consuming. (Rahman and Ng 2011) propose a local joint approach using ranking to decide whether a mention is discourse-new or linked to a previous entity. Finally, (Bengtson and Roth 2008) and (Durrett, Hall, and Klein 2013) use anaphoricity as a feature in the coreference model.

Conclusion and Perspectives

We have introduced a new structured model for jointly detecting anaphoricity and resolving coreference. Our experiments on gold mentions show that both anaphoricity detection and coreference resolution are improved in the joint model compared to non-joint and pipeline models, leading to results that are significantly higher than state-of-the-art. Our best model achieves a CoNLL score of 81.97.

The next step is to extend this model into a full end-to-end coreference system running on detected mentions. Our idea is to address another task, specific to detected mentions, consisting in detecting singleton and non-referential mentions. This was addressed by (Recasens, de Marneffe, and Potts 2013) with a local model, and we plan to integrate it in our joint, global model.

References

Bagga, A., and Baldwin, B. 1998. Algorithms for scoring coreference chains. In *LREC workshop on linguistics coreference*, volume 1, 563–566.

Bengtson, E., and Roth, D. 2008. Understanding the value of features for coreference resolution. In *EMNLP*, 294–303.

Bergsma, S., and Lin, D. 2006. Bootstrapping path-based pronoun resolution. In *COLING-ACL*, 33–40.

Björkelund, A., and Farkas, R. 2012. Data-driven multilingual coreference resolution using resolver stacking. In *EMNLP*, 49–55.

Björkelund, A., and Kuhn, J. 2014. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In *ACL*.

Bögel, T., and Frank, A. 2013. A joint inference architecture for global coreference clustering with anaphoricity. In Gurevych, I. e. a., ed., *Language Processing and Knowledge in the Web*, volume 8105 of *LNCS*. Springer. 35–46.

Cai, J., and Strube, M. 2010. End-to-end coreference resolution via hypergraph partitioning. In *COLING*, 143–151.

Chang, K.-W.; Samdani, R.; and Roth, D. 2013. A constrained latent variable model for coreference resolution. In *EMNLP*.

Chu, Y. J., and Liu, T. H. 1965. On the shortest arborescence of a directed graph. *Science Sinica* 14.

Crammer, K.; Dekel, O.; Keshet, J.; Shalev-Shwartz, S.; and Singer, Y. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research* 7:551–585.

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

Denis, P., and Baldridge, J. 2007. Joint determination of anaphoricity and coreference resolution using integer programming. In *HLT-NAACL*, 236–243.

Denis, P., and Baldridge, J. 2008. Specialized models and ranking for coreference resolution. In *EMNLP*, 660–669.

Denis, P., and Baldridge, J. 2009. Global joint models for coreference resolution and named entity classification. *Procesamiento del Lenguaje Natural* 42(1):87–96.

Durrett, G.; Hall, D.; and Klein, D. 2013. Decentralized entity-level modeling for coreference resolution. In *ACL*.

Edmonds, J. 1965. Optimum branchings. *Journal of Research of the National Bureau of Standards*.

Fernandes, E. R.; dos Santos, C. N.; and Milidiú, R. L. 2012. Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Joint Conference on EMNLP and CoNLL-Shared Task*, 41–48.

Finley, T., and Joachims, T. 2005. Supervised clustering with support vector machines. In *ICML*, 217–224.

Freund, Y., and Schapire, R. E. 1998. Large margin classification using the perceptron algorithm. In *Machine Learning*, 277–296.

Haghighi, A., and Klein, D. 2010. Coreference resolution in a modular, entity-centered model. In *HLT-NAACL*, 385–393.

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

Lassalle, E., and Denis, P. 2013. Improving pairwise coreference models through feature space hierarchy learning. In *ACL*.

Lee, H.; Peirsman, Y.; Chang, A.; Chambers, N.; Surdeanu, M.; and Jurafsky, D. 2011. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *CoNLL: Shared Task*, 28–34.

Luo, X.; Ittycheriah, A.; Jing, H.; Kambhatla, N.; and Roukos, S. 2004. A mention-synchronous coreference resolution algorithm based on the bell tree. In *ACL*, 135–142.

Luo, X. 2005. On coreference resolution performance metrics. In *HLT-EMNLP*, 25–32.

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

Ng, V., and Cardie, C. 2002a. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *COLING*, 1–7.

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

Ng, V. 2004. Learning noun phrase anaphoricity to improve coreference resolution: Issues in representation and optimization. In *ACL*, 151.

Ng, V. 2010. Supervised noun phrase coreference research: The first fifteen years. In *ACL*, 1396–1411.

Nicolae, C., and Nicolae, G. 2006. Bestcut: A graph algorithm for coreference resolution. In *EMNLP*, 275–283.

Poon, H., and Domingos, P. 2008. Joint unsupervised coreference resolution with markov logic. In *EMNLP*, 650–659.

Pradhan, S.; Moschitti, A.; Xue, N.; Uryupina, O.; and Zhang, Y. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, 1–40.

Rahman, A., and Ng, V. 2011. Narrowing the modeling gap: a cluster-ranking approach to coreference resolution. *JAIR* 40(1):469–521.

Recasens, M.; de Marneffe, M.-C.; and Potts, C. 2013. The life and death of discourse entities: Identifying singleton mentions. In *HLT-NAACL*, 627–633.

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

Stoyanov, V., and Eisner, J. 2012. Easy-first coreference resolution. In *COLING*, 2519–2534.

Stoyanov, V.; Cardie, C.; Gilbert, N.; Riloff, E.; Buttler, D.; and Hysom, D. 2010. Coreference resolution with reconcile. In *ACL*, 156–161.

Vilain, M.; Burger, J.; Aberdeen, J.; Connolly, D.; and Hirschman, L. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, 45–52.

Yu, C.-N. J., and Joachims, T. 2009. Learning structural svms with latent variables. In *ICML*, 1169–1176.