An Infinite Mixture Model for Coreference Resolution in Clinical Notes

Sijia Liu¹, Hongfang Liu, PhD², Vipin Chaudhary, PhD¹, Dingcheng Li, PhD²
¹University at Buffalo, the State University of New York, Buffalo, NY;
²Mayo Clinic, Rochester, MN

Abstract

It is widely acknowledged that natural language processing is indispensable to process electronic health records (EHRs). However, poor performance in relation detection tasks, such as coreference (linguistic expressions pertaining to the same entity/event) may affect the quality of EHR processing. Hence, there is a critical need to advance the research for relation detection from EHRs. Most of the clinical coreference resolution systems are based on either supervised machine learning or rule-based methods. The need for manually annotated corpus hampers the use of such system in large scale. In this paper, we present an infinite mixture model method using definite sampling to resolve coreferent relations among mentions in clinical notes. A similarity measure function is proposed to determine the coreferent relations. Our system achieved a 0.847 F-measure for i2b2 2011 coreference corpus. This promising results and the unsupervised nature make it possible to apply the system in big-data clinical setting.

Introduction

The rapid increase in the volume of electronic health records (EHRs) has created a huge opportunity for clinical research and practice [1, 2]. It is widely acknowledged that applying natural language processing (NLP) to process the unstructured clinical narratives complements structural EHR data [3]. However, the full utility of NLP has been hampered by the poor performance in relation detection tasks. For example, lacking coreference (linguistic expressions pertaining to the same entity/event) resolution may lead to misclassification in named entity recognition, patient risk prediction, cohort identifications, clinical decision support and other clinical applications. It is found that current state-of-the-art medical coreference resolution systems still run into trouble in cases requiring domain knowledge [4]. Hence, there is a critical need to advance the research for relation detection from EHRs.

Coreference resolution is a challenging problem in the area of natural language processing (NLP). It tries to resolve the coreferent relations among different nouns, noun phrases and pronouns. Coreferent relations are common in natural language discourses in both EHRs and in general domains. For example, in the sentence of “He is minimally responsive today which appears to be near his baseline and does not have any specific complaints”, the pronoun “which” refers to the phrase “minimally responsive”. This is an example of how a pronoun is coreferent to its antecedent noun phrases. In another example, the phrase “his anticoagulation” might be mentioned as “anticoagulation therapy” later in a discourse, which is an example of how different noun phrases can be coreferent. The real world objects mentioned in a discourse by noun phrases or pronouns are usually called entities, and the phrases or pronouns appearing in discourses are called mentions. Mentions may consist of several words in free texts and each of such words is called a token.

The coreference resolution has been studied for years in NLP. There are mainly three kinds of solutions: rule-based methods, supervised machine learning methods and unsupervised learning methods [3, 5]. Rule-based, also known as heuristics–based methods tend to incorporate linguistic knowledge and lexical patterns to resolve coreferent relations. Supervised machine learning methods [6] use gold standard annotations as the input of machine learning models to train classifiers that can fit the input data [6, 7]. Based on how the mentions are organized for the classifiers, the two most widely used models for this approach are mention-pair model [8-10] and entity-mention model [11, 12]. Unsupervised nonparametric Bayesian models attempt to make assumptions of the generative process in discourses and use Gibbs sampling [13, 14] or Expectation-Maximization [15] to infer the model parameters. Unsupervised Bayesian methods [16-18] do not require human efforts such as annotation and labeling in data preparation stages which is its advantage in trending big data applications.

In clinical domain, a coreference dataset annotated from clinical notes was released with mention annotations and gold standard results by i2b2 challenge organizers in 2011 [4]. The best two systems in the challenge achieved more than 0.9 F measures and both of them used supervised machine learning methods [19, 20]. Although there are some
strong rules which can help us identify the coreferent relations [21], it is difficult to design a rule-based system for clinical applications due to the variety of discourse styles and the complexity of the hidden information which are not observable directly from their contexts [21-23]. According to the competition results, supervised learning methods outperformed rule-based methods in i2b2 clinical corpus.

However, currently there is no system utilizing unsupervised models in clinical domain. The complexity in their inference phases makes these models difficult to be deployed into implemented applications. In addition, these models have some drawbacks if applied in clinical discourses directly. In general NLP, the distribution of mention frequency is quite different from clinical discourses. There are much more singleton mentions in clinical notes and in most cases there will be a very long chain of patient mentions and other non-personal entities may appear not more than twice. Thus, if we look at the discourse as a generative process, the probability to a new entity in clinical notes is much higher. Moreover, the vocabulary size of concepts in clinical corpora is also much larger than common English articles. This makes the posterior distribution sparser and harder to infer the latent variables in these models. However, nonparametric Bayesian models are still promising for clinical applications if we add more domain-knowledge as constraints in the sampling process.

Our study aims to propose an unsupervised model for clinical coreference resolution to achieve comparable performance to the state-of-art coreference resolution approaches. The training process of a coreference unsupervised Bayesian model involves maximizing the posterior probability among mentions in the given corpus. The challenge is how we can define a posterior probability so that if the posterior has been optimized, the coreference evaluation metrics are optimized as well. In this paper, we present an infinite mixture model to resolve the coreferent relations between annotated mentions. A similarity measurement between two mentions is proposed to evaluate the likelihood that two mentions are coreferent and then the similarity score is used to determine if there is an antecedent of the given mention. The paper is organized as follows: we first introduced our system pipeline. Then we proposed our infinite mixture model and evaluated the performance of our system on i2b2 dataset. Finally, the performance and error analysis along with some potential further improvements were discussed.

Methods

Our proposed coreference resolution system consists of several components and the pipeline is illustrated in Figure 1. The first component is the file preprocessor. It first reads both the free text of clinical records and corresponding annotated mentions, then extract lexical features, syntactic features and concept features of each mention. After we have the annotated mentions with features, we can then use definite sampling methods to infer entity indices of mentions. This can be achieved by Bayesian mixture models. After the entity indices are obtained, the mentions are chained according to entity indices for the evaluation.

Finite Mixture Model using Latent Dirichlet Allocation

We first adopt a commonly used probabilistic topic modeling method, Latent Dirichlet Allocation (LDA)[15], as a finite mixture model coreference resolution method. The conventional LDA assumes that each document is generated from a multinomial distribution of topics and a multinomial distribution of words yields the whole corpus. The parameters of these multinomial distributions are drawn from Dirichlet distributions with arbitrary hyperparameters. To adopt LDA to a coreference resolution system, several modifications to the model should be made. Firstly, the generative process of document has to be redefined as Table 1. Note that in the coreference LDA model each topic will only contain one type of mentions, which is reasonable because mentions with different types
or properties such as genders or numbers (singular or plural) should not be coreferent. To resolve the pronoun mentions, we simply match the pronoun with the most recent antecedent mention with a matched number (singular or plural) as the result. For example, the pronoun “those” will only match the most recent plural antecedent and all singular antecedents are skipped. In addition, because LDA takes single token as observed data, we only use the head word of each mention to train the model.

**Table 1 Generative Model of i2b2 Mention of finite mixture model using LDA**

For each document D:

Choose document-entity distribution parameter $\theta$ drawn from $\text{Dir}(\alpha)$;

For each mention $m$ in position $n$:

Choose an entity drawn from document-entity distribution $\text{Multinomial}(\theta)$

Choose a headword from multinomial entity-head word distribution

We conducted an experiment on LDA using our optimized head word extraction method. Unsurprisingly, the finite mixture model cannot infer most of the coreference entities correctly. Our LDA coreference resolution system is implemented based on the Java version of JGibbLDA [24]. The overall unweighted average of coreference evaluation metrics has only a 0.475 F measure, even lower than the baseline with 0.519 which outputs no coreference relations. There are several reasons for the failure of LDA as a coreference resolution method. The most important one is that although LDA can be used as an unsupervised clustering method, the strategy of which mentions should be put together makes it impossible to directly be used in coreference resolution. In conventional LDA, the co-occurrence of words is used to calculate the topic-word probability and the probability is proportional to the number of appearance of the word. This is to say, the more times two words appear together in a document, the more likely they will be put into one topic. This will cause errors if we would like to use LDA as a coreference resolution method, because the coreference relation is different from co-occurrence and in most of the cases, even two mentions appear together many times, they may not be from the same entity. For example, in clinical notes, a patient may have problem A and problem B, but apparently, problem A and problem B are from different entities so they should not coreferent but the inference of LDA may do that. Another issue is that LDA is a finite mixture model and the number of components has to be arbitrarily set before model training. And the number of components can significantly impact the clustering output. However, for the problem of coreference resolution, we do not know the number of entities until we get the result. Thus, the generative model of LDA has to be modified to solve coreference resolution problems.

**Infinite Mixture Model using Mention Similarity**

To solve the problems of LDA as a coreference resolution method, we adapted Hierarchical Dirichlet Process (HDP) [25] into our framework. Proposed as a topic modeling method as LDA, HDP can be regarded as an extension of LDA which replaces the finite topic multinomial distribution in LDA to an infinite mixture model: Dirichlet Process (DP) [26]. Therefore, the total number of topics in a corpus does not need to be assigned to an arbitrary value before training. Our model is inspired by Haghighi and Klein’s work [18] and a definite sampling inference algorithm is proposed to improve the performance on clinical corpus.

Given an assigned mention type from previous procedure, the entity will be generated from a modified Chinese

![Figure 2 A sample Chinese Restaurant Process illustration for clinical note](image-url)
Restaurant Process (CRP) from DP. A sample illustration of CRP in clinical note is shown in Figure 2. The first entity refers to the patient in the clinical note, and the second entity refers to the clinician who wrote the note. Each entity is regarded as a table in CRP, while each document is a restaurant which consists of an infinite number of tables, namely entities. Each mention is regarded as a customer who will find a table to sit in the restaurant. In conventional CRP, where the model will perform a “the riches get richer” manner because the probability of the customer assigned to a certain table is proportional to the number of existing customers on the table. In contrast, in our discourse generative model, the probabilities are calculated from the extracted feature vector of each mention and each entity candidates. This discourse generative model is thus an infinite mixture model.

A similarity score \( f(m_i, m_j) \) is calculated between pair-wise mentions. The weights of the features can be arbitrarily set according to the importance of each feature. Some of the mention properties have to be matched for any coreferent mentions, such as mention number and mention type. We call these properties hard constraints and others soft constraints. The combination of hard constraints is represented by a product of binary terms (0/1) in the similarity function. Through this, the similarity function can work in a rule-based manner for these strict rules. The similarity function of two mentions is defined as:

\[
f(m_i, m_j) = \prod_{k=N}^{N+N'} f_k(m_i, m_j) \cdot \sum_{k'=1}^{N'} w_{k'} f_{k'}(m_i, m_j),
\]

where \( m_i \) and \( m_j \) refer to any two mentions. Terms \( f_k(m_i, m_j) \) are hard constraints, all of which must be non-zeros to make the similarity score non-zero as well. Terms \( w_{k'} f_{k'}(m_i, m_j) \) are the weighted soft constraints, which are the main measure of similarity and \( w_{k'} \) is the weight of \( k' \)-th feature. Respectively, \( k \) and \( k' \) are the features index for hard constraints and soft constraints, while \( N \) and \( N' \) are the number of hard and soft constraints.

Several features are extracted for each mention before the calculation of pair-wise mention similarities. The features are: singular/plural of the head token; part-of-speech (POS) tag of the head token; token position of the first token of the mention; gender (male/female/neutral); if the mention is related to a person (true/false); if contains new entity indicators (“further”, “another”) and the Concept Unique Identifier (CUI) of the most specific concept within the mention.

The number property of each mention is extracted based on the POS tag of the head token. From the result of Penn TreeBank POS Tagger [27], the plural nouns are assigned the same POS tag as singular nouns but ending with “S” mark. For example, for nominal nouns, the POS tag is NN and the corresponding plural noun POS tag is NNS. From this rule, we can decide whether the head token is singular or plural. We can decide if a mention is a personal mention from both POS tag and the annotated mention types. CUIs from UMLS Metathesaurus [28] are used to identify each non-person mentions. For most of the mention strings with more than one token, there will be several concepts found within the mentions. We only use the most specific concept, which is assumed to have the maximum number of tokens as the returned CUI query result. For example, for the mention “total abdominal hysterectomy”, the UMLS dictionary lookup algorithm in cTAKES will return three terms, which are “total abdominal hysterectomy” (C0404079), “abdominal hysterectomy” (C0404077) and “hysterectomy” (C0020699). In this example, “C0404079” is returned.

The dependency relations within each mention are also considered when we choose the head word in each mention. In each sentence, the POS tagger and dependency parser are run to obtain the dependency relation within mentions in the sentence. The dependency parser we use is ClearTK Dependency Parser [29], which is implemented and included in cTAKES. For each word token, the dependency parser will return its parent. According to the parent information, we can construct a mention dependency tree, which is a subtree of the sentence dependency tree. Once a mention dependency tree is obtained, the root node of mention subtree is used as the head token of the mention. For example, in a mention “a CT scan on the chest”, from the dependency tree we will choose the token “scan” as the head token because it is the root. Similarly, in another example of “her CT scan”, the token “scan” is chosen as the head token as well. This is an improvement to the head word matching in [18] which simply used the last token of a mention as head token.

These extracted features are then used to calculate the similarity, taking the values according to Table 2. Feature 1 to 4 are soft constrains and 5 to 8 are hard constrains. As a result, both value of \( N \) and \( N' \) - are 4. In practice, the \( \beta \) is set
as 0.5 and $T$ is set to be one tenth of the total number of mentions in the document to make the exponential function decrease by a reasonable rate.

### Table 2 Set of features used in similarity functions

| ID | Features | Type | Definition | Values |
|----|----------|------|------------|--------|
| 1  | Token distance | Soft | The number of tokens between two mentions | $\begin{cases} \exp(-\beta|p_1 - p_2|), & |p_1 - p_2| \leq T \\ 0, & \text{otherwise} \end{cases}$ |
| 2  | CUI matching   | Soft | if the two CUI concepts match | 0, 1 |
| 3  | CUI not existing  | Soft | If either of the mentions does not have CUI extracted | 0, 1 |
| 4  | Head token matching | Soft | if the head tokens of the two mentions are the same | 0, 1 |
| 5  | Mention type matching | Hard | if the mention types match | 0, 1 |
| 6  | isPerson matching | Hard | if the mentions both refer to personal entities or both to non-personal entities | 0, 1 |
| 7  | Number matching | Hard | if the singular/plural forms match | 0, 1 |
| 8  | New entity indicator | Hard | if the later mention phrase contains a new entity indicator | 0, 1 |

### Inference by Definite Sampling

In this step, we first divide all mentions into three categories: personal mentions, pronoun mentions and non-personal mentions. The personal and pronoun mentions are simply the mentions with the annotated type of “person” and “pronoun”, respectively, in i2b2 dataset while non-personal mentions consist of types of test, problem and treatment.

Typically, the personal entities in clinical notes are limited among the following categories: the patient, the author who is likely to be a clinician, other attending physicians and patient’s family [21]. Personal pronouns like “I”, “she” and “him” are also annotated as person entities, thus we treat these pronouns as personal mentions. According to this observation, it is reasonable to apply a method exclusively for personal mentions. Accordingly, it is also safe to assume that most of the single third person pronouns, for example, “he”, “she”, “his” and “her”, refer to the patient [21], and we use this rule to directly assign the entity of patients to single third-person pronoun mentions in person category. In addition, descriptive mention phrases like “a 60 years old male” or “the female” are likely to refer the patient as well. So the mentions with token “patient”, “pt”, “male” or “female” are assigned to the patient entities as well. All other personal mentions are clustered using exact string matching.

### Table 3 Algorithm for definite sampling for non-pronoun mentions

```
Initialize: Length of the document $n$; Number of entities $K$: $K = 0$; Entity Assignment $\vec{e} = \{e_1, ..., e_n\}$
For each mention position $i = 1, ..., n$:
  For each antecedent position $j = 1, ..., i$:
    Calculate similarity between mention $i$ and $j$: $p(e_i = e_j | \vec{e}) \propto f(m_i, m_j)$
  End
  Update $e_i$ by maximum likelihood estimation:
    $e_i = \begin{cases} \arg \max_{e_j} p(e_i = e_j | \vec{e}), & \arg \max_{e_j} p(e_i = e_j | \vec{e}) > \text{Thres} \\ K, & K = K + 1 \end{cases}$
End
Return $\vec{e}$
```
For non-personal mentions, instead of directly assigning an entity index to personal mentions, we try to find the best antecedent for a mention with sampling strategies. Rather than using Gibbs sampling method to decide which of the antecedent mentions the current mention should corefer, we employ a definite sampling method for our CRP, derived from a similarity function to estimate how likely two mentions belongs to the same entity. The return value of similarity function is called similarity score.

After the similarity scores are calculated among all the antecedents, the definite sampling algorithm, which uses maximum likelihood estimation (MLE) to estimate the optimal entity, is shown in Table 3. Here we use the mention similarity defined before as the likelihood. The algorithm basically does pair-wise comparisons between two mentions and chooses the entity assignment maximizing the mention similarity (likelihood). Recall that we have hard constraints to make the matching of some of mention properties required. If any of these properties are mismatched, the similarity score will be zero. A positive threshold of similarity score will then be effective to filter out mentions with an all-zero similarity scores and our algorithm will assign a new entity index to it. In this manner new entities are generated. After we have iterated all the mentions in a document, the sampling step is done. As a result, for each mention we obtain an entity index. The generative process of our CRP is quite similar to the conventional infinite mixture model, except that the entity assignment probability is replaced from proportional to the element count of each cluster to proportional to mention similarity.

The pronouns in i2b2 dataset in the pronoun type are restricted to non-personal pronouns. All personal pronouns like “him” or “I” are annotated as the person type. Therefore, we only need to take care of non-personal pronouns like “it”, “this”, “those” or “which”. The number of pronoun can also be detected if the pronoun is “these” or “those”. Then we simply use the most recent mention that matches the hard constraints of number.

There are two common types of output format for the entity clusters widely used by the evaluation program of coreference resolution metrics: by chains and by pairs. The official evaluation script released with the i2b2 dataset requires a chain output format. Thus, the mentions with the same sampled entity index are output as a chain with its entity type without “pronoun”. The official documentation requires at each output chain has to include at least one mention which is not pronoun.

**Experimental Setup and Results**

Our proposed system is implemented using Apache cTAKES [30], which is a natural language processing system specifically designed for the extraction of information in clinical documents. cTAKES is an open source project implemented in Java based on Apache UIMA[31]. The aggregate analysis engine descriptor of UIMA makes it straightforward to design free text processing pipelines. Our analysis engine pipeline is illustrated in Figure 3. Both the markable concept files and raw text files are loaded by collection reader. In i2b2 dataset, each line contains only

![Figure 3 cTAKES analysis engine pipeline for mention feature extraction](image-url)
one sentence and thus the line number can be used as sentence index. Word tokens are separated by spaces. The lookup window annotator is used to improve the efficiency of the following UMLS dictionary lookup. It will annotate all the markable mentions in the clinical free text according to the concepts that are manually annotated by domain experts. As a result, the dictionary lookup function will only look up in the span of annotated mentions instead of each token in the clinical notes. The results of CUI from dictionary lookup are used as a feature while calculating mention similarities.

### Table 4 Statistics of training and testing dataset

|                | Number of documents | Number of mentions | Number of chains | Number of chained mentions | Data source                      |
|----------------|---------------------|--------------------|------------------|---------------------------|----------------------------------|
| Training       | 123                 | 12338              | 1182             | 5428                      | Pittsburg Progress               |
| Testing        | 493                 | 66345              | 7050             | 32123                     | Pittsburg Progress, Pittsburg Discharge, Beth Discharge, Partners Discharge |

The clinical coreference resolution system is evaluated on i2b2 2011 Track 1C dataset[4]. In Track 1C, the mentions are already manually annotated with spans and types provided for coreference resolution. The terminology we used in this paper is more commonly used thus is slightly different from the definition in i2b2 2011 Challenge. The terms of "mention" and "entity" are called "markable" and "chain", respectively. The Pittsburg Progress dataset is used as the training dataset. The system is developed on the training dataset and then tested on a larger testing dataset. All of our results presented are obtained in the complete i2b2 training set which contains Beth Discharge, Partners Discharge, Pittsburg Discharge and Pittsburg Progress. Some statistics of the training and testing dataset are shown in Table 4. Note that since our proposed system does not use any supervised techniques, the performances of different coreference evaluation metrics show only negligible differences in the training and testing dataset.

We performed several experiments to evaluate our proposed method. Firstly, we compared the F measures of the two baseline systems of different categories. The performance of the first baseline system in Table 5 is that we created a system that outputs no coreference chains. Namely, the result of the system is all the mentions are singletons and none of the mentions corefer to others. This system has an overall F measure of 0.519. The second baseline system is the exact string matching system. It uses the head token obtained by dependency parser and then only considering the mentions have the exact same head tokens as from the same entity. This system is implemented by increasing the weight of string matching term to a very large value in our proposed similarity function followed by changing the threshold of similarity measure to the value that only slightly smaller than that large weight. As a result, only the result of head token matching can impact the mention clustering. In practice, we set the mention clustering threshold to be 1.1, this optimizes our result in the way that either CUI matched concepts or string matched mentions can be clustered and the distance threshold is 0.1 for the token distance feature. It can be found that with this parameter set up our method’s result outperforms the other systems.

### Table 5 Comparison of the F measure of different models in i2b2 dataset

| Methods                  | Test   | Person | Problem | Treatment | Overall |
|--------------------------|--------|--------|---------|-----------|---------|
| Baseline                 | 0.166  | 0.593  | 0.249   | 0.306     | 0.519   |
| Exact string matching    | 0.634  | 0.667  | 0.727   | 0.815     | 0.765   |
| Infinite Mixture Model   | **0.719** | **0.734** | **0.826** | **0.826** | **0.847** |

The evaluation is obtained from the evaluation script provided by i2b2 2011 Challenge organizer. The system performance is measured by MUC [32], B-cubed [33], CEAF [34] and BLANC [35]. Each metrics has precision, recall and F-measure. The evaluation uses the unweighted average between MUC, B-cubed and CEAF as the final result. In the i2b2 dataset, the gold standard annotations contain the type of “pronoun”. As the result of coreference resolution, the “pronoun” type will finally merge into the other four entity types. Thus the evaluation metrics only contains four entity types without “pronoun”. Note that the overall precision/recall/F measure results are different from the average of all the categories in all the coreference measures. This is because in the overall evaluation, the
singleton and pronoun mentions are considered. In our system, the overall F measure of 0.847 is higher than the F measure of the highest category, i.e. problem with 0.826. Besides, the evaluation system only provides the overall F measure and there is no result of overall precision and recall.

**Table 6 Performance of infinite mixture model on i2b2 data**

| Category   | B3      | MUC      | BLANC     | CEAF     | Average |
|------------|---------|----------|-----------|----------|---------|
| Test       | 0.946/0.971/0.958 | 0.444/0.242/0.313 | 0.599/0.684/0.629 | 0.939/0.916/0.927 | 0.711   |
| Person     | 0.570/0.617/0.593 | 0.759/0.961/0.848 | 0.982/0.878/0.924 | 0.436/0.825/0.570 | 0.734   |
| Problem    | 0.934/0.940/0.937 | 0.716/0.599/0.652 | 0.773/0.832/0.800 | 0.924/0.903/0.913 | 0.826   |
| Treatment  | 0.936/0.960/0.948 | 0.745/0.596/0.663 | 0.770/0.837/0.800 | 0.910/0.877/0.893 | 0.826   |
| Overall    | 0.886/0.930/0.907 | 0.741/0.806/0.772 | 0.965/0.875/0.915 | 0.843/0.879/0.861 | 0.847   |

**Discussion**

Our proposed infinite mixture model with mention clustering method can achieve good results in clinical domain while only uses a limited number of features. The definite sampling algorithm also outperforms the conventional inference method according to the posterior probability when infer the coreferent relations. Although the current feature set is relatively small, it has the potential to be extended into a rich-feature system by incorporating more features into the mention similarity function. By utilizing our proposed model, the scope of methods to solve the coreference resolution problem can be greatly expanded. For example, since we already have a quantified weighted feature vector to represent mentions distance, more clustering methods can be then introduced to solve the mention clustering problem. We can also sample new mentions from the multivariate distribution of these features then improve the performance of the nonparametric Bayesian models. In previous solutions, only a few hard constraints like type, number and gender are considered as features in the generative process.

Compared with the performance by individual categories, we found that the test and person categories are the bottleneck in our system. For personal mentions, a rich-feature supervised machine learning classifier can achieve a 0.902 F measure in [20], which is significantly higher than the 0.734 F measure of our system. The classifier in [20] used several lexical features like text, tokens, character trigrams in patient mention identification. Although these features are not as interpretable as features which are used in Xu’s [19] and our system, the classifier can still be able to correctly identify most of the patient mentions. This observation indicates the variety of latent patterns among patient mentions in clinical notes. There are also some mentions of patients’ family members or other attending clinicians other that the authors of the notes. Our features for personal mentions may have difficulties when trying to solve these entities. The patient mentions can be easily annotated because there is only one patient in each clinical note, and the patient mentions are usually the majority in personal mentions. This makes the supervised learning in patient identification possible. Besides, the mentions which are represented by Protected Health Information (PHI) also make the problem more challenging. For example, the gender of the patient cannot be extracted directly from looking up the dictionary of first names. To improve our system’s performance, pre-training a classifier for patients using supervised methods can be considered. Meanwhile, we can still retain other categories as unsupervised. By doing this, a boost on both person category and overall performance can be expected.

For non-personal mentions, the complexity and variety of test mentions make the category hardest in the corpus. The MUC score of test in our system is much lower than those of other categories. It seems that our model is too simplified for test mentions. For example: the “follow-up echo” can refer to “another echocardiogram”. In this case even if we can detect that “echo” is an abbreviation, it remains challenging to decide if it should refer to “echocardiogram”. Usually, the word “another” may indicate that mention has not been mentioned before however in this example it has. For these reasons, the first chain is hard to be linked. In another example, the mention “the patient's abdominal exam” can refer to “the patient's evaluation”. Since “exam” and “evaluation” have the same meaning so they are possible to be clustered. However, for any systems which do not use supervised learning, it is hard for the system to detect that these two words have the same meaning. Even for supervised learning methods, if the training set does not contain this relation, it will not be able to detect this relation either. To correct these errors, more features should be analyzed according to how the entities might be mentioned by the note authors.
Except for specific improvements targeting the existing errors, more explorations can be made to further improve the proposed system. The hierarchy of UMLS concepts can also be used when calculating the likelihood. For example, if a parent-child relation is detected for two mentions, the mentions are still possible to be coreferent. Other medical related ontology like SNOMED can also be included as features. Besides, there can also be a spatial adjective detection method so that if the two mentions have the same kind of test in different part, these mentions may set to be not coreferent. The Gibbs sampling method can also be considered while updating the weights for feature vector so that the final weight will optimize the mixture similarity and then we may have a fully nonparametric Bayesian model.

**Conclusion**

We have described an infinite mixture model method for coreference resolution system in clinical domain. Unlike other solutions for the same task, we modeled the generative process of clinical discourse as Chinese Restaurant Process. The proposed infinite mixture model used definite sampling with maximum likelihood estimation of mention similarity to estimate the cluster of entities. The system can achieve an F measure of 0.847 on i2b2 2011 coreference dataset and the infinite mixture model is shown both effective and promising for further studies and applications.

**Acknowledgements**

We would like to thank the support from our colleagues, Yanshan Wang and Feichen Shen for their valuable suggestions on this research. This work was made possible by joint funding from National Institute of Health 1K99LM012021-01A1, R01GM102282, R01LM11369, R01LM011829, and R01 LM011934.

**References**

1. Tao, C., et al., *Time-Oriented Question Answering from Clinical Narratives Using Semantic-Web Techniques*, in *The Semantic Web – ISWC 2010*, P. Patel-Schneider, et al., Editors. 2010, Springer Berlin Heidelberg. p. 241-256.
2. Tate, A.R., et al., *Exploiting the potential of large databases of electronic health records for research using rapid search algorithms and an intuitive query interface*. Journal of the American Medical Informatics Association : JAMIA, 2014. 21(2): p. 292-8.
3. Zheng, J., et al., *Coreference resolution: a review of general methodologies and applications in the clinical domain*. Journal of biomedical informatics, 2011. 44(6): p. 1113-22.
4. Uzuner, O., et al., *Evaluating the state of the art in coreference resolution for electronic medical records*. Journal of the American Medical Informatics Association : JAMIA, 2012. 19(5): p. 786-91.
5. Elango, P., *Coreference Resolution: A Survey*.
6. Ng, V., *Supervised noun phrase coreference research: the first fifteen years*, in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics2010, Association for Computational Linguistics: Uppsala, Sweden. p. 1396-1411.
7. Rahman, A. and V. Ng. *Supervised models for coreference resolution*. in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009.
8. McCarthy, J.F. and W.G. Lehnert. *Using Decision Trees for Coreference Resolution*. in In proceedings of the fourteenth international joint conference on artificial intelligence.
9. Soon, W.M., H.T. Ng, and D.C.Y. Lim, *A Machine Learning Approach to Coreference Resolution of Noun Phrases*. COMPUTATIONAL LINGUISTICS, 2001. 27(4): p. 521-544.
10. Ng, V. *Unsupervised models for coreference resolution*. in Proceedings of the Conference on Empirical Methods in Natural Language Processing.
11. Yang, X., et al. *An Entity-Mention Model for Coreference Resolution with Inductive Logic Programming*. in Proceedings of ACL-08: HLT. Association for Computational Linguistics.
12. Luo, X., et al., *A mention-synchronous coreference resolution algorithm based on the Bell tree*, in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*2004, Association for Computational Linguistics: Barcelona, Spain. p. 135.
13. Heinrich, G., *Parameter estimation for text analysis*, 2005.
14. Griffiths, T.L., et al. *Integrating topics and syntax*. in Advances in neural information processing systems. 2004.
15. Blei, D.M., A.Y. Ng, and M.I. Jordan, *Latent dirichlet allocation*. the Journal of machine Learning research, 2003. 3: p. 993-1022.

16. Ng, V. *Unsupervised models for coreference resolution*. in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2008.

17. Haghighi, A. and D. Klein. *Coreference resolution in a modular, entity-centered model*. in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 2010.

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

19. Xu, Y., et al., *A classification approach to coreference in discharge summaries: 2011 i2b2 challenge*. Journal of the American Medical Informatics Association: JAMIA, 2012. 19(5): p. 897-905.

20. Rink, B., K. Roberts, and S.M. Harabagiu, *A supervised framework for resolving coreference in clinical records*. Journal of the American Medical Informatics Association: JAMIA, 2012. 19(5): p. 875-82.

21. Gooch, P. and A. Roudsari, *Lexical patterns, features and knowledge resources for coreference resolution in clinical notes*. Journal of biomedical informatics, 2012. 45(5): p. 901-12.

22. Jonnalagadda, S.R., et al., *Coreference analysis in clinical notes: a multi-pass sieve with alternate anaphora resolution modules*. Journal of the American Medical Informatics Association: JAMIA, 2012. 19(5): p. 867-74.

23. Chen, P., D. Hinote, and G. Chen, *A rule based solution to co-reference resolution in clinical text*. Journal of the American Medical Informatics Association: JAMIA, 2013. 20(5): p. 891-7.

24. Phan, X.-H., L.-M. Nguyen, and S. Horiguchi, *Learning to classify short and sparse text & web with hidden topics from large-scale data collections*, in *Proceedings of the 17th international conference on World Wide Web* 2008, ACM: Beijing, China. p. 91-100.

25. Teh, Y.W., et al., *Hierarchical dirichlet processes*. Journal of the american statistical association, 2006. 101(476).

26. Blei, D.M., M.I. Jordan, and others, *Variational inference for Dirichlet process mixtures*. Bayesian analysis, 2006. 1(1): p. 121-143.

27. Marcus, M.P., B. Santorini, and M.A. Marcinkiewicz, *Building a Large Annotated Corpus of English: The Penn Treebank*. Computational Linguistics, 1993. 19(2): p. 313-330.

28. Bodenreider, O., *The Unified Medical Language System (UMLS): integrating biomedical terminology*. Nucleic acids research, 2004. 32(Database issue): p. D267-70.

29. Bethard, S., P. Ogrren, and L. Becker. *ClearTK 2.0: Design Patterns for Machine Learning in UIMA*. in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*. 2014. Reykjavik, Iceland: European Language Resources Association (ELRA).

30. Savova, G.K., et al., *Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications*. Journal of the American Medical Informatics Association: JAMIA, 2010. 17(5): p. 507-513.

31. Ferrucci, D. and A. Lally, *UIMA: an architectural approach to unstructured information processing in the corporate research environment*. Nat. Lang. Eng., 2004. 10(3-4): p. 327-348.

32. Vilain, M., et al., *A model-theoretic coreference scoring scheme*, in *Proceedings of the 6th conference on Message understanding1995*, Association for Computational Linguistics: Columbia, Maryland. p. 45-52.

33. Bagga, A. and B. Baldwin. *Algorithms for Scoring Coreference Chains*. in *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*. 1998.

34. Luo, X., *On coreference resolution performance metrics*, in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* 2005, Association for Computational Linguistics: Vancouver, British Columbia, Canada. p. 25-32.

35. Recasens, M. and E. Hovy, *Blanc: Implementing the rand index for coreference evaluation*. Nat. Lang. Eng., 2011. 17(4): p. 485-510.