Using Domain Knowledge and Domain-Inspired Discourse Model for Coreference Resolution for Clinical Narratives

Correspondence to:

Prateek Jindal, Department of Computer Science, UIUC (email: jindal2@illinois.edu).

Address: 201 N. Goodwin Ave, Urbana, IL – 61801.

Authors:

Prateek Jindal, Dan Roth

Author Affiliations:

Department of Computer Science, UIUC. 201 N. Goodwin Ave, Urbana, IL – 61801.

Keywords:

Information Extraction, Electronic Health Records, Coreference Resolution, Natural Language Processing, Unsupervised Learning

ABSTRACT

Objective

This paper presents a coreference resolution system for clinical narratives. Coreference resolution aims at clustering all mentions in a single document to coherent entities.

Materials and Methods

We employ a knowledge-intensive approach for coreference resolution. The domain knowledge we use includes several domain specific lists, a knowledge intensive mention parsing and task informed discourse model.

Mention parsing allows us to abstract over the surface form of the mention and represent each mention using a higher level representation which we call the mention's Semantic Representation (SR). SR reduces the mention to a standard form and hence better support comparing and matching.

Existing coreference resolution systems tend to ignore discourse aspects and rely heavily on lexical and structural cues in the text. We break from this tradition and present a discourse model for “person” type mentions in clinical narratives which greatly simplifies the coreference resolution.

Results

We evaluated our system on 4 different datasets which were made available in the 2011 i2b2/VA coreference challenge. The unweighted average of F1 scores (over B-cubed, MUC and CEAF) varied from 84.2 to 88.1%. Our experiments show that domain knowledge proved to be very effective for different mention types for all the datasets.

Discussion

Our error analysis shows that most of the recall errors made by the system can be handled by further addition of domain knowledge. The precision errors, on the other hand, are more subtle and indicate the necessity to understand the relations in which mentions participate for building a robust coreference system.

Conclusion

This paper presents an approach that makes an extensive use of domain knowledge to significantly improve coreference resolution. On the acceptance of our paper in the journal, we would make our system and the knowledge sources developed publicly available.

INTRODUCTION

This paper presents the University of Illinois system for coreference resolution in clinical narratives. Coreference resolution aims at clustering all mentions in a single document to coherent entities. In the current approach we rely on the fact that the mentions have already been identified and classified into five types: tests, treatments, problems, people and pronouns and, moreover, that each cluster consists of mentions of a single type (with the exception that pronouns can be clustered with other types of mentions)

A lot of work has been done on coreference resolution but most of it has focused on news text. Existing state-of-the-art systems are expected to do quite poorly on clinical narratives because clinical narratives use a lot of medical terminology. Many features and resources employed by existing open-domain coreference systems are not sufficient for the clinical text as shown in a recent work by Bodnari et al. [1].

The primary contribution of this paper is the extensive use of background domain knowledge which we have acquired, in some cases in an automatic, unsupervised manner. We develop an approach that makes use of three types of knowledge:

1. *Canonicalization of surface forms of mentions*: To achieve this, we acquired several resources (a list of body parts, a list of anatomical terms of locations, hypernym-hyponym relations relevant to the medical domain, a list of relevant abbreviations, etc.) and developed a notion of equivalence classes that allow us to abstract over lexical items that are used in an equivalent way in the clinical narrative. We use our resources and equivalence classes as a way to standardize the mentions' representation.
2. *Clinical mention parsing*: A key contribution of our approach is in developing the idea of mention parsing as a way to abstract over the surface form of the mention and represent the mention using a higher level representation which we call the mention's Semantic Representation (SR). SR reduces the mention to a standard form and hence allows for better matching strategies.
3. *Domain specific discourse model*: Another key component of our approach is the development of domain-inspired discourse model for coreference resolution. In this paper we instantiate this model only for mentions of type “person”. We found that clinical narratives always discuss a single patient. Other than the patient, clinical narratives mention several doctors and, sometimes, a few family members of the patient. This model greatly simplifies the algorithm for coreference resolution. Specifically, it allows us to develop a two-layer algorithm for the coreference resolution of mentions of type “person”. Such layered architecture has been shown to be quite promising in recent work [2,3].

On the acceptance of our paper in the journal, we would make our system and the knowledge sources we developed publicly available.

BACKGROUND AND SIGNIFICANCE

Here we give a qualitative comparison of our system with the best coreference systems available. We discuss both in terms of algorithmic models and the features used.

Algorithmic Models: Culotta et al. [4] propose a first-order probabilistic model for coreference that enables features to be defined over sets of noun phrases. They report 45% error reduction on ACE (Automatic Content Extraction) coreference dataset over a comparable method that only considers features of pairs of noun phrases. In a later work, Bengtson and Roth [5] showed that a much simpler pairwise classification model for coreference resolution developed with a well-designed set of features can outperform systems built with complex models (like that of Culotta et al. [4]). Similar finding was reported by Haghighi and Klein [6] where they developed a deterministic pairwise coreference model which performed better than many state-of-the-art systems. They extracted semantic knowledge (in the form of compatibility lists) from unlabeled data sources. We also used a deterministic pairwise model for coreference resolution of *test*, *treatment* and *problem* mentions. However, the domain knowledge we develop and the way we use it are very different from the previous works.

Raghunathan et al. [2] propose a coreference architecture based on a sieve that applies tiers of deterministic coreference models one at a time from highest to lowest precision. We also use a layered architecture for coreference resolution for *person* mentions. However, our top layer is designed to give a high recall rather than high precision.

Features Used: All the coreference systems use some kind of string matching (like “Head Match”, “Modifier Match”, “Substring Match”) features to determine whether the two mentions corefer or not. Because mentions in medical narrative are typically more complex than those in the news domain, for effective matching of two mentions, we introduced the idea of mention parsing to abstract the mention to a higher level representation called as Semantic Representation (SR).

Coreference systems also rely on several syntactic and semantic features representing the context of a mention to determine coreference. For example, Culotta et al. [4] used a phrase structure grammar to parse the sentences. Raghunathan et al. [2] used breadth first traversal of syntactic trees for finding the best antecedent mention. As pointed out by Miyao et al. [7], most state-of-the-art parsers for English were trained with the Wall Street Journal (WSJ) portion of the Penn Treebank and high accuracy has been reported for WSJ text. However, these parsers rely on lexical information (specifically, lexical bigrams) to attain high accuracy. And it has been criticized that these parsers may overfit to WSJ text [8]. Miyao et al. [7] also show considerable performance improvements for a biological application after retraining the modern WSJ-trained parsers on biomedical corpus.

For deriving semantic features (like synonyms, antonyms, hypernyms, etc.), Bengtson and Roth [5] and Culotta et al. [4] used Wordnet (among other things). However, Wordnet has a poor coverage of proper nouns, in general, and clinical terms, in particular [9,10]. To aid coreference resolution for clinical text, we use several types of domain-specific knowledge like equivalence classes, hypernym-hyponym relations, abbreviations, lists of body parts, anatomical terms of location etc. The sources from where this knowledge was collected have been explained below in relevant sections.

MATERIALS AND METHODS

Coreference resolution aims at clustering the mentions within a single document based on the underlying referent entities. In this paper, we study coreference resolution for clinical narratives. The datasets used in our experiments were made available by the i2b2 team as part of 2011 i2b2/VA shared task on coreference resolution [11]. The input for the systems consist of patient reports (*.txt files) and mentions occurring in those reports (*.con files). Mentions have been classified into five different types: Test (TEST), Treatment (TRE), Problem (PROB), Person (PER) and Pronoun (PRON). *In the rest of the paper, we would use the symbols shown in parenthesis to denote the different mention types.* Given the entity mentions along with the types, the aim is to build coreference chains (*.chains files) for the first 4 types. The PRON mentions can corefer with the mentions of other types. So, there are no separate PRON chains. i2b2 challenge used unweighted average of B-cubed [12], MUC [13] and CEAF F1 [14] scores as the official metric for evaluation.

We observed from the documents that PER type behaves quite differently from the other three types. So, we used a different algorithm for generating PER coreference chains than the other 3 types as described in the following sections. In the rest of the document, we would refer to TEST, TRE and PROB concepts collectively as medical concepts.

COREFERENCE CHAINS FOR MEDICAL CONCEPTS (TEST, TRE and PROB)

As discussed earlier, our approach for coreference resolution is knowledge-intensive and relies heavily on domain knowledge, including abbreviations, hyponym-hypernym pairs, general mentions, equivalence classes and mention parsing. To determine the coreference chains for TEST, TRE and PROB categories, we employed the pairwise classification model for coreference (similar to Bengtson and Roth [5]). Bengtson and Roth [5] view coreference resolution as a graph problem where the set of mentions in a document correspond to nodes of the graph. The edges of this graph are generated by selecting the best antecedent (if it exists) for each mention. This is referred to as the "best-link" strategy. Selection of best antecedent is done by using a pairwise coreference function which has been learned over the training data. Finally, coreference chains are generated by taking the transitive closure of the coreferential mention pairs.

Unlike Bengtson and Roth, we employ a deterministic (rule-based) coreference function. We also do not follow the “best-link” left to right processing [4] but rather consider all possible pairs of mentions of a given type (with some exceptions as explained below) and determine whether the two mentions corefer or not. We also consider the pairs of PRON and non-PRON mentions (non-PRON refers to TEST, TRE and PROB). Three subsequent subsections would describe the coreference resolution for three different kinds of mention pairs:

1. Proper mentions
2. General mentions
3. Pronominal mentions

Proper Mentions

The primary way to determine whether such mention pairs corefer is *string matching*. However, it is very common that the coreferring expressions have very different surface forms. This makes simple *string matching* ineffective. To overcome this problem, we abstract the surface form of the mention to a higher level representation called as Semantic Representation (SR). While matching the two mentions, we compare their SRs for compatibility. Next, we describe the component of *mention processing* for converting the mention to its SR. After *mention processing*, we would describe how to determine whether the SRs of two mentions are compatible or not.

Mention Processing

Mention processing involves converting the mention into a higher level semantic representation. To perform this conversion, we use the concept of equivalence classes which is explained below. After equivalence classes, we would describe the rules for string normalization that we used to normalize the strings before parsing them. Finally, we would describe our procedure of mention parsing.

Equivalence Classes: One of the important components of mention processing is the use of semantic equivalence classes. Equivalence classes are used to address the fact that several different expressions can represent similar meanings. For example, consider the two mentions in the left part of Figure 1. One of the mentions uses the term “malignant neoplasm” while the other mention uses the term “cancer”. The two terms have same meaning. We group such similar terms in an equivalence class. Another such equivalence class is {dropsy, hydropsy, edema, swelling}. We derived our equivalence classes from UMLS [15], gold chains in the training data and from the abstracts of Wikipedia pages.

The members of the equivalence class act as the triggers of the class. The surface form of the given mention is searched for the members of the equivalence class. If any member of an equivalence class is found, that member is replaced with the id of the equivalence class.

Collection of equivalence classes from UMLS: The UMLS, or Unified Medical Language System, is a database that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems. MetaMap [16] is a configurable program which maps biomedical text to the UMLS Metathesaurus or, equivalently, discovers UMLS concepts referred to in text. When we process the mentions using MetaMap, MetaMap associates different spans of the mentions with UMLS concepts. Each UMLS concept serves as an equivalence class.

Collection of equivalence classes from the training data: Equivalence classes were obtained by a bootstrapping approach in which we ran our initial system (which didn’t have equivalence classes) on the training set and generated a list of recall errors made by the system. Some of the recall errors naturally suggested equivalence classes that should be incorporated into the system. Such equivalence classes were added to the system. This whole process was repeated a few times.

Collection of equivalence classes from Wikipedia: We read the Wikipedia abstracts for the common problems appearing in the training set. Many Wikipedia abstracts suggest possible equivalence classes. For example, Wikipedia page on Edema says, “Edema, formerly known as dropsy or hydropsy, …

produces swelling." From this abstract, we could derive the equivalence class {edema, dropsy, hydropsy, swelling}.

String Normalization: For effectively matching the two mentions, we normalize the surface form of the mentions in the following ways:

1. *Abbreviations*: Clinical narratives use a lot of abbreviations. A few examples are: mri (magnetic resonance imaging), copd (chronic obstructive pulmonary disease) etc. All the abbreviations were expanded to their full forms as a normalization step. Abbreviations were collected by parsing the Wikipedia pages on medical abbreviations. These abbreviations are located at: http://en.wikipedia.org/wiki/List_of_medical_abbreviations. Whenever an abbreviation has more than one expansion, we use all possible expansions to match the mention containing the abbreviation with other mentions. If we get a match for any of the expansions, then the mention-pair is considered coreferent.
2. *Converting hyponyms to hypernyms*: Hypernyms are abstractions of terms and could stand for multiple surface form (their hyponyms). Consequently, hyponyms to hypernyms conversion is useful for normalization. Some examples of such conversion are: chemotherapy → therapy, hemicolectomy → colectomy. The hyponym-hypernym pairs were collected from the training data in an unsupervised setting which doesn't require any human supervision or user feedback. The hyponym-hypernym pairs were generated in the following way: We require that hypernym string must be at least of length 4 (character-wise). For each document, we collected those mention pairs in which one of the members of the pair is a suffix of the other member. These mention pairs were tagged as hyponym-hypernym pairs if they occurred more frequently than a specified threshold. We also required that both the members of the pair should also occur in some coreference chain in the training data. Please see Appendix D for some examples of hypernym-hyponym pairs generated by us.

Mention Parsing: After normalizing the surface form of the mention as described above, we parse the mention to extract several components from it which are discussed below:

1. *Modifiers (Mod)*: Modifiers provide additional information about the concept under consideration.
2. *Body Parts (BP)*: For PROB mentions, BP indicates which part of the body (like heart, lung etc.) is affected by the problem. Similar interpretation holds for TRE and TEST mentions. A list of body parts is readily available from the web.
3. *Anatomical Terms of Location (AT)*: ATs (like right, anterior, proximal etc.) are used to describe the locations of structures in relation to other structures or locations in the body. A list of anatomical terms was collected from Wikipedia.
4. *Equivalence Classes (EC)*: Extraction of equivalence classes has already been described above. A phrase can appear in more than one equivalence class. If this is the case, we retain all the equivalence classes associated with the phrase.

5. *Remaining String (RS)*: RS refers to the remaining part of the surface form of the mention after the above 4 components have been extracted. We normalize RS using the Norm Lexical Tool [17] (which is part of the UMLS [15] knowledge sources).

The five components mentioned above constitute the Semantic Representation (SR) of the mention. Figure 1 shows the SRs obtained by parsing 4 different mentions. Please note that SW in Figure 1 refers to the stopwords.

Matching Mentions

After obtaining the SRs for the two mentions, we have to determine whether the two SRs match each other or not. We say that the two SRs match if all of the following conditions are satisfied:

1. Modifiers in the 2 mentions are not incompatible with each other. Two modifiers are considered to be incompatible if they are antonyms of each other. We use WordNet to determine the antonym relation between any two modifiers. If any antonym of the synsets corresponding to one modifier matches the synsets corresponding to another modifier, the two modifiers are considered to be incompatible. In addition, we also maintain a static list of "pairs of modifiers" which are not compatible with each other. This list was generated by manual inspection of some of the precision errors made by the system on the training data. For example, "poorly differentiated" and "well differentiated" form an incompatible pair.
2. If body parts are present in both the SRs, then they must be the same.
3. Anatomical terms in the 2 mentions must not be incompatible with each other. For example, "right" is not compatible with "left". Anatomical terms form a closed set with only a few members. It is very easy to determine the incompatible pairs from the list of anatomical terms using the definition of the terms. We have also provided a list of incompatible pairs of anatomical terms in Appendix E.
4. Let us call the string obtained by concatenating the RS and EC portions of a mention as "catRSEC" of a mention. catRSECs of two mentions must either be identical or catRSEC of one of the mentions must be a substring of the catRSEC of the other mention.

The left and right sides of Figure 1 show two mention pairs. Mentions in the pairs are matched to one another.

General Mentions

If a mention can refer to several different entities or concepts in different contexts, we call it a "general" mention. For example, the mention "surgery" can refer to several different operations like "lobectomy", "coronary artery bypass grafting", "hip replacement", etc. Mention pairs where at least 1 of the mentions is *general* are treated differently than proper mentions.

"treatment", "procedure", "surgery" etc. are some of the general mentions of type TRE. It is clear that string-matching operations are not sufficient for the mention pairs where one or both the mentions are general. For example, consider the text shown below (in italics):

*"She was taken urgently to the operating room, where she underwent [**embolectomy of the profunda superficial femoral vessels_TRE_1**] . The estimated blood loss was 200 cc. and she tolerated [**the procedure_TRE_1**] well . Postoperatively , she was again noted to have a cold right lower extremity with diminished pulses and was again taken to the operating room, where she underwent a right popliteal exploration and [**thrombectomy_TRE_2**]. She again tolerated [**the procedure_TRE_2]** well and did well postoperatively ."*

In the above text, first "procedure" refers to "embolectomy of the profunda superficial femoral vessels" and the second "procedure" refers to "thrombectomy".

Mention pairs where only the first mention[1] is general are not valid coreference pairs. For mention pairs where either only the second mention or both mentions are general, we generate a coreference pair if both the following conditions are satisfied:

1. *Number Compatibility*: Both the mentions must either be singular or plural. For example, "treatments" can't corefer with "treatment". All the non-general mentions are considered singular.
2. There must not be any intervening mention between the two mentions which has the same type as the mentions under consideration.

How to determine which mention is "general": To determine whether a mention is general, we maintain a static list of terms like "problem", "treatment", "surgery" etc. which indicate that a mention is general. These terms were collected by inspection of training data and by general knowledge.

Pronominal Mentions

In those mention pairs where only one of the mentions is of type PRON, we generate a coreference pair only if the PRON mention is the second mention. Pronominal coreference resolution is generally implemented by enforcing agreement constraints between the mentions [2]. The constraints that have been generally used are: number, gender, person and animacy. We found that such constraints are not very helpful for determining pronominal coreference for TEST, TRE and PROB mentions. This is because such mentions are mostly singular, neutral and non-animate. Moreover, personal pronouns have already been tagged as PER.

However, we found that a lot of mentions which were tagged as PRON are simply non-referential (or pleonastic) uses like clefts, extraposition etc. Such mentions cannot corefer with previous mentions. Filtering out such mentions in a preprocessing step reduces the number of incorrect coreference links. We constructed a set of simple rules to distinguish non-referential usage of pronouns from the referring ones. Table 1 shows a few example sentences of both types of usage (referring and non-referring) for pronoun "that".

[1] We call the mention that appears first in the document "first mention" and the mention which appears later the "second mention".

"that" as a referring mention	Her chest x-ray shows [a left lower lobe hematoma PROB] [that PROB] had greatly improved in the last few days .
	She was seen in the emergency room here on 08/31/04 where they performed [an abdominal CT TEST] [that TEST] showed a pelvic mass with multiple enlarged lymph nodes .
Non-referential uses of "that"	The results of [[these TEST] studies TEST] suggested [that] the patient 's symptoms were not due to subclavian steal phenomenon .
	On physical exam today , the patient did not have [any swelling PROB] , and he states [that] he did not feel [the swelling PROB] was present .

Table 1: This table shows the usage of pronoun "that" as a referring and a non-referring mention

We see that when "that" is used as a non-referential expression, either there is a verb (including auxiliary verbs) just before the occurrence of "that" or there is a noun or a pronoun just after the occurrence of "that". On the other hand, when "that" refers to some previous mention, it is generally followed by a verb (including auxiliary verbs like "was" etc.). Similar rules were obtained for other pronouns (For details, please refer to Appendix A). Given a mention pair, we identified whether the PRON mention was a non-referential usage or a possible reference. If the PRON mention was not filtered out, we generated a coreference pair if there were no non-PER mentions between the two given mentions.

Mention pairs where both mentions are of type PRON are treated like the case where only the second mention is of type PRON. The only additional constraint was to enforce the number agreement.

COREFERENCE CHAINS FOR PER

Discourse Model: One Patient, Several Doctors and a Few Family Members

We employ a domain-inspired discourse model for generating coreference chains for the class PER. Our discourse model can be specified as: One patient, several doctors and a few family members. The development of this model was based on the observation that clinical narratives only discuss a single patient. Other than the patient, multiple doctors are mentioned in the narrative, including the attending physician, doctors who are consulted or who have previously treated the patient or whom the patient will next be visiting, etc. Other than the patient and doctors, the clinical narratives sometimes mention a few family members like father, husband, wife, etc.

Employing an appropriate discourse model simplifies the process of coreference resolution significantly. The discourse model specified above readily yields a 2-layer algorithm for coreference resolution which is described below.

2-Layer Algorithm for Coreference Resolution

We employ a 2-layer algorithm for determining the PER coreference chains. In the first layer, we divide the PER mentions into 3 categories: (1) mentions corresponding to patient, (2) mentions corresponding to any of the doctors and (3) the rest of mentions. The coreference pairs are generated in the second

layer from within the categories obtained in the first layer since we know that coreference pairs do not cross the categories. We describe the 2 layers in detail below.

Design of the First Layer

We divide all the PER mentions into three categories (namely patient, doctors and the rest) based on the following criteria:

1. *Surface Form of the Mention*: A mention is added to the list of doctors if it has the tokens like “dr.”, “m.d.”, “cardiologist”, etc. Similarly, the mentions like “the patient”, “this patient”, etc. were added to the patient list.
2. *Context*: Table 2 shows common contexts in which patients and doctors appear. The mentions which appear in such common contexts were added to the appropriate list.
3. *Similarity*: We consider two mentions to be similar if the surface forms of the mentions have at least 1 token in common. Note that the common token can't be a person title like “mr.”, “mrs.”, etc. or a doctor title like “dr.”, “m.d.”, etc. If one of the mentions among a set of similar mentions has already been classified as belonging to the doctor list or the patient list, then all other mentions in the set are also assigned to the same list.
4. All other mentions, with the exception of pronouns, are put in a separate list. Such mentions generally refer to the patients' family members (e.g., “his father”, “his wife”).

Patients' Contexts	***Doctors' Contexts***
[patient] is a 61 year old male with a history of ...	He was seen by [doctor].
[patient] was diagnosed recently with pancreatic cancer after he ...	She will follow up with her pcp , [doctor] , at IVMC , after her discharge .
[patient] was admitted to the Retelk County Medical Center at that time and was treated with ...	cc : [doctor1] [doctor2] ...
DISCHARGE SUMMARY NAME : [patient]	She has been under the care of [doctor]

Table 2: This table shows the common contexts in which the mentions corresponding to patients and doctors appear.

The personal pronouns are categorized in the three lists (patient, doctors, rest) based on the following criteria:

1. The first person pronouns like I, me, my etc. are added to the doctor list. This is because a clinical narrative is generally dictated by a physician or physician's assistant.
2. The second person pronouns are added to the patient list or the doctor list based on the context in which they appear. See Table 3 for examples. If the context is not very clear, then the pronoun is assigned to the same list as any other second person pronoun in the vicinity for which the context is clear.
3. The third person pronouns like he, his, etc. are added to the patient list. This heuristic was found to be quite precise. The third person pronouns very rarely refer to doctors.
4. Pronouns like “who” are added to the doctor list only if there is some doctor mention preceding the pronoun within a margin of 2 words. Otherwise, the pronoun is added to the patient list.

Context	*Assigned List*
This is to notify you that your patient , AGACH , arrived in the Emergency Department ...	Doctor
Please call your primary care doctor for follow up next week .	Patient
If you have further chest pain , call your doctor .	Patient

Table 3: Second person pronoun can either refer to doctor or patient depending on the context

Design of the second layer

In this layer, we generate the actual coreference pairs as explained below:

1. Since our model assumes only one patient, all the mentions in the patient list are assigned to the same coreference chain.
2. From among the list of doctors, we generate a coreference pair between two mentions if any of the following two conditions are met:
    a. Lexical Match: The two mentions share at least one similar token (with the exception of person and doctor titles).
    b. Role Participation: The two mentions are separated by not more than 2 words and the first mention specifies some role like physician, pcp, cardiologist etc. and the second mention doesn't specify any such role (See Table 4 for examples)
3. The second person and first person pronouns (if any) in the doctor list are assigned to separate coreference chains.
4. For the rest of mentions, the coreference pairs are generated according to the lexical match condition.

Sentence	*Role*	*Doctor*
[His primary care physician] is [Dr. **NAME[ZZZ]].	primary care physician	Dr. **NAME[ZZZ]
[PCP] Name : [WHITE , ELVNO R]	PCP	WHITE , ELVNO R
She was seen by [her cardiologist] , [Dr. Clements] and had a Holter monitor on 2015-05-01 .	cardiologist	Dr. Clements

Table 4: This table shows a few example sentences where the doctors participate in some role.

RESULTS AND DISCUSSION

Datasets and Evaluation Metrics

We used the datasets provided by i2b2/VA team as part of 2011 coreference challenge. The data consists of clinical narratives from three different organizations: Partners HealthCare (Partners), Beth Israel Deaconess Medical Center (Beth) and University of Pittsburgh (Pitts). The data from University of Pittsburgh is divided into 2 parts, namely Discharge and Progress. All records have been fully de-identified and manually annotated for coreference. This gave us a total of 4 datasets. We used the same training and testing portion from these corpora as used in the 2011 i2b2 coreference challenge. The total number of documents in the training (testing) portion of Partners, Beth, PittsD and PittsP are 136

(94), 115 (79), 119 (77) and 122 (72) respectively. The total numbers of mentions in the training (testing) portion of these datasets are 17144 (11713), 24392 (15793), 12470 (8619) and 12338 (7742) respectively. We report the precision, recall and F1 scores for three standard evaluation metrics: (1) B-cubed [12] (2) MUC [13] and (3) CEAF [14]. We also report unweighted average of precision, recall and F1 scores of these three metrics.

Results on the Test Set

Error! Reference source not found. shows the precision, recall and F1 on the Test portion of all 4 datasets using 3 different metrics (B-cubed, MUC, CEAF). It also shows the unweighted average of precision, recall and F1 scores of individual metrics. In this table, the coreference chains for all the mention types (TEST, TRE, PROB, PER) are considered together. We see from the table that the unweighted average of F1 scores is the highest for Partners corpus. Unweighted average of both the corpora from Pittsburgh is roughly the same. Beth corpus has the lowest unweighted average of F1 scores. *The average precision, recall and F1 score across all the corpora is 0.900, 0.836 and 0.865 respectively.*

We also note from the table that the B-cubed metric gives much higher scores than the other 2 metrics. This is because of large number of singletons in the corpus which tend to inflate the B-cubed scores. The MUC metric, on the other hand, is not sensitive to the existence of singleton mentions [18].

	B-CUBED			***MUC***			***CEAF***			***Unweighted Avg.***		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Partners	0.936	0.977	0.956	0.909	0.773	0.836	0.906	0.800	0.850	0.917	0.850	0.881
Beth	0.917	0.975	0.945	0.886	0.691	0.776	0.881	0.739	0.804	0.894	0.802	0.842
Pitts Discharge	0.927	0.966	0.946	0.890	0.788	0.836	0.874	0.781	0.825	0.897	0.845	0.869
Pitts Progress	0.942	0.968	0.955	0.866	0.778	0.820	0.871	0.805	0.837	0.893	0.850	0.870

Table 5: This table shows the precision, recall and F1 on the Test portion of all 4 datasets using 3 different metrics (B-cubed, MUC, CEAF). The last column of the table shows the unweighted average of the F1 scores of three metrics.

Tables 6, 7, 8 and 9 in Appendix B show the F1 scores of our system for individual categories (TEST, TRE, PROB, PER) on the Test portion of different corpora using 3 different metrics (B-cubed, MUC, CEAF). The last column shows the unweighted average of the F1 scores of 3 metrics. We see from these tables that the PER category gets the maximum unweighted average of F1 scores among all the categories. Appendix F provides even more detailed results (which include precision and recall values as well) of our system on the test set.

Tables 10, 11, 12 and 13 in Appendix C compare our system with a strong baseline (based on general purpose state-of-the-art coreference system) across different categories for all the datasets. In these tables, we don't report the results for TEST category because we didn't see any statistically significant performance improvement for TEST category as a result of using domain knowledge. However, we observed statistically significant performance improvements for all other categories for all the datasets as Tables 10-13 in Appendix C show. All the improvements are statistically significant at the $p = 0.05$

level (i.e. 95% confidence) according to a paired bootstrap resampling test [19]. Among medical categories (i.e. TEST, TRE, PROB), PROB benefits the maximum from domain knowledge. Please refer to Appendix C for details.

Error Analysis

Here we present the error analysis on the test set. For the sake of this analysis, we selected a subset of the errors made by our system. We classify the errors into two types: precision errors and recall errors.

Precision errors: These errors correspond to the incorrect coreference pairs generated by our system. The major reasons for the precision errors are as follows:

1. *Temporal Issues (~22%)*: Many errors are due to temporal issues. For example, our system predicted two "white count" tests to belong to the same coreference chain in one of the documents. However, two "white count" tests were actually carried out at different times and, therefore, do not corefer.
2. *Similar Surface Forms (~62%)*: Different concepts may have similar surface forms which lead to precision errors. For example, in one of the documents, our system predicted "ct" and "ct myelography" to belong to the same coreference chain. However, the first "ct" was actually referring to "ct scan" which is different from "ct myelography". Similarly, in another document, our system wrongly predicted "oxygen" (treatment) supplied at different concentrations to be coreferrent.
3. *Different Individuals (~8%)*: Some of the errors are because of the fact that the two mentions actually correspond to different individuals. For example, in one document, one of the references to "heart attack" was associated with the patient while another reference to "heart attack" was associated with patient's mother.

Recall Errors: For every gold chain, we collected all those predicted chains which had one or more members of the gold chain. Ideally, all the members of the gold chain should come in a single predicted chain. But because of the recall errors, the gold chain members get split across different chains. The recall errors were primarily because of the following reasons:

1. Most of the recall errors are because of insufficient domain knowledge. For example, in one of the documents, two mentions "the mra of the intracranial circulation" and "mra of the head" belong to the same chain. However, our system failed to predict the coreference relation between these two mentions.
2. In the case of PER chains, the recall errors are primarily because of personal pronouns. For example, for some of the documents, it is difficult to decide who the narrator of the document is. For such documents, our system does not generate the coreference links between first person pronouns and the proper names which leads to recall errors.

Lessons Learned: Most of the recall errors can be handled by adding extra domain knowledge. To take care of the precision errors referred to above, we need to collect more information about a mention than we currently have. For example, it would be very helpful to know the relations [20,21] in which a

mention participates to determine coreference. Medication information [22] about TRE mentions would also be helpful.

DISCUSSION AND CONCLUSION

We participated in 2011 i2b2/VA coreference resolution challenge. All the top scoring teams (including ours) in the challenge were very close to one another in terms of performance on the coreference task. However, different teams used quite different approaches to do coreference resolution. Some of the systems were rule-based, others were supervised and the rest were hybrid. Our system had the maximum precision among all the submitted systems. However, its recall was 0.1 lower than the top performing system. This suggests that our system is more conservative in predicting coreference relation and in future, we would explore different strategies to increase the recall. Overall, our system lagged behind the top-scoring system just by 0.05 unweighted average F1 score.

Distinguishing characteristics of our approach to coreference resolution were to use (1) mention parsing to abstract over the surface form of the mention and (2) to develop the notion of discourse model for clinical narratives. Both these ideas are quite general and can be readily applied to coreference resolution in other domains as well.

Some of the participating systems in the i2b2/VA coreference challenge used a multi-pass sieve where more precise features are used before the less precise ones. We are hopeful that integrating such an approach into our system would help us to improve the recall of our system. To resolve the precision errors made by our system, we need to know the relations in which a mention participates. So, an interesting extension of our work would be to do relation identification and coreference resolution jointly.

FIGURE LEGEND

Figure 1: figure1_monoImage.eps

We have provided the figure files along with the manuscript.

ACKNOWLEDGMENTS

The authors would also like to thank the i2b2 National Center for Biomedical Computing for organizing the 2011 challenge and providing the data used in these experiments. We would also like to acknowledge the anonymous reviewers for their valuable suggestions.

FOOTNOTES

Competing Interests: None

Funding: This research was supported by Grant HHS 90TR0003/01 and by the Intelligence Advanced Research Projects Activity (IARPA) Foresight and Understanding from Scientific Exposition (FUSE) Program via Department of Interior National Business Center contract number D11PC2015. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the HHS, IARPA, DoI/NBC or the US government.

Contributorship Statement: Both the authors contributed to system design and drafting or revising the article. System development was primarily carried out by first author.

REFERENCES

(1) Bodnari A, Szolovits P, Uzuner Ö. MCORES: a system for noun phrase coreference resolution for clinical records. Journal of the American Medical Informatics Association 2012;**Forthcoming**.

(2) Raghunathan K, Lee H, Rangarajan S, et al. A multi-pass sieve for coreference resolution. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2010: 492-501.

(3) Lee H, Peirsman Y, Chang A, Chambers N, Surdeanu M, Jurafsky D. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. ACL HLT 2011 2011;73.

(4) Culotta A, Wick M, Hall R and McCallum A. First-order probabilistic models for coreference resolution. Proceedings of NAACL HLT. 2007: 81-88.

(5) Bengtson E and Roth D. Understanding the value of features for coreference resolution. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2008: 294-303.

(6) Haghighi A and Klein D. Simple coreference resolution with rich syntactic and semantic features. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3. Association for Computational Linguistics. 2009: 1152-1161.

(7) Miyao Y, Sætre R, Sagae K, Matsuzaki T and Tsujii J. Task-oriented evaluation of syntactic parsers and their representations. Proceedings of ACL-08: HLT. 2008: 46-54.

(8) Gildea D. Corpus variation and parser performance. Proceedings of EMNLP. 2001.

(9) McCrae J, Collier N. Synonym set extraction from the biomedical literature by lexical pattern discovery. BMC Bioinformatics 2008;**9**:159.

(10) Pantel P and Ravichandran D. Automatically labeling semantic classes. Proceedings of HLT/NAACL. 2004: 321-328.

(11) Uzuner Ö, Bodnari A, Shen S, Forbush T, Pestian J, South B. Evaluating the state of the art in coreference resolution for electronic medical records. Journal of the American Medical Informatics Association. 2012;**Forthcoming**.

(12) Bagga A and Baldwin B. Algorithms for scoring coreference chains. In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference. Citeseer. 1998: 563-566.

(13) Vilain M, Burger J, Aberdeen J, Connolly D and Hirschman L. A model-theoretic coreference scoring scheme. Proceedings of the 6th conference on Message understanding. Association for Computational Linguistics. 1995: 45-52.

(14) Luo X. On coreference resolution performance metrics. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2005: 25-32.

(15) UMLS

Available at: http://www.nlm.nih.gov/research/umls/. Accessed Oct 24, 2011.

(16) Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. 2010;**17**:229.

(17) The Specialist NLP Tools

Available at: http://lexsrv3.nlm.nih.gov/Specialist/Summary/lexicalTools.html. Accessed Sep 02, 2011.

(18) Kubler S and Zhekova D. Singletons and Coreference Resolution Evaluation. Proceedings of the International Conference on Recent Advances in NLP (RANLP). 2011: .

(19) Koehn P. Statistical significance tests for machine translation evaluation. Proceedings of EMNLP. 2004: 388-395.

(20) Uzuner O, Mailoa J, Ryan R, Sibanda T. Semantic relations for problem-oriented medical records. Artificial Intelligence in Medicine 2010;**50**:63-73.

(21) Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association 2011;**18**:552-556.

(22) Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. Journal of the American Medical Informatics Association 2010;**17**:514-518.