Using Knowledge and Constraints To Find the Best Antecedent

Prateek Jindal *Dan Roth*
Dept. of Computer Science, UIUC
201 N. Goodwin Ave., Urbana, IL - 61801
{jindal2, danr}@illinois.edu

Abstract

Coreference resolution is the problem of clustering mentions into entities and is very critical for natural language understanding. *This paper studies the problem of coreference resolution in the context of the newly emerging domain of Electronic Health Records (EHRs).* The commonly used "best-link" model for coreference resolution considers only the scores from a pairwise classifier in selecting the best antecedent. In this paper, we extend this model to include several constraints derived from surface-form of the mentions and the context in which they appear. Another major contribution of this paper is to show the use of domain-specific knowledge sources, mention parsing and clinical descriptors in deriving features which contribute to improved coreference resolution performance. We present experiments on 4 different clinical datasets illustrating that our approach outperforms a strong baseline and a state-of-the-art system by a wide margin.

Keywords: Natural Language Processing, Information Extraction, Coreference Resolution, Electronic Health Records, Knowledge Based Systems.

Proceedings of COLING 2012: Technical Papers, pages 1327–1342,
COLING 2012, Mumbai, December 2012.

1 Introduction

The HITECH (Health Information Technology for Economic and Clinical Health) Act, part of the 2009 economic stimulus package (American Recovery and Reinvestment Act) passed by the US Congress, aims at inducing more physicians to adopt *Electronic Health Records (EHRs)*. An EHR is an evolving concept defined as a systematic collection of electronic health information about individual patient. Ability to automatically extract information from EHRs lies at the heart of several applications.

This paper addresses the task of coreference resolution for EHRs. *Coreference resolution is the task of finding referring expressions in a text that refer to the same entity, i.e., finding expressions that corefer*. The set of coreferring expressions is called as a coreference chain. Consider the following text sampled from one of the EHRs in the corpus used by us:

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

In the above text, all the phrases which are shown in brackets refer to the same entity and hence form a coreference chain. It is clear that identifying such coreference chains requires a lot of medical knowledge. For example, we need to know that "mass" can refer to a "malignant histiocytoma".

Most of the work on coreference resolution has focussed on the news text. Several different architectures have been proposed for coreference resolution. Recently, entity-based models for coreference resolution have been proposed. Such approaches try to directly model the entities in the text and usually involve some kind of global inference and tend to be quite complex. However, most of the best results on coreference resolution were achieved with simpler architectures which use a pairwise classifier between mentions and a decoding strategy like "closest-first" or "best-link" to first find the best antecedent for every mention. This step is then followed up by an inference procedure in which coreference chains are formed (Chang et al., 2011; Pradhan et al., 2011).

In this paper, we extend the "best-link" model to include several constraints derived from surface-form of the mentions and the context in which they appear. Another contribution of this paper is to show the use of domain-specific knowledge sources (like UMLS[1], MetaMap), mention parsing and clinical descriptors (obtained from medical ontologies) in deriving the features which are helpful for coreference resolution. In clinical Information Extraction (IE), researchers often map clinical text to UMLS concepts (Zheng et al., 2012; Rink et al., 2012). But such mapping alone doesn't allow an IE system to exploit the useful information contained in the parent trees of the concepts. Clinical descriptors designed by us overcome this limitation. We use two medical ontologies, MeSH[2] and SNOMED CT[3] to design our descriptors.

We conducted experiments on four different clinical datasets. Our results show that knowledge sources help in improving the recall and constraints help to increase the precision of the system. Knowledge and constraints used by us helped us to achieve significant performance improvements over a strong baseline derived from existing state-of-the-art approaches.

[1] http://www.nlm.nih.gov/research/umls/
[2] http://www.nlm.nih.gov/mesh/meshhome.html
[3] http://www.ihtsdo.org/snomed-ct/

To summarize, the key contributions of our paper are as follows:

- This paper studies coreference resolution on the new and important domain of EHRs.
- This paper presents different knowledge sources which would be useful for Information Extraction in medical and clinical domains. We use medical ontologies (MeSH, SNOMED CT) to get clinical descriptors which encode useful information contained in the parent trees of the concepts.
- We propose a rich local model to find the best antecedent.
- We use *mention parsing* to obtain a semantic representation of the mentions. Similar technique can also be used for other domains.
- Our system outperforms a strong baseline on four different clinical datasets.

2 Task Description

Coreference resolution aims at clustering together textual mentions within a single document based on underlying referent entities. For our experiments, we used the datasets provided by i2b2 team as part of coreference challenge. *We use the same problem definition as was specified in the i2b2 coreference challenge.* Mentions have already been identified and classified into 4 types : test (TEST), treatment (TRE), problem (PROB) and pronoun (PRON). Coreference relation can exist only within the mentions of same type. However, PRON mentions can corefer with any other mention. Given the entity mentions along with the types, the aim is to build coreference chains for the first 3 types: TEST, TRE and PROB. Since PRON mentions can corefer with the mentions of other types, there are no separate pronoun (PRON) chains.

3 Coreference Model

In this paper, we view coreference resolution as a graph problem: Given a set of mentions and their context as nodes, generate a set of edges such that any two mentions that belong in the same equivalence class are connected by some path in the graph. We construct this entity-mention graph by finding out the best antecedent of each given mention (anaphor) such that the antecedent belongs to the same equivalence class as the anaphor. The "Best-Link" strategy (Ng and Cardie, 2002; Bengtson and Roth, 2008; Chang et al., 2011) for selecting the antecedent of a mention chooses as the antecedent that candidate which gets the maximum score according to a pairwise coreference function pc. We extend the "Best-Link" strategy by including several constraints in its objective function as shown below.

3.1 Decision Model: Constrained Best-Link

Given a document d and a pairwise coreference scoring function pc that maps an ordered pair of mentions to a value indicating the probability that they are coreferential, we generate a coreference graph G_d according to the following decision model:

For each mention m_i in document d, let B_{m_i} be the set of mentions appearing before m_i in d. Thus, $B_{m_i} = \{m_1, m_2, ..., m_{i-1}\}$. Let a be the highest scoring antecedent. Then, we have:

$$
\begin{aligned}
a &= \underset{m_j \in B_{m_i}}{\arg\max} \quad score_i(m_j) \\
&= \underset{m_j \in B_{m_i}}{\arg\max} \quad k_1 \cdot pc(m_j, m_i) - d(m_j, m_i) + \sum_{l=1}^{L} C_l(m_j, m_i)
\end{aligned} \tag{1}
$$

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

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

If $score_i(a)$ is greater than a threshold δ, then we add the edge (a, m_i) to the coreference graph G_d. Threshold parameter δ is chosen to be $\frac{k_1}{2}$. Value of $pc(m_j, m_i)$ lies between 0 and 1. The value of k_1 is chosen to be sufficiently greater than 1 so that the pairwise classifier is given preference over the distance term in choosing the best antecedent. But if the pc values of any two candidates are almost similar, then the antecedent which is closer to the anaphor gets the higher score because of the distance term in Equation (1). Thus, our decision model combines the advantages of both "best-link" and "closest-first" models which are generally used for coreference resolution. Setting $k_1 = \infty$ and $L = 0$ reduces our model to the standard "best-link" decision model.

p_l is the penalty associated with the l^{th} constraint. Thus, different constraints can have different penalties. Higher the penalty associated with the constraint, the stronger it is enforced. If $0 < p_l < \frac{k_1}{2}$, then the constraint is soft because violation of such constraint by a mention pair doesn't necessarily rule it out. But if $p_l > \frac{k_1}{2}$, then the constraint becomes hard.

The resulting graph produced by the decoding technique mentioned above contains connected components, each representing one equivalence class, with all the mentions in the component referring to the same entity. Equivalence classes are determined by taking the transitive closure of all the links.

3.2 Pairwise Coreference Function

We train 4 classifiers, one each for TEST, TRE, PROB and PRON classes. Each of these classifiers takes as input an ordered pair of mentions (a, m) such that a precedes m in the document, and produces as output a value that is interpreted as the conditional probability that a and m belong in the same equivalence class. For any mention-pair (a, m), the classifier is chosen based on the type of mention m.

For each mention m we select from m's equivalence class the closest preceding mention a and present the pair (a, m) as a positive training example to the classifier which corresponds to the type of mention m. For each m, we generate negative examples (a, m) for all mentions a that precede m and are not in the same equivalence class.

We learn the pairwise classifiers using LIBSVM package (Chang and Lin, 2011).

4 Baseline

In this section, we describe the baseline system used by us. We designed the baseline system based on the existing state-of-the-art coreference systems which use pairwise models (Bengtson and Roth 2008; Haghighi and Klein 2009). Baseline system uses the coreference model as described in the previous section. However, there are no constraints in the baseline system. The features used for training the pairwise classifier have been described below. All the features used by us take only two values: 1 (if the feature is active) or 0 (if the feature is not active).

4.1 Lexical Features

Lexical features indicate whether two strings share some property. These features are listed below:

- Both the mentions have identical surface forms (i.e. $extent_{m_i} == extent_{m_j}$).
- Surface form of one of the mentions is a proper substring of that of another.
- Both the mentions share the same head word.

4.2 Syntactic Features

We check for the presence of several syntactic constructs among the mentions and generate the following features which tell whether or not the given mention pair satisfies the constructs:

- *Apposition*: Two noun phrases (NPs) are appositive when they are placed side-by-side with one element serving to define or modify the other e.g. *In a recent examination, the patient was diagnosed with **[medulloblastoma]**, **[a malignant brain tumor]**.*
- *Predicate Nominative*: The predicate nominative is the noun following a linking verb that restates or stands for the subject e.g. ***[Coronary Arteriosclerosis]** is a **[heart disease]** which happens when the coronary arteries become narrowed.*
- *Relative Pronoun*: It is a pronoun that modifies the head of the antecedent NP e.g. *After discussion , an **[abdominal CT scan]** was obtained **[which]** revealed diffuse metastatic lesions of the ribs .*

4.3 Semantic Features

Some of the coreferential mention pairs have similar but not identical heads. To find out whether any two words are similar or not requires semantic knowledge. Wordnet has been extensively used as a source of semantic knowledge for general English text. We generate the following two features from Wordnet:

- *Wordnet-head-match*: We get the synsets of heads of both the mentions and see whether the heads share any common synset. For example, the words *hemorrhage* and *bleeding* share the same synset which refers to the *flow of blood from a ruptured blood vessel*.
- *Wordnet-head-hypernyms-match*: Some closely related words (like *epistaxis* and *hemorrhage*) do not share any common synset. However, if we consider the parents (or hypernyms) of the synsets of such words in the Wordnet hierarchy, we can see that the two words are similar. We take only the immediate hypernyms of the synsets. Inclusion of hypernyms which are more than 1 level above the synsets of the words leads to over-generalization. For example, we may get that *nausea* and *anemia* are coreferent which is actually not true.

4.4 Distance-Based Features

We used the following distance-based features:

- *Adjacent-Mentions*: This feature is active if there is no intervening mention between the given mentions which has the same type as the mentions under consideration.
- *Distant-Mentions*: This feature is active if the two mentions are separated by more than 2 sentences.

5 Using Domain-Specific Knowledge

One of the major limitations of the baseline system is that it lacks 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 system is not sufficient to

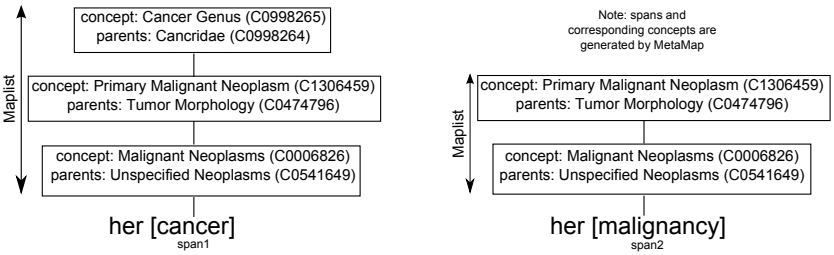

Figure 1: This figure shows the UMLS mappings for two mentions "her cancer" and "her malignancy". The terms "cancer" and "malignancy" have at least one common concept. Our matching procedure based on UMLS correctly predicts the given mentions to be coreferent.

address the ambiguity and variability that exists in medical terminology. To improve the performance of coreference resolution, we extended the baseline system by incorporating domain-specific knowledge into it.

5.1 UMLS and MetaMap

The UMLS (UMLS, 2012), or Unified Medical Language System, is a set of files and software that brings together many health and biomedical vocabularies. MetaMap (Aronson and Lang, 2010) is a configurable program which maps biomedical text to the UMLS Metathesaurus. We use the mapping provided by MetaMap to represent the mentions in a standard way which allows for effective matching of the mentions. We find the parents of concepts using the Web Service provided by UTS (UTS, 2012) (UMLS Technology Services).

Matching Mentions Using UMLS: We would refer to the surface forms of the two mentions by $s1$ and $s2$. First, we remove the stopwords from the given strings and then process the resulting strings using MetaMap and thus, get the mappings of the mentions to UMLS concepts. Next, we check whether any two spans (given by MetaMap) of $s1$ and $s2$ are equivalent. Two spans are considered equivalent if they share the same UMLS concepts (or parents of UMLS concepts). Whenever we find two equivalent spans, we remove them from $s1$ and $s2$. Finally, we check whether the resulting strings $s1$ and $s2$ match trivially. Two strings match trivially if they are identical or one of them is a substring of the other.

Consider Figure 1 for an example. This figure shows the UMLS mappings for two mentions "her cancer" and "her malignancy". "her" is considered as a stopword and is first removed from both the strings. Since the two spans "cancer" and "malignancy" share same UMLS concepts, they are equivalent. So, we remove "cancer" and "malignancy" from the two strings. The resulting strings are both empty and are considered to be matching.

Features Derived: Based on the matching procedure described above, we derive the following two features:

- *UMLS-Match*: In this feature, we do not consider the parents of the concepts during the matching
- *UMLS-Match-Parents*: In this feature, parents of the concepts are also considered during the matching

5.2 Mention Parsing

We parsed the mentions to extract the components like Modifiers, Body Parts and Anatomical Terms of location (ATs). We did not require exact match for these extracted components. We just specified that these components should not be incompatible with each other. The remaining portions of the surface forms of mentions were canonicalized and matching procedure described in Section 5.1 was used to determine whether they matched. Figure 2 shows an example where the structures obtained by parsing the two mentions are matching to one another.

Canonicalization referred to above involves the following two steps:

- *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[4], Medilexicon[5] etc. For ambiguous abbreviations, we considered all possible expansions. If a match was found using any of the expansions, then coreference pair was considered valid.
- *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 were found to be very 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 unannotated training data in an unsupervised setting.

Features Derived: Following feature was derived from mention parsing:

- *Mention-Parsing*: This feature is true for the mention pairs which match according to the mention parsing procedure described above.

5.3 Clinical Descriptors

MeSH[6] (Medical Subject Headings) is the National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms in a hierarchical structure that permits searching at various levels of specificity. We obtain MeSH descriptor for a concept in the following way:

1. First of all, we get all the paths from the concept to the root of MeSH hierarchy. In general, there can be more than 1 paths.
2. Then we construct one list which consists of the top 4 parents of all the paths obtained in Step 1.
3. The list obtained in step 2 is pruned where more preference is given to those parents which appear more frequently.
4. The final list obtained in step 3 is the MeSH descriptor of the concept.

Similar procedure is used to obtain the SNOMED CT[7] descriptor of a concept. SNOMED CT (SNOMED Clinical Terms) is yet another medical ontology which consists of the most comprehensive, multilingual clinical healthcare terminology in the world. SNOMED CT is owned,

[4] http://en.wikipedia.org/wiki/List_of_medical_abbreviations
[5] http://www.medilexicon.com/medicalabbreviations.php
[6] http://www.nlm.nih.gov/mesh/meshhome.html
[7] http://www.ihtsdo.org/snomed-ct/

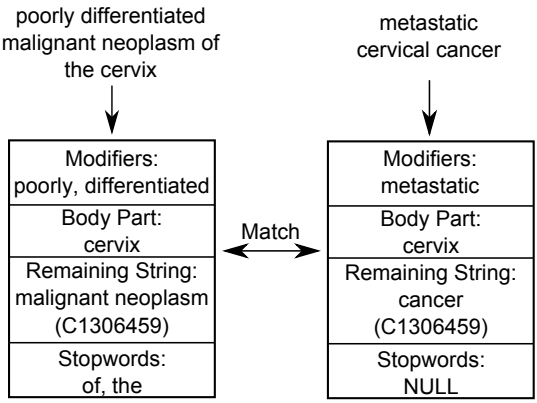

Figure 2: This figure shows the structures obtained by the mention parsing of two mentions shown on the top of the figure. Since the two structures match one another, we predict the two mentions to be coreferent.

maintained and distributed by the International Health Terminology Standard Development Organisation (IHTSDO). Figure 3 and Figure 4 show two different paths in MeSH and SNOMED CT parent trees for the same concept "Myocardial Infarction". We found that, in general, SNOMED CT gives much more paths (from concept to root of hierarchy) than MeSH. Some concepts in SNOMED CT have more than 300 possible paths to the root of hierarchy.

Features Derived: Based on the clinical descriptors described above, we derive the following two features:

- *MeSH-Match*: This feature is active if the Mesh descriptors of two concepts are the same
- *SNOMEDCT-Match*: This feature is active if the SNOMED CT descriptors of two concepts are the same

6 Description of Constraints

Constraints are used to model domain knowledge and they refer to those conditions which, if not satisfied, strongly indicate that the given mention pair is not coreferential. Features, on the other hand, can be more vague and don't necessarily provide such a strong clue. Constraints are applied only during the inference phase and not the learning phase. So, constraints can be added or removed without having to retrain the classifiers. Even if a particular constraint is not seen very often in the training data, it can still be very useful at the test time if the testing data contains cases where the constraint is applicable. This is a clear advantage of modeling constraints separately from the features. We divide the constraints in two categories depending on whether the constraint is derived from the surface form of the mentions or from the context in which the mentions occur. Constraints used by us are described in the following subsections. These constraints were obtained by the manual examination of small portion of training data.

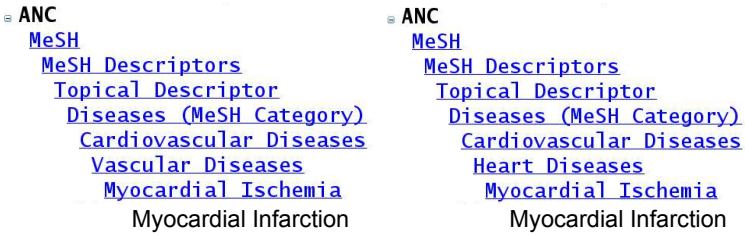

Figure 3: Figure showing two different paths in MeSH parent tree for the concept "Myocardial Infarction"

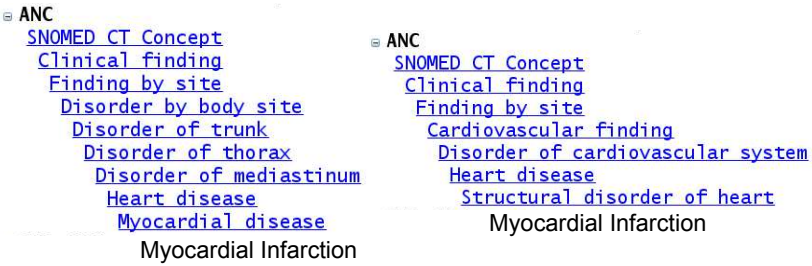

Figure 4: Figure showing two different paths in SNOMED CT parent tree for the concept "Myocardial Infarction"

6.1 Surface Form Constraints

Following surface form constraints were used by us:

- *Length Constraint*: Surface form of both the mentions must be at least 2 characters long.
- *Modifier Constraint*: Mentions should not have incompatible modifiers e.g. "small/large"
- *Body Parts Constraint*: If body parts (like chest, arm, head) are specified, they should not be incompatible.
- *Anatomical Terms Constraint*: If anatomical terms[8] (like proximal, anterior, dorsal) are specified, they should not be incompatible.
- *Popular Head Constraint*: Certain head words like "disease" occur very commonly in the dataset. Mentions which have same popular head are considered coreferential only if the classifier predicts the mentions to be coreferential even after removing the heads from the mentions.
- *Number Constraint*: Two mentions must agree in number.
- *Temporal Constraint*: If only one of the mentions contains the word "follow-up" or "repeat", then the mention pair is not considered coreferential because the two mentions refer to tests or treatments which have been done at different times.

[8] http://en.wikipedia.org/wiki/Anatomical_terms_of_location

6.2 Contextual Constraints

Following contextual constraints were used by us:

- *Family History*: If the left context of any mention (in a window of size 4) contains the phrase "family history", then the mention pair is not considered coreferential because one of the mentions refers to some family member of the patient and not the patient himself. Window size was determined using cross-validation on the training set.
- *Negation Constraint*: None of the mentions should be present in a negated form.
- *PRN Constraint*: Problem mentions which have "p.r.n." as the prefix can't participate in coreference relation because such mentions refer to hypothetical problems and not the real problems. For example, "p.r.n. headache" means "if the headache arises ...".
- *TEST Constraint*: We observed from the documents in the training data that the TEST mentions which appear under the heading "LABORATORY DATA" generally don't participate in coreference.

Other than the above mentioned constraints, following additional constraint was used to disallow coreference chains beginning with pronouns.

- In Equation (1), if m_j is a pronoun, then there must exist some mention m_k with $k < j$ such that m_k is a valid antecedent of m_j.

7 Experimental Setup

Datasets: For our experiments, we used the coreference datasets made available by i2b2 team as part of 2011 i2b2 challenge. The datasets consist of EHRs from three different organizations: Partners HealthCare (Part), Beth Israel Deaconess Medical Center (Beth) and University of Pittsburgh (Pit). The data from University of Pittsburgh is divided into 2 parts, namely Discharge and Progress records. All records have been fully de-identified and manually annotated for coreference. This gave us a total of 4 datasets. We would refer to these datasets as *Part*, *Beth*, *PitD* and *PitP* in the following discussion.

The total number of documents in the training set of *Part*, *Beth*, *PitD* and *PitP* are 136, 115, 119 and 122 respectively. Test set of *Part*, *Beth*, *PitD* and *PitP* contains 94, 79, 77 and 72 documents respectively. For more information about the datasets, please refer to Uzuner et al. (Uzuner et al., 2012). We used B-cubed (Bagga and Baldwin, 1998), MUC (Vilain et al., 1995) and CEAF (Luo, 2005) as the evaluation metrics in our experiments.

Choice of Parameters: We use cross-validation on the training data to determine the system parameters. In Equation (1), we set $k_1 = 100$. With this choice of k_1, distance term becomes significant only if the scores given by pairwise classifier for different mention pairs differ by less than 0.01. Since all our constraints are important to be enforced, we chose $p_l = 100$ in Equation (2) for all values of l. This choice of penalty parameters makes all the constraints hard.

8 Results

Table 1 compares the performance of four systems (1) Baseline (B), (2) Baseline + Knowledge (BK), (3) Baseline + Knowledge + Constraints (BKC) and (4) Baseline + Constraints (BC). We compare the performance of these systems for *Test*, *Treatment* and *Problem* categories on 4 different datasets, namely, *Part*, *Beth*, *PitD* and *PitP*. Table 1 reports precision (P), recall (R) and F1 scores for MUC evaluation metric. For B-cubed and CEAF Evaluation metrics, we only show the F1 scores because of space limitation. Please note that there are no separate scores for PRON category because there

are no separate PRON chains. PRON mentions are included within the TEST, TRE and PROB chains. Results shown in Table 1 are quite interesting and are explained below.

It is interesting to note that adding knowledge to the system always leads to higher recall values. On the other hand, addition of constraints always leads to higher precision values. Next, we note that different metrics behave differently in evaluating the performance of the systems. B-cubed metric gives higher F1 scores that CEAF metric which in turn gives higher F1 scores than MUC metric. This is because of the presence of large number of singletons in the corpora. B-cubed metric highly awards the correct prediction of singletons. MUC, on the other hand, is totally insensitive to singletons. CEAF is intermediate between B-cubed and MUC as far as singletons are concerned.

Next, we note the following major points about each category of mentions. For statistical significance tests, Bootstrap Resampling Test (Koehn, 2004) was used at $p = 0.05$.

1. ***Test***: For *Test* mentions, the best configuration is Baseline+Constraints (BC). For MUC metric, both BKC and BC performed the best for 2 corpora each. However, for B-cubed and CEAF evaluation metrics, BC performed the best for all the corpora. Hence, overall, we can say that BC is the best configuration for *Test* mentions. This is because of the fact that coreference for *Test* mentions (like "his ct scan", "a mammogram" etc.) can generally be easily predicted simply by looking at the surface forms. Also, many of the *Test* coreference chains are quite short with only 2-3 mentions which occur close to one another. So, knowledge is not so helpful for *Test* mentions.
2. ***Treatment***: For *Treatment* mentions, the best configuration is Baseline+Knowledge (BK). This is clearly evident from MUC metric. Only for *Beth* corpus, BKC performed better than BK but the difference is not statistically significant (67.8 vs 67.9). For B-cubed and CEAF evaluation metrics, the maximum F1 scores for *Treatment* category are quite close to Baseline scores and hence, the results are not statistically significant. Thus, B-cubed and CEAF metrics do not help much in predicting which system is better for *Treatment* mentions.
3. ***Problem***: For *Problem* mentions, the best system is Baseline+Knowledge+Constraints (BKC). This is clearly evident from B-cubed and CEAF Evaluation metrics. For MUC evaluation metric, BK performed better than BKC for 2 corpora. However, the difference in such cases is not statistically significant (69.1 vs 69.3 and 58.3 vs 58.4). Thus, we see that both, knowledge and constraints, benefit *Problem* mentions. This is due to the fact that *Problem* mentions, in general are quite long and complicated. *Problem* mentions generally occur with modifiers and have variegated surface forms. For example, "the patient's low potassium level" is coreferrent with "postoperatively hypokalemia".

Finally, in Table 2, we show the comparison of our system with a state-of-the-art system, Ware et al. (Ware et al., 2012), which used same test settings as ours. The numbers reported in Table 2 refer to the unweighted average of Bcubed, MUC and CEAF F1 scores computed across all the 4 corpora. We chose unweighted average for comparison because it was the official metric of i2b2 2011 shared task on coreference. We see from this table that our system consistently outperformed Ware et al.'s system for all categories of mentions.

9 Error Analysis

Table 3 shows the number of pairwise errors produced by our system on a portion of the test dataset. Rows indicate types of antecedent; columns are mention types. Each cell shows the number of precision/recall errors for that configuration. The total number of gold links is 2,252. We see that our system makes more precision errors than the recall errors. This is also confirmed by the results

	MUC Evaluation											
	B			BK			BKC			BC		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
						Test						
Part	30.4	84.8	44.8	29.3	88.8	44.0	32.4	85.8	47.0	33.8	83.8	**48.2**
Beth	13.2	70.2	22.2	13.9	77.8	23.6	16.3	71.7	**26.5**	16.0	67.5	25.9
PitD	29.4	82.7	43.4	28.9	86.3	43.3	30.4	81.0	44.2	30.9	79.2	**44.5**
PitP	25.1	79.3	38.1	25.7	86.0	39.5	28.6	80.2	**42.2**	27.4	74.4	40.1
						Treatment						
Part	58.1	81.4	67.8	57.7	86.0	**69.0**	58.1	83.1	68.4	58.1	79.4	67.1
Beth	57.9	79.2	66.9	57.4	82.7	67.8	58.5	81.0	**67.9**	58.5	78.0	66.9
PitD	51.0	72.1	59.7	51.7	77.4	**62.0**	52.4	74.7	61.6	51.3	70.9	59.5
PitP	55.8	72.2	63.0	55.4	76.5	**64.2**	55.9	74.7	64.0	56.0	71.2	62.7
						Problem						
Part	54.6	72.8	62.4	55.0	80.8	65.5	57.8	77.4	**66.2**	56.6	70.8	62.9
Beth	60.0	70.1	64.6	60.1	81.8	**69.3**	62.2	77.7	69.1	61.1	67.7	64.3
PitD	54.0	75.9	63.1	55.3	85.3	67.1	57.3	81.8	**67.4**	55.5	73.9	63.4
PitP	45.5	73.1	56.1	46.0	80.1	**58.4**	46.5	78.2	58.3	45.8	71.7	55.9

(a) MUC Evaluation

	B-Cubed Evaluation				CEAF Evaluation			
	B	BK	BKC	BC	B	BK	BKC	BC
	F1	**F1**	**F1**	**F1**	**F1**	**F1**	**F1**	**F1**
				Test				
Part	93.0	92.5	93.4	**93.8**	84.5	82.7	85.9	**87.4**
Beth	89.8	89.4	90.7	**90.9**	66.9	65.1	73.2	**74.4**
PitD	88.4	87.7	88.9	**89.3**	76.7	75.0	78.7	**79.8**
PitP	92.9	92.6	**93.2**	**93.2**	83.3	82.4	85.4	**85.6**
				Treatment				
Part	**92.8**	**92.8**	**92.8**	**92.8**	85.9	85.2	85.7	**86.0**
Beth	**92.5**	92.4	92.4	**92.5**	81.8	81.2	81.9	**82.1**
PitD	91.2	**91.5**	91.4	91.2	84.8	84.7	**84.9**	**84.9**
PitP	91.6	91.6	**91.7**	91.6	84.3	84.0	84.3	**84.4**
				Problem				
Part	92.3	92.3	**92.9**	92.6	84.9	84.8	**86.6**	86.1
Beth	92.2	92.3	**92.7**	92.3	83.7	83.8	**85.2**	84.3
PitD	90.5	90.6	**91.1**	90.8	83.1	83.6	**85.6**	84.5
PitP	93.7	93.6	93.8	**93.9**	85.3	85.1	**85.6**	**85.6**

(b) B-cubed and CEAF Evaluation

Table 1: This table compares the performance of four systems: (1) Baseline (B), (2) Baseline + Knowledge (BK), (3) Baseline + Knowledge + Constraints (BKC) and (4) Baseline + Constraints (BC). Part (a) of table reports Precision, Recall and F1 scores for MUC evaluation metric for TEST, TRE and PROB categories on 4 different datasets. Part (b) shows the F1 scores for B-cubed and CEAF evaluation metrics. For detailed discussion of the results, please refer to Section 8.

	Avg of B^3, MUC, CEAF F1		
	Test	**Treatment**	**Problem**
Ware et al.	68.4	79.4	80.8
This Paper	**69.1**	**80.7**	**81.6**

Table 2: This table shows the comparison of system presented in this paper with a state-of-the-art system, Ware et al. The numbers refer to the unweighted average of Bcubed, MUC and CEAF F1 scores computed across all the 4 corpora.

	TEST	TRE	PROB	PRON
TEST	92/60	-	-	22/12
TRE	-	187/186	-	57/14
PROB	-	-	320/293	70/21
PRON	32/10	47/13	97/37	20/7
Total	124/70	234/199	417/330	169/54

Table 3: This table shows the number of pairwise errors produced by our system on a portion of the test dataset. Rows indicate types of antecedent; columns are mention types. Each cell shows the number of precision/recall errors for that configuration. The total number of gold links is 2,252.

in Table 1. Error analysis of our system reveals that its precision can be improved by analyzing the context of the mentions more deeply. For example, it would be helpful to know the time (if mentioned) at which a particular test was conducted. It would be also beneficial to know whether a particular problem is mentioned in relation to the patient or one of his/her family members. On inspection, we found that our system made recall errors only on very difficult mention pairs. Predicting coreference relation among such mention pairs requires a lot of reasoning.

10 Related Work

For news text, several different architectures have been proposed for coreference resolution. Systems have been developed which allow for entity-level features or features over sets of noun phrases (Culotta et al., 2007). Such methods generally involve some kind of global inference which is difficult to implement and may also be intractable. Research (Finkel and Manning, 2008; Haghighi and Klein, 2007; Poon and Domingos, 2008) has also been carried out to explore how to reconcile pairwise decisions to form coherent clusters.

However, pairwise models with rich knowledge base have been shown to be very successful in both supervised and unsupervised setups (Bengtson and Roth, 2008; Haghighi and Klein, 2009). An important step in such models is to find the antecedent for each mention. For selecting the antecedent, "best-link" decoding strategy has been shown to give better results than "closest-first". In this paper, we extended the "best-link" strategy used by researchers by incorporating other factors like distance between mentions, several constraints etc. during the inference step.

There has been an increasing interest in knowledge-rich coreference resolution (Uryupina et al., 2011; Rahman and Ng, 2011; Bryl et al., 2010; Ng, 2010; Ponzetto and Strube, 2006; Bean and Riloff, 2004). Wikipedia is one of the most common knowledge resources that have been used by researchers. However, Wikipedia is not very good for clinical text because it doesn't have sufficient coverage of medical terms and also lacks precision. *In this paper, we used domain-specific knowledge sources like UMLS, MeSH and SNOMED CT to improve coreference resolution in clinical*

domain.

One of the earliest works in coreference resolution in clinical domain is that of Zheng et al. (Zheng et al., 2011). In this work, authors review recent advances in general purpose coreference resolution to lay the foundation for methodologies in the clinical domain. Later, Zheng et al. (Zheng et al., 2012) describe a simple pairwise classification technique for coreference resolution in clinical domain and got an overall B-cubed score of 0.69 and MUC score of 0.35. Bodnari et al. (Bodnari et al., 2012) and Jindal et al. (Jindal and Roth, 2012) also use a pairwise classification technique for clinical coreference resolution and use UMLS to get some of their semantic features. However, they don't use the concepts' parents information available in UMLS. Uzuner et al. (Uzuner et al., 2012) give a brief overview of several systems which participated in 2012 i2b2 coreference challenge. Most of the systems submitted in the challenge were rule-based. Rink et al. (Rink et al., 2012) used a multi-pass sieve architecture which is similar to the one developed by Raghunathan et al. (Raghunathan et al., 2010). Xu et al. (Xu et al., 2012) developed an effective strategy for pronoun resolution where they first determined the type of the pronoun and then chose the closest preceding concept of the same type as the antecedent. All these works assumed mentions' boundaries (along with their types) to be given just like ours.

Conclusion

Electronic Health Records are becoming increasingly important and their automatic analysis lies at the heart of several applications. This paper presented a system for coreference resolution for EHRs. In this paper, we proposed a rich model for selecting the best antecedent which involves inference using pairwise classifier scores and several constraints derived from surface-form of the mentions and the context in which they appear. We also showed the importance of domain-specific knowledge sources and clinical descriptors for achieving good performance in coreference resolution. While the knowledge sources used by us helped to improve the recall, constraints were helpful to increase the precision of system. Our experimental results show that different mention types benefit to different extent from knowledge and constraints. Our system consistently outperformed a strong baseline and a state-of-the-art system on four different datasets.

Acknowledgments

The authors would like to thank Ozlem Uzuner, Andreea Bodnari and Brett South for organizing the 2011 i2b2 challenge and providing the data used in these experiments. 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.

References

Aronson, A. and Lang, F. (2010). An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229.

Bagga, A. and Baldwin, B. (1998). Algorithms for scoring coreference chains. In *In LREC Workshop on Linguistics Coreference*, pages 563–566. Citeseer.

Bean, D. and Riloff, E. (2004). Unsupervised learning of contextual role knowledge for coreference resolution. In *Proc. of HLT/NAACL*, pages 297–304.

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

Bodnari, A., Szolovits, P., and Uzuner, Ö. (2012). Mcores: a system for noun phrase coreference resolution for clinical records. *J Am Med Info Assoc*, 19(5):906–912.

Bryl, V., Giuliano, C., Serafini, L., and Tymoshenko, K. (2010). Using background knowledge to support coreference resolution. In *Proceedings of ECAI 2010, August*.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

Chang, K., Samdani, R., Rozovskaya, A., Rizzolo, N., Sammons, M., and Roth, D. (2011). Inference protocols for coreference resolution. In *CoNLL Shared Task*, pages 40–44. ACL.

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

Finkel, J. and Manning, C. (2008). Enforcing transitivity in coreference resolution. In *Proceedings of 46th ACL-HLT Short Papers*, pages 45–48. ACL.

Haghighi, A. and Klein, D. (2007). Unsupervised coreference resolution in a nonparametric bayesian model. In *Annual meeting-Association for Computational Linguistics*, page 848.

Haghighi, A. and Klein, D. (2009). Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of EMNLP Volume 3*, pages 1152–1161. ACL.

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

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

Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of HLT and EMNLP*, pages 25–32. ACL.

Ng, V. (2010). Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of 48th ACL*, pages 1396–1411. ACL.

Ng, V. and Cardie, C. (2002). 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.

Ponzetto, S. and Strube, M. (2006). Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In *Proceedings of the NAACL*, pages 192–199. ACL.

Poon, H. and Domingos, P. (2008). Joint unsupervised coreference resolution with markov logic. In *Proceedings of EMNLP*, pages 650–659. ACL.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. (2011). Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. *CoNLL 2011*, page 1.

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., and Manning, C. (2010). A multi-pass sieve for coreference resolution. In *Proceedings of EMNLP*, pages 492–501. ACL.

Rahman, A. and Ng, V. (2011). Coreference resolution with world knowledge. In *Proceedings of the 49th ACL-HLT Volume 1*, pages 814–824. ACL.

Rink, B., Roberts, K., and Harabagiu, S. (2012). A supervised framework for resolving coreference in clinical records. *Journal of the American Medical Informatics Association*, 19(5):875–882.

UMLS (2012). http://www.nlm.nih.gov/research/umls/ (accessed aug 25, 2012).

Uryupina, O., Poesio, M., Giuliano, C., and Tymoshenko, K. (2011). Disambiguation and filtering methods in using web knowledge for coreference resolution. In *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference*.

UTS (2012). Umls technology services. https://uts.nlm.nih.gov/home.html (accessed aug 25, 2012).

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

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th MUC*, pages 45–52. ACL.

Ware, H., Mullett, C., Jagannathan, V., and El-Rawas, O. (2012). Machine learning-based coreference resolution of concepts in clinical documents. *J Am Med Info Assoc*, 19(5):883–887.

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

Zheng, J., Chapman, W., Crowley, R., and Savova, G. (2011). Coreference resolution: A review of general methodologies and applications in the clinical domain. *JBI*.

Zheng, J., Chapman, W., Miller, T., Lin, C., Crowley, R., and Savova, G. (2012). A system for coreference resolution for the clinical narrative. *J Am Med Info Assoc*.