A Study of Concept Extraction Across Different Types of Clinical Notes

Youngjun Kim MS[1], Ellen Riloff, PhD[1], John F. Hurdle, MD, PhD[2]
[1]School of Computing; [2]Department of Biomedical Informatics,
University of Utah, Salt Lake City, UT

Abstract

Our research investigates methods for creating effective concept extractors for specialty clinical notes. First, we present three new "specialty area" datasets consisting of Cardiology, Neurology, and Orthopedics clinical notes manually annotated with medical concepts. We analyze the medical concepts in each dataset and compare with the widely used i2b2 2010 corpus. Second, we create several types of concept extraction models and examine the effects of training supervised learners with specialty area data versus i2b2 data. We find substantial differences in performance across the datasets, and obtain the best results for all three specialty areas by training with both i2b2 and specialty data. Third, we explore strategies to improve concept extraction on specialty notes with ensemble methods. We compare two types of ensemble methods (Voting/Stacking) and a domain adaptation model, and show that a Stacked ensemble of classifiers trained with i2b2 and specialty data yields the best performance.

Introduction

Medical notes provide detail on patient encounters. Written by clinicians primarily for clinicians, they document (e.g., progress notes) or summarize (e.g., discharge summaries) patient care. They come in a variety of note types and are entered by health care professionals from varying backgrounds.

Information extraction from medical texts is a challenging problem of growing interest to both the natural language processing and medical informatics communities. Medical concept extraction (MCE) is one such task, which seeks to identify specific types of information such as medical problems, treatments, and tests. Previous research on this task has primarily focused on discharge summaries and progress notes [1-4], which we will refer to as **broad** medical texts because they describe a patient's overall care and their content can cover a diverse set of topics cutting across many areas of medicine. Most publicly available corpora of clinical medical notes consist of broad medical texts (e.g., *i2b2 Challenge Shared Tasks* [5-10] and *ShARe/CLEF eHealth Shared Tasks* [11-12]).

There has been relatively little research on medical concept extraction for more specialized clinical texts. Studies focused on Radiology and Pathology reports are an important exception, but we would argue that they also cover a broad set of clinical conditions. Broad medical texts have the advantage of being relatively well formatted, and they typically follow general documentation standards. In contrast, specialty notes conform to varying documentation standards, with little overlap between specialties. Patterson and Hurdle [13] and Friedman et al. [14] demonstrated that clinicians in different clinical domains use specific sublanguages. Still, given the general nature of broad medical notes, we speculate that their content could enrich MCE systems targeted at specialized note types and our work offers a practical way forward for clinical information extraction despite the common use of sublanguages.

Our research investigates methods for creating medical concept extraction systems that will perform well on specialty area notes. For this research, we created three new text corpora consisting of medical notes from three specialty areas: Cardiology, Neurology, and Orthopedics. We present an analysis of how they differ in content (semantic concepts and formatting) from each other and from i2b2 medical notes. We then examine a variety of information extraction (IE) models, and evaluate their performance on all of these data sets. The contributions of our work are twofold. First, we investigate how well MCE models perform on specialty notes when trained on a broad medical corpus and then when trained on the same type of specialty data. When training with a comparable amount of annotated data, we find that training with specialty texts outperforms training with broad medical texts. However, we achieve better performance for all three specialty areas by using a combination of both broad medical i2b2 data and specialty area data for training.

Second, we explore Voting and Stacked Learning ensembles to combine multiple MCE models. The ensemble architecture can be beneficial in two ways: (1) it can exploit multiple models that use different extraction techniques, and (2) it can exploit multiple models trained with different types of data (in our case, some trained on broad medical notes and some trained on specialty notes). To our knowledge, this is the first work that combines broad medical components and specialty area components in a single ensemble for MCE. Our results show that a stacked ensemble consisting of both types of components achieves the best balance of precision and recall.

Background

Medical concept extraction has been the focus of several shared tasks, such as the *i2b2 Challenge Shared Tasks* and the *ShARe/CLEF eHealth Shared Tasks* [8, 11-12]. Our work uses the annotated data set provided for the 2010 i2b2 Challenge [8]. In this challenge, machine learning approaches [15-16] showed superior results over hand-crafted rule-based systems. de Bruijn et al. [15] incorporated syntactic, orthographic, lexical, and semantic information (from various medical knowledge databases) and their system performed best in the i2b2 concept extraction challenge task with 83.64% recall, 86.88% precision, and 85.23% F_1 score. Jiang et al. [16] implemented an ensemble method to combine concept extraction models trained with local features and outputs from different knowledge databases. In 2013, Tang et al. [17] extended their work using clustering and distributional word representation features, achieving 84.31% recall, 87.38% precision, and an F_1 score of 85.82% on the i2b2 test set.

Our work is closely related to the classic task of Named Entity Recognition. In both newswire and biomedical texts, many types of supervised learning and sequential tagging methods have been used to extract specific types of entities [18-22]. In Clinical NLP (Natural Language Processing), several systems have been developed to process medical notes or biomedical texts. MedLEE [23] has been applied to chest radiology reports, discharge summaries, and operative reports to extract and encode medical information. MPlus [24] was used to extract medical findings, diseases, and appliances from chest radiograph reports. LifeCode [25] was developed to extract demographic and clinical information on emergency medicine clinical specialty and radiology reports.

Ensemble methods that combine multiple classifiers have been widely used for many NLP tasks. Voting strategies [26-28] and statistical approaches including stacked generalization [29-30] have generally shown better performance than individual classifiers. Our work is also related to supervised domain adaptation, which can be applied when some labeled data for the target domain is available. Many algorithms for efficient domain adaptation have been proposed, and domain adaptation-based models have been shown to improve performance for some tasks when limited annotated data is available for the target domain [31- 35].

Methods

Data Sets and Annotated Concepts

Our research starts with the medical concept extraction (MCE) task defined for the 2010 i2b2 Challenge [8]. This task involves extracting three types of medical concepts: *Problems* (e.g., diseases and symptoms), *Treatments* (e.g., medications and procedures), and *Tests*. The 2010 i2b2 corpus consists of 349 training documents and 477 test documents manually annotated by medical professionals. This test set contains 45,009 annotated medical concepts.

For our work, we created new text collections representing three specialized areas of medicine: Cardiology, Neurology, and Orthopedics. We annotated 200 clinical notes from the BLULab corpus[1] for each specialty area. Each specialty data set consists of different subtypes of notes. Table 1 shows the five most prevalent subtypes in each specialty data set.

Table 1. Five most prevalent note subtypes in each specialty area data set

Data	Note subtypes
Cardiology	Cardiology (surgery) discharge summary, Cardiology (surgery) consultation report, Cardiology operative report, Cardiology history and physical examination, Angio report
Neurology	Neurosurgery discharge summary, Neurosurgery transfer summary, Neurology consultation report, Neurology history and physical examination, Neurosurgery death summary
Orthopedics	Orthopedic (surgery) operative report, Trauma discharge summary, Orthopedic (surgery) discharge summary, Orthopedic surgery transfer summary, Orthopedics consultation report

[1] The BluLab corpus is a collection of de-identified clinical notes drawn from multiple clinical settings at the University of Pittsburgh. The dataset was available for research to investigators with local Institutional Review Board approval, but unfortunately the University of Pittsburgh has withdrawn the corpus for new studies. However interested researchers can collaborate with previously approved sites.

Two people with medical expertise manually annotated the specialty notes using the 2010 i2b2 Challenge guidelines. One annotator had previously annotated data for the official 2010 i2b2 Challenge data and the other annotator had equivalent medical knowledge. We measured their inter-annotator agreement on 50 documents annotated by both annotators during the pilot phase using Cohen's kappa [36] and their IAA was $\kappa = .67$. Each of the annotators then labeled 100 new documents for each specialty area, producing a total of 600 annotated specialty area texts. These texts contain 17,783 annotated concepts for Cardiology, 11,019 concepts for Neurology, and 12,769 concepts for Orthopedics.

Table 2 shows the number of annotated concepts of each type in the i2b2 test data and our three specialty data sets, as well as the average number of concepts per document. For example, the Cardiology data contains 7,474 *Problem* concepts and the average number of *Problem* concepts per text is 37, which is similar to the i2b2 data (39). However, the Neurology and Orthopedics data sets contain only 25 *Problem* concepts per document, on average. For *Treatment* concepts, the Neurology notes contain fewer than the i2b2 data but the Orthopedics notes contain more. The prevalence of *Test* concepts varies greatly: the i2b2 and Cardiology texts have many *Test* concepts per document, but they are much less common in the Neurology notes (11 per text) and Orthopedics notes (6 per text).

Table 2. The numbers of concepts in each data set

Categories	i2b2 Test		Cardiology		Neurology		Orthopedics	
	Total	Average	Total	Average	Total	Average	Total	Average
Problem	18,550	39	7,474	37	4,971	25	5,022	25
Treatment	13,560	28	5,706	29	3,815	19	6,494	33
Test	12,899	27	4,603	23	2,233	11	1,253	6
All Concepts	45,009	94	17,783	89	11,019	55	12,769	64
# Sentences	45,052	94	21,255	106	15,339	77	16,855	84

The last row of Table 2 compares the number of sentences in the data sets. The i2b2 test data contains 45,052 sentences (94 per file, on average). The Cardiology notes were generally longer with 106 sentences per text, while the Neurology and Orthopedics notes were generally shorter.

We also examined, qualitatively, the types of sections in each data set to gain more insight about content differences between specialist notes and the more general i2b2 notes. Table 3 shows the five most frequent section titles in each data set. Many section titles, such as 'Hospital course', are common across all of the data sets. However, we found section titles that are much more frequent in some types of specialty area notes. For example, sections related to 'Procedures' and 'Operations' occurred most frequently in Orthopedics notes. 'Consultation' sections were common in the Cardiology notes, but rare in the i2b2 notes.

Table 3. Five most frequent section titles in each data set

Data	Section Titles
i2b2 Test	Hospital course, History of present illness, Physical Examination, Past medical history, Allergies
Cardiology	Physical examination, Allergies, Past medical history, Social history, History of present illness
Neurology	Hospital course, Reason for admission, History of present illness, Discharge medications, Discharge instructions
Orthopedics	Hospital course, Procedures, Discharge instructions, Description of Operation, Complications

Although some of the same section titles occur in both broad medical notes and specialty notes, their contents can differ. For example, in the sections titled 'Procedures,' Orthopedics notes typically contain more detailed information than discharge summaries. Figure 1 illustrates an Orthopedics note that is similar to the ones in our collection.

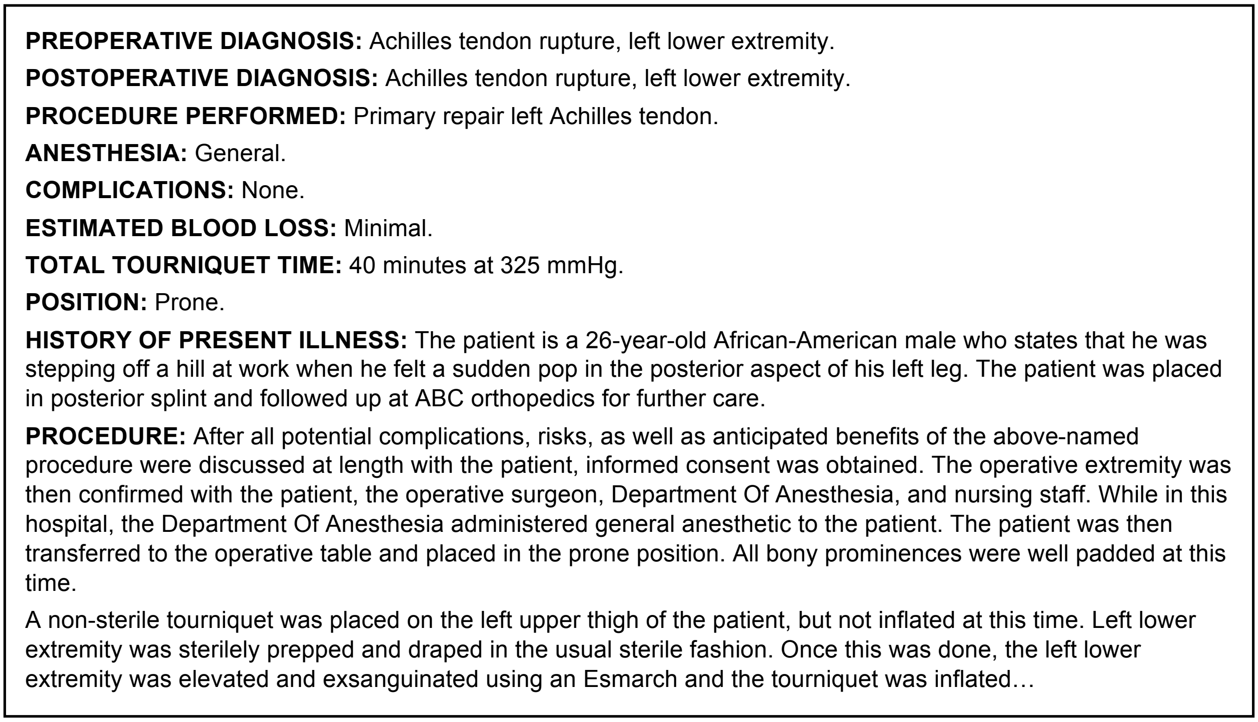

Figure 1. A sample Orthopedics note[2]

PREOPERATIVE DIAGNOSIS: Achilles tendon rupture, left lower extremity.

POSTOPERATIVE DIAGNOSIS: Achilles tendon rupture, left lower extremity.

PROCEDURE PERFORMED: Primary repair left Achilles tendon.

ANESTHESIA: General.

COMPLICATIONS: None.

ESTIMATED BLOOD LOSS: Minimal.

TOTAL TOURNIQUET TIME: 40 minutes at 325 mmHg.

POSITION: Prone.

HISTORY OF PRESENT ILLNESS: The patient is a 26-year-old African-American male who states that he was stepping off a hill at work when he felt a sudden pop in the posterior aspect of his left leg. The patient was placed in posterior splint and followed up at ABC orthopedics for further care.

PROCEDURE: After all potential complications, risks, as well as anticipated benefits of the above-named procedure were discussed at length with the patient, informed consent was obtained. The operative extremity was then confirmed with the patient, the operative surgeon, Department Of Anesthesia, and nursing staff. While in this hospital, the Department Of Anesthesia administered general anesthetic to the patient. The patient was then transferred to the operative table and placed in the prone position. All bony prominences were well padded at this time.

A non-sterile tourniquet was placed on the left upper thigh of the patient, but not inflated at this time. Left lower extremity was sterilely prepped and draped in the usual sterile fashion. Once this was done, the left lower extremity was elevated and exsanguinated using an Esmarch and the tourniquet was inflated…

Information Extraction Models

We developed four types of information extraction models that use a diverse set of extraction techniques.

Rules: We created a simple set of rules by harvesting information from the annotated training data. First, for each word in the training data we computed Prob(concept | word) and Prob(category | word). Next, we selected words that had frequency ≥ 3 and Prob(concept | word) ≥ .80. For each selected word, we chose the category with the highest probability and created a rule (e.g., *diabetes* → *Problem*). Given a new text, we then found all words that matched a rule and labeled them as concepts using the category assigned by the rule. When two or more labeled words were contiguous, we treated them as a single concept. For multi-word concepts, we calculated the average Prob(category | word) across the words in the concept. The category with the highest average probability was assigned to the concept.

MetaMap: We used a well-known knowledge-based system, MetaMap [37], that assigns UMLS Metathesaurus semantic concepts to phrases. We identified UMLS semantic type identifiers, using the UMLS Semantic Lexicon, that covered the types of medical concepts required for our task. We only used the final mappings of MetaMap to avoid generating nested terms because the i2b2 guidelines do not permit nested concepts. Table 4 shows the semantic types that we used for concept extraction[3]. We used MetaMap 2013v2 with the 2013AB NLM relaxed database.

Table 4. MetaMap semantic types used for concept extraction

Category	MetaMap semantic types
Problem	acab, anab, bact, celf, cgab, chvf, dsyn, inpo, mobd, neop, nnon, orgm, patf, sosy
Treatment	antb, carb, horm, medd, nsba, opco, orch, phsu, sbst, strd, topp, vita
Test	biof, bird, cell, chvs, diap, enzy, euka, lbpr, lbtr, mbrt, moft, phsf, tisu

[2] Excerpted from http://www.mtsamples.com/

[3] Refer to http://metamap.nlm.nih.gov/Docs/SemanticTypes_2013AA.txt for the mapping between abbreviations and the full semantic type names.

SVM: We trained a multi-class Support Vector Machine (SVM) classifier with a linear kernel using the LIBLINEAR software package [38]. We applied the Stanford CoreNLP tool [39] to our data sets for tokenization and part-of-speech (POS) tagging. We defined features for each medical concept's lexical string, POS tag, affix(es), orthographic features, and pairwise combinations of these features. We also extracted these features for the three words before and after the concept. To identify multi-word concepts, we reformatted the training examples with IOB tags (B: at the beginning, I: inside, or O: outside of a concept) and then trained the model to produce IOB labels as output. We trained a single SVM model to produce labels for all three concept types (*Problem*, *Treatment*, and *Test*).

CRF: We trained two types of sequential taggers using linear Conditional Random Fields (CRF) models [19]: models with forward transitions (CRF-fwd) and models with backward transitions by reversing the word sequence (CRF-rev) [40-41]. The CRF models used the same feature set as the SVM models. A single CRF model produces labels for *Problem*, *Treatment*, and *Test*.

We performed 10-fold cross validation on the i2b2 training set to optimize the parameters of the SVM and CRF classifiers for F_1 score maximization. For the SVMs, we tuned the cost parameter ($c = 0.1$) of LIBLINEAR. For the CRFs, we used Wapiti [42], a simple and fast discriminative sequence labeling toolkit. We set the size of the interval for the stopping criterion to be $e = .001$. For regularization, L_1 and L_2 penalties were set to 0.005 and 0.4 respectively. These parameter settings were kept the same throughout all of our experiments.

Ensemble Methods

We explored two types of ensemble architectures that have performed well for other NLP tasks: Voting ensembles and Stacked Learning ensembles [29]. Each ensemble consists of a set of MCE components. The general architectures of the Voting and Stacked ensembles are described below. In the Results section, we present experimental results for ensembles consisting of different mixtures of component systems.

Voting Ensemble: This ensemble collects the phrases labeled by a set of MCE components and outputs all phrases that received at least three votes (i.e., were labeled by at least three components). In the case of overlapping phrases, we choose the one with the highest confidence, based on the normalized confidence scores of the MCE models. For each MCE model, each confidence score was divided by the highest score produced by that model for normalization.

Stacked Learning Ensemble: This ensemble consists of a set of MCE components as well as a meta-classifier, which is a SVM classifier trained on the predictions of the individual MCE models. To create training instances for a document, we first aggregated all of the concept predictions into sets of unique predictions. For example, one aggregated prediction set might indicate that an instance of "*acute renal failure*" was labeled as a *Problem* by models M_1, M_3, and M_4. Each concept predicted by an MCE model was then compared with all concepts predicted by the other MCE models. For each pair of concepts, the following eight matching criteria are applied to create binary features:

- If the text spans match
- If the text spans partially match (any word overlap)
- If the text spans match and concept types match
- If the text spans partially match and the concept types match
- If the text spans have the same start position
- If the text spans have the same end position
- If one text span subsumes the other
- If one text spans is subsumed by the other

For an ensemble with k MCE models, $k \times 8$ features are defined so that each matching function is replicated for each MCE model. Each feature indicates whether Concept_1 and Concept_2 satisfy a specific matching function, given that Concept_1 was produced by a specific model. In addition, features are defined that count how many different models produced a predicted concept, and features are defined for predictions produced by just a single model (indicating which model produced the predicted concept). In a previous study [43], this type of Stacked ensemble architecture achieved performance comparable to the state-of-the-art on the i2b2 test data with 83.4% recall, 87.9% precision, and 85.6% F_1 score.

Results

We conducted an extensive set of experiments to evaluate the performance of each individual MCE model and Voting and Stacked Learning ensembles. We also experimented with models trained using the broad medical (i2b2) texts, using our specialty area texts, and using a mixture of both. We evaluated performance using the i2b2 test set as well as our three sets of specialty area notes: Cardiology, Neurology, and Orthopedics. The specialty area models (Sp) were trained and evaluated using 10-fold cross validation on our specialty notes data. A labeled phrase was scored as correct if it was assigned the correct concept type and its text span exactly matched the gold standard text span, disregarding articles and possessive pronouns (e.g., "*his*").

Performance of Individual MCE Models

Table 5 shows the performance of each MCE model based on Recall (Rec), Precision (Pr), and F$_1$ score (F). The **Rules (i2b2)** row shows results for the simple rules harvested from the i2b2 training data. Not surprisingly, these rules performed better on the i2b2 test set than on the specialty notes, but the scores were low across the board. The **Rules (Sp)** row shows results (averaged during cross-validation) for the rules harvested from the training folds and evaluated on the test folds for the specialty area data. These rules also performed poorly. The **MetaMap** row shows similarly low scores for MetaMap on all data sets. One reason for its low performance is that the concept and phrase boundary definitions of MetaMap's semantic categories are not perfectly aligned with i2b2's concept definitions.

The machine learning classifiers performed substantially better. The **SVM (i2b2)** row shows results for the SVM model trained on i2b2 data, which produced an F$_1$ score of 78.7% on the i2b2 test set but substantially lower F$_1$ scores on the specialty datasets. The **SVM (Sp)** row shows results for the SVMs trained on specialty area data. Performance substantially improved on the Orthopedics notes (from 43.5% to 52.6% F$_1$ score), but did not change much for the other specialty areas.

Table 5. Recall (Rec), Precision (Pr), and F$_1$ score (F) for individual MCE models

Model	i2b2			Cardiology			Neurology			Orthopedics		
	Rec	Pr	F	Rec	Pr	F	Rec	Pr	F	Rec	Pr	F
Rules (i2b2)	38.5	48.4	42.9	33.1	37.9	35.3	29.2	35.3	32.0	21.4	26.2	23.5
Rules (Sp)				32.6	38.6	35.3	30.9	33.0	31.9	26.8	27.9	27.3
MetaMap	36.0	47.3	40.9	31.1	40.0	35.0	29.4	34.6	31.8	22.6	26.3	24.3
SVM (i2b2)	80.6	76.9	78.7	64.5	59.4	61.8	59.4	55.7	57.5	45.7	41.6	43.5
SVM (Sp)				65.5	59.4	62.3	60.2	53.8	56.8	56.6	49.1	52.6
CRF-fwd (i2b2)	81.4	86.1	83.7	65.2	67.9	66.5	61.3	65.8	63.5	47.4	56.3	51.5
CRF-rev (i2b2)	**82.3**	**86.4**	**84.3**	65.8	68.0	66.9	61.7	65.7	63.6	48.2	55.8	51.7
CRF-rev (i2b2$_{180}$)	78.7	84.2	81.4	63.3	66.5	64.9	59.6	64.8	62.1	44.8	53.3	48.7
CRF-fwd (Sp)				63.8	69.3	66.4	59.2	64.6	61.8	55.4	62.3	58.6
CRF-rev (Sp)				65.2	69.1	67.1	60.5	64.6	62.5	56.0	60.6	58.2
CRF-rev (i2b2+Sp)				**68.7**	**70.3**	**69.5**	**64.6**	**66.8**	**65.7**	**59.3**	**62.5**	**60.9**

Both the CRF-fwd and CRF-rev models trained on i2b2 data performed better than the SVM models. Performance on the Cardiology and Neurology notes was similar when trained on specialty (Sp) data, but performance on the Orthopedics notes substantially improved. Since the i2b2 training data is much larger than our specialty area training data, we performed another experiment using only 180 randomly selected i2b2 training texts, to match the amount of specialty area training data (under 10-fold cross-validation, each fold trains with 180 documents). The performance of this model, shown in the CRF-rev (i2b2$_{180}$) row, is lower than when using all of the i2b2 training data. We can now see that training on specialty area data consistently performs better than training on i2b2 data when using comparable amounts of training data. The last row of Table 5 shows the results for training the CRF-rev model using all of the i2b2 training data as well as the specialty area training data. Performance improved for all three specialty areas by training with the combined data sets. The broad i2b2 data clearly provides added value. However, the F$_1$ scores for the three specialty areas ranges from 60.9% to 69.5%, which is substantially lower than the 84.3% F$_1$ score achieved for the i2b2 test set.

Performance of Voting and Stacked Ensembles

We also evaluated the performance of the Voting and Stacked ensemble architectures, which were populated with all five types of MCE components: Rules, MetaMap, SVM, CRF-fwd, and CRF-rev models. For both the Voting and Stacked architectures, we created three different types of ensembles: i2b2 ensembles consisting of MCE models trained on the i2b2 data, Sp ensembles consisting of MCE models trained on specialty data, and i2b2+Sp ensembles consisting of MCE models trained with i2b2 data and MCE models trained with specialty data. Consequently, the i2b2+Sp ensembles include nine different classifiers (two models each of Rules, SVM, CRF-fwd, CRF-rev, and one MetaMap model, because it does not use training data).

Table 6 shows the performance of these ensembles, as well as the EasyAdapt domain adaptation method [34], which we implemented as another point of comparison. For EasyAdapt, we used a CRF-rev classifier with the feature set augmented for broad medical (i2b2) notes as the source domain and specialty area notes as the target domain. For the sake of comparison, the first row of Table 6 displays again the results obtained for the best individual MCE model from Table 5, which was the CRF-rev classifier trained with both i2b2 and specialty data. Comparing the first two rows, we see that training a CRF-rev model with combined i2b2 and specialty area data outperforms the domain adaptation model on all three data sets.

For the Voting ensembles, the i2b2+Sp ensemble produced the best F_1 scores, but did not outperform the CRF-rev (i2b2+Sp) model. However, the Voting ensemble trained only on specialty notes (Sp) produced much higher precision than the CRF-rev model. A Voting ensemble appears to be an effective way to improve precision on specialty notes when a limited amount of annotated specialty data is available, although with some cost to recall.

Table 6. Recall (Rec), Precision (Pr), and F_1 score (F) for ensemble methods

Model	**Cardiology**			**Neurology**			**Orthopedics**		
	Rec	**Pr**	**F**	**Rec**	**Pr**	**F**	**Rec**	**Pr**	**F**
CRF-rev (i2b2+Sp)	68.7	70.3	69.5	**64.6**	66.8	65.7	**59.3**	62.5	60.9
EasyAdapt	66.1	69.5	67.8	62.0	65.4	63.7	57.7	62.0	59.8
Voting (i2b2)	61.0	73.0	66.5	56.2	70.4	62.5	40.7	64.2	49.8
Voting (Sp)	58.3	**77.8**	66.7	52.9	**74.3**	61.8	47.3	**73.0**	57.4
Voting (i2b2+Sp)	**69.4**	69.8	69.6	64.5	66.0	65.3	56.6	62.5	59.4
Stacked (i2b2)	65.7	69.0	67.3	61.8	66.9	64.3	47.8	57.6	52.3
Stacked (Sp)	63.4	73.9	68.2	57.4	70.9	63.4	52.3	70.2	60.0
Stacked (i2b2+Sp)	66.0	75.1	**70.2**	61.5	72.4	**66.5**	54.6	70.8	**61.6**

For Stacked Learning, every Stacked ensemble outperformed its corresponding Voting ensemble. The best Stacked ensemble (i2b2+Sp) included MCE models trained on i2b2 data as well as MCE models trained on specialty data, producing slightly higher F_1 scores than the CRF-rev models for all three specialty areas. Using a paired t-test to measure statistical significance, the F_1 score performance of the i2b2+Sp Stacked ensemble is significantly better than EasyAdapt and all of the Voting ensembles at the $p < .05$ significance level, but not significantly better than the CRF-rev (i2b2+Sp) model. However, the results show that the Stacked ensemble produces higher precision than the CRF-rev model (70% → 75% for Cardiology; 67% → 72% for Neurology; 63% → 71% for Orthopedics), with correspondingly smaller decreases in recall (69% → 66% for Cardiology; 65% → 62% for Neurology; 59% → 55% for Orthopedics).

Discussion and Analysis

The main conclusion of our research is that models trained with a combination of broad medical data and specialty data consistently perform better than models trained on either type of data alone when the amount of specialty data is limited. In addition, we find that a Stacked ensemble consisting of a diverse set of MCE models using different types of extractors achieves overall performance comparable to the best individual classifier in our experiments, but offers two advantages. First, the Stacked ensemble yields a recall/precision balance that favors precision, which may benefit applications that place a premium on high precision. Second, the Stacked ensemble can be easily augmented with additional components as new resources become available, because the meta-classifier automatically learns how to use them simply by re-training the meta-classifier component. In contrast, adding new components to Voting ensembles can require a change in voting strategies, and Voting ensembles do not provide a way to learn weights to optimally control the influence of different component models.

However, performance on all three types of specialty areas is much lower than performance on the broad medical (i2b2) texts. Clearly there is ample room for improvement for medical concept extraction from specialty area clinical notes and more work is needed on this topic. To better understand the strengths and weakness of our models, we manually inspected their output. We observed that our ensemble methods are particularly successful at identifying more accurate concept boundaries than the individual MCE models (e.g., identifying "*severe chest pain*" as a *Problem* concept instead of just "*severe*" or "*chest pain*"). We also analyzed the false negative errors by the CRF-rev models trained with i2b2 data and those trained with specialty data. Table 7 shows the results of this manual analysis, which were based on one test fold (20 notes) for each specialty area. The first row of Table 7 corresponds to errors due to unseen vocabulary. These concepts were misclassified when none of the words in a concept occurred in the training data. For example, the Cardiology concepts '*thoracoscopy'* and '*cardioplegia'* never appeared in the i2b2 training data. Unseen concepts accounted for roughly the same percentage of errors when training with i2b2 data or specialty data, but note that the i2b2 training set is roughly twice as large as each specialty area training set.

Table 7. False negative errors by CRF-rev (i2b2) and CRF-rev (Sp) models

Error types	Cardiology		Neurology		Orthopedics	
	i2b2	Sp	i2b2	Sp	i2b2	Sp
All unseen	22 (5%)	27 (6%)	31 (6%)	32 (6%)	53 (8%)	23 (4%)
At least one unseen word	138 (31%)	100 (21%)	186 (37%)	109 (21%)	355 (**51**%)	112 (19%)
At least one word rarely seen	70 (16%)	81 (17%)	70 (14%)	85 (17%)	94 (14%)	113 (19%)
All seen	213 (48%)	279 (56%)	211 (43%)	288 (56%)	184 (**27**%)	337 (58%)

The second row of Table 7 corresponds to false negatives for concepts containing at least one seen word and one unseen word. We see more false negatives in this category for the models trained with i2b2 data than the models trained with specialty data. For example, for the *Treatment* concept '*aortic crossclamping*', '*crossclamping*' never appeared in the i2b2 training data but it did appear in the Cardiology training data. This type of error was most common in the Orthopedics data (51% of the errors), which suggests that the Orthopedics notes contain many vocabulary terms that are not present in the i2b2 data.

The third row of Table 7 corresponds to false negatives for concepts containing all seen words, but at least one rarely seen word (frequency <= 3). For example, in the Cardiology data, the concepts '*psa data'* and '*r-wave'* were not identified by the i2b2 trained model. The model trained with Cardiology data could not extract '*nystatin'* and '*oximeter'*, even though they occurred (infrequently) in the Cardiology training data.

The last row of Table 7 corresponds to false negatives for concepts consisting entirely of words that occurred > 3 times in the training data. Many false negative errors fell into this category. Generally, there were more false negative errors of this type for the models trained with specialty data than those trained with i2b2 data, presumably because the vocabulary is more homogenous in the specialty areas so more words simply fall into the seen category.

Finally, we observed that many errors were due to incorrect phrase boundaries of medical concepts. For example, only the word "*hepatitis*" was labeled in the phrase "*hepatitis c*". We also witnessed some tricky errors due to contextual differences in the words surrounding medical concepts. For example, a *Treatment* concept '*lidocaine'* is often prescribed for usage on skin ("*treated with lidocaine jelly for pain control*"). However, in the Cardiology data, it is usually applied by infiltration ("*Lidocaine 20 cc was infiltrated into the tissues*").

Conclusion

We analyzed the differences in content between broad medical and specialty area notes, confirming prior research showing that specialty notes exhibit sublanguage behavior that requires rethinking the use of NLP tools developed on broad medical notes. We found that even though the CRF-rev (i2b2+Sp) and Stacked ensemble produce similar F_1 scores, they exhibit different behaviors with respect to the underlying recall and precision of their output. Consequently, our results suggest that the CRF-rev (i2b2+Sp) model may be preferable for applications where recall is more important than precision, while the Stacked ensemble may be preferable for applications where precision is more important than recall. Interestingly, Orthopedics specialty notes exhibit the most unique language when compared to other specialty notes or to broad medical texts. When a limited amount of annotated specialty area data is available, our research shows that training concept extractors with both broad medical data and specialty area data produces MCE models that achieve better performance on specialty notes than training with either type of data alone. In addition, our research found that a Stacked ensemble with a mixture of MCE components, including different types of MCE models as well as models trained on different types of data, achieves good performance and offers some advantages over other approaches. However, we also observed that MCE performance on specialty texts is substantially lower than state-of-the-art performance on broad medical texts. A promising direction for future work is to explore semi-supervised methods to exploit larger collections of specialty area notes for training.

Acknowledgments

This research was supported in part by the National Science Foundation under grant IIS-1018314 and the National Library of Medicine under grant R01-LM010981. We thank Jennifer Thorne, RN and Jenifer Williams, RN for their annotation work, and Dr. Stéphane Meystre for his comments and feedback.

References

1. Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. *J Am Med Inform Assoc.* 1994;1(2):142–60.
2. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A Simple Algorithm identifying negated findings and diseases in discharge summaries. *J Biomed Inform.* 2001;34:204–30.
3. Uzuner Ö, Sibanda TC, Luo Y, Szolovits P. A de-identifier for medical discharge summaries. *Artif Intell Med.* 2007;42:13–35.
4. Mykowiecka A, Marciniak M, Kupść A. Rule-based information extraction from patients' clinical data. *J Biomed Inform.* 2009;42:923–36.
5. Uzuner Ö, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge summaries. *J Am Med Inform Assoc.* 2008;15:14–24.
6. Uzuner Ö. Recognizing obesity and comorbidities in sparse data. *J Am Med Inform Assoc.* 2009;16:561–70.
7. Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc.* 2010;17(5):514–8.
8. 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.
9. Uzuner Ö, Bodnari A, Shen S, Forbush T, Pestian J, South BR. Evaluating the state of the art in coreference resolution for electronic medical records. *J Am Med Inform Assoc.* 2012;19:786–91.
10. Sun W, Rumshisky A, Uzuner Ö. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *J Am Med Inform Assoc.* 2013;20:806–13.
11. Suominen H, Salanterä S, Velupillai S, Chapman WW, Savova G, Elhadad N, et al. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. *Proc. ShARe/CLEF Evaluation Lab 2013.* 2013.
12. Kelly L, Goeuriot L, Suominen H, Schreck T, Leroy G, Mowery DL, et al. Overview of the ShARe/CLEF eHealth Evaluation Lab 2014. *Proc. ShARe/CLEF Evaluation Lab 2014.* 2014.
13. Patterson O, Hurdle JF. Document clustering of clinical narratives: a systematic study of clinical sublanguages. *AMIA Annu Symp Proc 2011*; 2011. p. 1099–107.
14. Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *J Biomed Inform.* 2002;35:222–35.
15. de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. Machine learned Solutions for Three Stages of Clinical In- formation Extraction: the State of the Art at i2b2 2010. *J Am Med Inform Assoc*. 2011;18(5):557–62.
16. Jiang M, Chen Y, Liu M, Rosenbloom ST, Mani S, Denny JC, Xu H. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J Am Med Inform Assoc*. 2011;18(5):601–6.
17. Tang B, Cao H, Wu Y, Jiang M, Xu H. Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. *BMC Med Inform Decis Mak*. 2013;13(S1):S1.

18. Collier N, Nobata C, Tsujii J. Extracting the names of genes and gene products with a hidden markov model. *Proc. 18th conference on Computational linguistics*; 2000. p. 201–7.
19. Lafferty JD, McCallum A, Pereira FC. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proc. ICML*; 2001. p. 282–9.
20. Collins M. Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. *Proc. ACL-02 on EMNLP*; 2002. p. 1–8.
21. Zhou G, Su J. Named entity recognition using an hmm-based chunk tagger. *Proc. ACL 2002*; 2002. p. 473–80.
22. McDonald R, Pereira F. Identifying Gene and Protein Mentions in Text Using Conditional Random Fields. *BMC Bioinformatics.* 2005;6:S6.
23. Friedman C, Alderson P, Austin J, Cimino J, Johnson S. A general natural language text processor for clinical radiology. *J Am Med Inform Assoc*. 1994;1(2):161–74.
24. Christensen LM, Haug PJ, Fiszman M. MPLUS: a probabilistic medical language understanding system. *Proc. ACL-02 Work. Nat. Lang. Process. Biomed. Domain*; 2002. p. 29–36.
25. Heinze DT, Morsch ML, Sheffer RE, Jimmink MA, Jennings MA, Morris WC, et al. LifeCode: A Deployed Application for Automated Medical Coding. *AI Magazine.* 2001;22(2):76-88.
26. Zhou G, Shen D, Zhang J, Su J, Tan S. Recognition of Protein/Gene Names from Text Using an Ensemble of Classifiers. *BMC Bioinformatics*. 2005;6(S1):S7.
27. Doan S, Collier N, Xu H, Duy PH, Phuong TM. Recognition of medication information from discharge summaries using ensembles of classifiers. *BMC Med Inform Decis Mak.* 2012;12:36.
28. Kang N, Afzal Z, Singh B, Mulligen EM , Kors JA. Using an ensemble system to improve concept extraction from clinical records. *J Biomed Inform*. 2012;45(3):423–8.
29. Wolpert DH. Stacked generalization. *Neural Networks*. 1992;5:241–59.
30. Džeroski S, Ženko B. 2004. Is combining classifiers with stacking better than selecting the best one? *Machine Learning*. 2004;54(3):255–73.
31. Florian R, Hassan H, Ittycheriah A, Jing H, Kambhatla N, Luo X, Nicolov N, Roukos S. Statistical Model for Multilingual Entity Detection and Tracking. *Proc. NAACL/HLT 2004*; 2004. p. 1–8.
32. Chelba C, Acero A. Adaptation of Maximum Entropy Capitalizer: Little Data Can Help a Lot. *Proc. EMNLP 2004*; 2004.
33. Foster G, Kuhn R. Mixture-Model Adaptation for SMT. *Proc. 2nd Workshop on Statistical Machine Translation*; 2007. p. 128–35.
34. Daumé H. Frustratingly easy domain adaptation. *Proc. ACL 2007*; 2007. p. 256–63.
35. Jiang J, Zhai C. Instance Weighting for Domain Adaptation in NLP. *Proc. ACL 2007*; 2007. p. 264–71.
36. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 1960;20(1):37–46.
37. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*. 2010;17(3):229–36.
38. 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.
39. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit, *Proc. 52nd ACL: System Demonstrations*; 2014. p. 55–60.
40. Kudo T, Matsumoto Y. Chunking with support vector machines. *Proc. NAACL-2001*; 2001. p. 1–8.
41. Finkel J, Dingare S, Manning CD, Nissim M, Alex B, Grover C. Exploring the Boundaries: Gene and Protein Identification in Biomedical Text. *BMC Bioinformatics.* 2005;6(S1):S5.
42. Lavergne T, Cappé O, Yvon F. Practical very large scale crfs. *Proc. 48th ACL-10*; 2010. p. 504–13.
43. Kim Y, Riloff E. A Stacked Ensemble for Medical Concept Extraction from Clinical Notes. *AMIA Jt Summits Transl Sci Proc. 2015*; 2015.