Exploiting Unlabeled Texts with Clustering-based Instance Selection for Medical Relation Classification

Youngjun Kim, MS[1], Ellen Riloff, PhD[1], Stéphane M. Meystre, MD, PhD[2]
[1]School of Computing, University of Utah, Salt Lake City, UT, USA;
[2]Medical University of South Carolina, Charleston, SC

Abstract

Classifying relations between pairs of medical concepts in clinical texts is a crucial task to acquire empirical evidence relevant to patient care. Due to limited labeled data and extremely unbalanced class distributions, medical relation classification systems struggle to achieve good performance on less common relation types, which capture valuable information that is important to identify. Our research aims to improve relation classification using weakly supervised learning. We present two clustering-based instance selection methods that acquire a diverse and balanced set of additional training instances from unlabeled data. The first method selects one representative instance from each cluster containing only unlabeled data. The second method selects a counterpart for each training instance using clusters containing both labeled and unlabeled data. These new instance selection methods for weakly supervised learning achieve substantial recall gains for the minority relation classes compared to supervised learning, while yielding comparable performance on the majority relation classes.

Introduction

Electronic health record (EHR) systems are becoming more prevalent in the U.S.[1] This growth has resulted in very large quantities of clinical patient data becoming available in electronic format, which holds tremendous potential for benefitting clinical research, quality improvement, and surveillance. A substantial proportion of patient-specific information in the EHR is only found in narrative, unstructured clinical notes.[2] Natural Language Processing (NLP) enables fast, scalable, and accurate extraction of structured and coded information from these clinical notes.[3] As part of this information extraction task, identifying semantic relations between concepts is essential to provide accurate and complete information about the concepts and their meaning. For example, extracting relations between mentions of a medication and mentions of allergy symptoms enables differentiation between situations when a medication causes the symptoms and situations when a medication is prescribed to alleviate symptoms.

Given a pair of medical concepts found in a sentence, a relation classification system must determine the type of relation that exists between the two concepts. Our research focuses on the relation classification task introduced in 2010 for the i2b2 Challenge Shared Tasks[4]. This task involves recognizing eight types of relations between pairs of three types of medical concepts: problems, treatments, and tests.

A key challenge of this task is the extremely skewed class distribution across relation types. For example, five types of relations are defined between problems and treatments plus a no relation category (*None*), but two of these categories (*None* and *TrAP* (treatment administered for problem)) account for 86% of the instances in the i2b2 Challenge data. Four relation types (*TrCP* (treatment causes problem), *TrIP* (treatment improves problem), *TrWP* (treatment worsens problem), and *TrNAP* (treatment not administered because of problem)) are distributed across the remaining 14% of the data. Each of these "minority" relations appears in just 2-6% of the data. Identifying these minority relations is extremely important from a practical perspective because they hold valuable information. For example, the dominant relations between problems and treatments are *TrAP* (administration of treatment) and *None* (no relation at all). In contrast, the minority relations (*TrCP, TrIP, TrWP, TrNAP*) represent situations where a treatment *causes*, *improves*, *worsens*, or is *contraindicated* for a problem, so they are arguably more important types of situations to recognize.

The most successful methods used for relation classification include various supervised machine learning algorithms.[4] Extremely skewed class distributions pose substantial challenges for supervised machine learning (ML) because only a small number of labeled examples are available for training. As a result, ML classifiers can achieve high accuracy for the dominant classes but often perform poorly with the minority classes. Manually annotating more data is not a viable solution because of the high cost of manual annotation by medical experts. Also, because the minority classes are relatively rare, each batch of new annotations would provide only a relatively small number of new examples. There is substantial cost for low reward.

Our research aims to improve relation classification in clinical texts with an emphasis on minority classes by exploiting large amounts of unlabeled clinical texts, which are readily available in abundant quantity. We present two new methods to selectively choose unlabeled instances for self-training in an iterative weakly supervised learning framework. Both methods apply a clustering algorithm to group instances into clusters based on similarity measures. The first method, called *Unlabeled Data Prototypes (UDP) Selection*, uses clusters containing only unlabeled instances and identifies one "prototype" instance from each cluster to use as additional training data. Intuitively, this method aims to identify the different types of instances that occur in the unlabeled data and selects a representative subset of them. The second method, called *Labeled Data Counterparts (LDC) Selection*, uses clusters containing both labeled and unlabeled data. For each labeled instance, this method identifies its closest counterpart in the cluster by selecting the unlabeled instance that is most similar to it. Intuitively, this method is designed to acquire a new set of training instances that mimic both the class distribution and semantic content (based on feature similarity) of the original training instances. Our experimental results show that these two instance selection methods produce classifiers that can identify minority relation classes more often and more accurately than traditional supervised learning or self-training.

Background

The relation classification task was defined as part of the Fourth i2b2/VA Shared Task Challenge[4] in 2010. Our research involves relation classification for pairs of medical concepts, assuming that the terms corresponding to the two concepts have already been identified. The task is to identify how medical problems relate to treatments, tests, and other medical problems in clinical texts. Many sentences contain multiple pairs of concepts, so the challenge includes identifying which pairs are related, as well as identifying the specific type of relation. This task has been typically cast as a supervised learning problem, where a classifier is trained with manually annotated data. Rink et al.[5] used supervised learning to produce the highest micro-averaged F_1 score, 73.7%, for this relation extraction task. Their system utilized external resources including Wikipedia, WordNet, and the General Inquirer lexicon[6] as part of their feature set. To improve recall, they set much lower weights to the pairs of non-related concepts (i.e., negative examples) when training an SVM (Support Vector Machine) classifier.

Previous work has presented micro-averaged F_1 scores, which assess performance over all of the positive instances regardless of which class they belong. However, micro-averaging obscures performance differences across the classes. For example, it is often possible for a system to achieve a high micro-averaged F_1 score by performing well on the majority class but recognizing few, if any, instances of the minority classes. Our research aims to shed light on the performance differences across relation classes, with the goal of comparing the ability of different methods to recognize the minority classes. So we will present macro-averaged F_1 scores in the rest of this manuscript.

The Rink et al. system reached macro-averaged scores of 51.7% recall, 55.8% precision, and 53.7% F_1 score (not officially reported in Rink et al.[5] but calculated by taking the average of the reported recall and precision of the different sub-classes). de Bruijn et al.[7] explored effective features also applicable to other clinical NLP tasks. In addition to supervised classification, they applied self-training on the provided unlabeled data. Their approach yielded a 73.1% micro-averaged F_1 score. The macro-averaged scores for their submission reached 43.7% recall, 66.8% precision, and 51.2% F_1 score. These results were calculated by the authors of this manuscript based on the output of de Bruijn et al.'s system[7]. Their subsequent research[8] using composite-kernel learning improved the accuracy of relation classification with a higher micro-averaged F_1 score of 74.2%. As an effort to overcome the class imbalance problem, they used down-sampling of negative examples before training the models. D'Souza and Ng[9] presented an ensemble approach exploiting human-supplied knowledge to set up individual classifiers. Their weighted-voting system outperformed a single classifier using the full set of features exploited by different members. Their best-scoring ensemble system produced 69.6% micro-averaged F_1 score. Note that their result is not directly comparable with the works described above because of different training data sizes.

In the biomedical and clinical domains, annotating data is especially expensive because of the need for domain experts. Consequently, most systems are trained with relatively small amounts of labeled text, even though much larger amounts of unlabeled text are readily available. Weakly supervised learning, also called semi-supervised learning, has been shown to benefit from training on both labeled and unlabeled data for other NLP tasks, including document classification[10], named entity recognition[11], and noun phrase chunking[12]. As a general framework, starting with a small set of initial labeled data, the learner outputs entities from unlabeled data with assigned entity types. Then, the detected entities are collected as new training instances for subsequent iterations. Iterative bootstrapping methods that use seeding heuristics to produce an initial set of training instances have also been a popular choice. Thelen and Riloff[13] showed that semantic lexicons could be learned with extraction patterns from unlabeled texts by

bootstrapping algorithms. Rosenberg et al.[14] used self-training to build object detection models and they pointed out that the choice of the initial seeds has a large effect on performance.

There has also been previous work on the relation classification task exploiting unlabeled data. Zhang[15] proposed a bootstrapping algorithm using random feature projection. Multiple classifiers were trained with randomly selected features from labeled data and they voted to assign labels to the unlabeled data. Mintz et al.[16] used Freebase[17], a large knowledge database, to train a learner with "distant" supervision. Sun et al.[18] presented a weakly supervised learning method with large-scale word clustering. They augmented the features derived from the word clusters to compensate for the absence of lexical features in labeled data. Related to medical relations, Wang and Fan[19] collected training data using a clustering algorithm. To minimize the manual annotations, the most representative instances with the highest average similarity to other members of each cluster were chosen for annotation.

Materials and Methods

Labeled Data Description

We used the i2b2/VA 2010 Shared Task corpus for our research, which consists of a training set of 349 annotated clinical notes and a test set of 477 annotated clinical notes. This test set contains 45,009 annotated medical concepts with 9,069 relations that occur in the same sentence. Relations were defined as follows[20]:

Medical problem—treatment (***Pr-Tr***) relations:

- Treatment *improves* medical problem (*TrIP*).
- Treatment *worsens* medical problem (*TrWP*).
- Treatment *causes* medical problem (*TrCP*).
- Treatment is *administered* for medical problem (*TrAP*).
- Treatment is *not administered* because of medical problem (*TrNAP*).
- Relation that does not fit into one of the above defined relationships (*NoneTrP*).

Medical problem—test (***Pr-Te***) relations:

- Test *reveals* medical problem (*TeRP*).
- Test *conducted to investigate* medical problem (*TeCP*).
- Relation that does not fit into one of the above defined relationships (*NoneTeP*).

Medical problem—medical problem (***Pr-Pr***) relations:

- Medical problem *indicates* medical problem (*PIP*).
- Relation that does not fit into PIP relationship (*NonePP*).

The test set contains 6,949 *Pr-Tr* pairs that occur in the same sentence, of which 3,463 are positive examples (participate in a relation) and 3,486 are negative examples (*NoneTrP*). *Pr-Te* relations include 3,620 positive and 2,452 negative examples (*NoneTeP*). *Pr-Pr* relations include 1,986 positive and 11,190 negative examples (*NonePP*). As seen in Figure 1, the class distributions across *Pr-Tr* and *Pr-Te* relation types are extremely skewed.

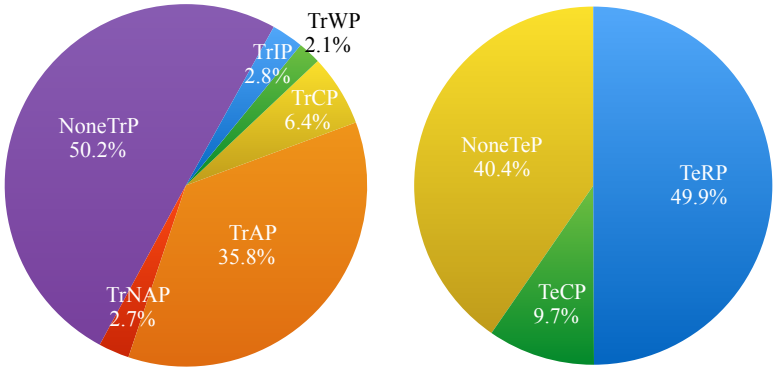


Figure 1. Distribution of treatment (*Pr-Tr)* and test (*Pr-Te)* relation types in the test set

Among *Pr-Tr* relations, four "minority" classes, *TrCP*, *TrIP*, *TrWP*, *TrNAP*, are distributed across 14% of the data. Each of these relations appears in just ~2-6% of the data. Among the *Pr-Te* relations, *TeCP* is the minority class, accounting for < 10% of the instances. Our goal is to improve relation classification with an emphasis on these minority classes by exploiting large amounts of unlabeled clinical texts. Since there is only one type of *Pr-Pr* relation (*PIP*), we focused exclusively on the *Pr-Tr* and *Pr-Te* relations in our efforts.

Unlabeled Data Preparation for Weakly Supervised Learning

For this research, we also used texts from the MIMIC II Clinical Database[21], which contains various types of clinical notes: discharge summaries, nursing progress notes, cardiac catheterization notes, ECG reports, radiology reports, and echocardiography reports. From this data set, we used 26,485 discharge summaries after filtering out notes with insufficient text content (<500 Bytes).

For weakly supervised learning preparation, we had to identify the medical concepts in our unlabeled data and classify the assertion of each medical problem concept. Assertion classification aims to determine the assertions of the problem concepts by assigning one of six categories: *present*, *absent*, *hypothetical*, *possible*, *conditional*, or *not associated with the patient*.[4] For concept extraction, we used our previous system consisting of a stacked learning ensemble[22, 23]. We slightly modified the feature set of the individual classifiers by adding skip-grams and word embedding features. For assertion classification of medical problems, we also added new word embedding features to our assertion classifier[24] and retrained the SVM model. As computed by the i2b2 Challenge evaluation script (class exact match), our stacked ensemble achieved 84.4% recall, 89.1% precision, and 86.7% F_1 score for concept extraction on the i2b2 test set. The assertion classifier reached 94.5% micro-averaged F_1 score. Using the predicted concepts assigned to the unlabeled data, we created a large set of relation pairs to generate feature vectors for weakly supervised learning and clustering.

We used CLUTO[25], a data clustering software that has been widely used in various tasks, to create clusters containing both labeled (i2b2 training) and unlabeled data: 517,689 pairs of *Pr-Tr* relations and 455,272 pairs of *Pr-Te* relations. The same feature vectors generated for SVM classification were re-used with the clustering algorithm. To determine the number of clusters, we use the root-mean-square standard deviation (RMSSD). RMSSD is a measure of homogeneity within clusters and large RMSSD values indicate that clusters are not homogeneous.[26] We ran a series of clustering processes with different numbers of clusters, *K,* and calculated the RMSSD for each *K*. We tried 20 different cluster sizes aimed at having the average number of members per cluster ranging from 40 to 800 (i.e. K = the number of instances × n, n = 1/800, 2/800,…, 19/800, 20/800). When we set n to 1/800 and 20/800 (= 1/40), we expected that on average 800 and 40 members would exist in each cluster, respectively. For each of the *Pr-Tr* and *Pr-Te,* we then detected the shift point (also known as the "Knee" point) of its RMSSD curve based on the Satopää et al.[27] method. The cluster sizes of 4,529 and 3,414 were identified as the Knee points for the *Pr-Tr* and *Pr-Te* relation clusters respectively. In the following paragraphs, we will describe our supervised classification models and then present the instance selection methods based on clustering unlabeled data.

Supervised Relation Classification

We created three supervised learning classifiers (one for each category of concept pairs: Pr-Tr, Pr-Te, and Pr-Pr) using a rich set of features. We applied the Stanford CoreNLP tool[28] for tokenization, lemmatization, part-of-speech (POS) tagging, and phrase chunking. We trained Support Vector Machine (SVM) classifiers with a linear kernel using the LIBLINEAR (Library for Large Linear Classification) software package[29]. The multi-class SVM classifiers use five types of features associated with a pair of concepts $<C_1, C_2>$:

- Assertion: For each medical problem concept, we create a feature for the assigned assertion type. Assertion categories are considered in a pre-defined order of precedence (e.g., *Possible* takes precedence over *Absent*.)[30]
- Context: To compensate for the absence of assertions for treatment and test concepts, we incorporated the *ConText* algorithm[31] at the sentence level to detect three types of contextual properties for each concept: *negation*, *hypothetical*, and *historical*. We also created a second set of *ConText* algorithm properties restricted to the six-word context window around C_1 and C_2 (three words on the left of C_1 and three words on the right of C_2).
- Distance: We created several features to represent the distance between concepts C_1 and C_2 by counting the number of words, concepts, and phrases (e.g., noun phrases and adjective phrases) between them. The number of concepts appearing before or after the pair was also measured. These features were designed to

help the classifiers distinguish between concept pairs that probably have a relation and distant pairs that probably have no relation between them.

- Lexical: Lexical features have been very effective for many NLP tasks. We create lexical features for the words contained in C_1 and C_2, the head words of C_1 and C_2, the words in a context window of size four around the concept pair (two words on the left of C_1 and two words on the right of C_2), and the words in between the two concepts. Also, we defined features for verbs that precede, follow, or occur between the concepts.
- Word Embedding: We used the *Word2Vec* software[32] to perform *K*-means clustering on the word embeddings. We created 1,000 clusters of semantically related words within the unlabeled data (i.e., MIMIC II Clinical Database[21]) with default parameters. Then, we used the cluster identifier of each word between the two concepts as a feature. We also used the cosine similarity of the word embedding vectors for the heads of C_1 and C_2.

We randomly selected 200 documents from the training set for development purposes. We tuned LIBLINEAR's parameters to maximize the micro-averaged F_1 score with the held-out development data. After experimenting with different values on the development data, we set the cost parameter *c* to 0.06 for *Pr-Tr*, and 0.02 for *Pr-Te* and *Pr-Pr*. Also, the weights of negative examples were set to 0.2 for *Pr-Tr* and *Pr-Te* and 0.3 for *Pr-Pr*. The lower the weight for instances with no relation, the higher recall was obtained on held-out data.

Although the classifiers showed good performance under the micro-averaged scoring metrics, performance on the minority classes was weak. As shown earlier, the class distributions are extremely skewed and the minority classes are relatively rare. To reduce the performance gap between the dominant classes and the minority classes, we also experimented with retraining the model by assigning higher weights to the minority classes to increase the importance of minority classes being classified correctly. It did not yield an increase in macro-averaged F_1 score and more detailed results will be reported in the results section. To improve performance across the different relation classes, we extend our methods to weakly supervised learning described in the following paragraphs.

Exploiting Unlabeled Data for Relation Classification

To take advantage of the large amounts of unlabeled clinical notes that are available, we explored an iterative weakly supervised learning framework. We developed two novel methods for instance selection that are specifically aimed at improving performance on minority classes. Our general framework involves the following steps: (1) a classifier is trained with supervised learning using the labeled training data, (2) the classifier is applied to the unlabeled data so that each unlabeled instance receives a predicted label, (3) a subset of the unlabeled instances is selected and then added to the set of labeled data (using the classifier's predictions as the labels), and (4) the classifier is retrained using the (larger) set of labeled data. This process repeats until a stopping criterion is met (e.g., for a fixed number of iterations or until no new instances can be labeled).

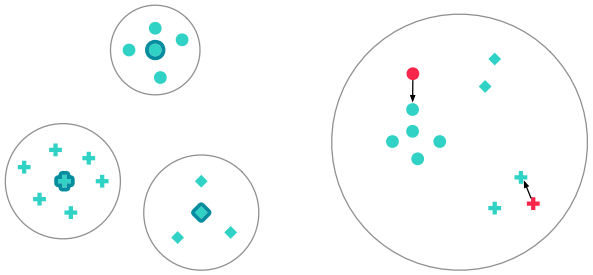
This paradigm is generally known as *self-training*, where the most common method for instance selection (step 3) sorts the instances based on the confidence scores produced by the classifier (i.e., confidence in the predicted labels) and then selects the most confidently labeled instances. This traditional self-training approach, however, tends to select instances of the dominant classes much more often than the minority classes because the classifier is more confident in its predictions for the dominant classes.

This issue motivated us to explore new methods for instance selection that try to create a diverse and representative set of new instances from the unlabeled data. Consequently, we developed two new methods for instance selection that first cluster the unlabeled data to identify groups of similar instances. Both methods generate clusters and assign labels to the instances in the same way. First, the labeled and unlabeled instances are combined into a single dataset and the clustering algorithm (described previously) is applied. We consider the instances with a high confidence score as candidates for selection. In each iteration, an instance can be selected when it is ranked in the top 25% per class. Then we assign a label to each cluster based on the most common relation type among these highly ranked instances. Each unlabeled instance selected from a cluster is assigned the relation type of the cluster.

The first instance selection method, called *Unlabeled Data Prototypes (UDP) Selection*, selects instances from clusters containing only unlabeled data. We compute the purity of each cluster and identify clusters where the highly confident cluster members have the same positive relation type (i.e., cluster purity = 1). We discard clusters with purity < 1 because the instances are similar but the classifier's predictions are inconsistent, so the predictions are suspect. The most representative instance from each cluster is then selected as additional training data, based on

average cosine similarity with other cluster members. We assumed that instances in these clusters are different from the training instances yet they are similar to each other in some way, so they represent some new type of information found in the unlabeled data. This method is illustrated in Figure 2(a). Green-colored instances represent unlabeled data.

Assuming that unlabeled data will be similar to labeled data when they co-exist in the same cluster, our second method, called *Labeled Data Counterparts (LDC) Selection*, selects instances from the clusters containing both labeled and unlabeled instances. For each instance labeled with a positive relation type, the unlabeled instance most similar to it in the same cluster is selected. Our intuition is that this approach will acquire new training instances that share features with the original labeled data and maintain the same class distribution. This method is illustrated in Figure 2(b). Red-colored instances represent labeled data and green-colored instances represent unlabeled data. In the next sections, we compare the performance of self-training with confidence-based instance selection against our new UDP and LDC instance selection methods.



(a) *UDP* (b) *LDC*

Figure 2. Clustering-based instance selection

Results

We have conducted an extensive set of experiments to evaluate the performance of supervised classifiers and weakly supervised learning with different instance selection methods. We evaluated performance with relation data from the i2b2 Challenge test set. We used the official i2b2 Challenge evaluation script to calculate micro-averaged measures. For macro-averaged measures, we created a new script to obtain average values for each relation type. The macro-averaged F_1 score is the harmonic mean of the macro-averaged recall and precision.

Supervised Learning Results

Table 1 shows the results produced with the supervised classifiers, which were trained to optimize for micro-averaged measures. This baseline supervised learning system was trained with the i2b2 training data and achieved micro-averaged scores of 74.9% recall, 73.7% precision, and 74.3% F_1 score.

Table 1. Results produced with the supervised classifier.

Relation type	Recall	Precision	F_1 score
ALL	74.9	73.7	74.3
Treatment	67.4	68.9	68.2
TrIP	31.8	63.6	42.4
TrWP	4.2	42.9	7.6
TrCP	52.3	59.5	55.6
TrAP	79.9	71.2	75.3
TrNAP	25.1	49.5	33.3
Test	82.9	81.5	82.2
TeRP	90.3	82.7	86.3
TeCP	45.1	71.4	55.3
PIP	73.2	67.9	70.4

Although our supervised classifiers achieve overall performance comparable to state-of-the-art relation classification systems, performance on the minority classes lags far behind the dominant classes. The F_1 score of *TrWP* was only 7.6% with a recall of 4.2%. Most of the *TrWP* instances were misclassified because of the very low prevalence of their cases. For example, a *TrWP* case from the test set, "*She has a known diagnosis of myelodysplasia that has become recalcitrant to Procrit*", the medical problem '*myelodysplasia*', the treatment '*Procrit*', and possibly a keyword '*recalcitrant*' never appeared in the training data. Based on macro-averaging, this system reached 50.2% recall, 63.6% precision, and 56.1% F_1 score.

We also experimented with decreasing the weights of negative examples to help increase recall on minority classes. This did not yield an increase in macro-averaged F_1 score because of a substantial drop in precision. Adjusting the importance of different relation types by assigning different weights also did not affect performance very much.

Comparing Supervised Learning and Weakly Supervised Learning Results

We evaluated the performance of self-training with traditional confidence-based instance selection (called **Self-training** below), and instance selection with our new *UDP* and *LDC* methods. We ran all of the weakly supervised learning methods for 20 iterations.

For self-training, we only selected positive examples (pairs of concepts with relations) from the unlabeled data to augment the labeled data. For each iteration, we added *K* newly labeled examples, where *K* = the number of positive examples in the original training data. Our intention was to be conservative in adding new examples and maintain the importance of labeled data. To keep the class distribution of the labeled data, we imposed that the number of newly labeled examples for each positive class should not exceed the number of examples in the original training data.

Table 2 shows results for each class and macro-averaged F_1 scores for the *Pr-Tr* and *Pr-Te* relations. For each relation type, the best results appear in boldface. We used paired t-tests to measure statistical significance.[33] Results that are significantly different from the supervised learning results at the 95% significance level are preceded by an asterisk (*). Self-training with confidence-based instance selection produced the best F_1 score on *TrCP* and *TrNAP* classes. For *TrWP* and *TeCP*, self-training's performance was significantly different than supervised learning.

Table 2. F_1 score of each method on the test set.

Relation type	Supervised	Self-training	*UDP*	*LDC*
Treatment	46.2	48.0	48.9	**49.7**
TrIP	42.4	46.0	***49.3**	*47.4
TrWP	7.6	*16.3	12.3	***19.2**
TrCP	55.6	**56.8**	55.5	53.1
TrAP	75.3	75.4	**75.8**	75.8
TrNAP	33.3	**35.4**	33.1	33.6
Test	72.0	72.6	72.8	**73.1**
TeRP	86.3	86.3	86.3	***86.7**
TeCP	55.3	*58.5	*59.2	***59.5**

Both the *UDP* and *LDC* instance selection methods produced higher macro-averaged F_1 scores than Self-training. The *UDP* method (third column of Table 2) produced the best F_1 score of 49.3% on the *TrIP* class. The F_1 scores for *TrIP* and *TeCP* were significantly higher than for supervised learning. The *LDC* method (last column of Table 2) produced the highest F_1 scores on most of the relation classes. It obtained the best macro-averaged F_1 scores for Treatment and Test. For *TrIP, TrWP*, *TeRP*, and *TeCP*, the performance of *LDC* method was significantly better than supervised learning.

Finally, we tried to combine the *UDP* and *LDC* methods. New instances were selected separately by the *UDP* and *LDC* methods and then the combined set of instances was added to the labeled data. However, this system produced an F_1 score of 58.3%, so did not outperform the *LDC* method on its own.

In another set of experiments, we performed ablation testing of the supervised learning system to evaluate the impact of each feature set based on micro-averaged and macro-averaged scores, separately. If some features have more impact for macro-averaged scores than micro-averaged scores, then our hypothesis is that they are especially important features for minority classes. The row header in Table 3 specifies the feature set that has been ablated. The columns named "Impact" in Table 3 present the F_1 score difference between the ablated classifier and the complete system. Every feature set contributed to the performance of the supervised classifiers. The macro-averaged F_1 score dropped the most when the lexical features were removed. This suggests that exploiting unlabeled data could be especially beneficial for the minority classes by bringing in new lexical features. The F_1 scores of *TrIP, TrNAP*, and *TeCP* decreased from 42.4%, 33.3%, and 55.3% to 28.9%, 22.2%, and 37.2% respectively without the lexical features.

Table 3. Features Contribution

	Macro-averaged		**Micro-averaged**	
Feature	**F_1 score**	**Impact**	**F_1 score**	**Impact**
All	56.1		74.3	
- Assertion	55.4	-0.7	73.8	-0.5
- Contextual	55.4	-0.7	74.2	-0.1
- Distance	55.2	-0.9	72.4	-1.9
- Lexical	48.2	-8.0	69.4	-4.9
- Word embedding	55.8	-0.3	73.8	-0.5

Analysis

We carried out an empirical analysis of self-training with confidence-based instance selection to better understand its limitations. After clustering the unlabeled data, we counted the number of instances selected from each cluster during the first iteration. We found that most instances were selected from a small subset of the clusters: about 10% of the clusters provided over 78% of the newly selected unlabeled instances. This shows that selecting instances based only on confidence scores tends to yield a relatively homogenous set of new instances that is low in diversity.

Table 4 displays the Recall, Precision, and F_1 results of *LDC* instance selection along with the total counts of true positives (TP) and the number and percentage of true positive gains (compared to supervised learning) in the rightmost column. The numbers in parentheses in the Recall, Precision, and F_1 columns indicate the difference between the supervised classifier and the LDC method. Results significantly different from supervised learning at the 95% significance level are preceded by an asterisk (*).

Table 4. Results of *LDC* with comparison to the supervised learning model

Relation type	**Recall**		**Precision**		**F_1 score**		**True positive**	**TP Gain (%)**	
Minority classes									
TrIP	*38.9	(+7.1)	60.6	(-3.0)	*47.4	(+5.0)	77	14	(+22.2)
TrWP	*11.9	(+7.7)	50.0	(+7.1)	*19.2	(+11.6)	17	11	(+183.3)
TrCP	*65.1	(+12.8)	*44.9	(-14.6)	53.1	(-2.5)	289	57	(+24.6)
TrNAP	23.6	(-1.6)	*58.4	(+9.0)	33.6	(+0.3)	45	-3	(-6.3)
TeCP	*57.7	(+12.6)	*61.4	(-10.0)	*59.5	(+4.2)	339	74	(+27.9)
Majority classes									
TrAP	*80.8	(+0.9)	71.3	(+0.2)	75.8	(+0.5)	2,009	23	(+1.2)
TeRP	*88.5	(-1.8)	*85.0	(+2.4)	*86.7	(+0.4)	2,682	-55	(-2.0)

Table 4 shows that most of the minority classes benefitted substantially from the *LDC* method. The largest percentage gain came for *TrWP* where LDC correctly identified 17 instances but the supervised learner only produced six true positives, resulting in a gain of 11 (183.3%). The majority classes also achieved slightly higher F_1 scores. The LDC method appears to be an effective way to improve recall on minority relation classes while maintaining good performance on the majority classes.

Conclusion

We showed that clustering-based instance selection from unlabeled text data could improve performance on minority classes for relation type classification between medical concepts. Experimental results show that our clustering-based methods outperformed supervised classification and traditional self-training from unlabeled texts. We believe that this approach offers a more robust solution for classification problems when the data has a highly skewed class distribution, acquiring manual annotations is expensive, but large quantities of unannotated text data are available.

Acknowledgments

The de-identified clinical records used in this research were provided by the i2b2 National Center for Biomedical Computing funded by U54LM008748 and were originally prepared for the NLP Shared Tasks organized by Dr. Özlem Uzuner and colleagues.

References

1. U.S. Department of Health and Human Services.
   Available from: http://www.hhs.gov/news/press/2013pres/05/20130522a.html
2. Pratt AW. Medicine, Computers, and Linguistics. Advanced Biomedical Engineering. 1973;3:97-140.
3. Meystre SM., Savova GK., Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008:128-144.
4. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA Challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc. 2011;18(5):552-6.
5. Rink B, Harabagiu S, Roberts K. Automatic extraction of relations between medical concepts in clinical texts. J Am Med Inform Assoc. 2011;18(5):594-600.
6. Stone PJ, Dunphy DC, Smith MS, Ogilvie DM. The General Inquirer: a computer approach to content analysis. The MIT Press; 1966. 651 p.
7. de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. Machine learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. J Am Med Inform Assoc. 2011;18(5):557-62.
8. Zhu X, Cherry C, Kiritchenko S, Martin J, de Bruijn B, Detecting concept relations in clinical text: Insights from a state-of-the-art model. J. Biomed. Inform. 2013;46(2):275-85.
9. D'Souza J, Ng V. Ensemble-based medical relation classification. In Proceedings of COLING; 2014. p. 1682-93.
10. Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In Proceedings of the 11th Annual Conference on Learning Theory, Madison, WI; 1998. p. 92-100.
11. Collins M, Singer Y. Unsupervised models for named entity classification. In Proceedings of EMNLP/VLC, College Park, MD; 1999. p. 100-10.
12. Pierce D, Cardie C. Limitations of co-training for natural language learning from large datasets. In Proceedings of EMNLP, Pittsburgh, PA; 2001. p. 1-9.
13. Thelen M, Riloff E. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing; 2002. p. 214-21.
14. Rosenberg C, Hebert M, Schneiderman H. Semi-supervised self-training of object detection models. In Seventh IEEE Workshop on Applications of Computer Vision; 2005.
15. Zhang Z. Weakly-supervised relation classification for information extraction. In Proceedings of the thirteenth ACM international conference on Information and knowledge management (CIKM '04). ACM, New York, NY; 2004. p. 581-8.
16. Mintz M, Bills S, Snow R, Jurafsky D. Distant supervision for relation extraction without labeled data. In Proceedings of the 47th ACL, Stroudsburg, PA, USA; 2009. p. 1003-11.

17. Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J. Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (SIGMOD '08). ACM, New York, NY, USA; 2008. p. 1247-50.
18. Sun A, Grishman R, Sekine S. Semi-supervised relation extraction with large-scale word clustering. In Proceedings of the 49th ACL/HLT, Stroudsburg, PA, USA; 2011. p. 521-9.
19. Wang C, Fan J. Medical relation extraction with manifold models. In Proceedings of the 52nd ACL, Baltimore, Maryland; 2014. p. 828-38.
20. 2010 i2b2 / VA Challenge Evaluation Relation Annotation Guidelines.
Available at https://www.i2b2.org/NLP/Relations/assets/Relation%20Annotation%20Guideline.pdf
21. Saeed M, Villarroel M, Reisner AT et al. Multi-parameter intelligent monitoring in intensive care II (MIMIC-II): A public-access ICU database. Critical Care Medicine. 2011;39(5):952-60.
22. Kim Y, Riloff E, Stacked generalization for medical concept extraction from clinical notes, In Proceedings of BioNLP 15; 2015. p. 61-70.
23. Kim Y, Riloff E, Hurdle JF. A Study of concept extraction across different types of clinical notes. In Proceedings of AMIA Annual Symposium; 2015. p. 737-46.
24. Kim Y, Riloff E, Meystre SM. Improving classification of medical assertions in clinical notes. In Proceedings of the 49th ACL/HLT; 2011. p. 311-6.
25. George Karypis. CLUTO a clustering toolkit. Technical Report 02-017, Dept. of Computer Science, University of Minnesota, 2002. Available at http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download
26. Liu Y, Li X, Xiong H, Gao X, Wu J. Understanding of internal clustering validation measures. In Proceedings of the 2010 IEEE International Conference on Data Mining (ICDM '10). IEEE Computer Society, Washington, DC; 2010. p. 911-6.
27. Satopää V, Albrecht J, Irwin D, Raghavan B. Finding a "Kneedle" in a haystack: detecting knee points in system behavior. In Proceedings of the 31st International Conference on Distributed Computing Systems Workshops (ICDCSW '11); 2011. p. 166-71.
28. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford CoreNLP natural language processing toolkit, In Proceedings of the 52nd ACL: System Demonstrations; 2014. p. 55-60.
29. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: A library for large linear classification, Journal of Machine Learning Research. 2008;9:1871-4.
30. 2010 i2b2 / VA Challenge Evaluation Assertion Annotation Guidelines.
Available at https://www.i2b2.org/NLP/Relations/assets/Assertion%20Annotation%20Guideline.pdf
31. Chapman WW, Chu D, Dowling JN. ConText: An algorithm for identifying contextual features from clinical text. In Proceedings of BioNLP 2007: Biological, translational, and clinical language processing, Prague, CZ; 2007.
32. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In Proceedings of International Conference on Learning Representations (ICLR). Scottsdale, Arizona; 2013.
33. Yeh A. More accurate tests for the statistical significance of result differences. In Proceedings of the 18th conference on Computational linguistics (COLING '00); 2000. p. 947-53.