Automating the Transformation of Free-Text Clinical Problems into SNOMED CT Expressions

Kevin J. Peterson, MS[1,3], Hongfang Liu, PhD[2]
[1] Division of Information Management and Analytics, Mayo Clinic, Rochester, MN
[2] Department of Health Sciences Research, Mayo Clinic, Rochester, MN
[3] Bioinformatics and Computational Biology Program, University of Minnesota, Minneapolis, MN

Abstract

An important function of the patient record is to effectively and concisely communicate patient problems. In many cases, these problems are represented as short textual summarizations and appear in various sections of the record including problem lists, diagnoses, and chief complaints. While free-text problem descriptions effectively capture the clinicians' intent, these unstructured representations are problematic for downstream analytics. We present an automated approach to converting free-text problem descriptions into structured Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) expressions. Our methods focus on incorporating new advances in deep learning to build formal semantic representations of summary level clinical problems from text. We evaluate our methods against current approaches as well as against a large clinical corpus. We find that our methods outperform current techniques on the important relation identification sub-task of this conversion, and highlight the challenges of applying these methods to real-world clinical text.

Introduction

As the healthcare industry increasingly embraces the promise of new data-driven approaches, the challenges of managing and organizing complex patient data become more pronounced.[1] Even before the deployment of the first electronic health record (EHR), healthcare organizations struggled to establish a structured, organized, and standard representation of patient data.[2] A major advance in this area came in the 1960s when Dr. Weed proposed orienting the data in the patient record around a list of current conditions, or the "problem list."[3] This emphasis on centralizing and enumerating relevant clinical problems enabled patient information to be consumed in a more systematic way, and helped to standardize physicians' interaction with the patient record.[4] A major advantage of this problem-oriented approach is that concise descriptions of clinical problems can summarize and emphasize sections of the larger clinical note narrative. These short phrases describing diagnoses and other patient issues are not limited to the problem list, however. "Summary level" descriptions of clinical problems are also found in diagnosis statements, chief complaints, and reasons for visit,[5,6] and all provide a concise way of expressing pertinent patient conditions.

There are a variety of ways in which these summary level problem descriptions are captured. Free-text is the most expressive form of these problem summaries, capable of capturing the clinical state directly as intended by the clinician.[7] While clinician-friendly, this unstructured representation presents significant problems for downstream analytics.[8] In contrast, a problem may be represented as codes chosen from a controlled terminology. This applies more structure, but limits expressiveness.[9] Even if codified capture is the goal, many systems still allow for free-text entry as a backup if the correct code cannot be readily found.[10,11] These competing representational priorities introduce a fundamental optimization problem in representing these entries – free-text maximizes usefulness for clinicians,[12,13] while structured and codified forms are more amenable to data analytics,[14] standardization activities,[15] and EHR secondary use.[5]

In this study we introduce a method to minimize this conflict between structured and unstructured forms by proposing a framework for converting free-text clinical problem descriptions to codified, structured formats using Natural Language Processing (NLP) techniques. The advantage of structured representations to downstream analytics primarily motivates this effort.[14] By leveraging deep learning methods, we aim to automatically translate text-based problems into Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) Expressions,[16] a structured representation capable of capturing the semantics of summary level problem descriptions in a computable way.

Background & Related Work

The selection and use of a controlled vocabulary to codify free-text clinical problems has been an active area of research.[17,6] In particular, SNOMED CT[18] has been shown both in principle and in practice to be an effective standard for capturing the semantics of these clinical conditions.[19–21] Generally, clinical problems can be represented using SNOMED CT concepts in one of two ways:[22]

- **Pre-Coordinated Concept:** A concept represented as an atomic unit with a single identifier.
  Example: `370221004`|*Severe asthma (disorder)*|
- **Post-Coordinated Concept:** A concept represented as the composition of multiple pre-coordinated concepts that in aggregate define the intended semantics.
  Example: `195967001`|*Asthma (disorder)*| + `24484000`|*Severe (severity modifier)*|

Although pre-coordination has the advantage of simplicity,[23] even summary level problem descriptions are often too expressive to be captured by a single, pre-coordinated SNOMED CT concept. Elkin et al. found that SNOMED CT could only represent 51.4% of problem list entries without composition compared to 92.3% with composition.[24] Liu also found that composition was necessary, observing that 53% of summary level data required two or more SNOMED CT concepts.[5]

Post-coordinated concepts can be readily represented in SNOMED CT via the SNOMED CT Compositional Grammar,[16] a formal specification for representing SNOMED CT post-coordinated expressions. For example,[†] "Severe asthma" can be represented via the following expression consisting of a main focal concept optionally qualified by attribute/value pairs:

```
195967001|Asthma (disorder)|:
        246112005|Severity (attribute)| = 24484000|Severe (severity modifier)|
```

Previous work on converting text to SNOMED CT expressions has focused on the identification and classification of the attribute relationships – for example, what (if any) SNOMED CT attribute best describes the relationship between "Severe" and "asthma." One general approach to this task is to iteratively learn the lexical patterns around how entities relate for a given relationship type.[25] Miñarro-Giménez et al. utilized this technique to fit extracted problem list concepts into learned SNOMED CT relationship patterns.[26] This work leveraged the fact that lexical patterns in pre-coordinated SNOMED CT terms are known and relatively predictable.[27]

Kate proposed a different approach to this task – not as a relation extraction task between two concepts within the context of a sentence, but as an attempt to identify if a relationship holds between the entire problem phrase and the concept, or "relation identification."[28] To illustrate the difference, take the "Severe asthma" example above. The Miñarro-Giménez et al. approach would attempt to find the relationship between "Severe" and "asthma," whereas Kate would take the entire phrase "Severe asthma" and attempt to determine which SNOMED CT concepts relate and how. Kate used a Support Vector Machine (SVM)[28] model trained separately for each relationship type to determine if a given relationship held between a concept and the full text entry. Our contributions in this study are focused on extending Kate's work in the following ways: First, we present an end-to-end process for converting free-text summary level problem descriptions to SNOMED CT expressions, enumerating the sub-tasks and incorporating additional NLP techniques such as dependency parsing. Next, we leverage deep learning techniques to increase relation identification performance as compared to the SVM model. Finally, we begin an initial evaluation of model performance in real-world scenarios using summary level problem descriptions extracted from clinical notes.

Methods

At a high level, our methods are broken into three sequential processing steps as shown in Figure 1. First, given a summary level problem description (in text form), the relevant biomedical concepts are recognized and extracted.

[†] See: https://confluence.ihtsdotools.org/display/DOCSTART/7.+SNOMED+CT+Expressions for more examples.

Next, one of the extracted concepts is chosen as the main semantic focal point, or the *focus concept* of the expression. Following this, the remaining concepts are attached to the focus concept by inferring their role in the expression as a whole. This relation identification task is a critical step in the formation of the SNOMED CT expression and constitutes the bulk of our contributions to this research area. These three steps are explained in detail below.

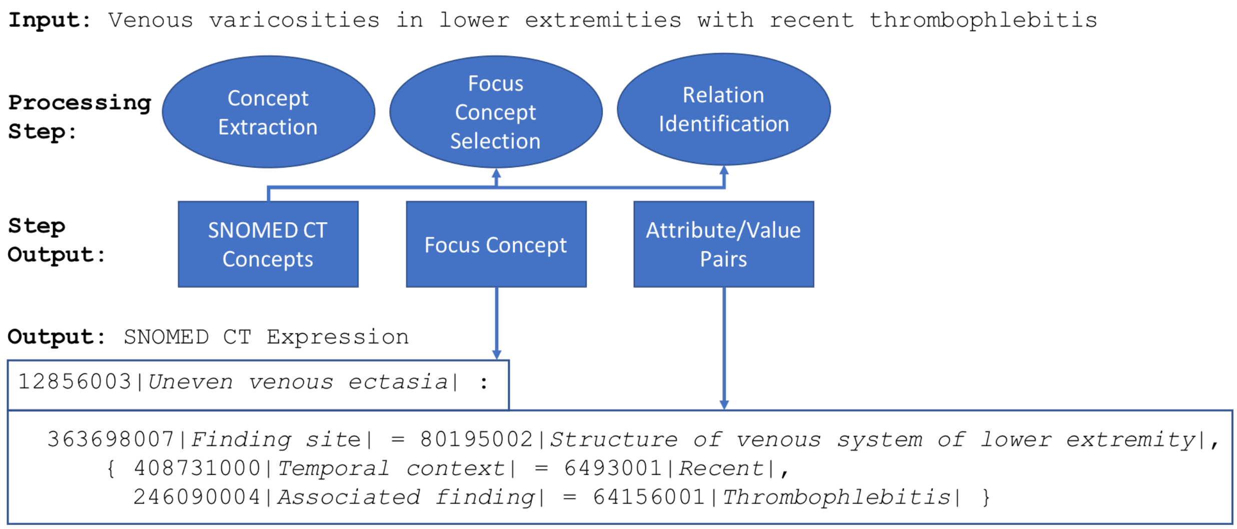

Input: Venous varicosities in lower extremities with recent thrombophlebitis
Processing Step:
Concept Extraction
Focus Concept Selection
Relation Identification
Step Output:
SNOMED CT Concepts
Focus Concept
Attribute/Value Pairs
Output: SNOMED CT Expression
12856003|Uneven venous ectasia| :
363698007|Finding site| = 80195002|Structure of venous system of lower extremity|,
{ 408731000|Temporal context| = 6493001|Recent|,
246090004|Associated finding| = 64156001|Thrombophlebitis| }

Figure 1: Overall steps to convert summary level problem text to a SNOMED CT expression. The processing steps are executed sequentially from left to right with arrows indicating where output from one step feeds into a subsequent step. The output of each processing step is shown linked to the portion of the SNOMED CT expression to which it contributes.

Concept Extraction

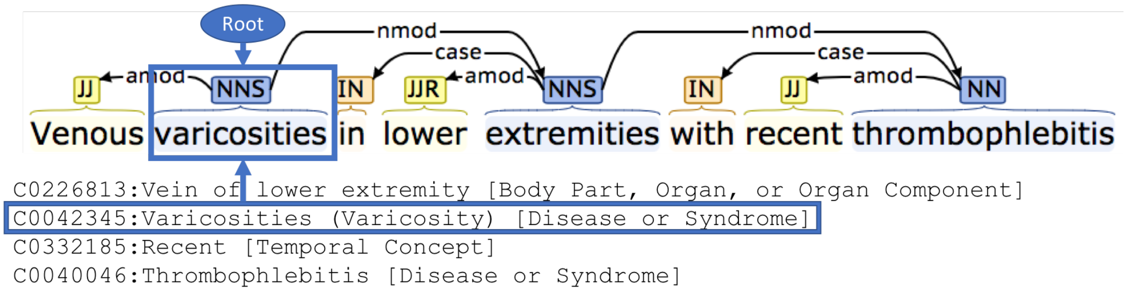
The first step in converting a free-text summary level problem description to a SNOMED CT expression is to extract a list of all relevant concepts from the text. To accomplish this, we leveraged MetaMap, a named-entity recognition tool developed by the National Library of Medicine (NLM) to extract Unified Medical Language System (UMLS) concepts from text.[29] An example output of the MetaMap application given the input problem "Venous varicosities in lower extremities with recent thrombophlebitis" is shown below:

```
C0226813:Vein of lower extremity [Body Part, Organ, or Organ Component]
C0042345:Varicosities (Varicosity) [Disease or Syndrome]
C0332185:Recent [Temporal Concept]
C0040046:Thrombophlebitis [Disease or Syndrome]
```

In this example, the extracted concepts are shown with their UMLS Concept Unique Identifiers (CUIs) and textual descriptions. As our goal is to construct SNOMED CT expressions, we must additionally map the UMLS concepts to SNOMED CT. By configuring MetaMap to match only on SNOMED CT terms, we ensured that each returned UMLS concept encompassed at least one SNOMED CT concept. If the UMLS concept included a single SNOMED CT concept, a direct mapping was made. There are, however, cases where multiple SNOMED CT concepts are incorporated into a UMLS concept.[30] In these cases, all matching SNOMED CT concepts were considered.

Focus Concept Selection

Choosing the focus concept given the list of extracted SNOMED CT concepts is the next step. Here we utilized dependency parsing to align the root node of the problem dependency tree with an extracted MetaMap concept, a technique inspired by Spasić's use of dependency trees to determine the semantic similarity of clinical terms.[31] First, dependency parsing was conducted on the input clinical problem description. Next, the word with the ROOT dependency was compared to all extracted MetaMap concepts. Finally, if one of the extracted MetaMap concepts was triggered by the root word, that concept was then chosen as the focus concept.[‡] Figure 2 shows the dependency parse with the root word *varicosities*, which is then matched to the relevant concept. We used the spaCy open-source NLP toolkit for dependency parsing along with specifically trained biomedical models from the scispaCy project.[32]


Root
nmod
case
amod
amod
nmod
case
amod
JJ
NNS
IN
JJR
NNS
IN
JJ
NN
Venous varicosities in lower extremities with recent thrombophlebitis

```
C0226813:Vein of lower extremity [Body Part, Organ, or Organ Component]
C0042345:Varicosities (Varicosity) [Disease or Syndrome]
C0332185:Recent [Temporal Concept]
C0040046:Thrombophlebitis [Disease or Syndrome]
```

Figure 2: Extracting the focus concept from summary level problem descriptions using dependency parsing. The focus concept is selected via alignment to the ROOT of the dependency parse.

Relation Identification

Identifying the relationships between the problem text and the extracted concepts is the next step. We build primarily on the relation identification task definition as described by Kate[28] and formalize it for our purposes as such: Given an input summary level problem description and a concept extracted via MetaMap, compute the appropriate SNOMED CT attribute (or relationship type) to connect them. Figure 3 outlines the high-level steps of the relation identification algorithm. First, the problem text and the text of each concept extracted via MetaMap are input into a classifier, the details of which are described further in the following sections. Next, the classifier outputs a probability for each SNOMED CT attribute type indicating the likelihood that a particular relationship holds between the problem text and the extracted concept. Finally, candidate relationships are pruned based on the stated domain and ranges of the SNOMED CT Machine Readable Concept Model (MRCM) Attribute Range Reference Set.[33] For example, if the extracted MetaMap concept of interest is 80195002|*Structure of venous system of lower extremity (body structure)*|, any SNOMED CT attribute with a range that is incompatible would be removed (such as 424226004|*Using device (attribute)*|, whose range is limited to children of 49062001|*Device (physical object)*|), as shown in Figure 3. After the pruning, the SNOMED CT attribute with the highest remaining probability is chosen.

A Deep Learning Approach to Relation Identification. Deep learning architectures have shown promise in a variety of NLP tasks,[34] and in this study we compare two popular models for the relation identification classifier. We first consider a Bidirectional Long Short-Term Memory (BiLSTM)[35,36] deep learning architecture. At its base level, a BiLSTM is a specialized type of Recurrent Neural Network (RNN),[37] an artificial neural network architecture that processes information sequentially, factoring in previous input at each current step. This makes RNNs specifically applicable to NLP tasks as text is processed much like a human would – reading words sequentially and inferring the semantics of the current word based on the previous ones.[38] The Long Short-Term Memory (LSTM) facet of the architecture allows for finer control over what information is retained and forgotten by employing more sophisticated feedback loops layered on top of the RNN framework.[35] The bidirectional extension to the LSTM completes our architecture, allowing context to be built not only forward but in the reverse direction as well.[39] In general, the LSTM family of models has shown promising results for NLP relationship extraction tasks.[40]

Convolutional Neural Networks (CNN)[41] are another deep learning architecture. Like LSTMs, CNN models also can recognize spatially related features of the data. For these models, input is filtered via sliding windows which are then pooled to create a subsampled representation of the input sequence. Although used heavily for image processing,

[‡] See https://metamap.nlm.nih.gov/Docs/MMI_Output_2016.pdf for details on how trigger words for MetaMap concepts were obtained.

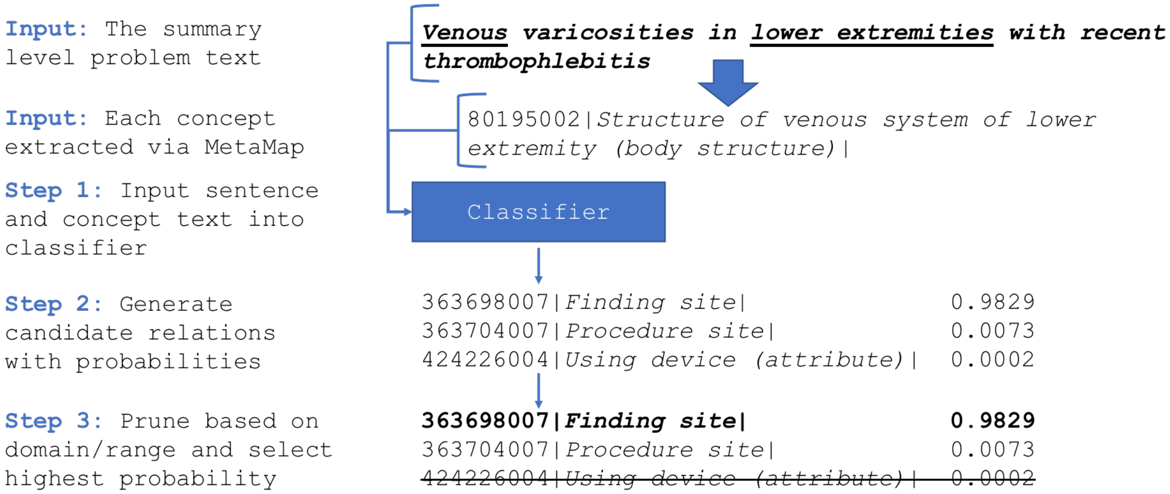

Input: The summary level problem text
Venous varicosities in lower extremities with recent thrombophlebitis
Input: Each concept extracted via MetaMap
80195002|Structure of venous system of lower extremity (body structure)|
Step 1: Input sentence and concept text into classifier
Classifier
Step 2: Generate candidate relations with probabilities
363698007|Finding site| 0.9829
363704007|Procedure site| 0.0073
424226004|Using device (attribute)| 0.0002
Step 3: Prune based on domain/range and select highest probability
363698007|Finding site| 0.9829
363704007|Procedure site| 0.0073
~~424226004|Using device (attribute)| 0.0002~~

Figure 3: Steps to identify relationships between concepts extracted from a free-text summary level problem description.

CNNs have shown promise in a variety of NLP related tasks including relationship classification.[42] In this study, we compare both models for our relationship identification task.

Both of our deep learning architectures use embedding models based on Bidirectional Encoder Representations from Transformers (BERT), a context-aware language model with state-of-the-art performance on a variety of NLP tasks.[43] For our experiments we leveraged Clinical BERT,[44] a pre-trained BERT model fine-tuned on a clinical text corpus.

Architecture. The high-level architecture and data flow for both the BiLSTM and CNN classifiers were similarly structured. First, two inputs are passed to the `Input` layer – the full problem text and the text of the extracted concept. Next, each input is passed to an `Embedding` layer to create a vector representation of the text input using the BERT model. The vectorized input is then processed by either a BiLSTM with 100 hidden units or a CNN with two convolutional layers. Both models were configured for 20% dropout to avoid overfitting. Finally, a fully-connected `Dense` layer with a softmax activation function is used to output the probabilities for each SNOMED CT attribute type.

Training. In concordance with Kate's approach,[28] stated concept relationships from SNOMED CT US Edition, September 2018 Release[18] were used to train the classifier. Training set construction began with all SNOMED CT stated concept-to-concept relationships excluding 116680003|*Is a*| relationships. The reasoning for excluding "Is a" relationships is that once the concepts of the expression are known (see the concept extraction step), any "Is a" relationship for these concepts can be directly inferred from the SNOMED CT hierarchy. Next, because our classifier inputs are text, we use the SNOMED CT concept labels for training. As SNOMED CT concepts may contain multiple labels, given each relationship we created training records for all possible pairs of source and target labels. Finally, we excluded relationship types with less than 125 instances, leaving a total of 1,526,043 training records and 78 relationship types available for training. All experiments below used this data set for training, and Experiments 1 & 2 used a held-out portion of this set for testing. We refer to this data set as the **SNOMED CT Relationship** data set.

Evaluation

System performance was measured for both the Focus Concept Selection and Relation Identification steps of the architecture. The Concept Extraction phase was not directly evaluated, as for this task MetaMap was used without modification (see Reátegui et al.[45] for a recent analysis of MetaMap performance on clinical text). Four experiments were conducted to evaluate model performance.

Experiment 1: First, both the BiLSTM and CNN models were compared against a baseline Naïve Bayes[46] model. All models were trained and tested on the same SNOMED CT Relationship data set. Data was prepared for this experiment by partitioning 25% of the SNOMED CT Relationship set for testing and 75% for training. The Naïve Bayes, BiLSTM, and CNN classifiers were all then trained and tested on the same data, with the exception of a further 20% of the training data being withheld from the BiLSTM and CNN models for validation. For testing, we recorded the F_1 scores for each individual attribute as well as overall averages for all classifiers.

Experiment 2: Next, we compared our results to previously reported results of Kate's SVM model.[28] We followed Kate's evaluation procedures in order to replicate his experiment using the best performing model from Experiment 1: For each attribute, 5000 relationships of the desired type were randomly selected from the SNOMED CT Relationship set along with an equal number of negative examples. Given this test set, the ability of the classifier to correctly determine whether or not the chosen relationship was present was recorded.

Experiment 3: To evaluate relation identification model performance in real-world scenarios and test generalizability to different data sets, we utilized a large data set of summary level clinical problems extracted from a Mayo Clinic corpus of over 14 million clinical documents. This corpus has been extensively analyzed by Liu et al. and is a rich source of diverse summary level problem descriptions.[5] Three trained annotators examined a random subset of 401 summary level clinical problem descriptions from this corpus. First, the annotators were asked to select the word or words that represented the focus concept of the problem. Next, the annotators we tasked with connecting the focus to relevant modifiers using one of twenty-one relationship types. These relationship types were chosen via analysis of common attributes across prominent clinical problem models including the Fast Healthcare Interoperability Resources (FHIR) - `Condition` resource,[47] Clinical Element Model (CEM) - `ClinicalAssert` model,[48] and openEHR - `Problem/Diagnosis` archetype.[49] For more information, see Goossen[50] for details regarding these models. Only relationships supported by $>$20 annotations in the test corpus are evaluated in this study. This test corpus was then used to evaluate both the Focus Concept Selection and Relation Identification parts of the pipeline. The three pair-wise Cohen's kappa inter-annotator agreement values for each annotator pair were 0.78, 0.85, 0.84 for the focus concept annotations and 0.76, 0.76, 0.82 for the relationship annotations.

Experiment 4: Finally, we evaluated how effective the dependency parse-based method is at locating the focus concept of the clinical problem. Using the test data set described in Experiment 3, we evaluated the ability to correctly identify the focus using three spaCy dependency parse models: (1) general-purpose English, (2) scispaCy biomedical, and (3) a custom model based on scispaCy biomedical fine-tuned with 150 manually-annotated clinical problem text examples.

RESULTS

Experiment 1: Overall results of the BiLSTM and CNN models compared to the Naïve Bayes baseline for the SNOMED CT relation identification task are shown in Table 1. While both deep learning models significantly outperformed the baseline, the BiLSTM model slightly outperformed the CNN model. Because of this result, downstream experiments focus on the BiLSTM model.

Table 1: Comparing overall SNOMED CT relation identification model performance.

	Accuracy	**F_1 score (macro avg)**	**F_1 score (weighted avg)**
Naïve Bayes (baseline)	0.720	0.460	0.665
CNN + Clinical BERT	0.886	0.822	0.880
BiLSTM + Clinical BERT	**0.888**	**0.851**	**0.888**

Experiment 2: Figure 4 shows the results of the BiLSTM classifier compared to the SVM classifier results as reported by Kate.[28] Note we did not evaluate `116680003`|*Is a*| relationships, so we do not fully correspond to Kate's results. Also, Kate reported two results for `363698007`|*Finding site*| based on two different domains/ranges. The F_1 score reported in Figure 4 represents the highest of the two scores.

Experiment 3: The results of the evaluation of the BiLSTM classifier against the annotated test corpus are shown in Table 2. This table contrasts two main data points: (1) the **Clinical Text F_1 score**, which measures the classifier's ability to predict the correct relationship type given the focus and a modifier in the clinical text corpus, and (2) the

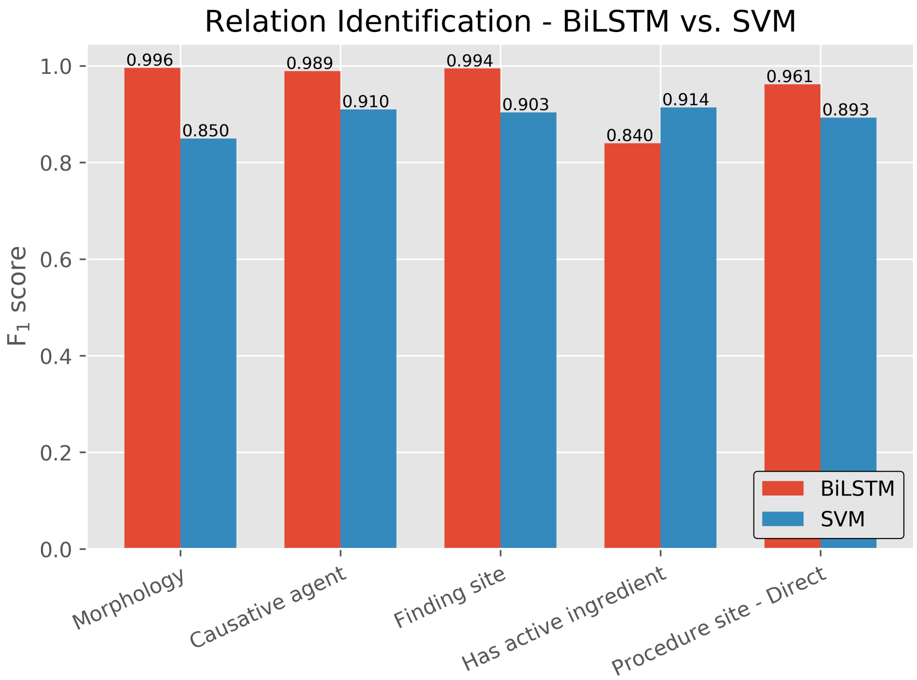


Figure 4: Comparing the BiLSTM + Clinical BERT model for SNOMED CT relation identification to the SVM model as reported by Kate[28] for five SNOMED CT attributes.

SNOMED CT Relationship F_1 score, which is the corresponding value derived from Experiment 1 for the given attribute. The $F_1\Delta$ value shows the difference between the two scores, illustrating the difference in performance when testing on relatively predictable SNOMED CT terms vs. real clinical text. The eight relationships supported by >20 annotations in the clinical text test dataset are displayed.

Table 2: Comparing BiLSTM relation identification scores using two different test data sets: real-world clinical text (Clinical Text F_1 score) and relationships from SNOMED CT (SNOMED CT Relationship F_1 score). The $F_1\Delta$ value equals Clinical Text F_1 score minus SNOMED CT Relationship F_1 score.

Attribute	**Clinical Text Precision**	**Clinical Text Recall**	**Clinical Text F_1 score**	**SNOMED CT Relationship F_1 score**	**F_1 Δ**
Severity	1.000	0.880	0.936	0.882	0.054
Laterality	0.990	0.950	0.970	0.999	-0.029
Clinical course	1.000	0.800	0.889	0.994	-0.105
Finding site	0.956	0.790	0.865	0.988	-0.123
Due to	0.528	0.487	0.507	0.830	-0.323
Has interpretation	0.577	0.714	0.638	0.992	-0.354
Following	0.733	0.306	0.431	0.837	-0.406
Associated with	0.579	0.134	0.218	0.738	-0.520

Experiment 4: Finally, Table 3 shows the effectiveness of using a dependency parse-based method for selecting the focus concept of a problem description. For each dependency parse model the accuracy is shown, where accuracy in this context reflects the number of times the model selected the same focus span as the human annotators over the 401 total entries in the test set.

Table 3: Evaluating the performance of the dependency parse-based method for selecting the focus concept of the clinical problem. Three different dependency parse models were evaluated.

Model	Accuracy
Default spaCy English (baseline)	0.68
ScispaCy Biomedical	0.75
ScispaCy Biomedical + fine-tuning	0.91

Discussion

Overall, both the CNN and BiLSTM significantly outperformed the Naïve Bayes classifier in identifying relationships between two SNOMED CT concepts, as shown in Table 1. This comparison provides evidence that a deep learning architecture is a viable approach and can outperform a simple Naïve Bayes baseline. The BiLSTM did also outperform the CNN model slightly. The difference was most evident in the F_1 macro avg score, which is important as this metric gives equal weight to each relationship type and disregards any SNOMED CT relationship class imbalance.

Figure 4 shows that a deep learning approach can outperform SVM classifiers at the relation identification task with SNOMED CT relationships. The BiLSTM scored higher for four of the five attributes tested, while the SVM outperformed the BiLSTM for one attribute: "Has active ingredient." It is worth noting that we cannot directly compare Kate's results[28] with our BiLSTM model beyond these five attributes listed in Figure 4. Overall F_1 scores are not comparable because Kate's model was trained and tested using the top 14 SNOMED CT attributes only, while our overall F_1 scores (see Table 1) are derived from a classifier trained and evaluated on 78 relationship types. Also, Kate's overall results factor in scores for `116680003`$|$*Is a*$|$ attributes – relationships that we omitted.

Table 2 highlights the challenges that come with applying these methods to clinical text. For several attributes, relation identification was significantly worse against real-world clinical problem descriptions compared to SNOMED CT relationships (as shown by $F_1\Delta$). These differences are not unexpected – the SNOMED CT text used for training is relatively structured,[27] but actual clinical problem descriptions are not. As shown, performance degradation for several relationship classes is pronounced, with a highly negative $F_1\Delta$ score signifying low model generalizability from the SNOMED CT text corpus to the clinical text.

Finally, the focus concept selection results in Table 3 show not only that a dependency parse-based method of focus concept selection is an effective technique, but that using a domain-specific pre-trained model does boost performance noticeably. Even more, these results indicate that even minimal fine-tuning (150 manual annotations) can have a fairly large impact on overall performance.

Conclusion

The goal of this work was to present an end-to-end system for converting unstructured summary level problem descriptions into SNOMED CT expressions. Our contribution focused primarily on introducing a new deep learning method for relation identification between concepts and problem phrases. We show that our method outperforms current approaches to identifying relationships between clinical phrases and SNOMED CT concepts, a fundamental part of building SNOMED CT expressions. We also show that a model trained exclusively on SNOMED CT stated relationship text does not transfer to clinical text without performance degradation.

Limitations

Our study has several limitations. First, there is no available gold-standard test set for evaluating the full conversion of text-based problem descriptions to SNOMED CT expressions – we must evaluate individual steps of the pipeline independently. Furthermore, such a test set is challenging to construct as there may exist more than one way syntactically to represent the same conceptual expression. Also, it has been shown that codification of problems by physicians is subject to considerable variation.[51] All these factors together make quantitative evaluation of this task difficult.

Acknowledgment. This study was funded by grant NCATS U01TR02016. We thank Donna Ihrke, Luke Carlson, and Sunyang Fu for assisting in the test corpus annotation and guideline development.

References

[1] Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst. 2014;2(1):3.

[2] Luo JS. Electronic medical records. Primary Psychiatry. 2006;13(2):20–23.

[3] Weed LL. Special article: medical records that guide and teach. N Engl J Med. 1968;278(12):593–600.

[4] Wright A, Sittig DF, McGowan J, Ash JS, Weed LL. Bringing science to medicine: an interview with Larry Weed, inventor of the problem-oriented medical record. Journal of the American Medical Informatics Association. 2014;21(6):964–968.

[5] Liu H, Wagholikar K, Wu STI. Using SNOMED-CT to encode summary level data–a corpus analysis. AMIA Summits on Translational Science Proceedings. 2012;2012:30.

[6] Fung KW, McDonald C, Srinivasan S. The UMLS-CORE project: a study of the problem list terminologies used in large healthcare institutions. Journal of the American Medical Informatics Association. 2010;17(6):675–680.

[7] Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc. 2011;18(2):181–186.

[8] Hodge CM, Narus SP. Electronic problem lists: a thematic analysis of a systematic literature review to identify aspects critical to success. Journal of the American Medical Informatics Association. 2018;25(5):603–613.

[9] Van Vleck TT, Wilcox A, Stetson PD, et al. Content and structure of clinical problem lists: a corpus analysis. In: AMIA Annu Symp Proc. vol. 2008. American Medical Informatics Association; 2008. p. 753.

[10] Price CM, C de C Williams A, Smith BH, Bottle A. Implementation of Patient-Reported Outcomes (PROMs) from specialist pain clinics in England and Wales: experience from a nationwide study. European Journal of Pain. 2019;23(7):1368–1377.

[11] Jaffe IS, Chiswell K, Tsalik EL. A decade on: systematic review of ClinicalTrials.gov infectious disease trials, 2007-2017. In: Open Forum Infectious Diseases; 2019. p. 1–9.

[12] Zelingher J, Rind DM, Caraballo E, Tuttle MS, Olson N, Safran C. Categorization of free-text problem lists: an effective method of capturing clinical data. In: Proceedings of the Annual Symposium on Computer Application in Medical Care. American Medical Informatics Association; 1995. p. 416.

[13] Wright A, Maloney FL, Feblowitz JC. Clinician attitudes toward and use of electronic problem lists: a thematic analysis. BMC Med Inform Decis Mak. 2011;11(1):36.

[14] Wang Y, Kung L, Byrd TA. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change. 2018;126:3–13.

[15] Holmes C. The problem list beyond meaningful use: part I: the problems with problem lists. Journal of AHIMA. 2011;82(2):30–33.

[16] International Health Terminology Standards Development Organization (IHTSDO). Compositional Grammar Specification and Guide, v2.3.1; 2016.

[17] Campbell JR, Payne T. A comparison of four schemes for codification of problem lists. In: Proc Annu Symp Comput Appl Med Care. American Medical Informatics Association; 1994. p. 201.

[18] SNOMED International. SNOMED CT Sept 2018; 2018. Accessed: 2019-01-01. https://www.nlm.nih.gov/healthit/snomedct/us_edition.html.

[19] Wasserman H, Wang J. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. In: AMIA Annu Symp Proc. vol. 2003. American Medical Informatics Association; 2003. p. 699.

[20] Agrawal A, He Z, Perl Y, et al. The readiness of SNOMED problem list concepts for meaningful use of electronic health records. Artificial intelligence in medicine. 2013;58(2):73–80.

[21] Campbell JR, Xu J, Fung KW. Can SNOMED CT fulfill the vision of a compositional terminology? Analyzing the use case for problem list. In: AMIA Annu Symp Proc. vol. 2011. American Medical Informatics Association; 2011. p. 181.

[22] Elkin PL, Tuttle M, Keck K, Campbell K, Atkin G, Chute CG. The role of compositionality in standardized problem list generation. Studies in Health Technology and Informatics. 1998;52:660–664.

[23] Schulz S, Schober D, Raufie D, Boeker M. Pre-and postcoordination in biomedical ontologies. In: OBML 2010 Workshop Proceedings; 2010. p. L1–L4.

[24] Elkin PL, Brown SH, Husser CS, et al. Evaluation of the content coverage of SNOMED CT: ability of SNOMED

Clinical Terms to represent clinical problem lists. In: Mayo Clinic Proc. vol. 81. Elsevier; 2006. p. 741–748.

[25] Brin S. Extracting patterns and relations from the world wide web. In: International Workshop on The World Wide Web and Databases. Springer; 1998. p. 172–183.

[26] Miñarro-Giménez JA, Martínez-Costa C, López-García P, Schulz S. Building SNOMED CT post-coordinated expressions from annotation groups. Studies in Health Technology and Informatics. 2017;235:446–450.

[27] López-García P, Schulz S. Structural patterns under X-rays: is SNOMED CT growing straight? PLoS ONE. 2016;11(11):e0165619.

[28] Kate RJ. Towards converting clinical phrases into SNOMED CT expressions. Biomedical Informatics Insights. 2013;6:BII–S11645.

[29] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: AMIA Annu Symp Proc. American Medical Informatics Association; 2001. p. 17.

[30] Fung KW, Hole WT, Nelson SJ, Srinivasan S, Powell T, Roth L. Integrating SNOMED CT into the UMLS: an exploration of different views of synonymy and quality of editing. Journal of the American Medical Informatics Association. 2005;12(4):486–494.

[31] Spasić I, Corcoran P, Gagarin A, Buerki A. Head to head: semantic similarity of multi–word terms. IEEE Access. 2018;6:20545–20557.

[32] Neumann M, King D, Beltagy I, Ammar W. ScispaCy: fast and robust models for biomedical natural language processing. arXiv preprint arXiv:190207669. 2019;.

[33] Benson T. Principles of Health Interoperability HL7 and SNOMED. Springer Science & Business Media; 2012.

[34] Lopez MM, Kalita J. Deep learning applied to NLP. arXiv preprint arXiv:170303091. 2017;.

[35] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation. 1997;9(8):1735–1780.

[36] Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing. 1997;45(11):2673–2681.

[37] Rumelhart DE, Hinton GE, Williams RJ, et al. Learning representations by back-propagating errors. Cognitive Modeling. 1988;5(3):1.

[38] Cheng J, Dong L, Lapata M. Long short-term memory-networks for machine reading. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing; 2016. p. 551–561.

[39] Goldberg Y. A primer on neural network models for natural language processing. Journal of Artificial Intelligence Research. 2016;57:345–420.

[40] Li F, Zhang M, Fu G, Ji D. A neural joint model for entity and relation extraction from biomedical text. BMC Bioinformatics. 2017;18(1):198.

[41] Kim Y. Convolutional neural networks for sentence classification. In: EMNLP; 2014. p. 1746–1751.

[42] Zeng D, Liu K, Lai S, Zhou G, Zhao J. Relation classification via convolutional deep neural network. In: Proceedings of the 25th International Conference on Computational Linguistics (COLING); 2014. p. 2335–2344.

[43] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018;.

[44] Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. arXiv preprint arXiv:190403323. 2019;.

[45] Reátegui R, Ratté S. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. BMC Medical Informatics and Decision Making. 2018;18(3):74.

[46] Rish I. An empirical study of the Naive Bayes classifier. In: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence. vol. 3; 2001. p. 41–46.

[47] Bender D, Sartipi K. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. In: Proc. of the 26th IEEE International Symposium on Computer-Based Medical Systems. IEEE; 2013. p. 326–331.

[48] OpenCEM. OpenCEM Browser;. Accessed: 2019-06-06. http://www.opencem.org/.

[49] Kalra D, Beale T, Heard S. The openEHR foundation. Studies in Health Technology and Informatics. 2005;115:153–173.

[50] Goossen WT. Detailed clinical models: representing knowledge, data and semantics in healthcare information technology. Healthcare Informatics Research. 2014;20(3):163–172.

[51] Rothschild AS, Lehmann HP, Hripcsak G. Inter-rater agreement in physician-coded problem lists. In: AMIA Annu Symp Proc. vol. 2005. American Medical Informatics Association; 2005. p. 644.