Identifying Condition-action Statements in Medical Guidelines: Three Studies using Machine Learning and Domain Adaptation

Hossein Hematialam
UNC Charlotte
Wlodek W. Zadrozny (wzadrozn@uncc.edu)
UNC Charlotte

Research Article

Keywords: medical guidelines, clinical guidelines, condition-action sentences, natural language processing, NLP, classification, transfer learning, domain adaptation, BioBERT, transformer, UMLS

Posted Date: May 11th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-500521/v1

RESEARCH

Identifying condition-action statements in medical guidelines: Three studies using machine learning and domain adaptation

Hossein Hematialam[1] and Wlodek W. Zadrozny[1,2*]

*Correspondence:
wzadrozn@uncc.edu
[1]Department of Computer Science, UNC Charlotte, Charlotte, NC, USA
Full list of author information is available at the end of the article

Abstract

Background: Medical guidelines provide the conceptual link between a diagnosis and a recommendation. They often disagree on their recommendations. There are over thirty five thousand guidelines indexed by PubMed, which creates a need for automated methods for analysis of recommendations, i.e., recommended actions, for similar conditions.

Results: This article advances the state of the art in text understanding of medical guidelines by showing the applicability of transformer-based models and transfer learning (domain adaptation) to the problem of finding condition-action and other conditional sentences. We report results of three studies using syntactic, semantic and deep learning methods, with and without transformer-based models such as BioBERT and BERT. We perform in depth evaluation on a set of three annotated medical guidelines. Our experiments show that a combination of machine learning domain adaptation and transfer can improve the ability to automatically find conditional sentences in clinical guidelines. We show substantial improvements over prior art (up to 25%), and discuss several directions of extending this work, including addressing the problem of paucity of annotated data.

Conclusion: Modern deep learning methods, when applied to the text of clinical guidelines, yield substantial improvements in our ability to find sentences expressing the relations of condition-consequence, condition-action and action.

Keywords: medical guidelines; clinical guidelines; condition-action sentences; natural language processing; NLP; classification; transfer learning; domain adaptation; BioBERT; transformer; UMLS

1 Introduction

Clinical decision-support systems (CDSSs) typically address two major tasks: diagnosis — determining "what is true" about a patient; and recommendation — determining "what to do (or not)" for the patient. Medical guidelines provide the conceptual link between a diagnosis and a recommendation. For example, they may include sentences such as this:

> "*In the population aged 18 years or older with CKD and hypertension*, `initial (or add-on) antihypertensive treatment should include an ACEI or ARB to improve kidney outcomes`"

The italics show the diagnosis part, i.e., a *condition*, and the courier font a recommendation, i.e., an *action*. This article focuses on automated identification of

condition-action sentences in medical guidelines. We present results of three studies, which use different text analytics techniques, and show that:

- Modern deep learning techniques using attention-based models give substantial improvements in accuracy (11%) and F1-score (25%) over earlier machine learning methods.
- Transfer learning can potentially be used on text of medical guidelines in new domains, even with small amounts of available training data. Namely, training on two guideline documents produces results better than hand-coded rules, and comparable to standard machine learning methods (using syntactic and semantics features), even though they only have 445 words in common, and their distributions are completely different.

The three studies use, respectively, syntactic, semantic and deep learning methods, evaluated on a set of three annotated medical guidelines. Our main contribution is to show the applicability of the recently developed techniques, namely neural network transformers and transfer learning, to this particular problem, and in comparing them with alternatives based on older machine learning techniques.

In another contribution, we have released two annotated guidelines used in these experiments, adding to the one previously published data set of [1].

1.1 Motivation

However, before we proceed with the details of our contributions and experiments, let us spend a few moments on the motivation of this work.

There were, in 2017, over twenty thousand clinical practice guidelines indexed by PubMed [1] , with over 1,500 appearing every year [2]. Our current, April 2021, search shows 35,000+ articles on PubMed are indexed as "guideline"/"practice guideline."

Such guidelines may disagree on their recommendations, as documented in prior work, including ours, [3, 4, 5]. Controversies over prostate screening (PSA), breast cancer screening, hypertension, and other treatment and prevention guidelines are well-known. In addition, clinical recommendations are often in conflict when managing comorbidities [6].

Notice that the disagreements focus on actions, i.e., what to do in particular situations (conditions of the patient). For example, for what ages and breast conditions should mammography be recommended.

We believe patient outcomes would be improved, overtreatment would be reduced, and better processes for creation of treatment guidelines could be established, if we could better reason about individual guidelines and guidelines corpora. In particular, it is natural to imagine decision support systems for healthcare professionals [7] accessing properly indexed and contextualized condition-action statements. Therefore, we should understand whether, and how, such condition-action statements can be automatically extracted from texts.

1.2 Organization of the article and brief description of the studies

After establishing some preliminaries and discussing related work in Section 2, we will methodically describe each of our set of experiments. In each study, multiple machine learning models are evaluated. Our studies progress through different methods

[1] https://pubmed.ncbi.nlm.nih.gov/

of machine learning, starting with learning patterns based on part-of-speech (Study 1), adding syntactic and semantic information (Study 2), and several experiments in transfer learning using deep learning methods (Study 3).[2]

In these studies, we discuss on *condition-action* (CA), *condition-consequence* (CC), and *action* (A) sentences. The class consisting of CC and CA classes will be abbreviated as CCA; we will refer to the CCA classes as *conditional* sentences. In a few instances, for comparison with other works, we will also discuss class CCA+A. With this naming convention, we now can summarize the topics of studies:

Study 1. (Described in Section 4.) Identifying conditional and condition-action statements using domain-independent *syntactic features*, part-of-speech (POS) tags.

Study 2. (Described in Section 4.3.) Identifying conditional and conditional statements using both domain-independent features and UMLS *semantic types*.

Study 3. (Described in Section 5.3.) Experimenting with *deep learning, domain adaptation, and machine learning transfer*.

- Experiment 1. Identifying condition-action and conditional statements using pretrained transformers models.
- Experiment 2. Identifying condition-action and conditional statements using pretrained transformers models and features from Study 1 and Study 2.
- Experiment 3. Repeating Experiments 1 and 2 by training classifiers on two guidelines (rhinosinusitis+hypertension), and testing them on the third guideline (asthma).

All these experiments use three clinical guidelines: asthma, rhinosinusitis, and hypertension.

The data and its preparation are discussed in Section 3. Afterwards, we describe the experiments in Sections 4–5. In particular, Section 5.3 describes several experiments in domain adaptation and transfer learning. Discussion and Conclusions follow in Sections 6 and 7.

2 Preliminaries and related work

This article focuses on a specific text analytics problem, namely on finding sentences with condition-action recommendations in medical guidelines. For this purpose, we use a variety of classification techniques ranging from traditional methods such as logistic regression and random forest, to new deep learning and transfer learning methods introduced in the last few years.

Its motivation, as stated earlier, comes from two directions: clinical decision support and natural language processing. As such it belongs to the broad category of applications of artificial intelligence in medicine, and in particular, in clinical decision support.

The specific techniques of supervised machine learning used in this article comprise both the classical approaches such as logistic regression and random forest, and more recently introduced deep learning methods involving domain adaptation

[2]Study 1, reported earlier in [8], is cited here to enable full comparison with Study 2 and Study 3.

and transfer learning. All of these are applied to the task of classifying sentences as to whether they express a condition and action or not.

Therefore, in the preliminaries we need to cover both topics: natural language processing of medical text, as well as the newer machine learning techniques, including their applications to medical text, also in the context of clinical decision support.

2.1 Five decades of automated analysis of medical texts

Text analysis of medical records is already mentioned in a 1975 article by N.Sager [9, 10], and included extracting information to populate relational databases. Over the five decades of research in this space, many new techniques have been developed and applied to medical texts.

Even though no system can claim to "really" understand natural language, significant progress has been made in the last ten years [11]. In several domains, such as data extraction, classification and question answering, automated systems have dramatically improved their performance, in some cases performing better than humans. This progress is chiefly due to the unmatched pattern recognition and memorization capabilities of deep neural networks (see, e.g., [12] for an overview).

If we focus on medical texts, we see that modern NLP methods have been applied to clinical decision support, e.g., [13], to clinical trials [14], to automatic extraction of adverse drug events and drug related entities [15], and to other areas [16, 17].

2.2 Analysis of medical guidelines

A recent overview of research on clinical guidelines and its applications can be found in [18, 19]. However, over the past few decades many problems and approaches have been tried to represent and execute clinical guidelines over patient-specific clinical data. They include document-centric models, decision trees and probabilistic models, and "Task-Network Models"(TNMs) [20], which represent guideline knowledge in hierarchical structures containing networks of clinical actions and decisions that unfold over time. A general-purpose architecture for syntax-semantic translation of medical guidelines sentences, using classical NLP techniques, and based on GATE [21], has been recently proposed in [22]. A methodology for using linguistic patterns in guideline formalization to aid human modellers and reduce the human modelling effort has been proposed in [23]. A method to identify activities to be performed during a treatment which are described in a guideline document appears in [24], used relations of the UMLS Semantic Network [25].

Most related to our work has been a proposal [1] for a rule-based method to identify conditional activities in guideline documents. In their experiment with document-specific rules, they achieved a recall of 75%, a precision of 88%, and 81% F-score on the same chapter of asthma guidelines which is used in our research.

Similarly, the use of specific heuristic patterns has been shown to lead to a relatively high 85.54% accuracy, in identifying recommendation statements in the hypertension guideline [26]. Ensemble learning was applied [27] to the same set of three guidelines as used in this article, achieving $\tilde{8}$0-84% accuracy. Part of the ensemble was a deep learning module, but it was the weakest overall performer. These results were obtained on the same guidelines as in our experiments, but at different granularity, namely on classes of combined action and conditional sentences (CCA+A).

As we show later in Section 6.1, our current methods provide about 5% – 11% improvement over these results.

Our own earlier work [8] reported lower results than [1]. The difference was due to our using of completely automated feature selection when training on an annotated corpus, and not relying on manually created extraction rules. In addition, the results in [1] demonstrate recalls on specific patterns. Thus, if applied to all activities in their annotated corpus, their recall was shown to be 56%, and on our annotated corpus, it was 39%. As we show later in this article in Sections 5 and 6.1, we can achieve the F1 scores of 81% and higher, using completely automated methods; even purely transfer-based methods can produce a 67% F1 score and 68% recall.

2.3 Deep learning methods, domain adaptation and transfer learning

There are plenty of overviews of deep learning methods, e.g. [28, 29]. In our experiments we use pretrained and transformer models such as BERT [30] and BioBERT [31], which are relatively well-known. However, we feel the need to discuss the concepts of *transfer learning* and *domain adaptation*, which are a focus of some experiments reported in this article.

We will start by observing that the two concepts overlap and are often used interchangeably. In particular, in natural language processing, as observed by [32], transfer learning is sometimes referred to as domain adaptation. Wikipedia tries to make a distinction. It explains that the basic idea of *domain adaptation* is to learn a model on a dataset, in a way that would make it applicable in other, related situations. For example, adapting spam filtering models from one set of users to another. [3] On the other hand, transfer learning "focuses on storing knowledge gained while solving one problem and applying it to a different but related problem," for example, "knowledge gained while learning to recognize cars could apply when trying to recognize trucks".[4]

For the purpose of this article, we adopt the definition from a 2015 survey of the topic [33]: "*domain adaptation* is a subcategory of transfer learning. In domain adaptation, the source and target domains all have the same feature space (but different distributions); in contrast, *transfer learning* includes cases where the target domain's feature space is different from the source feature space or spaces" (our emphasis). We can contrast with this the traditional machine learning which generally assumes that the data is in the i.i.s. form (independent and identically distributed), and that from sampled, labeled data we can train a good model for test data.

Domain adaptation and transfer learning are very active areas of research [34]. We also observe their growing importance for natural language processing [35, 36] and in clinical NLP. For example, [37] argue that "researchers in clinical NLP should treat domain adaptation, transfer learning, etc. as a first-class problem rather than a niche area", and [38] as 'worth exploring'.

In our case, Experiment 1 of Study 3 (Section 5.1), where we used pretrained deep learning models to find conditional sentences, is an example of domain adaptation. We also perform two experiments in transfer learning in Study 3. In Experiments 2 (Section 5.2), we add features from Study 1 and 2, which are not used for training

[3] https://en.wikipedia.org/wiki/Domain_adaptation

[4] https://en.wikipedia.org/wiki/Transfer_learning

of the deep learning models. In Experiment 3 (Section 5.3), we train on a data set combining two guideline documents, and test on a different document, showing earlier (Section 3) that the two frequency distributions of words are very different.

3 The data

In this section we discuss the data. First, we discuss the sources of our data and its basic statistical information. Next, to show that our experiments fall under the category of transfer learning, we discuss the differences between feature distributions used in the experiment of Study 3 (Section 5).

3.1 The dataset of three annotated guidelines

We used three medical guidelines documents to create gold standard datasets and (in 2017) made them publicly available[5]. To the best of our knowledge, these three annotated documents are the only such dataset, besides the original annotations from [1] . We annotated three sets of guidelines to create gold standards to measure the performance of our condition-action extracting models. The sets of guidelines are: hypertension [39], Chapter 4 of asthma [40], and rhinosinusitis [41]. Chapter 4 of the asthma guidelines was selected for comparison with prior work of Wenzina and Kaiser [1]. Each sentence was annotated by one domain expert and us, with the disagreements of less than 10 percent. We have annotated the guidelines for the conditions, consequences, modifiers of conditions, and type of consequences.

Our data preparation process proceeded as follows: we started by converting the guidelines from PDF or HTML to the text format, editing sentences only to manage conversion errors, most of which were bullet points. Tables and some figures posed a problem, which were simply treated as unstructured text.

The next step, the annotation of the guidelines text, focused on determining whether there were condition statements in the candidate sentences or not. The instruction to the annotators was to try to paraphrase candidate sentences as sentences with "if condition, then consequences." If the transformed/paraphrased sentence conveyed the same meaning as the original, we considered it to be a condition-consequence sentence, and we could annotate the condition and consequence parts. For example, the sentence *"Beta-blockers, including eye drops, are contraindicated in patients with asthma"* from [40] can be paraphrased to *"If patients have asthma, then beta-blockers, including eye drops, are contraindicated"*. The paraphrased sentence conveys the same meaning. So, it became a condition-consequence sentence in our dataset. On the other hand, for example, we cannot paraphrase *"Further, the diagnostic criteria for CKD do not consider age-related decline in kidney function as reflected in estimated GFR"* from [40] to an if-then sentence; it, therefore, belongs to the category no condition (nor action).

We annotated the type of sentences based on their semantics: We classified them into four classes: condition-action (CA); condition-consequence (CC), which includes effect intention, and event; action (A); and no condition nor action (NC), which includes all other sentences. Examples of sentences we are trying to find are shown in Table 1.

[5] `https://data.world/hematialam/condition-action-data`

Table 1 Examples of classified sentences and their classes/types.

Type	Example
Condition-Action	*Timely referral is indicated if chronic or recurrent symptoms severely affect the patient's productivity or quality of life.*
Condition-Consequence	*Most patients with uncomplicated viral URIs do not have fever.*
Action	*Adjustment is necessary for fluticasone and mometasone and may also be necessary for alternative devices.*
No condition (nor action)	*"Further, the diagnostic criteria for CKD do not consider age-related decline in kidney function as reflected in estimated GFR"*

Table 2 contains the basic statistical information about the three guidelines. The numbers do not add up, because certain types of sentences were omitted from the annotation process (see [8]). This is, among others, due to the fact that they require a model that crosses the sentence boundaries to be interpreted. For example, *"The most effective therapy is intranasal steroids."*

Table 2 Statistical information about annotated guidelines. `Words` **– the total number of words in the document.** `Avg Length` **– average number of words per sentence (applies to all sentences).** `CA` **condition-action (recommendation);** `CC` **condition-consequence;** `A` **action;** `NC` **no condition (nor action).**

Guidelines	`Words`	`Avg Length`	`Sentences`	`CA`	`CC`	`A`	`NC`
Asthma	3621	16	224	38	7	8	117
Rhinosinusitis	19870	27	726	97	39	15	610
Hypertension	8182	34	238	63	14	1	200

In all experiments, except for Experiment 3 in Study 3, described in Section 5.3, the data was split 75% for training and 25% for testing. In Study 3, Experiment 3 tests domain adaptation/transfer learning, and therefore, the rhinosinusitis and hypertension guidelines were used for training and the asthma guidelines for testing.

3.2 The data from the perspective of domain adaptation and transfer learning

We plan to argue that our experiments in Section 5 prove the applicability of transfer learning to detection of sentences with conditions and actions (separately or jointly). For this argument, we need to establish that the feature distribution of training data is different than the test data, and that their feature sets are different.

We will be using the following feature sets in Study 3 (Section 5) where we are experimenting with deep learning, domain adaptation, and machine learning transfer.

— Experiment 1: Identifying conditional statements using pretrained transformers models. Here, the feature set consists of the vectors from BioBERT and other transformer models. Clearly, the distribution of words in the BioBERT training data (trained on a large set of diverse biomedical texts) is different than in the selected three guidelines.

— Experiment 2: Identifying conditional statements using pretrained transformers models and features from Study 1 and Study 2. Here, the features are the sum of BioBERT vectors and the features from Study1 and (separately) the features from Study 2. And, for the reason of vocabulary distribution alone, it can be viewed as a case of domain adaptation, as defined earlier in Section 2.3.

— Experiment 3: transfer learning. It consists of repeating Study 2, Experiments 1 and 2 by training classifiers on two guidelines (rhinosinusitis+hypertension) and testing them on the third guideline (asthma).

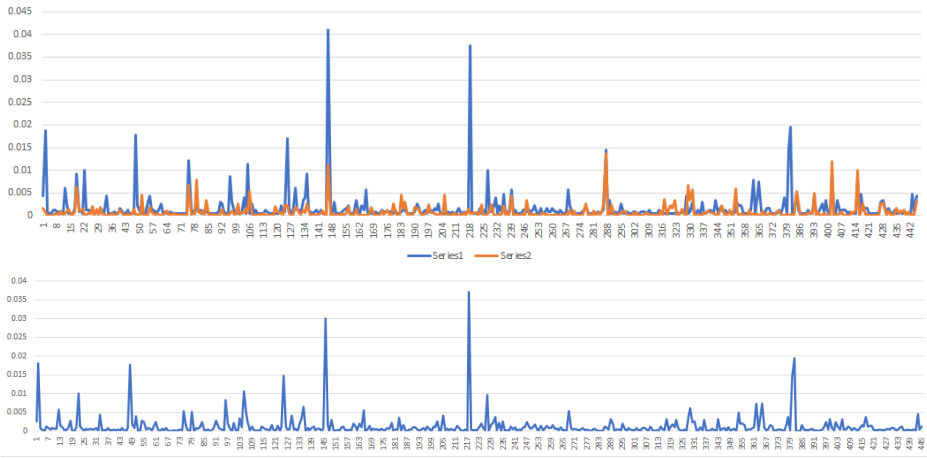

Figure 1 Top panel: the frequency distribution on 445 words-in-common in the training data (hypertension+rhinosinusitis; Series 1) and test data (asthma; Series 2). The bottom panel shows the (large) difference between the two distributions.

To establish the applicability of the concept of transfer learning to Experiment 3 (Study 3), we need to pay attention to feature distribution. In Experiment 3, the combined rhinosinusitis and hypertension guidelines have the vocabulary of 2719 words (training set), while the asthma guidelines (test set) has the vocabulary of 661 words (test set). So, intuitively, these distributions should be different, and this difference is apparent in Figure 1.

This visual observation is indeed confirmed by the Kolmogorov-Smirnov test (K-S test), showing the difference with the significance level better than 0.001 on the total vocabulary. Even the distributions on the 445 words-in-common are very different according to the K-S test, with the significance level of about 0.025. The restricted distributions on the training and testing data set based on the K-S test with the significance level of about 0.025. This is supported by another test on these two restricted distributions. The Kullback–Leibler relative entropy (K-L divergence) is high in both directions: training–testing has the value of 0.383, and for testing–training we get 0.481. Thus, clearly, however we look at the distributions, they are very different.

4 Using syntactic and semantic features (Studies 1 and 2)

Our baseline, Study 1, recognized condition-action statements by applying classical machine learning algorithms using a combination of domain independent syntactic and semantic features, and extending our earlier results [8]. It also extended (and

improved) the results of [1] by using additional datasets, and it proved that finding sentences with conditions and actions does not have to be tailored to a specific document, nor hand-coded in the form of regular expressions.

4.1 The feature set

We now briefly summarize these results, as they are given here for context and completeness. To provide the required preliminaries for Study 2 (Section 4.3) and Study 3 (Section 5), it is useful to discuss the features and methods based on [8].

The features, consisting of part-of-speech tags and syntactic patterns, were extracted from the sentences in the guidelines using the CoreNLP [42] shift-reduce constituency parser. More specifically, they consisted of POS tags and sequences POS tags (a modified algorithm, subsuming the one in [8], is shown in Section 4.3). For example, we have the following set of features, derived from the parse tree shown in Figure 2 for the sentence

If there is no response to treatment the drug should be discontinued.

Example 4.1 Example list of features:

```
['IN', 'NN', 'VBZ', 'JJR', 'IN', 'CD', ',', 'VBP', 'DT', 'NN', '.', 'IN-NN',
'IN-NN-VBZ', 'NN-VBZ', 'NN-VBZ-JJR', 'VBZ-JJR', 'VBZ-JJR-IN', 'JJR-IN', 'JJR-IN-CD',
'IN-CD', 'IN-CD-,', 'CD-,', 'CD-,-VBP', ',-VBP', ',-VBP-DT', 'VBP-DT', 'VBP-DT-NN',
'DT-NN', 'DT-NN-.', 'NN-.']
```

Classifier	class	precision	recall	F1-score	Accuracy
Random Forest	CA	0.95	0.42	0.58	0.90
Gradient Boosting	CA	0.81	0.52	0.63	0.90
Logistic Regression on RF&GB	CA	0.66	0.54	0.59	0.88

Table 3 Classification results on annotated guidelines using only POS tags and their combinations, focusing on the detection of condition-action sentences. These baseline results are the core of Study 1.

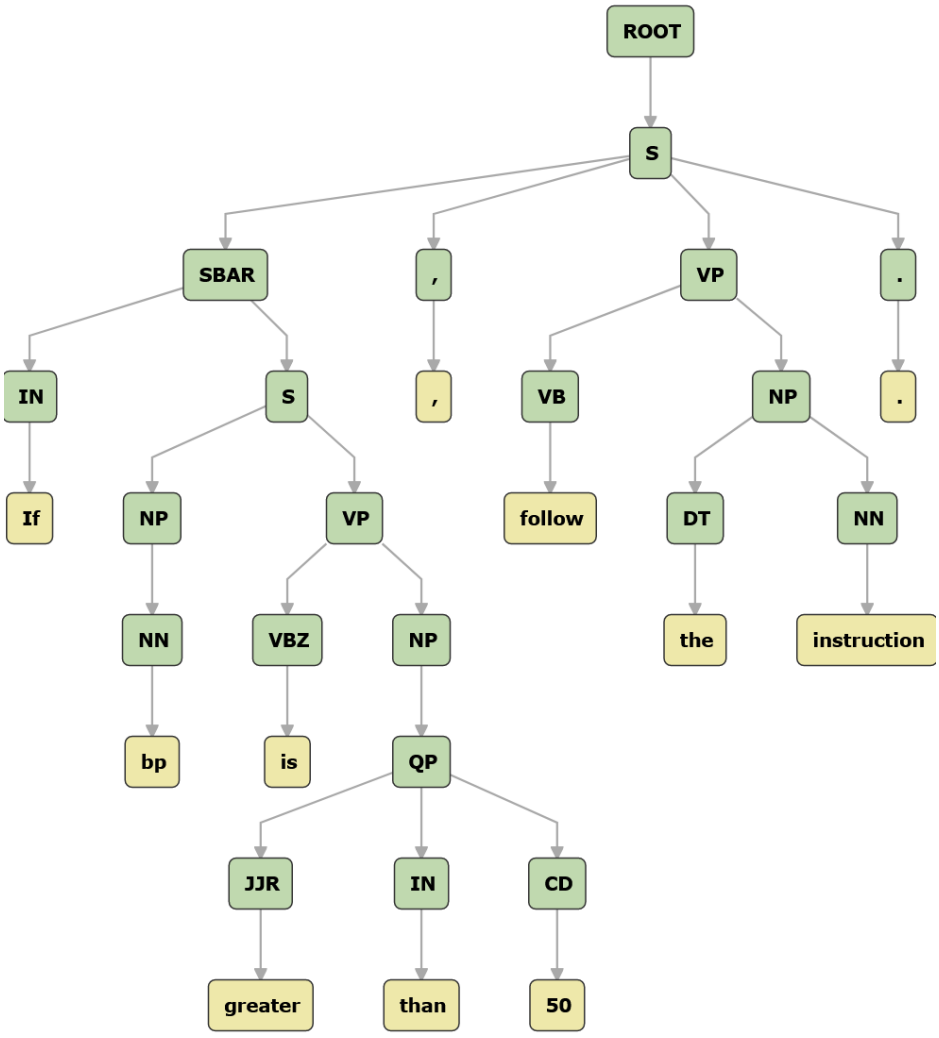

Figure 2 Example of the parse tree showing the part-of-speech (POS) features used in Study 1, and syntactic structures used in Study 2.

4.2 Evaluation measures and baseline results

We will be using the following *evaluation measures* to report results of our experiments: precision (P), recall (R), F1-measure, and accuracy (A).

$$P = \frac{TP}{TP + FP}, \qquad R = \frac{TP}{TP + FN}, \qquad F1 = \frac{2 * P * R}{P + R}, \qquad A = \frac{TP + TN}{N}$$

In the above definitions, TP is the number of items (e.g. condition expressions) that are correctly classified to a category, FP is the number of items that are misclassified, and FN is the number of items misclassified to other categories (e.g. condition as action or no-condition). N is the total number of items to be classified.

Classifier	class	precision	recall	F1-score	Accuracy
Random Forest	CCA	0.72	0.44	0.56	0.85
Gradient Boosting	CCA	0.62	0.44	0.52	0.83
Logistic Regression on RF&GB	CCA	0.63	0.52	0.57	0.84

Table 4 Summary of classification results on annotated guidelines, using domain independent syntactic features (Study 1), and based on [8]. The CA and CC classes are combined and shown as CCA.

In our evaluation we used the `scikit-learn` implementation of the algorithms. The results for extracting condition-action (CA) statements are shown in Table 3. We achieved recall of 52%, precision of 81% for the class CA and F1-score of 63%.

Table 4 shows combined results on the CC and CA classes, shown as CCA. Later, in Section 5, we show further improvements applying the three algorithms mentioned here to the vector representations produced by the deep learning models. In addition, we also show strong improvements on the combined classes CC, CA and A, used in [26, 27].

4.3 Identifying conditional statements using semantic types (Study 2)

Study 2 extends Study 1 by adding semantic features, that is, it is applying the domain knowledge to the process of finding condition-action sentences. For each candidate, we extract POS tags and syntactic tags at the phrase or clause level (see, Figure 2 and Example 4.2). We also use MetaMap[6]to extract UMLS semantic types of entities. We add these features to the features from Study 1.

Algorithm 1

Input: `Sentence`
Output: `Sentence type`
1: Parse the `Sentence`.
2: **if** There is a modifier tagged as `IN`, `WRB`, `WP`, or `TO` **then**
3: Extract linguistic (syntactic and semantic) features from the sentence
4: Using the extracted features, detect the sentence in one of the two categories: CA or CCA
5: **return** `Sentence type`
6: **end if**
7: **return** NC

Algorithm 1 shows the steps and the preconditions of the extraction process; the process requires the existence of specific syntactic modifiers. Here, `IN` denotes a preposition or a subordinating conjunction, `WRB` stands for a Wh-adverb, `WP` for a Wh-pronoun. and `TO` for the preposition 'to'.

In all experiments, after creating the lists of features for each sentence, we used Random Forest classifier and Gradient Boosting classifier to classify sentences in our data sets. We also used the combined output probabilities from the first two classifiers to create features for a logistic regression classifier (see Tables 5 and 5).

We used the features created by the transformer models in three types of experiments:

1 We evaluate the performance of four classifiers, logistic regression, random forest, gradient boosting, and ensemble model using logistic regression on the output probabilities from random forest and gradient boosting. We only use the vectors created by the transformers as input to these four classical models.

[6] `https://metamap.nlm.nih.gov/`

2 We merged the features from Study 1 with the vectors from the transformer models. We evaluate the performance of the classifiers mentioned above in identifying conditional statements.
3 We used a combination of vectors and features from Study 2 to identify conditional statements using the classifiers mentioned earlier.

Example 4.2 shows the type of features that are added to the syntactic features shown earlier in Example 4.1. Note that the two sentences are different.

Example 4.2 Additional features for the sentence:
If bp is greater than 50, follow the instruction.

```
--- ['SBAR_IN-S-,', 'QP_JJR-IN-CD']
--- ['SBAR-VP-.', 'QP']
--- ['if-[gngm]', 'than-[fndg]']
--- ['if_[fndg]', 'if_[gngm]', 'than_[fndg]']
```

In this experiment, for the class CA (condition-action) we obtained a precision value of 88%, recall of 56%, and F1 score of 68% using gradient boosting and similar results using random forest and logistic regression, as shown in Table 5. Thus adding semantic features leads to at least 5% improvement (compared to Table 3).

We achieved a precision of 75%, a recall of 57 %, and F1-score of 65% for the combined class CCA of conditional statements; see Table 6, showing improved results (over 10%), as compared to Table 4.

Table 5 Classification results on annotated guidelines, focusing on condition-action sentences and using all features (semantic and syntactic); Study 2.

Classifier	class	precision	recall	F1-score	Accuracy
Random Forest	CA	1.0	0.52	0.68	0.92
Gradient Boosting	CA	0.88	0.56	0.68	0.91
Logistic Regression on RF&GB	CA	0.77	0.60	0.67	0.90

Table 6 This table shows the classification results on combined condition-consequence and condition-action classes using all features (semantic and syntactic); Study 2

Classifier	class	precision	recall	F1-score	Accuracy
Random Forest	CCA	0.81	0.48	0.60	0.87
Gradient Boosting	CCA	0.72	0.54	0.62	0.86
Logistic Regression on RF&GB	CCA	0.75	0.57	0.65	0.87

5 Deep learning and transfer methods (Study 3)

This study uses pretrained deep neural language representation models. As before, we use these models to identify various types of conditional statements, including condition-action (CA) and condition-consequence (CC) statements.

Transfer learning through pretrained language models is a very common method in NLP. Typically, a deep learning model for target tasks are partially pretrained to create a language model and, then fine-tuned on the supervised dataset. There are many well-known such language models (representations) that provide similar capabilities, but not necessarily similar performance.

In our experiments, we used the following models: BERT [43], XLNet [44], DistilBERT [45], BioBERT `v. 1.1` [31], SciBERT `scivocab-uncased` [46], as well as BlueBERT `base-PubMed` and `base-PubMed+MIMIC-III` [47].

In contrast with the standard practice, because of the small size of our available data (discussed in Section 3), we could not retrain the deep learning models. Instead, we used the representations they produce to train a collection of 'standard' classifiers, such as logistic regression and random forest. In other words, for any sentence of the guidelines, a deep learning model produced a vector. A learning algorithm views each dimension as a feature, and assigns importance to individual features or their collections. The result of this learning process is classifier.

In Experiment 1 of Study 3, we use pretrained deep learning models to find conditional sentences, which is an example of domain adaptation. We also perform two experiments in transfer learning in Study 3. In Experiment 2 we add features from Study 1 and 2, which are not used for training of the deep learning models (although, in principle, they may be *latently* present in their neural representations [48, 49, 50, 51]).

In Experiment 3 of that study, we train on a data set combining two guideline documents and test on a different document. We show that earlier in Section 3, the two frequency distributions of words are very different.

As recommended in [43], for the BERT-based models, the final hidden states corresponding to `[CLS]` token were used as the aggregate sequence representation for classification tasks. For the XLNet model, we used the final hidden states corresponding to the last token. This method provides us, for each transformer model, a sentence representation as a tensor of shape (1,768), i.e., a vector of 768 parameters/features.

5.1 Deep learning. Study 3, Experiment 1

The results of two experiments using pretrained transformer models as a source of features for logistic regression are shown in Tables 7 and 8. We only show results using logistic regression, which is overall the best classifier in this context. The two experiments provide a baseline for the conditional classes CA and CCA. We can see that XLNet is the weakest overall performer, and BioBERT the strongest.

Table 7 This table illustrates the classification results on identifying condition-action statements (CA) using transformer embeddings as features (Study 3, Experiment 1). In this table, we only report results from the logistic regression classifier, but use different vectorized representations of sentences coming from the models in the first column.

Model	Classifier	precision	recall	F1-score	Accuracy
BERT	Logistic Regression	0.78	070.	0.74	0.92
DistilBERT	Logistic Regression	0.80	0.80	0.80	0.93
XLNet	Logistic Regression	0.60	0.56	0.58	0.87
BioBERT	Logistic Regression	**0.89**	**0.82**	**0.85**	**0.95**
SciBERT	Logistic Regression	0.75	0.72	0.73	0.91
BlueBert	Logistic Regression	0.80	0.74	0.77	0.93
BlueBERTMIMIC	Logistic Regression	0.75	0.72	0.73	0.91

We view this experiment as our first rudimentary experiment in domain adaptation. We note the large discrepancy in size of the data (30K words vs. 20B words for BioBERT) and a potential for noise from the vocabulary mismatch. However, despite these potential problems, transformer-based models give a substantial boost in performance compared with Tables 5 and 6 from Study 2 (and Study 1).

Table 8 This table illustrates the classification results on identifying conditional statements (CCA) using transformer embeddings as features (Study 3, Experiment 1). In this table, we only report results from Logistic Regression classifier.

Model	Classifier	precision	recall	F1-score	Accuracy
BERT	Logistic Regression	0.74	0.79	0.77	0.91
DistilBERT	Logistic Regression	0.80	0.78	0.79	0.92
XLNet	Logistic Regression	0.61	0.64	0.62	0.85
BioBERT	Logistic Regression	0.85	0.79	**0.82**	**0.93**
SciBERT	Logistic Regression	0.70	0.74	0.72	0.89
BlueBERT	Logistic Regression	0.81	0.72	0.76	0.91
BlueBERTMIMIC	Logistic Regression	0.78	0.72	0.75	0.91

5.2 Deep learning. Study 3, Experiments 2

In Experiment 2, we add to transformer vectors the syntactic and semantic features from Study 1 and Study 2. As before, the data was split 75% for training and 25% for testing. We use the transformer models as the source of vectors, and each vector consists of 768 numerical features. To these vectors we add the features from Study 1 and Study 2, and then apply logistic regression as the learning mechanism. We can view this experiment as another case of domain adaptation, as the new feature set adds UMLS concepts, and the distribution of common features is different as well. Also, BioBERT was not trained on identifying conditional sentences, so the task is new as well.

The results are better than in the earlier experiments in Study 2, reported in Tables 5 and 6. However, we can see from Tables 9 and 10 that these additional features *decrease* the performance of BioBERT (perhaps because some of them are implicitly encoded in BioBERT vectors). Interestingly enough, they improve the performance of other models, even though they are known to also encode syntactic and semantic information (as shown in previously cited [48, 49, 50, 51]).

Table 9 This table illustrates the classification results on identifying condition-action statements (type CA) using different features (Study 3, Experiment 2). In this table, we only report results from the Logistic Regression classifier

Model	transformer vectors		adding features from Study 1		adding features from Study 2	
	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy
BERT	0.74	0.92	0.81	0.94	0.82	0.94
DistilBERT	0.80	0.93	0.74	0.92	0.80	0.94
XLNet	0.58	0.87	0.67	0.90	0.67	0.90
BioBERT	**0.85**	**0.95**	0.79	0.93	0.81	0.94
SciBERT	0.73	0.91	0.78	0.93	0.76	0.93
BlueBERT	0.77	0.93	0.76	0.93	0.79	0.94
BlueBERTMIMIC	0.73	0.91	0.76	0.93	0.80	0.94

Table 10 This table illustrates the classification results on identifying conditional statements (type CCA) using different features (Study 3, Experiment 2). In this table, we only report results from the Logistic Regression classifier

Model	Study 3		features from Study 1 and Study 3		features from Study 2 and Study 3	
	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy
BERT	0.77	0.91	0.77	0.91	0.81	0.93
DistilBERT	0.79	0.92	0.77	0.92	0.78	0.92
XLNet	0.62	0.85	0.63	0.85	0.66	0.70
BioBERT	**0.82**	**0.93**	0.80	0.92	0.80	0.93
SciBERT	0.72	0.89	0.77	0.91	0.79	0.92
BlueBERT	0.76	0.91	0.79	0.92	0.81	0.93
BlueBERTMIMIC	0.75	0.91	0.71	0.90	0.73	0.90

5.3 Extracting conditional statements using transfer learning

Based on the 2015 survey of the topic [33], we defined *transfer learning* as focused on cases where the target domain's feature space is different from the source feature space or spaces.

In Study 3, Experiment 3 tests the applicability of transfer learning by using the rhinosinusitis and hypertension guidelines for training, and the asthma guidelines for testing. As observed earlier in Section 3.2, the training and testing data sets have different vocabularies, very different distributions (established by Kolmogorov-Smirnov test, and by K-L divergence), and even very different distributions on the common vocabulary as shown in Figure 1, and confirmed by the K-S test.

As we can see in Tables 11 and 12, we get results comparable to Study 2, which shows that out of the box transfer learning on unseen documents, and with completely different distribution of features, can perform on the level of classical algorithms trained under the i.i.s. (independent and identically distributed) assumption with 75%-25% train-test split.

Table 11 Study 3. Experiment 3. On the class of conditional sentence (CCA), 72% F1 and 87% accuracy (A) shows applicability of machine learning transfer; it beats results of Study 2 Table 6 of 65%. Syntactic and semantic features from Study 2 were used in the first and third experiment.

Model	Classifier	P	R	F1	A
All Study 2 features	Random Forest	1.0	0.11	0.20	0.72
	Gradient Boosting	0.82	0.34	0.48	0.77
	Logistic Regression on RF&GB	0.85	0.32	0.47	0.77
BioBERT (only)	Logistic Regression	0.85	0.62	**0.72**	**0.87**
	Random Forest	0.56	0.11	0.19	0.74
	Gradient Boosting	0.58	0.47	0.52	0.77
	Logistic Regression on RF&GB	0.56	0.49	0.52	0.76
BioBERT + all Features	Logistic Regression	0.91	0.47	0.62	0.85
	Random Forest	0.75	0.07	0.12	0.75
	Gradient Boosting	0.62	0.44	0.52	0.78
	Logistic Regression on RF&GB	0.64	0.47	0.54	0.79

Table 12 Study 3. Experiment 3. On the class of condition-action (CA) sentences the 67% F1 score shows the applicability of transfer learning to this class, closely matching the 68% F1 score of Table 5 (and comparable accuracy). Syntactic and semantic features from Study 2 were used in first and third experiment.

Model	Classifier	P	R	F1	A
All Study 2 features	Random Forest	1.0	0.03	0.05	0.78
	Gradient Boosting	0.89	0.21	0.34	0.82
	Logistic Regression on RF&GB	0.89	0.21	0.34	0.82
BioBERT (only)	Logistic Regression	0.65	0.68	**0.67**	**0.85**
	Random Forest	0.50	0.29	0.37	0.78
	Gradient Boosting	0.50	0.50	0.50	0.78
	Logistic Regression on RF&GB	0.51	0.50	0.51	0.78
BioBERT+all Features	Logistic Regression	0.71	0.53	0.61	0.85
	Random Forest	0.53	0.21	0.30	0.78
	Gradient Boosting	0.61	0.53	0.56	0.82
	Logistic Regression on RF&GB	0.57	0.42	0.48	0.80

6 Discussion and future work

In this section we summarize our work from the point of view of comparison with prior art, and then discuss possible directions of future work. The overall improvements are summarized in Tables 13 and 14. They show respectively the 22% and 25% gains in F1-measure.

Table 13 This table summarizes the improvements in classification results on identifying condition-action statements

Experiment	Classifier	Features	F1	F1-gain	A	A-gain
Study 1	GB	POS tags	0.63	0	0.84	0
Study 2	R	Semantic + Syntactic	0.68	+5%	0.86	+2%
Study 3 Ex. 1	LR	BioBERT vectors	0.85	+17%	0.95	+9%

Table 14 This table summarizes the improvements in classification results in identifying conditional statements

Experiment	Classifier	Features	F1	F1-gain	A	A-gain
Study 1	GB	POS tags	0.57	0	0.84	0
Study 2	R	Semantic + Syntactic	0.65	8%	0.87	3%
Study 3 Ex. 1	LR	BioBERT vectors	0.82	17%	0.93	6%

6.1 Comparison with the prior art

Our results show that significant improvements of the prior art are possible using domain adaptation and transfer methods. We start with the pioneering work of Wenzina and Kaiser [1], who proposed a heuristic-based information extraction method for identifying condition-action statements.

The authors calculate a score for statements based on the appearances of trigger words ("if" and "should") and sequences of semantic types from UMLS. They achieved a recall of 75%, a precision of 88%, and an 81% F1-score on the same chapter of asthma guidelines as the one used in our research. Their results only demonstrate recall on activities with specific patterns — the appearance of the trigger words "if" or/and "should."

However, if we consider all activities in their annotated corpus, the recall drops to 56%. Furthermore, we disagree with some of their annotations. We believe there are more condition-action statements in the chapter of asthma guidelines. If we apply their approach to our annotated corpus, which we used in our experiments, their recall will be 39%. In the experiments reported in this article, we achieved precision of 89%, recall of 82%, and an 85% F1 score on identifying condition-action statements.

Hussain et al. [26] used only the hypertension guideline annotations from our gold standard dataset to develop a heuristic model for identifying medical recommendations. In their study, they considered all condition-action, condition-consequence, and action statements as recommendations. They achieved 85.54% accuracy in detecting recommendations using 10 heuristic patterns identified manually by authors.

Hussain and Lee [27] proposed two methods to detect *recommendations* from clinical practice guidelines. They were defined in [26, 27] as the combined classes CCA+A, i.e., condition-consequence, condition-action and action.

In their experiment, first they used the TF-IDF vectors of preprocessed sentences as features for machine learning models. Second, they added aspects (UMLS concepts) of the tokens to the sentences and used the TF-IDF vectors of the modified sentences. They trained and evaluated their models on hypertension and rhinosinusitis guideline annotations from our gold standard dataset. They achieved approximately 80% accuracy for the first experiment and 84% accuracy for the second one. Although deep learning was used as a part of an ensemble learning model [27], it was the weakest overall performer.

In contrast, using the transfer method described earlier, with BioBERT vectors as features for a logistic regression classifier, we achieved, as shown in Table 15, a 91

% accuracy in detecting *recommendations*. They are defined in [26, 27] as consisting of the combined classes CCA+A, i.e., condition-consequence, condition-action and action. However, we should note that accuracy is perhaps a less informative measure than F1 in cases of imbalanced classes. For example, in the top row of Table 7, we see the 79% accuracy of random forest with an abysmal 5% F1 score.

Table 15 This table illustrates the classification results on identifying recommendations, defined in [27] and [26], as CCA+A. This experiment uses as features the embeddings from the transformer models, as previously shown in Study 3, Example 1.

Model & Classifier	data	features	precision	recall	F1-score	Accuracy
Logistic Regression	Hypertension	BioBERT vectors	0.94	0.75	0.83	0.91
Logistic Regression	Hypertension & rhinosinusitus	BioBERT vectors	0.77	0.65	0.71	0.87
Logistic Regression	Hypertension & rhinosinusitus	BlueBERT vectors	0.88	0.64	0.74	0.89
Logistic Regression	Hypertension & rhinosinusitus & Asthma	BioBERT vectors	0.79	0.73	0.76	0.90
Logistic Regression	Hypertension & rhinosinusitus & Asthma	SciBERT vectors	0.81	0.75	0.78	0.91
Heuristic model [26]	Hypertension	heuristic patterns	-	-	-	0.86
Ensamble learner[27]	Hypertension	TF-IDF from sentences	-	-	-	0.80
Ensamble learner [27]	Hypertension & rhinosinusitis	TF-IDF from sentences and concepts	-	-	-	0.84

6.2 Open issues and future work

The open issues and possible extensions of this work can go in several directions. The most obvious next step, after identifying conditional sentences, is to extract the specifics: conditions, actions and consequences. A discourse-oriented direction of analysis would allow us to find conditions, actions and consequences spread over paragraphs or sections of texts. Combining discourse analysis with the extraction of specific entities and events should result in improved accuracy of both classification and extraction, and would open the possibility of applications e.g. to analysis of electronic health records (EHR). Finally, creating more annotated guidelines would ameliorate the problem of the paucity of data mentioned in Section 3.

We also would like to note that the detection of sentences with condition-consequence, condition-action etc. is applicable to completely different domains, e.g. to business process management [52]. However, in our discussion we focus on biomedical literature and applications.

Regarding the lack of annotated guidelines, from Experiment 3 in Study 3 (Section 5.3), we see that transfer learning is possible in this space. Among other things, this means that the process of annotating new sets of guidelines could be accelerated by an automated preprocessing step, and then the human annotators would only focus on borderline cases in the spirit of active learning [53].

Fine-grained discourse analysis is a difficult and complex task [54, 55], but one that is necessary. For example, none of the reported experiments can account for the case when condition and action do not appear in the same context, as in the sentence:

The most effective therapy is intranasal steroids.

The same is true for building computational models of medical discourse, where the perspective can range from finding connectives [56, 57] to fine-grained information and event extraction (e.g. [58, 59, 60, 61]).

However, with progress in this space enabled by deep learning, better text analysis and other semantic techniques (e.g. [62, 63, 64, 65, 66, 67], and this article), we can imagine, in future work, that after analyzing condition-action sentences, a system could extract rules for clinical decision support and convert them into a formal representation. This, in turn, would allow a comparison with rules generated from cases using unsupervised or supervised machine learning. It could also produce formal rules which could be directly programmed into clinical decision support systems; such formalisms already exist (e.g. [68, 69]). An automated process, like envisioned here, should make it possible to understand effectiveness of treatments with a high level of granularity, resulting in faster data-driven updates to clinical guidelines and improved treatments.

7 Conclusions

In this article we showed that modern deep learning methods, when applied to the text of clinical guidelines, yield substantial improvements in our ability to find sentences expressing the relations of condition-consequence, condition-action and action.

As shown in a series of experiments, a combination of machine learning domain adaptation and transfer can improve the ability to automatically find conditional sentences in clinical guidelines. We showed substantial improvements over prior art (+5% minimum, +25% maximum), and discussed several directions of extending this work, including addressing the problem of paucity of annotated data.

In summary, we presented three studies using syntactic, semantic and deep learning methods, and performed in-depth evaluation on a set of three annotated medical guidelines. Despite the limitation of having only a small set of annotated data, we showed the applicability of the recently developed techniques, namely neural network transformers and transfer learning to the problem of detection of conditional sentences.

Ethics approval and consent to participate
n/a

Consent for publication
All the material presented in this article was created and is owned by the authors. The authors hereby express their consent for publication.

Availability of data and material
The data is publicly available `https://data.world/hematialam/condition-action-data`; the code will be made public upon the publication.

Funding
This project was not funded by any grant.

Acknowledgements
n/a

Competing interests
The authors declare that they have no competing interests.

Author's contributions
HH and WZ wrote the main manuscript text. The reported experiments were performed by HH. Both authors reviewed the manuscript.

Author details
[1]Department of Computer Science, UNC Charlotte, Charlotte, NC, USA. [2]Department of Computer Science, School of Data Science, UNC Charlotte, Charlotte, NC, USA.

References

1. Wenzina R, Kaiser K. Identifying Condition-Action Sentences Using a Heuristic-based Information Extraction Method. In: Process Support and Knowledge Representation in Health Care. Springer; 2013. p. 26–38.
2. Catillon M. Medical knowledge synthesis: A brief overview. NBER: Cambridge, MA, USA; 2017.
3. CDC. Breast Cancer Screening Guidelines for Women. Centers for Disease Control and Prevention; 2017. Available from: https://www.cdc.gov/cancer/breast/pdf/BreastCancerScreeningGuidelines.pdf.
4. Hematialam H, Garbayo L, Gopalakrishnan S, Zadrozny WW. A Method for Computing Conceptual Distances between Medical Recommendations: Experiments in Modeling Medical Disagreement. Applied Sciences. 2021;11(5):2045.
5. Zadrozny W, Hematialam H, Garbayo L. Towards Semantic Modeling of Contradictions and Disagreements: A Case Study of Medical Guidelines. Proc 12th International Conference on Computational Semantics (IWCS); arXiv preprint arXiv:170800850. 2017;.
6. Bilici E, Despotou G, Arvanitis TN. The use of computer-interpretable clinical guidelines to manage care complexities of patients with multimorbid conditions: A review. Digital health. 2018;4:2055207618804927.
7. Musen MA, Middleton B, Greenes RA. Clinical decision-support systems. In: Biomedical informatics. Springer; 2014. p. 643–674.
8. Hematialam H, Zadrozny W. Identifying Condition-Action Statements in Medical Guidelines Using Domain-Independent Features. arXivorg, http://arxivorg/abs/170604206. 2017 jun;.
9. Sager N. Computerized discovery of semantic word classes in scientific fields. New York University; 1977.
10. Grishman R. Directions in artificial intelligence: Natural language processing. Courant Institute of Mathematical Sciences, New York University; 1975.
11. Zhou M, Duan N, Liu S, Shum HY. Progress in Neural NLP: Modeling, Learning, and Reasoning. Engineering. 2020;6(3):275–290.
12. Smith NA. Contextual word representations: putting words into computers. Communications of the ACM. 2020;63(6):66–74.
13. Seneviratne O, Das AK, Chari S, Agu NN, Rashid SM, Chen CH, et al. Enabling Trust in Clinical Decision Support Recommendations through Semantics. SeWeBMeDa at ISWC; 2019.
14. Chen X, Xie H, Cheng G, Poon LK, Leng M, Wang FL. Trends and Features of the Applications of Natural Language Processing Techniques for Clinical Trials Text Analysis. Applied Sciences. 2020;10(6):2157.
15. Ju M, Nguyen NT, Miwa M, Ananiadou S. An ensemble of neural models for nested adverse drug events and medication extraction with subwords. Journal of the American Medical Informatics Association. 2020;27(1):22–30.
16. Nogales A, García-Tejedor Á, Monge D, Vara JS, Antón C. A survey of deep learning models in medical therapeutic areas. Artificial Intelligence in Medicine. 2021;p. 102020.
17. Nadif M, Role F. Unsupervised and self-supervised deep learning approaches for biomedical text mining. Briefings in Bioinformatics. 2021;22(2):1592–1603.
18. Gatta R, Vallati M, Fernandez-Llatas C, Martinez-Millana A, Orini S, Sacchi L, et al. Clinical Guidelines: A Crossroad of Many Research Areas. Challenges and Opportunities in Process Mining for Healthcare. In: International Conference on Business Process Management. Springer, Cham; 2019. p. 545–556.
19. Gatta R, Vallati M, Fernandez-Llatas C, Martinez-Millana A, Orini S, Sacchi L, et al. What Role Can Process Mining Play in Recurrent Clinical Guidelines Issues? A Position Paper. International Journal of Environmental Research and Public Health. 2020;17(18):6616. Available from: https://www.mdpi.com/1660-4601/17/18/6616.
20. Peleg M, Tu S, Bury J, Ciccarese P, Fox J, Greenes RA, et al. Comparing computer-interpretable guideline models: a case-study approach. Journal of the American Medical Informatics Association. 2003;10(1):52–68.
21. Cunningham H. GATE: A framework and graphical development environment for robust NLP tools and applications. In: Proc. 40th annual meeting of the association for computational linguistics (ACL 2002); 2002. p. 168–175.
22. Schlegel DR, Gordon K, Gaudioso C, Peleg M. Clinical Tractor: A Framework for Automatic Natural Language Understanding of Clinical Practice Guidelines. In: AMIA Annual Symposium Proceedings. vol. 2019. American Medical Informatics Association; 2019. p. 784.
23. Serban R, ten Teije A, van Harmelen F, Marcos M, Polo-Conde C. Extraction and use of linguistic patterns for modelling medical guidelines. Artificial intelligence in medicine. 2007;39(2):137–149.
24. Kalyan KS, Sangeetha S. SECNLP: A survey of embeddings in clinical natural language processing. Journal of Biomedical Informatics. 2020;101:103323.
25. McCray AT. The UMLS Semantic Network. In: Proceedings/the... Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care. American Medical Informatics Association; 1989. p. 503–507.
26. Hussain M, Hussain J, Sadiq M, Hassan AU, Lee S. Recommendation Statements Identification in Clinical Practice Guidelines Using Heuristic Patterns. In: 2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD). IEEE; 2018. p. 152–156.
27. Hussain M, Lee S. Information Extraction from Clinical Practice Guidelines: A Step Towards Guidelines Adherence. In: International Conference on Ubiquitous Information Management and Communication. Springer; 2019. p. 1029–1036.
28. Zhang A, Lipton ZC, Li M, Smola AJ. Dive into Deep Learning. 2020. URL https://d2l ai. 2020;.
29. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Nets. In: Advances in Neural Information Processing Systems 27 (NIPS 2014); 2014. p. 2672–2680.
30. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).

Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171–4186. Available from: https://www.aclweb.org/anthology/N19-1423.
31. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–1240.
32. Pan SJ, Yang Q. A survey on transfer learning. IEEE Transactions on knowledge and data engineering. 2009;22(10):1345–1359.
33. Sun S, Shi H, Wu Y. A survey of multi-source domain adaptation. Information Fusion. 2015;24:84–92.
34. Ramponi A, Plank B. Neural Unsupervised Domain Adaptation in NLP—A Survey. arXiv preprint arXiv:200600632. 2020;.
35. Ruder S. Neural transfer learning for natural language processing. Ph.D. Thesis, NUI Galway; 2019.
36. Ruder S, Peters ME, Swayamdipta S, Wolf T. Transfer learning in natural language processing. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials; 2019. p. 15–18.
37. Laparra E, Bethard S, Miller TA. Rethinking domain adaptation for machine learning over clinical language. JAMIA open. 2020;3(2):146–150.
38. Rosenthal S, Das S, Hsueh PYS, Barker K, Chen CH. Efficient goal attainment and engagement in a care manager system using unstructured notes. JAMIA Open. 2020;3(1):62–69.
39. James PA, S O, BL C, et al. 2014 evidence-based guideline for the management of high blood pressure in adults: Report from the panel members appointed to the eighth joint national committee (jnc 8). JAMA. 2014;311(5):507–520. Available from: +http://dx.doi.org/10.1001/jama.2013.284427.
40. British Thoracic Society, Network SIG, et al. British guideline on the management of asthma. Thorax. 2008;63:iv1.
41. Chow AW, Benninger MS, Brook I, Brozek JL, Goldstein EJ, Hicks LA, et al. IDSA clinical practice guideline for acute bacterial rhinosinusitis in children and adults. Clinical Infectious Diseases. 2012;54(8):e72–e112.
42. Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations; 2014. p. 55–60. Available from: http://www.aclweb.org/anthology/P/P14/P14-5010.
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. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. Xlnet: Generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:190608237. 2019;.
45. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:191001108. 2019;.
46. Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. arXiv preprint arXiv:190310676. 2019;.
47. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. arXiv preprint arXiv:190605474. 2019;.
48. Chi EA, Hewitt J, Manning CD. Finding Universal Grammatical Relations in Multilingual BERT. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020. p. 5564–5577.
49. Jawahar G, Sagot B, Seddah D. What does BERT learn about the structure of language? In: ACL 2019-57th Annual Meeting of the Association for Computational Linguistics; 2019. p. 3651–3657.
50. Rosa R, Mareček D. Inducing syntactic trees from BERT representations. arXiv preprint arXiv:190611511. 2019;.
51. Luo Z. Have Attention Heads in BERT Learned Constituency Grammar? arXiv preprint arXiv:210207926. 2021;.
52. Vo NPA, Manotas I, Popescu O, Cerniauskas A, Sheinin V. Recognizing and Splitting Conditional Sentences for Automation of Business Processes Management. arXivorg 210400660. 2021;.
53. Settles B. Active learning. Synthesis lectures on artificial intelligence and machine learning. 2012;6(1):1–114.
54. Brown G, Brown GD, Brown GR, Gillian B, Yule G. Discourse analysis. Cambridge university press; 1983.
55. Stede M. Discourse processing. Synthesis Lectures on Human Language Technologies. 2011;4(3):1–165.
56. Hahn U, Romacker M. Text structures in medical text processing: empirical evidence and a text understanding prototype. In: Proceedings of the AMIA Annual Fall Symposium. American Medical Informatics Association; 1997. p. 819.
57. Gopalan S, Devi SL. BioDCA Identifier: A System for Automatic Identification of Discourse Connective and Arguments from Biomedical Text. In: Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016); 2016. p. 89–98.
58. Hishiki T, Collier N, Nobata C, Okazaki T, Ogata N, Sekimizu T, et al. Developing NLP tools for genome informatics: An information extraction perspective. Genome Informatics. 1998;9:81–90.
59. Kovačević A, Dehghan A, Filannino M, Keane JA, Nenadic G. Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. Journal of the American Medical Informatics Association. 2013;20(5):859–866.
60. Kilicoglu H, Rosemblat G, Fiszman M, Shin D. Broad-coverage biomedical relation extraction with SemRep. BMC bioinformatics. 2020;21:1–28.
61. Tian Y, Shen W, Song Y, Xia F, He M, Li K. Improving Biomedical Named Entity Recognition with Syntactic Information. BMC Bioinformatics. 2020;21.
62. Kurfalı M. Labeling Explicit Discourse Relations Using Pre-trained Language Models. In: International Conference on Text, Speech, and Dialogue. Springer; 2020. p. 79–86.
63. Kanerva J, Ginter F, Pyysalo S. Dependency parsing of biomedical text with BERT. BMC bioinformatics. 2020;21(23):1–12.
64. Cho H, Lee H. Biomedical named entity recognition using deep neural networks with contextual information. BMC bioinformatics. 2019;20(1):1–11.
65. Hong S, Lee JG. DTranNER: biomedical named entity recognition with deep learning-based label-label

transition model. BMC bioinformatics. 2020;21(1):53.
66. Wang J, Li M, Diao Q, Lin H, Yang Z, Zhang Y. Biomedical document triage using a hierarchical attention-based capsule network. BMC Bioinformatics. 2020;21(13):1–20.
67. Bai T, Gong L, Wang Y, Wang Y, Kulikowski CA, Huang L. A method for exploring implicit concept relatedness in biomedical knowledge network. BMC bioinformatics. 2016;17(9):53–66.
68. Minutolo A, Esposito M, De Pietro G. A pattern-based approach for representing condition-action clinical rules into DSSs. In: Innovations and Advances in Computer, Information, Systems Sciences, and Engineering. Springer; 2013. p. 777–789.
69. Pandey SR, Ma J, Lai CH. A supervised machine learning approach to generate the auto rule for clinical decision support system. Trends in Medicine. 2020;20(3):1–9.

Figures

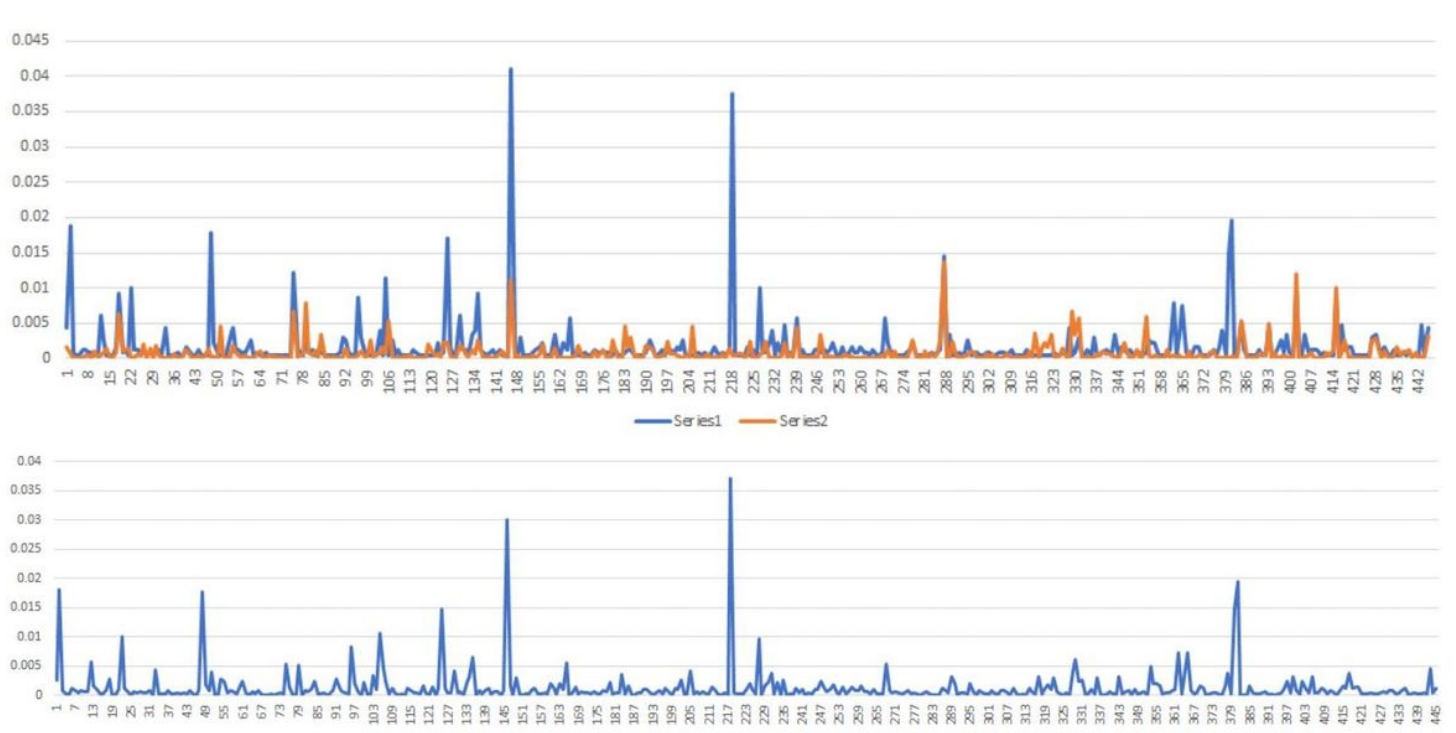

Figure 1

Top panel: the frequency distribution on 445 words-in-common in the training data (hypertension+rhinosinusitis; Series 1) and test data (asthma; Series 2). The bottom panel shows the (large) difference between the two distributions.

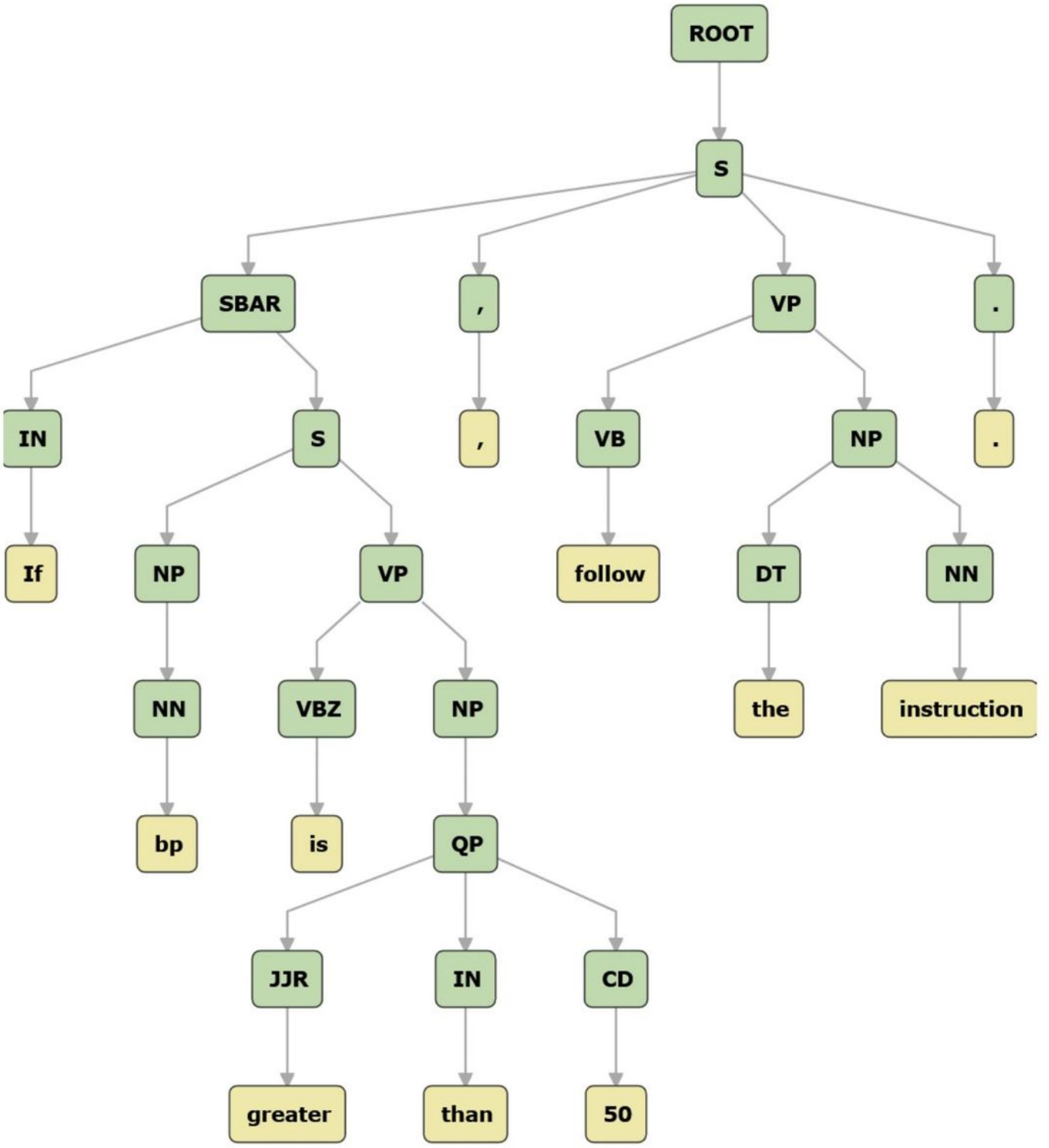

Figure 2

Example of the parse tree showing the part-of-speech (POS) features used in Study 1, and syntactic structures used in Study 2.