Joint Inference for End-to-End Coreference Resolution for Clinical Notes

Prateek Jindal
Yahoo! Inc.
2021 S. First St
Suite 110
Champaign, IL, USA
pjindal@yahoo-inc.com

Dan Roth
Department of Computer Science, UIUC
201 N. Goodwin Ave
Urbana, IL, USA
danr@illinois.edu

Carl A. Gunter
Department of Computer Science, UIUC
201 N. Goodwin Ave
Urbana, IL, USA
cgunter@illinois.edu

ABSTRACT

Recent US government initiatives have led to wide adoption of Electronic Health Records (EHRs). More and more health care institutions are storing patients' data in an electronic format. These EHRs contain valuable information which can be used in important applications like Clinical Decision Support (CDS). So, Information Extraction (IE) from EHRs is a very promising research area. This paper presents a robust method for end-to-end coreference resolution for clinical narratives. For our experiments, we used the datasets provided by i2b2/VA team as part of i2b2/VA 2011 shared task on coreference resolution. One part of this data was annotated according to ODIE guidelines and another part was annotated according to i2b2 guidelines. We designed a global inference strategy for end-to-end coreference resolution which jointly determines the mention types and coreference relations between them. This technique avoids the problem of error-propagation which is common in pipeline systems. For pronominal resolution, we developed different strategies for resolving different pronouns. We report the best results to date on both ODIE and i2b2 data. We got the best results for both types of cases: (1) where gold mentions are already given and (2) for end-to-end coreference resolution. ODIE and i2b2 data are annotated quite differently. Best results on both types of data proves the robustness of our algorithm.

Categories and Subject Descriptors

G.1.6 [**Numerical Analysis**]: Optimization—*Constrained optimization, Integer programming*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Discourse, Text analysis*; I.5.4 [**Pattern Recognition**]: Applications—*Text processing*; J.3 [**Life and Medical Sciences**]: Medical information systems

BCB'14, September 20–23, 2014, Newport Beach, CA, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2894-4/14/09 ...$15.00.
http://dx.doi.org/10.1145/2649387.2649437.

General Terms

Algorithms, Languages, Performance

Keywords

Natural Language Processing, Health Informatics, Coreference Resolution, Joint Inference, Integer Programming

1. INTRODUCTION

This paper presents the method for end-to-end coreference resolution for clinical narratives. End-to-end coreference resolution involves determining the mentions and also the coreference relations between them. Typically, a pipeline approach is used for end-to-end coreference resolution where the mentions are first determined and then the coreference relations are found among them. Named entity types and other attributes of mentions are generally used while determining the coreference relations among them. Such an approach has limitations because there may be some errors in the first phase where the attributes of the mentions are determined. These errors are propagated to the next stages and it is not possible to correct such errors later on. To overcome this problem, we present a flexible architecture in this paper which doesn't make hard decisions on mention types while performing mention detection. Instead a joint inference procedure makes the final decisions.

Another major contribution of this paper is in pronominal resolution. Quite often, we find in coreference resolution literature 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 used by us for pronominal resolution is quite general and will be useful for coreference resolution on other types of text as well.

We tested our approach on the data that was made available by i2b2/VA team in 2011 shared task on coreference resolution. Some of this data (say, $data^{ODIE}$) was annotated according to ODIE guidelines and the rest of the data (say, $data^{i2b2}$) was annotated according to i2b2 guidelines. The shared task involved two different scenarios. In the first scenario, gold mentions were already given and participants were supposed to identify the coreference chains. In the second scenario, only free text was given and participants were supposed to find both the mentions and the coreference chains. The second scenario is referred to as end-to-end coreference resolution.

Using our approach for end-to-end coreference resolution, we got the best results on both $data^{ODIE}$ and $data^{i2b2}$. We also report the best results on both these corpora for the case where gold mentions are already given.

The rest of this paper is organized as follows. Section 2 describes the related work. In Section 3, we give a brief description of the datasets that we used. We describe our coreference strategy from Section 4 to Section 8. Specifically, Section 6 describes our coreference classifier and the various features that we used in it. Section 8 describes the joint inference procedure for coreference resolution. We describe our results in Section 9 and Section 10. Finally, we give some directions for future work in Section 11.

2. RELATED WORK

Coreference resolution is a very important task to understand the semantics of the text and to extract meaningful information from it. i2b2/VA organized a challenge on coreference resolution for clinical narratives in 2011 [38]. A lot of teams from around the world participated in the challenge. Most of the teams focused on the task where gold mentions (along with types) were already given and the aim was to simply recognize the coreference chains. Specification of the mentions along with their types simplifies the problem of coreference resolution considerably. However, for the real-world applications, what we really want is the capacity for end-to-end coreference resolution. Therefore, in this paper, we focus on end-to-end coreference resolution.

In 2011 i2b2/VA challenge, three teams participated in end-to-end coreference resolution. Cai et al. [8] proposed a weakly supervised algorithm which performs classification and clustering steps together with the help of a global inference procedure. Their inference procedure uses mention types as one of the features. These mention types are still determined in a pipeline fashion. Both Lan et al. [22] and Grouin et al. [14] used rule-based systems to find the coreferential pairs where the mention types were used in a pipeline fashion. Named-entity types have been shown to be important features for coreference resolution in the news domain also. But there also, researchers primarily take to pipeline approach. The well-known problem with pipeline based systems is that of error-propagation i.e., the errors made in earlier stages get passed on to the later stages. Our own previous works [16, 17] were also based on pipeline approaches.

First of all, Pascal and Baldridge [12] proposed to model coreference relations jointly with named entity types. However, they used the hard constraint that all the mentions in a coreference chain must have the same type. Considering the fact that named entity tagger may not give perfect distributions, this constraint is too restrictive. Therefore, in this paper, we soften this constraint by introducing a penalty parameter which determines the degree to which this constraint is enforced.

Features commonly used for pronominal resolution [31, 30] 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 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 be noted that researchers [31, 30, 5, 10] 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.

3. DESCRIPTION OF DATASETS

For our experiments, we used the datasets provided by i2b2 team as part of coreference challenge. The data consists of three types of text files: (1) '*.txt' files contain the plain clinical narratives, (2) '*.con' files contain the concepts found in the corresponding .txt files and (3) '*.chain' files contain the coreference chains.

The data provided in the challenge came from three different institutions: (1) Partners HealthCare (`PHC`), (2) Beth-Israel Deaconess Medical Center (`BIDMC`) and (3) Mayo Hospital (`MAYO`). The data from Mayo institution was annotated according to `ODIE` guidelines [34] whereas the data from other two institutions was annotated according to `i2b2` guidelines. We describe the characteristics of both `ODIE` and `i2b2` data below in more detail. All records have been fully de-identified and manually annotated for coreference.

1. ODIE: `ODIE` annotation specifies the following types of mentions:

    - people
    - procedure
    - disease or syndrome
    - sign or symptom
    - anatomical site
    - laboratory or test result
    - indicator reagent diagnostic aid
    - organ or tissue function
    - others

    Mayo data has 2 types of reports: 'clinical' (`MayoC`) and 'pathology' (`MayoP`). The training sets of `MayoC` and `MayoP` contain 28 and 30 documents respectively. The test sets of these contain 19 and 20 documents respectively.

2. I2b2: `i2b2` annotation specifies the following types of mentions:

    - problem
    - test
    - treatment
    - person
    - pronoun

    The total number of documents in the training sets of `PHC` and `BIDMC` are 136 and 115 respectively. Test sets of `PHC` and `BIDMC` contain 94 and 79 documents respectively. For more information about the datasets, please refer to Uzuner et al. [38] and Bodnari et al. [6].

4. OUR COREFERENCE METHOD

The commonly used coreference models scan the document from left to right and identify the antecedent (if it exists) for each anaphor. To select the antecedent, either the Best-Link strategy [5, 10, 28] or Closest-First strategy [36] is used. In this paper, we don't predict the antecedents for the anaphors one-by-one. Instead, we jointly predict the antecedents for all the anaphors in a single global inference step. To perform the global inference, first of all, we identify the mentions. Our mention detector outputs the distributions over the types of mentions. Then we use a pairwise classifier to determine the probability of coreference relation for all mention pairs. The probability distributions output by mention classifier and pairwise coreference classifier are then given as input to an inference mechanism which makes the final assignments regarding mention types and coreference relations. *This is a clear advantage over the pipeline systems which have to make hard assignments a priori.*

5. MENTION IDENTIFICATION

To perform end-to-end coreference resolution, we first identify mention boundaries using a `CRF` model [21]. `CRF` model used `BIO` encoding for representing chunks and was implemented using MALLET toolkit [24]. CRF model used the following features: surface forms of words, part-of-speech labels, shallow parse labels and features derived from MetaMap [3]. We also used conjunction of these features.

For the mentions which were identified by the `CRF` module, we determined the distribution over mention types using an `SVM` classifier [9]. `SVM` classifier used the following features: Tokens of the mention, full-text of the mention (after normalization), bi-grams, headword, suffixes of headword and features derived from MetaMap, UMLS [37], MeSH [25] and SNOMED CT [35].

6. COREFERENCE CLASSIFIER

Coreference classifier (pc) takes an ordered pair of mentions as input and maps it to a value indicating the probability that they are coreferential. For each mention m_i in document d, let B_{m_i} be the set of mentions appearing before m_i in d. Thus, $B_{m_i} = \{m_1, m_2, , m_{i-1}\}$. Then, we define the variables q_{ji} as follows:

$$q_{ji} = pc(m_j, m_i) \quad \forall (j < i) \tag{1}$$

q_{ji} variables will be later used in an inference procedure to determine the coreferential pairs. Value of q_{ji} variables lies between 0 and 1. We divide the features used in coreference classifier into three categories: (1) Baseline Features, (2) Features using domain-specific knowledge and (3) Context-based features. All these types of features are described in the following subsections.

6.1 Baseline Features

Baseline features refer to those features which are typically used for coreference resolution. These features are further subdivided into following 3 categories.

6.1.1 Lexical Features

Similar to Bengtson and Roth [5], we used the following lexical features: (a) Exact (or extent) match, (b) Substring relation and (c) Head match.

6.1.2 Syntactic Features

For syntactic features, we used Apposition and Predicate Nominative as described by Raghunathan et al. [31].

6.1.3 Semantic Features

Similar to Bengtson and Roth [5], we used WordNet to check whether given mentions are synonyms or hypernyms of one another.

6.2 Features using domain-specific knowledge

In medical terminology, 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 which is explained below. Importance of using knowledge has been emphasized in other domains as well[32, 7, 33].

6.2.1 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[41] etc. For ambiguous abbreviations, we considered all possible expansions.

6.2.2 Converting Hyponyms to Hypernyms

During preprocessing, we converted some of the common hyponyms to the corresponding hypernyms. Examples of such conversions are: chemotherapy → therapy, hemicolectomy → 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. Appendix A shows some examples of hyponym-hypernym pairs that we generated.

6.2.3 Mapping to Biomedical Vocabularies

We used MetaMap [3] and MetamorphoSys tools to map the mentions to concepts in biomedical vocabularies like UMLS [37]. 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.

6.3 Description of Discourse (or Context-Based) Features

We used the following discourse related features in our classifier:

1. ***Length feature***: This feature tells whether the surface form of the mention has only 1 character.

2. ***Compatible Body Parts***: If body parts (like *chest*, *arm*, *head*) are specified, they should not be incompatible.

3. ***Compatible Anatomical Terms***: If anatomical terms [1] (like *proximal*, *anterior*, *dorsal*) are specified, they should not be incompatible. Appendix B gives a list of incompatible anatomical terms.

4. ***Number Agreement***: Two mentions should agree in number.

5. ***Temporal Agreement***: Certain words like *follow-up* or *repeat* convey the 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 can't be coreferential.

6. ***Section feature***: 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 don't 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.

7. ***Value Constraint***: *Test* mentions generally have a value associated with them. If any two *Test* mentions don't have the same value, then they can't be coreferential.

8. ***Assertion Constraint***: We implemented an algorithm for finding the assertion status (like *present*, *absent* etc.) of *problem* mentions as described by Xu et al. [42]. We also used the dictionaries released by the authors [26] in our implementation. Two mentions can't be coreferential if they don't have the same assertion status.

7. PRONOMINAL COREFERENCE RESOLUTION

In the datasets that 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 (like *tests*, *treatments* and *problems*). Next two subsections describe our overall strategy for pronominal resolution. In Appendix C, we describe the relative contribution of different pronouns to the overall performance of coreference resolution.

7.1 Determining Anaphoricity

First of all, we determine whether the given pronoun is anaphoric or not. Ng and Cardie [27] 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 [29]. 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.

7.2 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.

7.2.1 which and that

Referential cases of pronouns "which" and "that" behave quite similarly. So, we use the same strategy for determining their antecedents. Both these pronouns (*which* and *that*) 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.

7.2.2 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 pronouns "this" and "it" can be separated from them by one or more medical mentions. Thus, antecedent of these pronouns (*this* and *it*) is not necessarily in the same sentence.

To determine the antecedent of pronouns "this" and "it", we trained an SVM classifier to identify whether pronoun under consideration is being used as a *test*, *treatment* or *problem*. Thus, this classifier has 3 possible outputs: `TEST`, `TRE` or `PROB`. Following features were used for training this classifier: (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. Antecedent should either be in the preceding sentence or if it is in the same sentence, it should be separated from pronoun by some conjunction (like *and*, *but*, *although* etc.).

2. Antecedent should have the same type (`TEST`, `TRE` or `PROB`) as the pronoun (as given by SVM classifier).

8. JOINT INFERENCE STRATEGY

In this section, we describe the joint inference procedure that we used. Assume that there are N mentions in a document. Also, assume that each mention has K possible types. We introduce indicator variable m_{ij} (for all values of i and j) which would be equal to 1 if and only if i^{th} mention is of j^{th} type. The probability with which i^{th} mention takes j^{th} type is denoted by p_{ij}. Let x_{ij} be the cost associated

$$
\begin{aligned}
\min \sum_{i=1}^{N}\sum_{j=1}^{K} & ((-\log_{10} p_{ij}) m_{ij}) \\
& + \sum_{i=1}^{N}\sum_{\substack{j=1\\ j<i}}^{N} [\{(-\log_{10} q_{ji}) c_{ji}\} + \{-\log_{10}(1-q_{ji})(1-c_{ji})\}] \\
& + \frac{1}{2}\rho \sum_{i=1}^{N}\sum_{\substack{j=1\\ j<i}}^{N}\sum_{k=1}^{K} (1-w_{jik})
\end{aligned}
\tag{2}
$$

subject to:

$$\sum_{j=1}^{K} m_{ij} = 1 \tag{3}$$

$$\sum_{\substack{j=1\\ j<i}}^{N} c_{ji} \leq 1 \quad \forall i \tag{4}$$

$$w_{jik} \Leftrightarrow 1 - c_{ji} \geq |m_{jk} - m_{ik}| \quad \forall i \forall j \forall k \tag{5}$$

$$m_{ij}, c_{ji}, w_{jik} \in \{0, 1\} \tag{6}$$

Figure 1: Final Optimization Problem for Coreference Resolution

with assigning j^{th} type to i^{th} mention. It is given by the following equation:

$$x_{ij} = -\log_{10} p_{ij} \tag{7}$$

Now, for i^{th} mention, there are $(i-1)$ mentions which are preceding it. These $(i-1)$ mentions are possible candidates which can serve as the antecedent for i^{th} mention. We introduce an indicator variable c_{ji} to indicate that j^{th} mention is the antecedent for i^{th} mention. Assume that the probability that j^{th} mention is the antecedent for i^{th} mention is given by q_{ji}. Let y_{ji} be the cost associated with assigning j^{th} mention as the antecedent for i^{th} mention. It is given by the following equation:

$$y_{ji} = -\log_{10} q_{ji} \tag{8}$$

Let y_{ji}^{C} be the complementary cost of not assigning j^{th} mention as the antecedent of i^{th} mention. Then y_{ji}^{C} is given by the following equation:

$$y_{ji}^{C} = -\log_{10} q_{ji}^{C} = -\log_{10}(1 - q_{ji}) \tag{9}$$

Next, we want to impose the constraint that all the mentions (other than pronouns) which are in the same coreference chain should have the same type. We formulate this constraint as a soft constraint in our inference procedure. Let ρ be the cost associated with violating this constraint for any coreference pair. Let w_{jik} be the indicator variable which indicates that if j^{th} mention is chosen as the antecedent for i^{th} mention, then j^{th} mention agrees with i^{th} mention as far as k^{th} type is concerned. Mathematically, it can be described as follows:

$$w_{jik} \Leftrightarrow 1 - c_{ji} \geq |m_{jk} - m_{ik}| \quad \forall i \forall j \forall k \tag{10}$$

Now, consider the following equation:

$$v_{ji} = \frac{1}{2}\sum_{k=1}^{K}(1 - w_{jik}) \tag{11}$$

It can be easily verified that v_{ji} would be equal to 1 if and only if j^{th} mention has the same type as i^{th} mention. Otherwise, it would be equal to 0.

Figure 1 shows the final optimization problem. In this figure, Equation (2) represents the objective of optimization problem. It includes the costs described by Equations (7), (8), (9) and also the penalty associated with violating the constraint that coreferring mentions should have the same type. Equation (3) enforces the constraint that each mention can have only one unique type. Equation (4) enforces the constraint that any mention can have at most one antecedent. Equation (5) is same as Equation (10). Finally, Equation (6) expresses the fact that m_{ij}, c_{ji} and w_{jik} are all indicator variables.

9. RESULTS

In this section, we will compare our system with previous state-of-the-art approaches. Specifically, we would compare with the following systems:

1. `LIMSI` - Grouin et al. [14]

2. `CITY` - Phil Gooch [13]

3. `HITS` - Cai et al. [8]

4. `MSRA` - Xu et al. [42]

5. `OPEN` - Yang et al. [43]

6. `BRAND` - Anick et al. [2]

	MayoC	MayoP
LIMSI	79.6	67.0
CITY	77.9	61.8
HITS	81.7	67.5
THIS PAPER	**81.8**	**70.5**

Table 1: This table shows that we get best results on both 'clinical' (MayoC) and 'pathology' (MayoP) sections of ODIE corpus for the case where gold mentions are already given.

7. IIS - Dai et al. [11]
8. UIUC - Jindal and Roth [19]

All the above systems participated in i2b2 2011 coreference challenge and have been briefly described by Uzuner et al. [38]. We used B-cubed [4], MUC [39] and CEAF [23] as the evaluation metrics in our experiments. The official metric of i2b2 coreference challenge was the unweighted average of F1 scores of these 3 metrics.

We report the scores for both the scenarios: (1) when gold mentions are given and (2) for end-to-end coreference resolution. For evaluation, we used the official evaluation script provided by challenge organizers. As noted before, we have two types of data: $data^{ODIE}$ and $data^{i2b2}$. $data^{ODIE}$ consists of a set of clinical narratives from Mayo Institution and is further subdivided into two categories, namely (1) Clinical reports (MayoC) and (2) Pathology reports (MayoP). $data^{i2b2}$ consists of a set of clinical narratives from two different institutions namely, (1) Partners HealthCare (PHC) and (2) Beth-Israel Deaconess Medical Center (BIDMC). In the following, we will report the scores for different subdivisions of $data^{ODIE}$ and $data^{i2b2}$ separately.

9.1 When Gold Mentions Are Given

In this subsection, we will consider the case where the gold mentions are already given and the system has to only identify coreference chains. For this case, coreference relation can exist only within the mentions of same type. However, pronoun mentions can corefer with any other mention.

9.1.1 ODIE Data

Table 1 shows a comparison of our system with previous state-of-the-art approaches on both sections (*clinical* and *pathology*) of ODIE dataset. The numbers shown in this table correspond to average F1 score across all the ODIE categories ("anatomical site", "procedure", etc.). From Table 1, we see that we get the best results on both *clinical* and *pathology* sections of ODIE dataset.

9.1.2 i2b2 Data

Table 2 shows a comparison of our system with previous state-of-the-art approaches [8, 14, 42, 43, 2, 13, 11, 19, 40] on i2b2 dataset. Just like for Table 1, the numbers shown in Table 2 correspond to average F1 score across all the i2b2 categories ("test", "treatment" etc.). From Table 2, we see that we get the best results on both corpora (PHC and BIDMC).

9.2 End-to-End Coreference Resolution

In this subsection, we will present the results for end-to-end coreference resolution. For end-to-end coreference resolution, mentions and their types are not known in advance.

	PHC	BIDMC
MSRA	86.9	85.9
OPEN	85.2	84.7
CITY	84.2	82.6
BRAND	82.0	81.0
HITS	84.8	82.4
IIS	81.0	80.0
LIMSI	83.8	78.8
UIUC	83.0	78.7
THIS PAPER	**87.4**	**86.0**

Table 2: This table shows that we get best results on both corpora (PHC and BIDMC) for the case where gold mentions are already given.

	MayoC	MayoP
LIMSI	62.9	58.0
HITS	49.9	50.1
THIS PAPER	**64.4**	**63.3**

Table 3: This table shows that we get best results on both 'clinical' (MayoC) and 'pathology' (MayoP) sections of ODIE corpus for end-to-end coreference resolution.

	PHC	BIDMC
THIS PAPER	**80.5**	**78.9**

Table 4: This table shows our results on both PHC and BIDMC corpora for end-to-end coreference resolution. No other team reported end-to-end results on these corpora in i2b2 coreference challenge.

9.2.1 ODIE Data

Table 3 compares our results with previous best approaches for end-to-end coreference resolution on both 'clinical' and 'pathology' sections of ODIE dataset. The numbers in Table 3 correspond to average F1 score across all the ODIE categories. This table shows that we get the best results on both sections of ODIE dataset.

9.2.2 i2b2 Data

In i2b2 2011 coreference challenge, none of the teams reported end-to-end results for i2b2 dataset. In Table 4, we give the results of our system for end-to-end coreference resolution on this dataset.

10. DISCUSSION

In Table 5, we show the performance of our system for individual categories for 'clinical' section (MayoC) of ODIE data. This table reports precision, recall and F1 score for B-cubed, MUC and CEAF evaluation metrics. It also reports the unweighted average of F1 scores of these 3 metrics. From this table, we can see that average F1 score is about 70% for 'Disease or Syndrome' and 'Sign or Symptom' categories. For 'Anatomical Site' and 'Procedure' categories, average F1 score is 41.2% and 57.3% respectively. Thus, we see that our system is not performing as well on 'Anatomical Site' and 'Procedure' categories as on other 2 categories. The MUC score for 'Anatomical Site' and 'Procedure' categories

	B-CUBED			MUC			CEAF			Average
	P	R	F1	P	R	F1	P	R	F1	F1
Disease or Syndrome	88.5	87.7	88.1	52.3	43.7	47.6	86.4	63.7	73.3	69.7
Sign or Symptom	87.9	91.1	89.5	47.1	47.1	47.1	83.4	68.1	75.0	70.5
Anatomical Site	80.1	64.1	71.2	29.2	14.6	19.4	65.8	21.9	32.9	41.2
Procedure	85.0	80.6	82.7	32.6	19.2	24.1	84.1	53.1	65.1	57.3
Overall	87.4	85.6	86.5	47.1	34.9	40.1	83.5	55.4	66.6	64.4

Table 5: This table shows the performance of our system for individual categories for end-to-end coreference resolution on the test portion of 'clinical' section (MayoC) of Mayo ODIE data.

	B-CUBED			MUC			CEAF			Average
	P	R	F1	P	R	F1	P	R	F1	F1
Test	95.2	96.2	95.7	45.8	52.8	49.1	94.0	92.7	93.4	79.4
Treatment	91.6	93.3	92.4	57.0	60.5	58.7	87.5	80.5	83.8	78.3
Problem	94.0	93.2	93.6	62.3	54.6	58.2	90.5	81.7	85.9	79.2
Overall	95.1	95.4	95.2	58.6	57.3	57.9	90.6	85.4	87.9	80.4

Table 6: This table shows the performance of our system for individual categories for end-to-end coreference resolution on the test portion of PHC corpus.

reveals that the recall for these categories is quite low. Thus, our system can perform even better if we manage to improve the recall for 'Anatomical Site' and 'Procedure' categories. This will be the subject of future work.

Table 6 shows the performance of our system for individual categories for PHC corpus. This table reports the precision, recall and F1 score for B-cubed, MUC and CEAF evaluation metrics. It also reports the average F1 score of these 3 metrics. From this table, we see that the average F1 score for all three categories, namely, 'test', 'treatment' and 'problem' is about 79%. Thus, we performed quite well on all the categories for PHC corpus. It can also be seen from Table 6 that in general, both precision and recall values are quite high. So, our system doesn't suffer from either poor recall or poor precision.

From Table 5 and Table 6, we see that our system gives better performance on i2b2 corpus than on ODIE corpus. This is partly because of the fact that we had much more training data for i2b2 corpus than for ODIE corpus. One interesting research direction for future can be to examine whether we can use training data with i2b2 annotations to improve the performance on ODIE data.

11. FUTURE WORK

Following are some directions for future work:

1. Some previous works [18, 20, 15] have leveraged sentence structure to jointly predict mention types. Such a technique can potentially be integrated in our current approach.

2. There are several privacy concerns in obtaining and annotating clinical notes. So, it is highly desirable to develop techniques for automatically filtering out sensitive information from clinical notes.

3. Since annotating clinical notes is an expensive process, it is highly desirable to develop unsupervised methods for clinical coreference resolution.

4. In this paper, we saw two different systems of annotations - ODIE and i2b2. A good coreference resolution system should be able to perform well on several different (but related) systems of annotations without being explicitly trained on them. This is the subject for future work.

12. CONCLUSION

In this paper, we proposed a flexible architecture to determine the mention types and coreference chains jointly. We also presented an effective strategy for performing pronominal resolution which is based on the fact that different pronouns behave differently. We report the best results to date on both ODIE and i2b2 corpora for the case where gold mentions are already given. For end-to-end coreference resolution, we report the best results on ODIE data.

13. ACKNOWLEDGMENTS

We would like to thank Ozlem Uzuner for organizing the clinical NLP shared tasks. We would also 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.

14. REFERENCES

[1] AnatomicalTerms. http://en.wikipedia.org/wiki/anatomical_terms_of_location (accessed may 10, 2014), 2014.

[2] P. Anick, P. Hong, N. Xue, and Y. Yang. Coreference resolution for electronic medical records. In *Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data*, 2011.

[3] A. Aronson and F. Lang. An overview of metamap: historical perspective and recent advances. *Journal of*

the American Medical Informatics Association, 17(3):229, 2010.

[4] 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.

[5] 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.

[6] 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.

[7] V. Bryl, C. Giuliano, L. Serafini, and K. Tymoshenko. Using background knowledge to support coreference resolution. In *Proceedings of the 19th European Conference on Artificial Intelligence (ECAI 2010), August*, 2010.

[8] J. Cai, E. Mujdricza-Maydt, Y. Hou, and M. Strube. Weakly supervised graph-based coreference resolution for clinical texts. In *Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data.*, 2011.

[9] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[10] 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.

[11] H. Dai, C. Chen, C. Wu, P. Lai, R. Tsai, and W. Hsu. Coreference resolution of medical concepts in discharge summaries by exploiting contextual information. *Journal of the American Medical Informatics Association*, 19(5):888–896, 2012.

[12] P. Denis, J. Baldridge, et al. Global joint models for coreference resolution and named entity classification. *Procesamiento del Lenguaje Natural*, 42:87–96, 2009.

[13] P. Gooch and A. Roudsari. Lexical patterns, features and knowledge resources for coreference resolution in clinical notes. *Journal of Biomedical Informatics*, 2012.

[14] C. Grouin, M. Dinarelli, S. Rosset, G. Wisniewski, and P. Zweigenbaum. Coreference resolution in clinical reports - the limsi participation in the i2b2/va 2011 challenge. In *Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data*, 2011.

[15] P. Jindal. *Information extraction for clinical narratives*. PhD thesis, University of Illinois at Urbana-Champaign, 2014.

[16] P. Jindal and D. Roth. Using knowledge and constraints to find the best antecedent. In *COLING*, pages 1327–1342, 2012.

[17] P. Jindal and D. Roth. End-to-end coreference resolution for clinical narratives. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence (IJCAI)*, pages 2106–2112. AAAI Press, 2013.

[18] P. Jindal and D. Roth. Extraction of events and temporal expressions from clinical narratives. *Journal of biomedical informatics (JBI)*, 46:S13–S19, 2013.

[19] 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 (JAMIA)*, 20(2):356–362, 2013.

[20] P. Jindal and D. Roth. Using soft constraints in joint inference for clinical concept recognition. In *EMNLP*, pages 1808–1814, 2013.

[21] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[22] M. Lan, J. Zhao, K. Zhang, H. Shi, and J. Cai. Comparative investigation on learning-based and rule-based approaches to coreference resolution in clinic domain: A case study in i2b2 challenge 2011 task 1. In *Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data. i2b2. Boston, MA, USA*, 2011.

[23] 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. Association for Computational Linguistics, 2005.

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

[25] MeSH. http://www.nlm.nih.gov/mesh/meshhome.html (accessed may 10, 2014), 2014.

[26] MicrosoftLists. http://research.microsoft.com/en-us/projects/ehuatuo/default.aspx (accessed may 10, 2014), 2014.

[27] 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. Association for Computational Linguistics, 2002.

[28] V. Ng and C. Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 104–111. Association for Computational Linguistics, 2002.

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

[30] 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.

[31] 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.

[32] A. Rahman and V. Ng. Coreference resolution with world knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 814–824. Association for Computational Linguistics, 2011.

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

[34] G. K. Savova, W. W. Chapman, J. Zheng, and R. S. Crowley. Anaphoric relations in the clinical narrative: corpus creation. *Journal of the American Medical Informatics Association*, 18(4):459–465, 2011.

[35] SNOMEDCT. http://www.ihtsdo.org/snomed-ct/ (accessed may 10, 2014), 2014.

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

[37] UMLS. http://www.nlm.nih.gov/research/umls/ (accessed may 10, 2014), 2014.

[38] 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. *Journal of the American Medical Informatics Association*, 19(5):786–791, 2012.

[39] 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.

[40] H. Ware, C. Mullett, V. Jagannathan, and O. El-Rawas. Machine learning-based coreference resolution of concepts in clinical documents. *Journal of the American Medical Informatics Association*, 19(5):883–887, 2012.

[41] Wikipedia. http://en.wikipedia.org/wiki/main_page (accessed may 10, 2014), 2014.

[42] Y. Xu, J. Liu, J. Wu, Y. Wang, Z. Tu, J. 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.

[43] H. Yang, A. Willis, A. de Roeck, and B. Nuseibeh. A system for coreference resolution in clinical documents. In *Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data*, 2011.

APPENDIX

A. HYPONYM-HYPERNYM PAIRS

Some examples of hyponym-hypernym pairs generated by us are as follows:

1. Adenocarcinoma, carcinoma
2. Birthweight, weight
3. Brachytherapy, therapy
4. Chemotherapy, therapy
5. Cystoprostatectomy, prostatectomy
6. Cytopathology, pathology
7. Empiricvancomycin, vancomycin
8. Gastrojejunostomy, jejunostomy
9. Guidewire, wire
10. Hemicolectomy, colectomy
11. Hemilaminectomy, laminectomy
12. Hemodialysis, dialysis
13. Hepatosplenomegaly, splenomegaly
14. Ischemiccardiomyopathy, cardiomyopathy
15. Ketoacidosis, acidosis
16. Levalbuterol, albuterol
17. Lymphadenopathy, adenopathy
18. Methemoglobin, hemoglobin
19. Orhydronephrosis, hydronephrosis
20. Osteoarthritic, arthritic
21. Osteochondromatosis, chondromatosis
22. Periampullary, ampullary
23. Peripancreatic, pancreatic
24. Plasmapheresis, pheresis
25. Radiotherapy, therapy
26. Serratiaurosepsis, sepsis
27. Thromboembolus, embolus
28. Urosepsis, sepsis

B. INCOMPATIBLE ANATOMICAL TERMS

Following is a list of incompatible pairs of anatomical terms:

1. ipsilateral, contralateral
2. superficial, deep
3. visceral, parietal
4. axial, abaxial
5. rostral, caudal
6. anterior, posterior
7. dorsal, ventral
8. left, right
9. proximal, distal

C. DETAILED PRONOMINAL RESOLUTION ANALYSIS

In this appendix, we discuss the relative contribution of different pronouns to the overall performance of coreference resolution. For the experiments mentioned in this appendix, we did not include the discourse-based features described in Section 6.3. We would refer to our system which doesn't perform pronominal resolution as BK (where B stands for *Baseline* and K stands for *Knowledge*).

In Table 7, we show the performance improvement corresponding to each pronoun individually for `PHC` corpus. The first column in this table shows the performance of the

	BK	**BK+which**	**BK+this**	**BK+that**	**BK+it**
Test (TEST)					
MUC	38.0	49.6	38.9	41.3	38.6
B3	95.6	95.2	95.4	95.4	95.5
CEAF	87.3	87.9	87.3	87.5	87.4
Avg	73.7	77.5	73.8	74.7	73.8
Treatment (TRE)					
MUC	76.2	77.6	75.7	76.3	76.0
B3	95.8	95.5	95.6	95.8	95.8
CEAF	87.4	87.6	87.3	87.5	87.4
Avg	86.5	86.9	86.2	86.5	86.4
Problem (PROB)					
MUC	71.6	73.8	72.0	72.6	71.8
B3	95.6	95.3	95.4	95.5	95.5
CEAF	87.4	87.8	87.4	87.5	87.4
Avg	84.9	85.6	84.9	85.2	84.9
Overall (OVERALL)					
MUC	68.7	71.6	68.7	69.6	68.8
B3	96.3	96.5	96.2	96.3	96.2
CEAF	87.1	87.8	87.1	87.3	87.1
Avg	84.0	85.3	84.0	84.4	84.1

Table 7: This table shows the F1 scores in all the metrics for each of the pronouns individually. These results correspond to test portion of PHC corpus.

	BK	**BK+which**	**BK+which+this**	**BK+which+this+that**	**BK+All**
Test (TEST)					
MUC	38.0	49.6	50.1	52.7	53.2
B3	95.6	95.2	95.1	95.0	94.9
CEAF	87.3	87.9	87.8	87.9	87.9
Avg	73.7	77.5	77.6	78.5	78.7
Treatment (TRE)					
MUC	76.2	77.6	77.0	77.1	76.9
B3	95.8	95.5	95.3	95.3	95.2
CEAF	87.4	87.6	87.5	87.5	87.5
Avg	86.5	86.9	86.6	86.6	86.5
Problem (PROB)					
MUC	71.6	73.8	74.1	75.0	75.2
B3	95.6	95.3	95.1	95.1	95.0
CEAF	87.4	87.8	87.7	87.8	87.8
Avg	84.9	85.6	85.7	86.0	86.0
Overall (OVERALL)					
MUC	68.7	71.6	71.6	72.4	72.5
B3	96.3	96.5	96.4	96.5	96.5
CEAF	87.1	87.8	87.8	88.1	88.1
Avg	84.0	85.3	85.3	85.7	85.7

Table 8: This table shows the F1 scores on PHC corpus as pronouns are added to our system in a cumulative fashion.

BK system. Then next 4 columns show the performance of *BK* system as the capability to resolve one of the pronouns (*which*, *this*, *that* or *it*) was added to it. We see from this table that different pronouns give different performance improvements. Pronoun 'which' gives the maximum performance improvement of 1.3 F1 points. 'which' is followed by 'that' which gives a performance improvement of 0.4 F1 points. Pronoun 'it' gives only a small improvement of 0.1 F1 points and pronoun 'this' did not give any noticeable improvement. It is also interesting to note that none of the pronouns lead to a degradation in the performance.

In Table 8, we show the cumulative performance of the *BK* system as the ability to resolve different pronouns (*which*, *this*, *that* and *it*) is added to it. The results shown in this table are quite consistent with the results shown in Table 7. We see that addition of pronouns 'which' and 'that' gives the performance improvement of 1.3 and 0.4 F1 points respectively. Addition of pronouns 'this' and 'it' did not give any noticeable performance improvements.