End-to-End Coreference Resolution for Clinical Narratives

Prateek Jindal and **Dan Roth**
University of Illinois at Urbana-Champaign
{jindal2, danr}@illinois.edu

Abstract

Coreference resolution is the problem of clustering mentions into entities and is very critical for natural language understanding. This paper studies the problem of coreference resolution in the context of the important domain of clinical text. Clinical text is unique because it requires significant use of domain knowledge to support coreference resolution. It also has specific discourse characteristics which impose several constraints on coreference decisions. We present a principled framework to incorporate knowledge-based constraints in the coreference model. We also show that different pronouns behave quite differently, necessitating the development of distinct ways for resolving different pronouns. Our methods result in significant performance improvements and we report the best results on a clinical corpora that has been used in coreference shared tasks. Moreover, for the first time, we report the results for end-to-end coreference resolution on this corpora.

1 Introduction

This paper addresses the task of coreference resolution for clinical narratives. *Coreference resolution* is the task of finding referring expressions in a text that refer to the same entity, i.e., finding expressions that corefer. Consider the following text sampled from the corpora we used:

> *This 63-year-old man had **[malignant fibrous histiocytoma of duodenum]**, discovered in 02/95. Other than **[a mass in the duodenum]**, the patient was also diagnosed with anemia. A **[leiomyosarcoma]** was resected after embolization of the splenic artery. However, **[it]** could not be completely excised; moreover **[the tumor]** metastasized to the liver as was discovered on follow up scan in 06/95.*

In the above text, all the phrases which are shown in brackets refer to the same entity and hence form a coreference chain. It is clear that identifying such coreference chains requires a lot of medical knowledge. For example, we need to know that *mass* can refer to a *malignant histiocytoma*. To address this need, we need to use domain-specific knowledge sources. While the literature on coreference resolution heavily discusses the need to incorporate background knowledge [Ratinov and Roth, 2012; Rahman and Ng, 2011], there has been very limited success in doing it. The first contribution of this paper is that it provides a principled way for incorporating knowledge-based constraints into the coreference resolution process and also exhibits its significant contribution to performance.

Best-link strategy has been successfully used in coreference resolution on different types of datasets [Xu *et al.*, 2012b; Bengtson and Roth, 2008]. Chang et al. [2011] proposed a variation of *best-link* strategy where they also incorporated several constraints in its objective function. In this paper, we use the coreference model similar to that of Chang et al. [2011]. However, our constraints are quite different from that of Chang et al. Many of our constraints are obtained from the context in which the mentions appear. Unlike Chang et al., we received significant improvements by using the constraints. This shows that the context in which the mentions appear is quite important.

The second contribution of this paper is in pronominal resolution. Quite often, we find in coreference resolution literature [Bengtson and Roth, 2008; Raghunathan *et al.*, 2010; Poon and Domingos, 2008; Chang *et al.*, 2011] that researchers use the same model for resolving all kinds of pronouns. We, however, found that different pronouns behave quite differently. So, we developed separate modules for finding the antecedents of different kinds of pronouns. The method that we used for pronominal resolution is quite general and will be useful for coreference resolution on other domains as well.

For our experiments, we used the datasets provided by i2b2/VA challenge organizers in 2011 shared task on coreference resolution for clinical narratives. *To the best of our knowledge, we report the best results on this corpora.* Finally, we also report, for the first time, results on end-to-end coreference resolution on *i2b2* corpora.

To summarize, the key contributions of our paper are as follows: **(1)** This paper shows that well-informed constraints can give significant performance improvements in coreference resolution, **(2)** It exhibits the distinct behavior of different pronouns and makes use of it in resolving their coreference, and **(3)** It reports state-of-the-art results on coreference resolution on clinical text, including an end-to-end system.

2 Task Description

Coreference resolution aims at clustering together textual mentions within a single document based on underlying referent entities. For our experiments, we used the datasets provided by i2b2 team as part of coreference challenge. We address the task of coreference resolution in two different settings as explained below.

In the first setting, we use the same problem definition as was specified in the Task 1C of i2b2 coreference challenge. In this setting, mentions have already been identified and classified into 4 types: test (TEST), treatment (TRE), problem (PROB) and pronoun (PRON). Coreference relation can exist only within the mentions of same type. However, PRON mentions can corefer with any other mention. Given the entity mentions along with the types, the aim is to build coreference chains for the first 3 types: TEST, TRE and PROB. Since PRON mentions can corefer with the mentions of other types, there are no separate PRON chains. In the following, we will use the term *"medical mentions"* to collectively refer to mentions of type TEST, TRE and PROB.

In the second setting, we perform end-to-end coreference resolution for clinical notes. In this setting, the input consists of clinical notes in free-text format and the aim is to build coreference chains for the medical concepts. To perform end-to-end coreference resolution, we first identify mention boundaries and then classify the mentions into 4 types: TEST, TRE, PROB and PRON. Then coreference chains are found in a way similar to that of first setting.

In next few sections, we describe our approach for coreference resolution when the mentions are already given (i.e. according to first setting). In §9, we describe our approach for end-to-end coreference resolution.

3 Coreference Model

We view coreference resolution as a graph problem: Given a set of mentions and their context as nodes, generate a set of edges such that any two mentions that refer to the same entity are connected by some path in the graph. We construct this entity-mention graph by finding out the best antecedent of each given mention (anaphor) such that the antecedent refers to the same entity as the anaphor. For finding the best antecedent for *medical mentions*, we use a variant of *Best-Link* strategy. The *Best-Link* strategy [Ng and Cardie, 2002b; Bengtson and Roth, 2008; Chang *et al.*, 2011] for selecting the antecedent of a mention chooses that candidate as the antecedent which gets the maximum score according to a pairwise coreference function pc. We extend the *Best-Link* strategy by including a distance term and several constraints in its objective function as discussed below. For finding the best antecedent for *pronominal mentions*, we use a different approach which will be explained in §6.

3.1 Decision Model: Constrained Best-Link

Given a document d and a pairwise coreference scoring function pc that maps an ordered pair of mentions to a value indicating the probability that they are coreferential, we generate a coreference graph G_d according to the following decision model: For each mention m_i in document d, let $B_{m_i} = \{m_1, m_2, ..., m_{i-1}\}$ be the set of mentions appearing before m_i in d. Let a be the highest scoring antecedent. Then, we have:

$$\begin{aligned} a &= \arg\max_{m_j \in B_{m_i}} score_i(m_j) \\ &= \arg\max_{m_j \in B_{m_i}} pc(m_j, m_i) - \frac{d(m_j, m_i)}{k} + \sum_{l=1}^{L} C_l(m_j, m_i) \end{aligned} \quad (1)$$

In the above equation, $d(m_j, m_i)$ refers to a normalized distance between m_j and m_i which takes values between 0 and 1. C_l refers to the l^{th} constraint and is defined as follows (for all values of l):

$$C_l(m_j, m_i) = \begin{cases} 0 & \text{if } l^{th} \text{ constraint is satisfied} \\ -p_l & \text{otherwise} \end{cases} \quad (2)$$

where p_l is the penalty associated with the l^{th} constraint. Thus, different constraints can have different penalties. The higher the penalty associated with the constraint, the stronger it is enforced. If $score_i(a)$ is greater than a threshold δ, then we add the edge (a, m_i) to the coreference graph G_d. The value of $pc(m_j, m_i)$ lies between 0 and 1. The value of k is chosen to be sufficiently greater than 1 so that the pairwise classifier is given preference over the distance term in choosing the best antecedent. But if the pc values of any two candidates are almost similar, then the antecedent which is closer to the anaphor gets the higher score because of the distance term in Equation (1). Thus, our decision model combines the advantages of both "best-link" and "closest-first" models which are generally used for coreference resolution. Setting $k = \infty$ and $L = 0$ reduces our model to the standard "best-link" decision model.

The resulting graph produced by the decoding technique mentioned above contains connected components (determined by transitive closure) with all the mentions in the component referring to the same entity.

3.2 Pairwise Coreference Function

Coreference function pc in Equation (1) consists of 3 different classifiers, one each for TEST, TRE and PROB classes. Each of these classifiers takes as input an ordered pair of mentions (a, m) such that a precedes m in the document, and produces as output a value that is interpreted as the conditional probability that a and m belong in the same equivalence class.

4 Description of Features

In this section, we describe the features used by pairwise classifiers. We divide the features into two main categories as described in the following two subsections.

4.1 Baseline Features

Baseline features refer to those features which are typically used for coreference resolution. These features are subdivided into the following 3 categories: **(1) Lexical Features:** Similar to Bengtson and Roth [2008], we used the following lexical features: (a) Exact (or extent) match, (b) Substring

relation and (c) Head match. **(2) Syntactic Features:** For syntactic features, we used *Apposition* and *Predicate Nominative* as described in Raghunathan et al. [2010]. **(3) Semantic Features:** Similar to Bengtson and Roth [2008], we used WordNet to check whether given mentions are synonyms or hypernyms of one another.

4.2 Features Using Domain-Specific Knowledge

In medical text, the same concept can be represented in several different ways. For example, *headache*, *cranial pain* and *cephalgia* all refer to the same concept. Similarly, *Atrial Fibrillation*, *AF* and *AFib* also refer to the same concept. The baseline features are not sufficient to address the ambiguity and variability that exists in medical terminology. To improve the performance of coreference resolution, we used several types of domain-specific knowledge as explained below. The importance of using knowledge has been emphasized in other domains as well [Rahman and Ng, 2011; Bryl *et al.*, 2010; Ratinov and Roth, 2012].

Expanding the abbreviations Clinical narratives use a lot of abbreviations. A few examples are: *MRI* (Magnetic Resonance Imaging), *COPD* (Chronic Obstructive Pulmonary Disease) etc. Abbreviations were expanded to their full forms as a normalization step. We collected abbreviations from several sources like training data, Wikipedia[1] etc. For ambiguous abbreviations, we considered all possible expansions.

Converting Hyponyms to Hypernyms During preprocessing, we converted some of the common hyponyms to the corresponding hypernyms. Examples of such conversions are: chemotherapy $\rightarrow$ therapy, hemicolectomy $\rightarrow$ colectomy. Such conversions are quite helpful because it is a common practice in clinical documents to refer to some of the problems and treatments introduced earlier in the document with their more general names later on. These hyponym-hypernym pairs were collected from the training data.

Mapping to Biomedical Vocabularies We used MetaMap [Aronson and Lang, 2010] and MetamorphoSys tools to map the mentions to concepts in biomedical vocabularies like UMLS[2], MeSH[3] and SNOMED CT[4]. Such mapping helps us to determine whether any two mentions are equivalent or not. For example, *cancer* and *malignancy* both map to same UMLS concept namely *Primary Malignant Neoplasm*. From such mapping, we can infer that *cancer* and *malignancy* can be coreferential to one another even though they are lexically quite different.

5 Description of Constraints

Although our model allows for both hard and soft constraints, we used only hard constraints in the current work. These constraints allow us to override the decision of the pairwise classifier, where appropriate. Following is a list of constraints we used.

- *Length Constraint*: Surface form of both the mentions must be at least 2 characters long.
- *Body Parts Constraint*: If body parts (like chest, arm, head) are specified, they should not be incompatible. For example, "pain" in *chest* and "pain" in *leg* cannot be coreferential.
- *Anatomical Terms Constraint*: If anatomical terms[5] (like proximal, anterior, dorsal) are specified, they should not be incompatible. For example, "pain" in *right hand* and "pain" in *left hand* cannot be coreferential.
- *Temporal Constraint*: Certain words like *follow-up* or *repeat* convey temporal information about the mentions. For example, the word *repeat* in the mention *repeat chest x-ray* indicates that *chest x-ray* is being done for the second time. If two mentions refer to tests or treatments which were done at different times, then they cannot be coreferential.
- *Section Constraint*: Clinical reports often specify different sections like *History of Present Illness*, *Laboratory Data*, *Medications on Discharge* etc. We developed an algorithm for finding and normalizing the section headings. If a mention appears in either *Family History* section or *Social History* section in a clinical report, we do not consider it for coreference. This is because such mentions generally describe the problems associated with family members of the patient and not the patient himself/herself.
- *Value Constraint*: `TEST` mentions generally have a value associated with them. If any two `TEST` mentions do not have the same value, then they cannot be coreferential.
- *Assertion Constraint*: We implemented an algorithm for finding the assertion status (like *present*, *absent* etc.) of `PROB` mentions as described by Xu et al. [Xu *et al.*, 2012a]. Two mentions cannot be coreferential if they do not have the same assertion status.

6 Pronominal Coreference Resolution

In the medical corpora we worked with, pronominal resolution is primarily limited to 4 types of pronouns: (1) which (2) that (3) this and (4) it. Other pronouns like these, those, whichever etc. hardly participate in coreference relation in our datasets. Also, personal pronouns like he, she, him, you, yourself etc. refer to persons and hence are not relevant to us because we are interested in forming coreference chains for only medical mentions (`TEST`, `TRE` and `PROB`).

Features commonly used for pronominal resolution [Raghunathan *et al.*, 2010; Poon and Domingos, 2008] include distance, number agreement, gender agreement, entity type, grammatical person (first, second and third) etc. However, many of these features are not very helpful in our case. For example, all the medical mentions

[1] http://en.wikipedia.org/wiki/List_of_medical_abbreviations
[2] http://www.nlm.nih.gov/research/umls/
[3] http://www.nlm.nih.gov/mesh/meshhome.html
[4] http://www.ihtsdo.org/snomed-ct/
[5] http://en.wikipedia.org/wiki/Anatomical_terms_of_location

have neuter gender. So, *gender agreement* is not helpful. Similarly, *grammatical person feature* is also not helpful because it is relevant only for personal pronouns. It should also to be noted that researchers [Raghunathan *et al.*, 2010; Poon and Domingos, 2008] commonly use the same technique for resolving different types of pronouns. However, in our experiments, we found that different pronouns behave very differently and therefore, we designed separate modules for finding the antecedent for different types of pronouns. Next two subsections describe our overall strategy for pronominal resolution.

6.1 Determining Anaphoricity

We first determine whether the given pronoun is anaphoric or not. Ng and Cardie [2002a] have previously shown the benefits of predicting anaphoricity. To identify non-referential cases for pronoun *it*, we implemented the heuristics mentioned by Paice and Husk [1987]. To determine the anaphoricity for the remaining pronouns (*this*, *that* and *which*), we learned a classifier with the following features: (a) Pronoun under consideration (*this*, *that* or *which*), (b) Part-of-Speech tag of pronoun and (c) Number of tokens in the immediate noun phrase encompassing the pronoun.

6.2 Finding the Antecedent

In the previous step, we filtered out the pronouns which were non-referential. For the remaining pronouns, we need to find the best antecedent. Depending on the pronoun under consideration, we used different techniques for finding the antecedent as described below.

which and that Referential cases of pronouns *which* and *that* behave quite similarly. Therefore, we use the same strategy for determining their antecedents. Both these pronouns are often used as a relative pronoun and they mark the beginning of a dependent clause. We select the closest medical mention in the associated independent clause as the antecedent for such pronouns. However, if there is any intervening noun phrase between the pronoun and the closest medical mention, then we leave such a pronoun as a singleton and mark its antecedent as `NULL`. It should be clear from the above description that we restrict the antecedent of pronouns *which* and *that* to come from the same sentence.

this and it For pronouns *which* and *that*, we could simply select the closest medical mention (subject to some constraints) as the antecedent. However, the antecedent of *this* and *it* can be separated from them by one or more medical mentions. Thus, antecedent of these pronouns is not necessarily in the same sentence.

To determine the antecedent of pronouns *this* and *it*, we trained an SVM classifier to identify whether the pronoun under consideration is being used as a test, treatment or problem. Thus, this classifier has 3 possible outputs: `TEST`, `TRE` or `PROB`. The features used for training this classifier are: (a) Pronoun under consideration (*this* or *it*), (b) Verb in the associated clause, (c) Is pronoun acting as a subject or an object, (d) Is there a preposition in the path from pronoun to its associated verb and (e) Part-of-Speech of pronoun.

Finally, we selected the closest medical mention which satisfied the following criteria as the antecedent for pronouns *this* and *it*:

1. The antecedent should either be in the preceding sentence or, if it is in the same sentence, it should be separated from pronoun by a conjunction (and, but, although etc.).
2. The antecedent should have the same type (`TEST`, `TRE` or `PROB`) as the pronoun (as given by SVM classifier).

7 Experimental Setup

Datasets: For our experiments, we used the coreference datasets made available by the i2b2 team as part of 2011 i2b2 shared task. The datasets consist of EHRs from two different organizations: Partners HealthCare (`Part`) and Beth Israel Deaconess Medical Center (`Beth`). All records have been fully de-identified and manually annotated for coreference.

The total number of documents in the training set of `Part` and `Beth` are 136 and 115 respectively. The test set of `Part` and `Beth` contains 94 and 79 documents respectively. Total number of mentions in `Part` and `Beth` datasets are 28857 and 40185 respectively. For more information about the datasets, please refer to Uzuner et al. [2012] or Bodnari et al. [2012]. We used B-cubed [Bagga and Baldwin, 1998], MUC [Vilain *et al.*, 1995] and CEAF [Luo, 2005] as the evaluation metrics in our experiments. We also report the unweighted average of F1 scores of these 3 metrics because it was the official metric in i2b2 coreference challenge.

Choice of Parameters: We use cross-validation to determine the system parameters. In Equation (1), we set $k = 100$. With this choice of k, distance term becomes significant only if the scores given by pairwise classifier for different mention pairs differ by less than 0.01. Threshold parameter δ is chosen to be 0.5. As far as constraints are concerned, we decided to formulate all our constraints as hard constraints. To formulate all our constraints as hard constraints, we chose $p_l = 1.0$ in Equation (2) for all values of l.

8 Results

Table 1 compares the performance of four systems as described below:

1. *Baseline (B)*: Baseline system uses only the baseline features described in §4.1. It does not perform pronominal resolution and also does not use any constraints.
2. *Baseline + Knowledge (BK)*: This system uses all the features described in §4. In other aspects, it is similar to *Baseline* system.
3. *Baseline + Knowledge + Pronouns (BKP)*: This system adds pronominal resolution to *BK* system.
4. *Baseline + Knowledge + Pronouns + Constraints (BKPC)*: This is the final system. It adds the ability to deal with constraints to the *BKP* system.

	B			BK			BKP			BKPC		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Test (TEST)												
MUC	29.7	52.8	38.0	-	-	-	39.0	82.7	53.0	57.8	66.0	**61.6**
B3	94.4	96.8	95.6	-	-	-	92.7	97.6	95.1	96.2	96.7	**96.4**
CEAF	81.7	93.8	87.3	-	-	-	82.4	94.6	88.1	93.1	94.9	**94.0**
Avg			73.6						78.7†			**84.0**†
Treatment (TRE)												
MUC	74.4	76.2	75.3	-	-	-	73.0	79.9	**76.3**	73.0	79.9	**76.3**
B3	95.9	95.9	**95.9**	-	-	-	94.7	96.2	95.4	94.7	96.2	95.4
CEAF	86.6	89.4	88.0	-	-	-	86.7	89.5	**88.1**	86.7	89.5	**88.1**
Avg			86.4						**86.6**†			**86.6**
Problem (PROB)												
MUC	72.8	66.4	69.5	69.7	73.5	71.6	69.9	81.2	75.1	74.9	76.8	**75.8**
B3	96.6	94.8	**95.7**	95.4	95.8	95.6	93.8	96.3	95.0	95.1	95.7	95.4
CEAF	87.4	87.9	87.7	84.9	90.1	87.4	85.3	90.5	87.8	89.0	89.8	**89.4**
Avg			84.3			84.9†			86.0†			**86.9**†

Table 1: This table compares the performance of four systems: *B*, *BK*, *BKP* and *BKPC* on Part dataset (see §8). Average F1 scores in this table show that the performance of coreference resolution is significantly improved by adding knowledge, pronominal resolution and constraints to the system. The dagger sign (†) indicates statistically significant performance improvement over the previous step.

In Table 1, we compare the performance of these 4 systems for TEST, TRE and PROB categories on Part corpus. We don't show the detailed results for Beth corpus because of space limitations. But it follows a very similar trend. Table 1 reports precision (P), recall (R) and F1 scores for MUC, B-cubed and CEAF evaluation metrics. It also shows the average F1 score of these three metrics. Please note that there are no separate scores for PRON category because there are no separate PRON chains. PRON mentions are included within the TEST, TRE and PROB chains.

It is interesting to note that adding knowledge to the system always leads to higher recall values. On the other hand, addition of constraints always leads to higher precision values. Next, we note that different metrics behave differently in evaluating the performance of the systems. It can be seen that the B-cubed metric gives the highest scores. Even for *Baseline* system, B-cubed metric gives about 95% F1 score. This is because of the fact that the corpora that we used contain a very large number of singletons. B-cubed gives very high scores because it highly awards the correct prediction of singletons. MUC, on the other hand, is totally insensitive to singletons. CEAF is intermediate between B-cubed and MUC as far as singletons are concerned. From this discussion, we can see that B-cubed metric is not very discriminative for our corpora. But MUC and CEAF are quite good for comparing the performance of different systems. Average F1 score shown in Table 1 is the official metric used in i2b2 shared task and is a good indicator of the performance of the system. Next, we note the following major points about each category of mentions.

Test: We do not use the features derived from domain-specific knowledge sources for TEST mentions because coreferring mentions of TEST type tend to have similar surface forms. So, knowledge-based features are not helpful for TEST mentions. In Table 1, we see from average F1 score that constraints and pronominal resolution are very helpful for TEST mentions. The average F1 score jumps from 73.6 to 78.7 on adding pronouns to baseline system. On further addition of constraints, average F1 jumps from 78.7 to 84.0 (an increase of 5.3 F1 points).

Treatment: Just as in the case for TEST mentions, we don't use features derived from domain-specific knowledge sources for TRE mentions as well. Average F1 score shows that pronominal resolution gives small improvement of 0.2 F1 points for TRE mentions.

Problem: For PROB mentions, both knowledge and constraints were used. The average F1 scores in Table 1 show that PROB mentions benefit significantly from knowledge, pronouns and constraints. Average F1 score goes from 84.3 to 84.9 on adding knowledge to baseline system. It further increases to 86.0 and then to 86.9 on adding pronominal resolution and constraints respectively.

The improvements obtained by adding knowledge, pronominal resolution and constraints shown in Table 1 are statistically significant at $p = 0.05$ (indicated by dagger sign (†)) according to Bootstrap Resampling Test [Koehn, 2004]. The only exception to this is the TRE category which didn't get significant improvement by the addition of constraints.

Finally, in Table 2, we compare our system with several other state-of-the-art systems for coreference resolution in the medical domain. The numbers reported in Table 2 refer to the unweighted average of B-cubed, MUC and CEAF F1 scores. We chose unweighted average for comparison because it was the official metric of i2b2 2011 shared task on coreference. For both Part and Beth corpora, our system outperformed all other systems. Xu et al. [Xu *et al.*, 2012b] got the highest scores in i2b2 2011 shared task on coreference. We can see from Table 2 that for TEST and

	Avg of B^3, MUC, CEAF F1		
	TEST	TRE	PROB
	Part Corpus		
Xu et al.	82.6	85.7	86.8
Jindal & Roth	76.1	84.4	84.0
Dai et al.	79.7	81.6	80.5
Gooch & Roudsari	80.5	84.3	83.5
This Paper	**84.0**†	**86.6**†	**86.9**
	Beth Corpus		
Xu et al.	78.0	83.9	**86.8**
Jindal & Roth	65.5	83.0	84.0
Dai et al.	75.6	80.2	79.7
Gooch & Roudsari	78.4	81.7	81.5
This Paper	**79.2**†	**84.4**†	**86.8**

Table 2: This table compares our final system with several other state-of-the-art systems on both Part and Beth corpora. For both these corpora, our system outperformed all other systems. The dagger sign (†) indicates statistically significant performance improvement over Xu et al.'s system. *Thus, we report the best results on shared task corpora.*

	Part Corpus			Beth Corpus		
	P	**R**	**F1**	**P**	**R**	**F1**
Test (TEST)						
MUC	48.5	50.8	49.6	31.4	38.0	34.4
B3	95.8	96.2	96.0	96.2	97.0	96.6
CEAF	94.1	93.1	93.6	93.3	92.4	92.9
Avg			79.7			74.6
Treatment (TRE)						
MUC	59.2	63.3	61.2	58.1	58.9	58.5
B3	91.7	93.8	92.7	92.0	92.6	92.3
CEAF	87.4	81.8	84.5	83.8	78.5	81.1
Avg			79.5			77.3
Problem (PROB)						
MUC	62.8	56.8	59.7	61.4	57.4	59.4
B3	93.8	93.5	93.6	92.4	92.5	92.4
CEAF	90.5	82.2	86.2	88.8	78.8	83.5
Avg			79.8			78.4

Table 3: This table shows the performance of our final system for end-to-end coreference resolution. For detailed discussion, please refer to §9.

TRE categories, we got significantly higher scores than Xu et al. for both Part and Beth corpora. In particular, we improved over Xu et al.'s score by 1.3 and 0.7 F1 points respectively for TEST and TRE categories (when averaged over both Part and Beth corpora). This improvement is statistically significant at $p = 0.05$ (indicated by dagger sign (†)) according to Bootstrap Resampling Test. For PROB category, we got an improvement of 0.1 F1 points for Part corpus. But it is not statistically significant. For PROB category in Beth corpus, our score is similar to that of Xu et al. As far as other systems [Jindal and Roth, 2012; Dai *et al.*, 2012; Gooch and Roudsari, 2012] are concerned, our scores are much higher than theirs for all mention categories and all the differences are statistically significant at $p = 0.05$. *Thus, we report the best results on the i2b2 shared task corpora.*

9 End-to-End Coreference Resolution

In this section, we describe our approach for end-to-end coreference resolution. To perform end-to-end coreference resolution, we first identify mention boundaries along with mention types. We used a CRF model [Lafferty *et al.*, 2001] to perform mention detection. CRF model used BIO encoding for representing chunks and was implemented using MALLET toolkit [McCallum, 2002]. The features used by CRF model include surface forms of words, part-of-speech labels, shallow parse labels and features derived from MetaMap. We also used conjunction of these features. Once we have the mentions along with their types, we perform coreference resolution in the same way as described in §3 to §6.

For evaluating the end-to-end coreference resolution system, we used the script provided by i2b2 2011 challenge organizers. Table 3 shows the performance of our final system for end-to-end coreference resolution. It reports precision (P), recall (R) and F1 scores for MUC, B-cubed and CEAF evaluation metrics. It also shows the average F1 score of these 3 metrics. This table shows the results for TEST, TRE and PROB categories on both Part and Beth corpora. As far as we know, *end-to-end results have not been reported previously on both these corpora*. By comparing the average F1 scores in Tables 1 and 3, we notice that the scores of our final system are about 5-8% lower for end-to-end task than the task where gold mentions were given. The decrease in performance is because of errors made in mention detection. However, it is very encouraging to see that the performance on the end-to-end task is still quite high. For example, on Part dataset, the average F1 score is higher than 79% for all the categories (TEST, TRE and PROB). This is much higher than the best result of 63.4% F1 in CoNLL 2012 shared task on coreference [Pradhan *et al.*, 2012]. Zheng et al. [2012] performed end-to-end coreference resolution on ODIE corpus. However, their average F1 score is quite low (50.9%).

Conclusion

This paper studied the problem of coreference resolution on clinical narratives. We reported the best results on the datasets used by us. This is attributed to better pronominal resolution and the use of constraints. Our successful use of constraints highlights the importance of the context in which the mentions appear. Our method for pronominal resolution is quite general and would benefit other types of text as well. Finally, we also report results on end-to-end coreference resolution.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable suggestions. This research was supported by Grant HHS 90TR0003/01. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the HHS or the US government.

References

[Aronson and Lang, 2010] A.R. Aronson and F.M. Lang. An overview of metamap: historical perspective and recent advances. *J Am Med Info Assoc*, 17(3):229, 2010.

[Bagga and Baldwin, 1998] A. Bagga and B. Baldwin. Algorithms for scoring coreference chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566. Citeseer, 1998.

[Bengtson and Roth, 2008] E. Bengtson and D. Roth. Understanding the value of features for coreference resolution. In *Proceedings of the Conference on EMNLP*, pages 294–303. Association for Computational Linguistics, 2008.

[Bodnari *et al.*, 2012] A. Bodnari, P. Szolovits, and Ö. Uzuner. Mcores: a system for noun phrase coreference resolution for clinical records. *Journal of the American Medical Informatics Association*, 19(5):906–912, 2012.

[Bryl *et al.*, 2010] V. Bryl, C. Giuliano, L. Serafini, and K. Tymoshenko. Using background knowledge to support coreference resolution. In *Proceedings of ECAI*, 2010.

[Chang *et al.*, 2011] K.-W. Chang, R. Samdani, A. Rozovskaya, N. Rizzolo, M. Sammons, and D. Roth. Inference protocols for coreference resolution. In *CoNLL Shared Task*, pages 40–44, Portland, Oregon, USA, 2011. Association for Computational Linguistics.

[Dai *et al.*, 2012] H.J. Dai, C.Y. Chen, C.Y. Wu, P.T. Lai, R.T.H. Tsai, and W.L. Hsu. Coreference resolution of medical concepts in discharge summaries by exploiting contextual information. *JAMIA*, 19(5):888–896, 2012.

[Gooch and Roudsari, 2012] P. Gooch and A. Roudsari. Lexical patterns, features and knowledge resources for coreference resolution in clinical notes. *JBI*, 2012.

[Jindal and Roth, 2012] P. Jindal and D. Roth. Using domain knowledge and domain-inspired discourse model for coreference resolution for clinical narratives. *Journal of the American Medical Informatics Association*, 2012.

[Koehn, 2004] P. Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, volume 4, pages 388–395, 2004.

[Lafferty *et al.*, 2001] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, pages 282–289. Morgan Kaufmann Publishers Inc., 2001.

[Luo, 2005] X. Luo. On coreference resolution performance metrics. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32. ACL, 2005.

[McCallum, 2002] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[Ng and Cardie, 2002a] V. Ng and C. Cardie. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. ACL, 2002.

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

[Paice and Husk, 1987] C.D. Paice and G.D. Husk. Towards the automatic recognition of anaphoric features in english text: the impersonal pronoun it. *Computer Speech & Language*, 2(2):109–132, 1987.

[Poon and Domingos, 2008] H. Poon and P. Domingos. Joint unsupervised coreference resolution with markov logic. In *Proceedings of the Conference on EMNLP*, pages 650–659. Association for Computational Linguistics, 2008.

[Pradhan *et al.*, 2012] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. *EMNLP-CoNLL 2012*, page 1, 2012.

[Raghunathan *et al.*, 2010] K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and C. Manning. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501. Association for Computational Linguistics, 2010.

[Rahman and Ng, 2011] A. Rahman and V. Ng. Coreference resolution with world knowledge. In *Proceedings of ACL-HLT-Volume 1*, pages 814–824. ACL, 2011.

[Ratinov and Roth, 2012] L. Ratinov and D. Roth. Learning-based multi-sieve co-reference resolution with knowledge. In *EMNLP*, 2012.

[Uzuner *et al.*, 2012] O. Uzuner, A. Bodnari, S. Shen, T. Forbush, J. Pestian, and B.R. South. Evaluating the state of the art in coreference resolution for electronic medical records. *JAMIA*, 2012.

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

[Xu *et al.*, 2012a] Y. Xu, K. Hong, J. Tsujii, I. Eric, and C. Chang. Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. *Journal of the American Medical Informatics Association*, 19(5):824–832, 2012.

[Xu *et al.*, 2012b] Y. Xu, J. Liu, J. Wu, Y. Wang, Z. Tu, J.T. Sun, J. Tsujii, I. Eric, and C. Chang. A classification approach to coreference in discharge summaries: 2011 i2b2 challenge. *Journal of the American Medical Informatics Association*, 19(5):897–905, 2012.

[Zheng *et al.*, 2012] J. Zheng, W.W. Chapman, T.A. Miller, C. Lin, R.S. Crowley, and G.K. Savova. A system for coreference resolution for the clinical narrative. *Journal of the American Medical Informatics Association*, 2012.