Exploring Semi-Supervised Coreference Resolution of Medical Concepts using Semantic and Temporal Features

Preethi Raghavan*, **Eric Fosler-Lussier***, **and Albert M. Lai**†
*Department of Computer Science and Engineering
†Department of Biomedical Informatics
The Ohio State University, Columbus, Ohio, USA
{raghavap, fosler}@cse.ohio-state.edu, albert.lai@osumc.edu

Abstract

We investigate the task of medical concept coreference resolution in clinical text using two semi-supervised methods, co-training and multi-view learning with posterior regularization. By extracting semantic and temporal features of medical concepts found in clinical text, we create conditionally independent data views; co-training MaxEnt classifiers on this data works almost as well as supervised learning for the task of pairwise coreference resolution of medical concepts. We also train MaxEnt models with expectation constraints, using posterior regularization, and find that posterior regularization performs comparably to or slightly better than co-training. We describe the process of semantic and temporal feature extraction and demonstrate our methods on a corpus of case reports from the New England Journal of Medicine and a corpus of patient narratives obtained from The Ohio State University Wexner Medical Center.

1 Introduction

The clinical community creates and uses a variety of semi-structured and unstructured electronic textual documents that include medical reports such as admission notes, progress notes, pathology reports, radiology reports and hospital discharge summaries. The documents, collectively termed *clinical narratives*, account for various medical conditions, procedures, diagnoses and assessments in a patient's medical history. Researchers have investigated ways in which clinical text can be automatically processed for enabling access to relevant information for physicians and health researchers (Embi and Payne, 2009). One application is to support patient recruitment into clinical trials (research studies that try to answer scientific questions to find better ways to prevent, diagnose, or treat a disease) by matching patient characteristics against eligibility criteria (Raghavan and Lai, 2010). While there has been significant efforts to move to structured data collection, clinical narratives remain a critical data source for these tasks.

Extracting structured information from unstructured clinical text using natural language processing (NLP) is complicated by the distinct clinical reporting sub-language characterized by incomplete sentences and domain specific abbreviations (Friedman et al., 2002). The large number of clinical narratives generated per patient, over the years, along with redundant information within and across narratives, further adds to the complexity of using information structured using NLP. There is a tendency to copy and edit parts of an old clinical narrative whenever a new one is created, thus leading to redundant information in clinical narratives of a patient. Furthermore, since different types of clinical narratives are created for different purposes, certain narratives may summarize information from various other, at times older, clinical narratives. All of this makes the task of automatically processing unstructured clinical narratives significantly difficult. However, the ability to resolve medical concept coreferences helps deal with redundant information within and across clinical narratives and thus produce a unique list of medical concepts in the patient's clinical history.

We investigate the task of resolving references to

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 731–741,
Montréal, Canada, June 3-8, 2012. ©2012 Association for Computational Linguistics

the same medical concept in the clinical narratives of a patient using supervised and semi-supervised methods. Our main contributions are as follows:

1. Since manual coreference annotation of patient narratives is a slow and expensive process and publicly available datasets are difficult to acquire, we study the application of semi-supervised methods, co-training and using expectation constraints with posterior regularization, to medical concept coreference resolution (MCCR).
2. We work with the hypothesis that if two medical concepts have the same meaning and have occurred at the same time, there is a very high probability that they corefer. Based on this hypothesis, we explain extraction of semantic and temporal feature sets that are effectively used for MCCR.
3. We propose a method to associate medical concepts with time durations centered around admission and discharge dates of the patient using CRFs.
4. With the help of corpora created from the New England Journal of Medicine (NEJM) and actual patient narratives obtained from the medical center, we demonstrate that the semi-supervised methods perform comparably with supervised learning for pairwise MCCR using a MaxEnt classifier.

2 Related Work

Free-text reports form a significant portion of the information content in a patient's medical record. There is great need for tools that can structure the information in clinical text for use in various studies studies such as clinical trials, quality assessment of healthcare delivery in institutions, and public health research. Researchers have been investigating ways in which clinical free-text can be structured to transform the information content in a clinical narrative into a representation suitable for computational analysis (Ananiadou et al., 2004). Medical NLP systems like Mayo's cTakes (Savova et al., 2010), IBM's MedKAT,[1] and MedLEE (Chiang et al., 2010), have components specifically trained or designed for the clinical domain, to support tasks such as named entity recognition. Previous attempts at learning temporal relations between medical events in clinical text include work by Jung et al. (2011) and Zhou et al. (2006). Gaizauskas et al. (2006) learn the temporal relations *before, after, is_included* between events from a corpus of clinical text much like the event-event relation tlink learning in Timebank (Pustejovsky et al., 2003). A comprehensive survey of temporal reasoning in medical data is provided by Zhou and Hripcsak (2007). Chapman et al. (2011) discuss barriers to NLP development in the clinical domain.

Coreference resolution is a well-studied problem in computational linguistics (Ng, 2010; Raghunathan et al., 2010). Supervised machine learning algorithms have been previously used for noun phrase coreference resolution with fairly good results (Soon et al., 2001; Raghunathan et al., 2010). Recently, the i2b2 challenge[2] on coreference resolution examined coreference resolution in clinical data. The problem addressed in our paper is similar to the task described in the i2b2 challenge.[3] Besides the i2b2 challenge, there has not been significant work in MCCR. This may be due to various privacy concerns and the efforts required to anonymize and annotate massive amounts of patient narratives. Zheng et al. (2011) review heuristic-based, supervised and unsupervised methods for coreference resolution in the context of the clinical domain. He (2007) studied coreference resolution in discharge summaries, treating coreference resolution as a binary classification problem and investigated critical features for coreference resolution for entities that fall into five medical semantic categories commonly appearing in discharge summaries. However, we focus on feature extraction to determine the similarity between medical concepts, both in terms of meaning and time of occurrence, for resolving coreferences within and across all types of clinical narratives.

A disadvantage of supervised machine learning approaches is the need for an unknown amount of annotated training data for optimal performance. Researchers then began to experiment with weakly supervised machine learning algorithms such as co-training (Blum and Mitchell, 1998). Muller et al. (2002) investigate the practical applicability of co-training for the task of building a classifier for coreference resolution and observed that the results were

[1] https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/OHNLP

[2] https://www.i2b2.org/NLP/Coreference/

[3] https://www.i2b2.org/NLP/Coreference/assets/CoreferenceGuidelines.pdf

mostly negative for their dataset.

Ganchev et al. (2010) propose a posterior regularization framework for weakly supervised learning to derive a multi-view learning algorithm. Multi-view methods typically begin by assuming that each view alone can yield a good predictor. Under this assumption, we can regularize the models from each view by constraining the amount by which we permit them to disagree on unlabeled instances. In the proposed approach, they train a model for each view, and use constraints that the models should agree on the label distribution.

We investigate the applicability of these two weakly supervised methods to the task of MCCR using semantic and temporal views. Savova et al. (2011) discuss the creation of a corpus for coreference resolution in the clinical narrative. We annotate a corpus of clinical narratives to tag medical concepts, temporal relations, and coreference information. We use this corpus as a gold standard to evaluate the proposed approach to resolving coreferences between medical concepts in clinical text.

To summarize, we study the problem of intra and cross-narrative coreference resolution on longitudinal patient data using relatedness between medical concepts in terms of semantics and time. Further, we importantly demonstrate that this task gives us reasonable results even when modeled as a semi-supervised problem. Creating annotated clinical corpora is tedious, time consuming, and costly, as it requires experts with medical domain knowledge. Thus, the ability to train semi-supervised models with limited labeled data for MCCR would be of tremendous value.

3 Problem Description

Coreference resolution in clinical text refers to the problem of identifying all medical concepts that refer to the same medical concept. Medical concepts are medical entities, events or states associated with the patient's medical condition and healthcare. These include medical conditions, drugs administered, diseases, procedures and lab tests as well as normal health situations like pregnancy affecting the patient's health. The task of MCCR is similar to noun phrase coreference resolution. However, medical concepts are not restricted to noun phrases. For instance, the actions *cauterize* and *cauterization* are both considered medical concepts.

To make the task of identifying medical concepts from clinical text more deterministic, any contiguous group of words that have a direct or close match in the Unified Medical Language System (UMLS) Metathesaurus[4] is considered a medical concept. The UMLS includes a large Metathesaurus of concepts and terms from many biomedical vocabularies and a lexicon which contains syntactic, morphological, and orthographic information for biomedical and common words in the English language.

Problem Formulation. Consider a corpus of clinical narratives, where multiple clinical narratives are associated with each patient. If P_i, $i \in \{1, 2, ..., n\}$ where n is the number of patients in corpus, then for each P_i, we have a set of associated clinical narratives. Each clinical narrative in turn has a set of medical concepts. Thus, each P_i has a set of associated medical concepts, $M = \{M_1, M_2, M_3, ..\}$ that occur within each clinical narrative as well as across clinical narratives for that P_i. We study the problem of MCCR of all medical concepts in M for each P_i.

4 Semantic and Temporal Features

We extract features based on semantic and temporal relatedness for each pair of medical concepts. Semantic relatedness measures closeness between medical concepts in terms of their meaning. This is quantified by measuring distance between medical events in the UMLS Metathesaurus graph structure (Xiang et al., 2011). Temporal relatedness measures the closeness between medical concepts in terms of when they occurred. This is achieved by first, learning to assign every medical concept to a time-bin, and then using the time-bin as a feature for learning to resolve coreferences. Extracting semantic and temporal features helps identify conditionally independent views of the data for co-training classifiers. As previously noted by Nigam and Ghani (2000), it is hard to identify conditionally independent views for real-data problems. However, we believe there are no natural dependencies between the semantic and temporal feature sets. While semantic features help identify synonymous medical concepts, that alone may not guarantee coreference. Medical con-

[4] https://uts.nlm.nih.gov/home.html

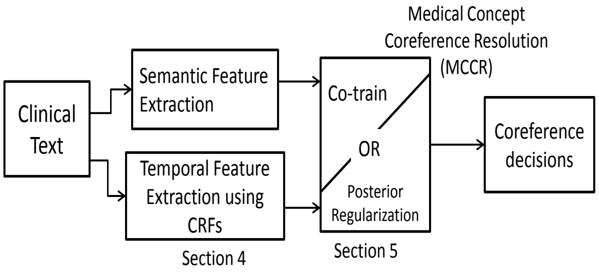

Figure 1: MCCR pipeline: Extract semantic and temporal features from clinical text to train MaxEnt classifiers for medical concept coreference resolution using 1) Co-training or 2) Posterior Regularization

cepts that are similar in meaning, but dissimilar in terms of their time of occurrence, most probably do not corefer. Similarly, medical concepts that occur during the same time duration but are dissimilar in terms of meaning, most probably do not corefer.

Semantic Relatedness. We leverage the UMLS to derive a semantic relatedness score between medical concepts. The UMLS codifies concepts found in various medical vocabularies (e.g., ICD[5] and SNOMED-CT[6]) and includes relationships between various concepts. The medical concepts and their relationships are modeled in a graph structure. We use the k-Neighborhood decentralization method (kDLS) (Xiang et al., 2011) to index and transitively traverse associated relations between concept unique identifiers (CUIs) in the UMLS graph. The UMLS uses semantic relations to mark the available links between two concepts. Around 2,404,937 CUIs and 15,333,246 links between them are seen in the full UMLS graph structure. The kDLS method is shown to outperform both breadth-first and depth-first search in terms of speed and various other measures in finding important information, such as reachability, distance, and a summary of paths, between two concepts in the UMLS graph structure. The relation between two concepts M_j (denoted by x) and M_k (denoted by y) is measured as follows.

$$R(x,y) = \sum_{p \in D_{(x,y)}} \frac{1}{\gamma^{length(p)-1}} + \sum_{q \in D_{(y,x)}} \frac{1}{\gamma^{length(q)-1}}$$

where $D(x,y)$ is the set of paths from x to y and $D(y,x)$ is the set of paths from y to x obtained using the kDLS method, excluding paths with length equal to 1. In order to make the measurement between a medical concepts unbiased against the available links in the UMLS that directly connect them, the paths with length being 1 between them are not counted. Each path's contribution to the relation score $R(x,y)$ is determined by its length and γ. γ is varied between 1 to 50; if γ is set to 1, then all paths contribute equally to R irrespective of their lengths. When γ increases, more weight will be placed on the short paths as opposed to the long paths. Xiang et al. (2011) observe several fold enrichment values when γ is varied between 5 and 15.

Besides traversing the UMLS graph structure using the kDLS method to obtain a similarity score between medical concepts, we also measure similarity between medical concepts by taking into account the surrounding context. We do so by measuring the KL-divergence between the sentences to which the medical concepts belong. In order to avoid the possibility of an empty set when calculating the intersection of the probability distributions, we use a smoothing method that makes the probability distributions sum to 1 (Brigitte, 2003).

Another important semantic feature is the type of relation between the medical concepts. This feature is calculated by first computing the stemmed word overlap between the medical concepts and deriving features based on exact and partial matches between the word stems of the medical concepts. If there is no exact or partial match between the concepts, we query the UMLS to check if the stem of one of the medical concepts occurs in the UMLS definition or atoms of the other medical event. An atom is the smallest unit of naming within the UMLS. A medical concept in UMLS represents a single meaning and contains all atoms in the UMLS that express that meaning in any way, whether formal or casual, verbose or abbreviated. All of the atoms within a concept are synonymous.

Besides the described features, we also include the UMLS semantic category of each medical concept and the WordNet[7] similarity score between sentences containing the medical concept.

Temporal Relatedness. Clinical text is frequently characterized by temporal expressions co-

[5] http://www.cdc.gov/nchs/icd.htm

[6] http://www.ihtsdo.org/snomed-ct/

[7] http://wordnet.princeton.edu/

occurring with medical concepts (Zhou and Hripcsak, 2007). For instance, *two days ago*, fever started *4 days before* rash, *July 10th, 2010* etc. The ability to associate medical concepts with temporal expressions helps order medical concepts and determine potential temporal overlap between them. This in turn could be a powerful discriminatory feature in MCCR. Consider the medical concept *chest pain* that occurs multiple times in a clinical narrative. If these mentions of *chest pain* have occurred at the same time, there is a possibility that they all refer to the same instance of the medical concept *chest pain*.

Instead of relying on implicit temporal references that may or may be evident from the clinical narrative, we focus on temporal expressions that are found in most clinical narratives. We do so by leveraging structural properties of clinical narratives such as section information and explicit temporal information such as admission and discharge dates, to learn to assign medical concepts to time periods we refer to as time-bins.

We now proceed to explain the process of assigning medical concepts to time-bins using CRFs. Clinical narratives are usually formatted with a structured header with information that includes the patient admission and discharge date. Clinical narratives are also typically divided into sections. Sections represent a logical, and at times, temporal grouping of information in the narrative. Sections such as "history of present illness," "physical examination," "review of systems," "impression," and "assessment plan" tend to occur in a certain order within each clinical narrative. Thus, section transitions may indicate a temporal pattern for medical concepts across those sections. For example, "past medical history" (before admission), followed by "findings on admission" (on admission), followed by "physical examination" (after admission). Sections of certain types may also exhibit certain temporal patterns. A "history of present illness" section may start with diseases and diagnoses 30 years ago and then proceed to talk about them in the context of a medical condition that happened few years ago and finally describe the patient's condition on admission. Given the temporal patterns within sections and at section transitions, it works well to treat the list of medical concepts from each clinical narrative as a sequence (considering them in narrative order) and learning to label them with a corresponding time-bin. We define the following sequence of time-bins centered around admission and discharge, {*way before admission, before admission, on admission, after admission, after discharge*}.

We model the problem of assigning medical concepts to time-bins as a sequence labeling task using a CRF where we predict labels from the set {*way before admission, before admission, on admission, after admission, after discharge*} as a sequence Y predicted from the detected medical concepts X. CRFs use two types of features in classification, state features and transition features. State features consider relating the label y (time-bin) of a single vertex (medical concept) to features corresponding to a medical concept x, and are given by,

$$S(x, y, i) = \sum_j \lambda_j s_j(y, x, i)$$

Transition features consider the mutual dependence of labels y_{i-1} and y_i (dependence between the time-bins of the current and previous medical event in the sequence) and are given by,

$$T(x, y, i) = \sum_k \mu_k t_k(y_{i-1}, y_i, x, i)$$

Above, s_j is a state feature function, and λ_j is its associated weight and t_k is a transition function, and μ_k is its associated weight. In contrast to the state function, the transition function takes as input the current label as well as the previous label, in addition to the data.

Example state features include indicator features based on verbs patterns in the same sentence as that of the medical concept, last verb before the medical concept, and type of clinical narrative. We also include position of medical event in the narrative as well as within each section, the temporal expressions and dates co-occurring with the medical concept as features and the difference between these dates and the admission date on each clinical narrative. Example transition features include section transitions based on the sections under which the medical concept occurs, UMLS relatedness score between the previous and current medical concept, difference in verb patterns between the previous and current medical concept, difference in dates (if any) between the dates co-occurring with the previous and current medical concept.

In order to enable feature extraction for this learning task, we use the following heuristic-based algorithm to automatically identify sections and associate medical concepts with them.

1. Extract lines that are all upper-case, and longer than a word, from all narratives in corpus. They mostly correspond to section titles.
2. Derive the stem of each word in the title using a Porter stemming algorithm[8] and sort stemmed titles by frequency. If two or more words in the title overlap, they are considered the same. This gives us a candidate set of section titles.
3. When parsing a clinical narrative, and encountering a stemmed ngram matching a section title from the frequent list, all subsequent sentences are associated with that section until a new section title is encountered. If an exact match is not found, we allow partially matching ngrams to be considered as section titles.

Along with the time-bin that are learned using the process described above, dates and temporal expressions extracted from the annotations in our corpus are also used as temporal features. The list of features extracted for the task of MCCR include the following:

1. Verb pattern in the sentence in which the medical concept occurs.
2. Last verb before the medical concept in the same sentence.
3. Type of clinical narrative.
4. Section under which the medical concept is mentioned.
5. Position of the medical concept.
6. Dates that fall in the same sentence as the medical concept.
7. Difference between admission date and the date in the same sentence as the clinical narrative.
8. The learned time-bin of each medical concept. We also derive features based on the overlapping in time-bins for the medical concept pair and the nature of time-bin (past, present, future).
9. Difference in verb patterns in the sentences of the medical concept pair.
10. Difference in dates between the medical concept pair.
11. UMLS relatedness score between the medical concept pair and all the UMLS related and other features described previously in the semantic relatedness section.

When applying CRFs to the problem of assigning medical concepts to time-bins, an observation sequence is medical concepts in the order in which they appear in a clinical narrative, and the state sequence is the corresponding label sequence of time bins. Thus, given a sequence of concepts in narrative order $\{M_1, M_2, M_3, ..\}$, we learn a corresponding label sequence of time-bins {*way before admission, before admission, on admission, after admission, after discharge*}. The learned label sequence is now used as part of the temporal feature set in co-training and posterior regularization for MCCR.

5 Weakly Supervised Learning

5.1 Co-training

We co-train two MaxEnt classifiers, one each on the semantic features f_s and temporal features f_t of the data, to classify pairs of medical concepts as *corefer* or *no-corefer* in a semi-supervised fashion. We use the co-training algorithm proposed by Blum and Mitchell (1998).

The assumption here is that each feature set contains sufficient information to train a model for classification of medical concepts. Consider the concept pair, {*renal inflammation, posterior uveitis*} that corefer. The semantic view for this concept pair may not strongly indicate coreference. The "UMLS relation type" feature indicates that the two concepts are not similar in meaning. However, both concepts are mapped to the same time-bin *after admission*. Thus, the time-bin along with features extracted based on explicit temporal expressions co-occurring with the medical concepts indicate a coreference between the pair of medical concepts. Similarly, the semantic view is confident about confident about the coreference of certain medical concept pairs which do not occur in the same time-bin. The classifiers trained on each view complement each other in the learning process. Thus, we can leverage the predictions made by each classifier on the unlabeled dataset to augment the training data of both classifiers.

The co-training algorithm is shown in Table 1. We set a threshold for an unlabeled sample to be added

[8] http://tartarus.org/martin/PorterStemmer/

```
Function coTrain
Repeat till all unlabeled data is labeled.
    1. Train classifier $c_1$ on $t_{fs}$ to obtain model $m_1$
    2. Train classifier $c_2$ on $t_{ft}$ to obtain model $m_2$
    3. Use $m_1$ to classify a subset of unlabeled data
       and update the training data as,
       $t_{fs}$.subset = $\{u_{subset1}$, predicted label$\}$
       iff classifier confidence $>$ 1/number of labels
    4. Use $m_2$ to classify a subset of unlabeled data
       and update the training data as,
       $t_{ft}$.subset = $\{u_{subset2}$, predicted label$\}$
       iff classifier confidence $>$ 1/number of labels
    5. $t_{fs}$ = $t_{fs}$ + $t_{ft}$.subset +
       $\{u_{subset1}$, predicted label$\}$
    6. $t_{ft}$ = $t_{ft}$ + $t_{fs}$.subset +
       $\{u_{subset2}$, predicted label$\}$
```

Table 1: Co-training algorithm for the binary pairwise classification task of MCCR (Blum and Mitchell, 1998).
c = classifier, *u* = unlabeled data.
$u_{subset1}$, $u_{subset2}$ = subsets of unlabeled data.
$u_{subset1}$ and $u_{subset2}$ are mutually exclusive.
F = $\{f_s, f_t\}$ is the features space divided into conditionally independent semantic and temporal feature sets.
t_{fs} = $\{f_s, l\}$ training data consisting of semantic features of a medical concept pair along with class label.
t_{ft} = $\{f_t, l\}$ training data consisting of temporal features of a medical concept pair along with class label.

into the labeled pool. An unlabeled sample is labeled in a particular iteration, if *classifier confidence* $>$ *1/number of labels*. In the next iteration, randomly pick a subset of unlabeled samples and label all samples in this subset. This could include samples that have already been labeled in previous iterations. A label is assigned in a subsequent iteration if: the sample was previously labeled OR if *classifier confidence* $>$ *threshold*. The parameters in this algorithm are the number of iterations, the pool size of examples selected from the unlabeled set in each iteration and the number of labeled examples added at each iteration to the labeled data pool. Similar to Blum and Mitchell (1998), we update the pool size by $2p + 2n$ in each iteration, where p is the number of medical pairs that corefer and n is the number of medical concept pairs that do not corefer.

5.2 MaxEnt with Posterior Regularization

The next semi-supervised learning method applied to MCCR is MaxEnt with posterior regularization using expectation constraints (Ganchev et al., 2010). This method incorporates prior knowledge directly on the output variables during learning. The prior knowledge is expressed as inequalities on the expected value under the posterior distribution of user-defined constraint features. Thus, posterior regularization incorporates side-information into unsupervised estimation in the form of constraints on the model's posteriors. It is similar to the EM algorithm during learning, but it solves a problem similar to Maximum Entropy inside the E-Step to enforce the constraints.

Posterior regularization is used to derive a multi-view learning algorithm while specifying constraints that the models should agree on the label distribution. We train MaxEnt models based on two views of the data, semantic and temporal. This method starts by considering the setting of complete agreement where there is a common desired output for the two models and each of the two views is sufficiently rich to predict labels accurately. The search is restricted to model pairs p_1, p_2 that satisfy $p_1(y|x) \approx p_2(y|x)$, where p_1 and p_2 each define a distribution over labels. The product distribution $p_1(y_1)p_2(y_2)$ is considered and constraint features are defined such that the proposal distribution $q(y_1, y_2)$ will have the same marginal for y_1 and y_2. There is one constraint feature defined for each label y given by, $\phi_y(y_1, y_2) = \delta(y_1 = y)\delta(y_2 = y)$, where $\delta(.)$ is the 0-1 indicator function. The constraint set $Q = q : Eq[\phi] = 0$ requires that the marginals over the two output variables are identical $q(y_1) = q(y_2)$. An agreement between two models is defined as $agree(p_1, p_2) = argmin\ KL(q(y_1, y_2)||p_1(y_1)p_2(y_2))$ | Eq [ϕ] = 0.

In the semantic feature set, we convert the following feature (described in Section 4) into expectation constraints. The type of relation between the pair of medical concepts, is derived from matching the word stems and querying the UMLS definition and atoms of the medical concepts. Based on the relation between the medical concepts (i.e., partial match, complete match, UMLS definition match, UMLS atom match, and no match), we indicate the probability of label distribution coref and no-coref. If the relation turns out to be no match, there is a high probability that the medical concepts do not corefer. In the temporal feature set, we convert the features based on time-bins of the medical concepts in the pair into expectation constraints.

Class(time-bin)	Precision	Recall
after discharge	96.05	62.53
before admission	94.02	92.44
on admission	33.25	75.16
way before admission	50.42	66.72
after admission	93.62	99.14

Table 2: Sequence tagging of medical concepts with time-bins using CRFs.

6 Experimental Setup

6.1 Corpus Annotation

Annotation of clinical text is a time consuming and costly process. Many annotation efforts have used physicians to annotate the data. Instead, we use annotators that are students or recently graduated students from diverse clinical backgrounds with varying levels of clinical experience. In spite of this diversity, the annotation agreement across our team of annotators is high; all annotators agreed on 89.5% of the events and our overall inter-annotator Cohen's kappa statistic (Conger, 1980) for medical events was 0.865. The annotators mark medical concepts, coereference chains and temporal expressions in the clinical narratives and the NEJM case reports. They also map each medical concept to a UMLS CUI.

6.2 Feature Extraction

The first step involves extraction of semantic and temporal features for the annotated medical concepts, as described in Section 4 from both corpora. The semantic relatedness scores are computed using the kDLS (Xiang et al., 2011) method to calculate the relationship between concepts in the UMLS with value of γ set to 7. The type of relation between medical concepts is derived by matching word stems in each medical concept using the Lucene[9] implementation of the Porter stemming algorithm. We query the latest release (UMLS 2011AB) of the UMLS Metathesaurus for finding a match between medical concept and the UMLS definition or UMLS atoms. The WordNet similarity score is computed using Java API for WordNet Searching (JAWS).[10]

Explicit temporal expressions annotated in the corpora are included in our temporal feature set. Medical concepts in the NEJM are mostly described temporally relative to the patient's admission. Temporal expressions like "2 years before admission" and "3 weeks before admission" are common. Hence, we use a heuristic-based algorithm to associate medical concepts with explicit temporal expressions in the NEJM corpus. The algorithm parses case reports and identifies the temporal expressions anchored to admission. All medical concepts following such a temporal expression are anchored to it until a new temporal expression is encountered. Over 88% of the medical concept-temporal expression associations done with the algorithm above is accurate when compared against the NEJM gold standard.

Class	NEJM		Clinical Narratives	
	Precision	Recall	Precision	Recall
coref	79.24	94.53	74.81	88.33
no-coref	86.71	90.62	83.92	94.86

Table 3: Supervised learning for MCCR.

As described in Section 4, we apply sequence tagging using a CRF to assign medical concepts in clinical narratives to time-bins. We use the implementation of CRF in Mallet,[11] trained by Limited-Memory BFGS for our experiments. We use the Stanford POS tagger[12] to identify verbs and derive verb patterns. The dataset for the task of assigning medical concepts to time-bins consisted of 1613 medical concepts. We used a 60-40 train-test split to train a CRF using a sequence of medical concepts and observed an overall accuracy of 92%. The precision and recall values for each time-bin class is indicated in Table 2. The percentage of medical concepts that fall under "way before admission" and "on admission" are less than 5%, affecting the learning accuracy of those classes. When modeled as a multiclass classification task using MaxEnt, we achieve around 86% accuracy.

7 MCCR Results and Discussion

We perform the following experiments for pairwise MCCR: 1) Supervised learning with a MaxEnt classifier, using the combined semantic and temporal feature set, 2) Co-training two MaxEnt models, 3) Training MaxEnt models with using posterior regularization.

[9] http://lucene.apache.org/

[10] http://lyle.smu.edu/~tspell/jaws/

[11] http://mallet.cs.umass.edu/

[12] http://nlp.stanford.edu/software/tagger.shtml

Class	NEJM		Clinical Narratives	
Co-train	Precision	Recall	Precision	Recall
coref	70.32	82.54	69.26	87.31
no-coref	82.54	84.85	71.15	89.44
PR	Precision	Recall	Precision	Recall
coref	76.63	90.41	74.81	84.25
no-coref	80.35	89.21	78.93	87.46

Table 4: Co-training and posterior regularization (PR) for MCCR using semantic and temporal feature sets.

We use the MaxEnt classifier available in Mallet for 1) and 2) and the the Mallet implementation of MaxEnt models with posterior regularization for 3).

The NEJM corpus has 722 medical concepts, 12576 candidate pairs of medical concepts including 137 pairs that corefer. We include all 12576 pairs in our experiments. The clinical narrative corpus has 1613 medical concepts. The candidate pairs and coreference chains for each patient is as follows. Patient 1 has 241001 candidate pairs, 29 coreference chains. Patient 2 has 149604 candidate pairs, 9 coreference chains. Patient 3 has 6,446,521 candidate pairs, 20 coreference chains. From all the candidate pairs in the clinical narrative corpus, 1025 pairs corefer. We randomly sample the no-coref instances to restrict the corpus size to 1 million candidate pairs of medical concepts.

The results for all 3 experiments for both corpora is shown in Tables 3, 4. We also train-test a supervised MaxEnt classifier on a 60-40 split of the entire corpus. This gives us a precision of 74.81% and 88.33% recall (coref) for the binary classification task of pairwise MCCR in the clinical narratives corpus. In the both the semi-supervised experiments, we use an initial labeled pool size of 30 where 12 medical concept pairs that corefer (p) and 18 that do not corefer (n). The growth size is each iteration of co-training is $2p + 2n$. At each iteration, confidently labeled examples are added to the training set from the previous iteration. The co-training algorithm is run until all unlabeled instances become labeled. The parameters in the posterior regularization implementation include the regularization penalty for each step and the number of iterations. We use the default values (maxIterations=100, pGaussianPriorVariance=0.1, qGaussianPriorVariance=1000) suggested on the Mallet toolkit page (Bellare et al., 2009). Co-training two MaxEnt models based on independent semantic and temporal views of the data results in 69.26% precision and 87.31% recall (coref), whereas training MaxEnt models with expectation constraints gives us 74.81% precision and 84.25% recall (coref), on the corpus of clinical narratives.

Posterior regularization does better than co-training and the performance of both the semi-supervised methods is comparable to if not as good as the supervised classifier trained on a 60-40 split of the corpus. Thus, our results indicate that the use of semantic and temporal features is effective for MCCR in clinical text. It is clear from the co-training and posterior regularization results that treating MCCR as a semi-supervised problem works.

8 Conclusions

We investigated the task of MCCR in clinical text using supervised and semi-supervised learning methods. We create annotated corpora of clinical text with case reports from the NEJM and narratives obtained from The Ohio State University Wexner Medical Center. We work with the hypothesis that determining semantic and temporal similarity between medical concepts helps resolve coreferences. In order to test this hypothesis, we describe the process of semantic and temporal feature extraction from clinical text. We demonstrate the effectiveness of the extracted features in a supervised binary classification task for MCCR with MaxEnt classifiers (using the combined feature set) as well as using semi-supervised methods of co-training MaxEnt classifiers and training MaxEnt models using posterior regularization (using two independent views of the data - semantic view and temporal view). Thus, we show that MCRR can be performed using semi-supervised learning with semantic and temporal views of the data.

Acknowledgments

The project described was supported by the National Center for Research Resources, Grant UL1RR025755, KL2RR025754, and TL1RR025753, and is now at the National Center for Advancing Translational Sciences, Grant 8KL2TR000112-05, 8UL1TR000090-05, 8TL1TR000091-05. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Sophia Ananiadou, Carol Freidman, and Juníchi Tsujii. 2004. Introduction: named entity recognition in biomedicine. *J. of Biomedical Informatics*, pages 393–395.

Kedar Bellare, Gregory Druck, and Andrew McCallum. 2009. Alternating projections for learning with expectation constraints. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 43–50.

Avrim Blum and Tom M. Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *COLT'98*, pages 92–100.

Bigi Brigitte. 2003. Using Kullback-Leibler distance for text categorization. In *Proceedings of the 25th European conference on IR research*, ECIR'03, pages 305–319.

Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Guergana K Savova Leonard W D'Avolio, and Ozlem Uzuner. 2011. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. In *JAMIA*.

Jung-Hsien Chiang, Jou-Wei Lin, and Chen-Wei Yang. 2010. Automated evaluation of electronic discharge notes to assess quality of care for cardiovascular diseases using Medical Language Extraction and Encoding System (MedLEE). *JAMIA*, pages 245–252.

A.J. Conger. 1980. Integration and generalization of kappas for multiple raters. In *Psychological Bulletin Vol 88(2)*, pages 322–328.

Peter J Embi and Philip Payne. 2009. Clinical research informatics: challenges, opportunities and definition for an emerging domain. *Journal of the American Medical Informatics Association*, 16(3):316–327.

Carol Friedman, Pauline Kra, and Andrey Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35(4):222–235.

Rob Gaizauskas, Henk Harkema, Mark Hepple, and Andrea Setzer. 2006. Task-oriented extraction of temporal information: The case of clinical narratives. In *Proceedings of the Thirteenth International Symposium on Temporal Representation and Reasoning*, TIME '06, pages 188–195.

Kuzman Ganchev, Joo Graa, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, pages 2001–2049.

Tian Ye He. 2007. *Coreference Resolution on Entities and Events for Hospital Discharge Summaries*. EECS, Cambridge, MA, MIT. M.Eng.

Hyuckchul Jung, James Allen, Nate Blaylock, Will de Beaumont, Lucian Galescu, and Mary Swift. 2011. Building timelines from narrative clinical records: initial results based-on deep natural language understanding. In *Proceedings of BioNLP 2011 Workshop*, BioNLP '11, pages 146–154.

Christoph Muller, Stefan Rapp, and Michael Strube. 2002. Applying co-training to reference resolution. In *ACL*, pages 352–359.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the ACL*, pages 1396–1411.

Kamal Nigam and Rayid Ghani. 2000. Analyzing the effectiveness and applicability of co-training. In *CIKM'00*, pages 86–93.

James Pustejovsky, Jos M. Castao, Robert Ingria, Roser Sauri, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering'03*, pages 28–34.

Preethi Raghavan and Albert M. Lai. 2010. Leveraging natural language processing of clinical narratives for phenotype modeling. In *PIKM'10*, pages 57–66.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 492–501, Stroudsburg, PA, USA. Association for Computational Linguistics.

Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin Kipper Schuler, and Christopher G. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *JAMIA*, pages 507–513.

Guergana K. Savova, Wendy Webber Chapman, Jiaping Zheng, and Rebecca S. Crowley. 2011. Anaphoric relations in the clinical narrative: corpus creation. *JAMIA*, 18(4):459–465.

Wee Meng Soon, Hwee Tou Ng, and Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, pages 521–544.

Yang Xiang, Kewei Lu, Stephen L James, Tara B Borlawsky, Kun Huang, and Philip R O Payne. 2011. k-neighborhood decentralization: A comprehensive solution to index the UMLS for scale knowledge discovery. In *Journal of Biomedical Informatics*.

Jiaping Zheng, Wendy Webber Chapman, Rebecca S. Crowley, and Guergana K. Savova. 2011. Coreference resolution: A review of general methodologies and applications in the clinical domain. *Journal of Biomedical Informatics*, 44(6):1113–1122.

Li Zhou and George Hripcsak. 2007. Temporal reasoning with medical data - a review with emphasis on medical natural language processing. *Journal of Biomedical Informatics*, pages 183–202.

Li Zhou, Genevieve B. Melton, Simon Parsons, and George Hripcsak. 2006. A temporal constraint structure for extracting temporal information from clinical narrative. *Journal of Biomedical Informatics*, pages 424–439.