Coreference analysis in clinical notes: a multi-pass sieve with alternate anaphora resolution modules

Siddhartha Reddy Jonnalagadda,1 Dingcheng Li,1 Sungghan Sohn,1 Stephen Tze-Inn Wu,1 Kavishwar Wagholikar,1 Manabu Torii,2 Hongfang Liu1

ABSTRACT
Objective This paper describes the coreference resolution system submitted by Mayo Clinic for the 2011 i2b2/VA/Cincinnati shared task Track 1C. The goal of the task was to construct a system that links the markables corresponding to the same entity.

Materials and methods The task organizers provided progress notes and discharge summaries that were annotated with the markables of treatment, problem, test, person, and pronoun. We used a multi-pass sieve algorithm that applies deterministic rules in the order of preciseness and simultaneously gathers information about the entities in the documents. Our system, MedCoref, also uses a state-of-the-art machine learning framework as an alternative to the final, rule-based pronoun resolution sieve.

Results The best system that uses a multi-pass sieve has an overall score of 0.836 (average of B3,M U C ,B l a n c ,a n d Zheng3 performed a comprehensive methodological review of coreference resolution in general English and argued that it may be possible to apply the same techniques in the clinical domain. They grouped existing methods largely into the following three types:

1. Heuristics-based approaches based on linguistic theories and rules3–9
2. Supervised machine learning approaches with binary classification of markable mention/entity pairs10–15 or classification by ranking markables16–17
3. Unsupervised machine learning approaches, such as non-parametric Bayesian models18 or expectation-maximization clustering.

In the current work, we employed a multi-pass sieve framework to exploit a heuristic-based approach along with a supervised machine learning method, specifically factorial hidden Markov models (FHMMs). Raghunathan et al20 developed a multi-pass system that applies tiers of resolution models one at a time. Each tier (sieve) consists of similar deterministic rules and builds on outputs of previously applied sieves. Sieves yielding a higher precision are arranged first in the system. Some of the sieves include: pronoun, head match, appositive, and demonym. Lee et al21 applied this system in the general domain for the CoNLL-2011 shared task for coreference analysis and produced the best performance. Their work is based on the ‘method of successive approximation’ for learning that was successfully used previously for named entity classification,22, 25 machine translation,24 and dependency parsing.25

Li et al23 used FHMMs for pronominal anaphora resolution.27 FHMMs are an extension of traditional hidden Markov models,20 where the hidden state at each time step t (ie, word ot) is expanded to contain more than one random variable. Their state-of-the-art anaphora resolution system uses features, such as part of speech, gender, grammatical number (singular/plural), and concept class. Figure 1 shows Li et al’s model, where the hidden states h at each word are factored into three components: coreference features cr, part-of-speech tags pos, and an operation variable op to control reference. This factorization allows the learning of complex hidden states even with limited training data.

The 2011 i2b2/VA/Cincinnati challenge28 focuses on coreferential relations between common, clinically relevant classes in medical text. These classes include problem, treatment, test, person, and pronoun. Coreferring mentions are to be paired together, and the pairs are to be linked to form a chain that represents the entity being referenced. The aim of the challenge is to produce coreferential chains of these mentions at document level (ie, coreference relations are made across paragraphs or sections within the same document, but not across documents).
This paper describes our coreference resolution system, MedCoref, developed by Mayo Clinic natural language processing (NLP) program for Track 1C. We developed a multi-pass sieve system in Java along the same lines for clinical notes by adapting the existing sieves and adding additional sieves, and then integrated these sieves with FHMM anaphora resolution. Additionally, we performed a thorough study on pronominal coreference resolution considering the two approaches. The code for our system, MedCoref, is available at https://sourceforge.net/projects/ohnlp/files/MedCoref under the unrestrictive open-source Apache v2 license. This enables hospital systems to use our system that leverages the benefits of the Stanford coreference resolution system combined with adaptations suitable for clinical narratives and allows them to adapt the system to their environment.

DATA
The Track 1C data consist of three sets from three different institutions: Partners HealthCare, Beth Israel Deaconess Medical Center, and the University of Pittsburgh. The data from the University of Pittsburgh contain two types of notes: discharge notes and progress notes. All protected health information is fully de-identified. In the training set, gold standard markables and chains are manually annotated. The training set contains a total of 492 notes (Partners: 136, Beth: 115, Pittsburgh: 119 discharge and 122 progress notes) and the test set contains a total of 322 notes (Partners: 94, Beth: 79, Pittsburgh: 77 discharge and 72 progress notes).

METHODS
The markables for coreference analysis could be classified (as per ACE guidelines, see http://projects.ldc.upenn.edu/ace/docs/English-Entities-Guidelines_v6.6.pdf) into proper mentions (proper names), nominal mentions (noun phrase whose head is a common noun), and pronoun mentions. In coreference analysis research and broader NLP research, deterministic hierarchical systems that apply rules in the order of precision are shown to be effective. On the other hand, NLP tasks, such as clinical concept extraction (mention detection), are additionally handled through machine learning approaches.

Figure 2 shows the system architecture. The eight sieves are analogous to inclusion criteria where at least one of them needs to be satisfied. The two filters are similar to exclusion criteria where even when one is matched, the mention pairs are not linked. Set-up C uses a rule-based pronoun sieve as the final step. Set-up A uses the FHMM-based sieve that is unaware of the mention clusters. For set-up B, we merge the chains in set-up A and C. In general English, rule-based systems were shown to be the most effective for coreference resolution. We investigated whether this is true for clinical narratives, that is, we resolved the coreference in proper mentions and nominal mentions using the initial sieves. The final sieve resolved the pronominal coreference using the information gathered about the entities (clusters of mentions). Alternatively, we used the system of Li et al. to resolve pronominal coreference. We not only compared the performance of the pronominal coreference methodologies individually (in addition to the other chains), but we also compared individual methodologies against chains merged from both methods.

Relationship detection order
The different sieves used in the system, according to the order of application, are displayed in figure 2. The mentions in each document are ordered by their appearance. For each sieve, the
For instance, mentions of a problem or treatment could be related to different persons because of the information recorded in the ‘family medical history’ section. As such, a non-chronic problem that a patient had previously or a test underwent previously as recorded in the ‘history of present illness’ section does not have a relationship with the current problem or test. Similarly, a treatment in the ‘current medications’ section need not be related to another one in the ‘discharge medications’ section. Clinical notes are often divided into sections, or segments, such as ‘history of present illness’ or ‘past medical history.’ Those sections can be helpful in identifying coreferred pairs. Intuitively, two mentions associated with the same term appearing in two sections, ‘history of present illness’ and ‘diagnosis,’ have a higher probability of being a coreferred pair than two mentions associated with ‘family history’ and ‘diagnosis.’

We adapted SecTag developed by Denny et al.35 to associate each sentence in the clinical notes to section headers. The sections that the mentions belong to, the class of the mention (eg, problem, treatment, etc), and a list of chronic problems are used to create the rule-based filter that rejects the relationships detected by sieves. The rules are part of the open-source code and some of them are represented in table 1.

**Vicinity filter**

Unlike proper mentions, nominal mentions in the same document could refer to completely different entities, as their primary role is to describe a closer antecedent proper mention. For example, consider the sentences in box 1. The ‘pathology’ in the second sentence and ‘pathology’ in the final sentence refer to different tests.

Hence, we designed a second filter that rejects relationships if the mentions only contain a list of stop terms compiled by us as part of the MedTagger project (see online supplementary file).

**Sieves**

Sieve 1 accepts mentions that match exactly when aligned to the right and the antecedent has a higher number of words. Since two mentions with the same name in a clinical document need not corefer, we found it helpful to perform a right-aligned match. This will be useful in scenarios, such as that shown in box 2 where the first and the last ‘echocardiograms’ are different.

Sieve 2 accepts a pair when the mention is a relative pronoun that is governed by the antecedent as detected by rules based on part-of-speech tags (the pair immediately following each other or intercepted only by a verb). The part-of-speech tags are assigned by the OpenNLP POS tagger trained for clinical text.36 It also accepts mention pairs where one of them is an abbreviation of another as detected using the abbreviation list assembled from the Unified Medical Language System (UMLS; version 2011AA) using the tool present in Liu et al.6 The medical domain favors brevity. Recognizing abbreviations is important for medical coreferential relationships are tested for each pair of mentions starting from the last appearing (probable) mention. For each mention, a probable antecedent is searched for starting from the closest mention. The assumption is that in narrative text, since there are less intervening words that could disturb the relationship. Such an assumption makes sense for the clinical narratives, which is a sublanguage that typically does not contain complex or nested sentences.35

**Section filter**

In general English, if two mentions have the same surface text, more than 95% of the time the mentions corefer.20 However, in clinical narratives, this might not be the case for several reasons.
language processing and information retrieval systems. Expanding acronyms to their full names can be helpful in coreference resolution. For example, detecting a coreferred pair, such as ‘CHF’ and ‘the failure’ is more difficult than detecting a coreferred pair such as ‘congestive heart failure’ and ‘the failure’ through language processing. If we expand an abbreviation to its corresponding full name, its coreferred mentions, if any, can be detected using traditional coreference techniques.

Sieves 5 through 6 are implemented in the same fashion as by Raghunathan et al. They take into account: (a) head match—whether the head of the mention matches the head of one of the mentions in the antecedent entity; (b) compatible modifiers—whether all the noun and adjective modifiers of the mention are present in a single mention of the antecedent entity; and (c) word inclusion—whether all the words in the mention are present among the words in the mentions from the antecedent entity. Sieve 5 requires all the above three conditions to match. Sieve 4 uses head match and word inclusion. Sieve 5 uses head match and compatible modifiers. Sieve 6 uses a relaxed head match where the head of the mention is present in any part of the mentions in the antecedent entity and word inclusion.

For sieves 2 and 7, we used synonyms and other relationships extracted from the UMLS. From each document, we extracted three sets of mentions by semantic type: tests, problems, and relationships of types: synonym (for sieve 2), and parent CUIs for a mention were connected to the CUIs for each in-set mention, using the UMLS MRREL table. We only considered relationships of types: synonym (for sieve 2), and parent—child and narrow—broad (sieve 7).

Sieve 7 uses the Porter Stemmer algorithm to stem words constituting the mentions and the open class words, such as prepositions and articles, are dropped. The mention pair is accepted as a coreferred pair if (1) the stems of the headwords are the same, and (2) the remaining stemmed words in one of the mentions are all in the other mention. For example, the mention ‘shortness of breathing’ (stemmed to ‘short breath’) is mapped to the antecedent mention ‘very short breath’ (stemmed to ‘very short breath’). Such an approach had been used previously in the general domain by Yang et al. and Zitouni et al. Stemming headwords is important because some medical terms refer to the same thing although they have different forms. The second criterion is based on the assertion that the modifiers of a noun phrase (eg, adjectives, prepositions, numbers, possessives, proper nouns, non-finites, and quantifiers) carry important information for coreference resolution. Two noun phrases with the same head string may refer to distinct entities if their modifiers do not match.

Sieve 8 for set-up C is a rule-based pronoun sieve. As shown in figure 2, the seven sieves that were applied prior to this sieve collect information about entities referred by the mentions in the same document. Based on the collected information pertaining to the entity’s grammatical number (singular/plural depending on the pronouns already in the mentions in the entity and part-of-speech tags), gender (male/female depending on the pronouns in the entity and the markable class), and animacy (person/object based on the pronouns and markable type), each pronoun is assigned to an antecedent entity when each of these features match.

FHMIs is sieve 8 for set-up A. Three adjustments to Li et al.’s model were made for the i2b2/VA/Cincinnati shared task. First, due to speed concerns, we do not incrementally copy or (coreference) features for words that are not mentions. This modification would have little impact because, as described in Li et al.’s paper, in a first-order HMM employed in the current system, neighboring words that are separated by more than one word are assumed independent. A corollary to this first adjustment is that the dependencies between neighboring words no longer exist—only those between neighboring mentions are kept. The model described in figure 1 is still applicable to the current system, where the i-th observation $o_i$ is now a mention rather than a word.

Second, we divide pronouns (as represented in the pos (part-of-speech) model) from the OpenNLP POS tagger into finer categories, such as non-personal pronouns or first/second/third-person pronouns. Training the pronoun resolution model at this granularity yields intuitive empirical information that the system may make use of, based on the discourse context of ‘clinical notes.’ Second-person pronouns refer to the patient in most cases. First-person pronouns are ambiguous and refer to either the patient or the care provider.

In the original model, an entity $e$ (a subvariable of $cr$) was a named entity from some recognition algorithm, such as the input concepts given in Track 1C. This is not useful given the first adjustment, and $e$ is therefore modified to represent the mention’s concept type (ie, problem, treatment, test, and person).

Evaluation metrics
A mention pair identified as belonging to the same entity is a true positive when that is confirmed by the gold standard; otherwise, it is a false positive. When a mention pair that belongs to the same entity as per the gold standard is not linked by the system, it is a false negative. A true positive or false positive occurs when at least one sieve detects the link, and vice
versa. Most metrics capture the notion of correctness through precision—the ratio of true positives among all the system outputs—and the notion of completeness through recall—the recall of true positives among the total number of positives. An F score represents the overall performance as a harmonic mean of the precision and recall.

In the i2b2/VA/Cincinnati challenge on coreference resolution, system performance was measured using MUC, B-CUBED (B), entity-based CEAF, and BLANC, similar to Semeval-2010 and CoNLL-2011. For official evaluation, an average of B-CUBED, CEAF, MUC, and MELA (mention, entity, and link average) was used, without including BLANC. Scores of the three metrics were averaged with equal weights, in the same manner as CoNLL 2011.

The reported performance measures were calculated using the Python script provided by the challenge organizers.

RESULTS

To evaluate performance of the machine learning-based pronoun sieve on each of the four training corpus parts in Track 1C, FHMM probability models were trained on the other three corpus parts using relative frequency estimation. For the test corpus, we combined all four training models.

The FHMM was evaluated using the ratio of the number of correctly resolved relationships over the total number of relationships, consistent with Li et al. and Ge et al. Table 2 shows the accuracy results on the training corpus.

The performance as per the evaluation metrics defined above of the different sieves and set-ups on the development set are shown in table 3.

After the initial right exact-match sieve, the recalls (for all metrics) gradually increased, with a slighter decrease in precision for proper and nominal mentions (sieve 2 contains relative pronouns). Altogether, the average F score does not increase substantially. Hence, we might conclude that the sieves contribute toward improving the system gradually.

However, as shown by the p values measured using the Student's t test between each sieve and its successive one in table 3, we conclude that the sieves contribute toward improving the system gradually.

DISCUSSION

Table 4 shows some of the true positives, false positives, and false negatives of the system.

The examples for true positives illustrate how the rules worked as we intended. Table 4 refers each true positive to the corresponding sieve defined in the Methods section and figure 2. The false positives occur mainly because of the lack of knowledge of semantics. For example, in the first false positive example, the ‘baseline creatinine’ and ‘creatinine’ are the same kind of tests conducted at different instances and so are considered different as per the definition of this task. Within our framework, it is possible to create a new filter that rejects such mention pairs based on the domain knowledge that a baseline test is different from a test conducted at a later instance. The false positives also occur because of the insufficient gathering of the context. For example, in the last example of the true positives, the mention ‘he’ is compatible with the mention ‘the patient’ and hence they are linked. In the last example of the false positives, although the mentions ‘this’ and ‘his prealbumin’ are grammatically compatible, the second mention refers to time and the first mention refers to test. When an aggregate system extracted the named entities of type time (in addition to test), this situation would be resolved.

The orders of the sieves themselves were adapted from the Stanford coreference system for general English. When we added a few completely novel sieves such as using UMLS relationships other than exact synonyms, stemming, and bag of words match, we added them at the end (right before pronominal resolution). However, one could independently investigate in the future by altering the order.

The performance of the machine learning-based pronoun sieve is 10% less than the corresponding performance for general English. This might be attributed to the distinguishing features of the clinical notes and an improvement in this performance might need the addition of features specific to these notes. Such features would take into account the differences between the various semantic types, medical specialties, and types of notes.

The purpose of the initial sieves is to gather global information about the entities in the document. After addition of the rule-based pronoun sieve, the average F score increases by 10.7%.
On the other hand, the performance increment after addition of the state-of-the-art machine learning-based pronoun sieve is only 7.7%. We believe that this is because our machine learning-based sieve learns features based on mentions and is unaware of the global properties of the entity (mention cluster) itself. Others such as Mitkov observe, ‘Machine learning algorithms for pronoun resolution do not necessarily perform better than the traditional rule-based approaches.’

While we have used supervised machine learning for clinical information extraction tasks, such as named entity recognition, association extraction, and drug adverse effect extraction, machine learning-based systems are still used sparingly at an enterprise level by Mayo Clinic and other organizations, such as Regenstrief Institute. Systems trained using supervised machine learning algorithms are often sensitive to the distribution of data, and a model trained on clinical notes from an institution may perform poorly on those from another. For example, Wagholikar et al. showed recently that a machine learning model for concept extraction trained on the i2b2/VA/Cincinnati corpus achieved a significantly lower F score when tested on the Mayo Clinic corpus. Other researchers recently reported this phenomenon for part-of-speech tagging. Such poor performances will then be cascaded to higher-level tasks, such as coreference resolution and semantic role labeling. Besides the inherent challenges pertaining to the peculiar sublanguage, the difficulty in applying machine learning to clinical NLP may be attributed to the difficulty in developing a corpus annotation standard across institutions and use cases, preparing large annotated corpora conforming to the standard, and limitation of data sharing in the domain. Hence, we chose the hybrid approach, where the deterministic framework allows experts to add rules or modify existing ones while taking advantage of machine learning techniques where possible.

With the i2b2/VA/Cincinnati shared task test corpus, the accuracy of MedCoref remained consistent (F score of 0.84). The best system, which is a machine learning system, has an F score of 0.92. The minimum F score for the task was 0.58. Our system scored at the median (exact median F score of 0.85) among the 20 teams that participated and ranked at 11. These results, including ours, are inflated to some degree, since it is not an end-to-end evaluation where named entities, such as treatment and problem, need to be automatically extracted from text. In addition, it would be sound but incomplete to evaluate NLP systems using a rather homogeneous corpus, especially for the case of clinical narratives that seem to have entirely different characteristics depending on where they originate. In practice, the portability and adaptability of a system is an important concern for clinical NLP applications.

There are two other rule-based systems in the competition besides ours (Hinote et al and Gooch et al). The rest used rules for preprocessing (Yang et al), deciding the order of the machine learning components (Rink et al) or used completely supervised approaches (Anick et al, Cai et al, Xu et al, etc.). Experts, through adding more rules (such as semantic clues like dates and locations by Hinote et al and Wikipedia abbreviations by Gooch et al), could further improve and locally customize MedCoref. This could readily accommodate using outputs from machine learning systems as well. Our system is flexible in

Table 4 Example outputs of the system

<table>
<thead>
<tr>
<th>Sentence 1</th>
<th>Sentence 2</th>
<th>Sieve</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positives</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The diagnosis, therefore, was relapsed C difficile colitis.</td>
<td>An abdominal CAT scan revealed thickened bowel wall and thumb printing,</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>primarily involving the cecum and right colon greater than the left,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>consistent with C difficile colitis.</td>
<td></td>
</tr>
<tr>
<td>2. Chronic pleural effusion.</td>
<td>Briefly, the patient has a history of chronic obstructive pulmonary disease,</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>ethanol abuse, chronic pleural effusions, and chronic renal insufficiency.</td>
<td></td>
</tr>
<tr>
<td>The patient is an 85-year-old white male with a history of ischemic bowel</td>
<td>With intravenous hydration the BUN and creatinine fell to 12/1.9 which is</td>
<td>3</td>
</tr>
<tr>
<td>status post recent admission for urosepsis and C difficile colitis.</td>
<td>within normal limits for this patient.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Initially treated with intravenous ceftriaxone, gentamicin, and Flagyl for</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>presumed sepsis, either with urine or bowel source.</td>
<td></td>
</tr>
<tr>
<td>Patient is a 28 year old gravida IV, para 2 with metastatic cervical cancer</td>
<td>Given the patient’s history of cervical cancer, the pleural effusion was</td>
<td>5</td>
</tr>
<tr>
<td>admitted with a question of malignant pericardial effusion.</td>
<td>felt most likely to be malignant.</td>
<td></td>
</tr>
<tr>
<td>The patient was alert and oriented throughout the admission; however,</td>
<td>The patient was alert and oriented throughout the admission; however,</td>
<td>6</td>
</tr>
<tr>
<td>by personality, he is somewhat cantankerous and demanding of the nurses.</td>
<td>by personality, he is somewhat cantankerous and demanding of the nurses.</td>
<td>7</td>
</tr>
<tr>
<td>False positives</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The patient has chronic renal insufficiency with baseline creatinine 1.8–2.</td>
<td>Creatinine had risen to 4.3 on admission presumed secondary to sepsis and</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>dehydration.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The patient has a history of ischemic bowel status post SMA Percutaneous</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Transluminal Coronary Angioplasty with recent admission for gram negative</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rod urosepsis complicated by C difficile colitis.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Continue to take the antibiotics as directed.</td>
<td>3</td>
</tr>
<tr>
<td>The patient had a PICC line placed and will continue a 8 week course of</td>
<td>18. Heparin Lock Flush (Porcine) 100 unit/ml Syringe Sig: Two (2) ML</td>
<td>4</td>
</tr>
<tr>
<td>antibiotics.</td>
<td>Intravenous DAILY (Daily) as needed: 10 ml NS followed by 2 ml of 100</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Units/ml heparin (200 units heparin) each lumen Daily and PRN.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>His prealbumin is up slightly from last week’s level of &lt;7 to 11 this week.</td>
<td>5</td>
</tr>
<tr>
<td>False negatives</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Chronic pleural effusion.</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>1. Colitis.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>She also received Cisplatin 35 per meter squared on 06/19 and Iflex and</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Mesna on 06/18.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

J Am Med Inform Assoc 2012;19:867—874. doi:10.1136/amiajnl-2011-000766
accommodating additional components and integrating different technologies and it is suitable for practical use.

CONCLUSION

We designed a multi-pass sieve system for coreference resolution in clinical notes. We demonstrated that, using relatively simple rules, basic part-of-speech information, and semantic type properties, an effective coreference resolution system could be designed. Pronominal coreference resolution is shown to be more accurate when an entity-centered approach is used rather than a mention-centered approach. The source code of the system described in this paper is available at https://sourceforge.net/projects/ohnlp/files/MedCoref.

Acknowledgments

This research was evaluated using the gold standard developed as part of the 2011 i2b2/VA/Cincinnati challenge.

Contributors

The seven authors are justifiably credited with authorship, according to the authorship criteria. SJ, HL: conception, design, development, analysis and interpretation of data, drafting of the manuscript, final approval given; DL: acquisition of data, analysis and interpretation of data, final approval given; SS, SW, KW: development, critical revision of the manuscript, final approval given; MT: critical revision of the manuscript, final approval given.

Funding

This work was funded by National Science Foundation ABI 0845223 and National Institute of Health R01LM009959A1 to Dr Hongfang Liu.

Competing interests

None.

Provenance and peer review

Not commissioned; externally peer reviewed.

Data sharing statement

The code and the accompanying data are available as open-source at https://sourceforge.net/projects/ohnlp/files/MedCoref.

REFERENCES


