Distant Supervision for Treatment Relation Extraction by Leveraging MeSH Subheadings

Tung Tran, M.S[a], Ramakanth Kavuluru, Ph.D[a,b]

[a]*Department of Computer Science, University of Kentucky*
[b]*Division of Biomedical Informatics, Department of Internal Medicine, University of Kentucky*

Abstract

The growing body of knowledge in biomedicine is too vast for human consumption. Hence there is a need for automated systems able to navigate and distill the emerging wealth of information. One fundamental task to that end is *relation extraction*, whereby linguistic expressions of semantic relationships between biomedical entities are recognized and extracted. In this study, we propose a novel distant supervision approach for relation extraction of binary *treatment* relationships such that high quality positive/negative training examples are generated from PubMed abstracts by leveraging associated MeSH subheadings. The quality of generated examples is assessed based on the quality of supervised models they induce; that is, the mean performance of trained models (derived via bootstrapped ensembling) on a gold standard test set is used as a proxy for data quality. We show that our approach is preferable to traditional distant supervision for *treatment* relations and is closer to human crowd annotations in terms of annotation quality. For treatment relations, our generated training data performs at 81.38%, compared to traditional distant supervision at 64.33% and crowd-sourced annotations at 90.57% on the *model-wide* PR-AUC metric. We also demonstrate that examples generated using our method can be used to augment crowd-sourced datasets. Augmented models improve over non-augmented models by more than two absolute points on the more established F1 metric. We lastly demonstrate that performance can be further improved by implementing a classification loss that is resistant to label noise.

Keywords: relation extraction, medical treatment relation, distant supervision, MeSH subheadings

Email addresses: `tung.tran@uky.edu` (Tung Tran, M.S), `ramakanth.kavuluru@uky.edu` (Ramakanth Kavuluru, Ph.D)

1. Introduction

The growing body of knowledge in the biomedical domain, constituting over 27 million articles indexed by PubMed[1] as of 2018, is too vast for human consumption. These articles span academic journals, books, and other resources covering a wide range of topics including medicine, nursing, dentistry, pharmacy, biology, and healthcare. In order to leverage this wealth of information, there has been intense research focus on creating high precision systems for information retrieval and question-answering. These efforts, under the broad theme of knowledge discovery, rely on being able to intelligently recognize and capture semantic relations as conveyed in natural language — hence the importance of relation extraction systems. In this study, we focus on the *binary* relation extraction of *treatment* relations. The task of binary relation extraction is simple: given some textual input, extract (*subject*, *predicate*, *object*) triples where *subject* and *object* are entities and *predicate* is a class of semantic relation. For example, (*insulin*, *treats*, *diabetes type 1*) is a triple, or *semantic predication*, that can be extracted from the sentence "Insulin is prescribed for the treatment of Diabetes Type 1." The difficult nature of this task becomes obvious when we consider that such relationships can be expressed in a variety of complex yet valid ways.

The *treats* predicate is an important predicate in the medical domain, alongside *causes*, and warrants special attention. In this study, we propose a method to generate quality examples for distantly-supervised learning of treatment relation extraction. The proposed method builds on the concept of distant supervision originally proposed by Mintz et al. [23] — henceforth referred to as *traditional* distant supervision (TDS). TDS considers any pair of entities in the same sentence to be a positive example so long as they participate as part of a known semantic predication in an existing knowledgebase. The proposed distant supervision method, referred to as MeSH Subheadings based Distant Supervision (MSDS), relies on MeSH indexing to approximate concept relationships. Specifically, we look for PubMed abstracts for which there exists both the *Therapeutic Use* and *Therapy* Medical Subject Headings (MeSH) subheadings; these subheadings (also known as qualifiers) inform

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

us respectively that there is an entity corresponding to a drug or physical agent being used *and* that there is also an entity corresponding to a disease for which a therapy is specified.

Although MeSH indexing does not provide explicit concept linkage, the intuition is that articles with both *Therapy* and *Therapeutic Use* subheadings are more likely to convey treatment than articles with only one of them or without any of them. Furthermore, we use the *headings* (also known as descriptors) associated with these *subheadings* (also known as qualifiers) as a concept-level filter when considering candidate entity pairs given all mentioned entities identified by NLM's MetaMap [3] concept identification and mapping tool. MSDS relies on the fact that each MeSH term with a heading and subheading is associated with a descriptor unique identifier (DUI) and a qualifier unique identifier (QUI) respectively. For example, the MeSH term "Type 1 Diabetes Mellitus/drug therapy" is annotated with a DUI of D003922 representing Diabetes Mellitus Type 1 and QUI of Q000188 representing *Drug Therapy*. DUIs can be mapped to Concept Unique Identifiers (CUIs) by referencing UMLS Metathesaurus and therefore matched to concept mentions identified in the article by MetaMap. If the MeSH term "Insulin/administration & dosage" is incidentally also indexed for the same article, it is reasonable to infer that the article discusses the treatment of diabetes type 1 through insulin administration. Therefore, sentences in the article where both entities occur together are considered positive examples for treatment.

An important appeal of distant supervision is the fact that there are no costs in terms of money and labor. Compared to human-annotated datasets, quality is usually compromised for quantity. MSDS is capable of generating an abundance of training data without compromising as much on quality when compared with TDS. In fact, the quality of examples extracted by MSDS is closer to human crowd-sourced annotations than TDS. We demonstrate this by comparing models trained using data generated by MSDS to models trained on TDS; the models are evaluated using an adjudicated "gold standard" dataset curated by Dumitrache et al. [7]. Moreover, we demonstrate that examples obtained using MSDS can be used to augment crowd-sourced data for improved performance at no additional cost in human annotations. Lastly, we show that using a modified loss function resistant to noisy labels can improve the performance of models trained on data generated by our method. As presented, MSDS is limited to treatment-type relations while TDS is readably generalizable to other relation types; we discuss this limitation and ways of extending MSDS to other major relation types including *causes*, *prevents*, and *diagnose* in Section 5.

2. Background

Abacha and Zweigenbaum [1] introduced the MeTAE (Medical Texts Annotation and Exploration) platform that allows for the extraction and annotation of medical entities and relationships from medical text. The rule-based method begins with seed concept pairs that are linked by the “may treat” relation according to the UMLS Metathesaurus. The pair consists of a “problem” concept and a “treatment” concept. For each concept pair, very focused queries are submitted to the PubMed Central database[2]; these queries target articles where the *problem* concept exists as the major-focused heading of a Therapy (TH) subheading and the *treatment* concept exists as an unbounded MeSH heading. The queries are designed according to the following pattern: “⟨problem⟩TH[MAJR] and ⟨treatment⟩/MH”[3]. The text of the returned articles are sentence-segmented and each constituent sentence is sent to the MetaMap tool [3] for concept identification and mapping. Only sentences containing both concepts of the seed pair are kept for further pattern construction in the form of regular expressions. MSDS is similar to MeTAE in that both are designed to exploit the MeSH indexing of PubMed to pinpoint pairs of entities that are related by a *treatment* semantic relation; however, there are major differences in terms of both motivation and execution. We highlight the differences between our method and MeTAE as follows.

- MeTAE leverages MeSH subheadings to curate sentences for manual *pattern* construction; the hand-crafted patterns are later used for rule-based relation extraction. On the other hand, MSDS leverages MeSH subheadings to automatically generate distantly-supervised examples, without human intervention, for the purpose of training supervised relation extraction models.

- MSDS utilizes MeSH subheadings in a more precise manner in that we leverage not only the *Drug Therapy* subheading but also the *Therapeutic Use* subheading, the lat-

[2] https://www.ncbi.nlm.nih.gov/pmc/

[3] Note that ⟨problem⟩ and ⟨treatment⟩ are placeholders for the queried concept pair

ter of which allows us to better filter for entities representing a treatment or drug. To show why this is advantageous, we offer the following example. Suppose we want to extract positive treatment examples from the sentence "A 15-year-old female adolescent developed **drug hypersensitivity syndrome** 4 weeks after starting **minocycline** therapy for **acne vulgaris**." Also, suppose the related abstract contains MeSH terms including "Acne Vulgaris/drug therapy", "Minocycline/therapeutic use", and "Drug Hypersensitivity/etiology." If entities are only bounded by the *Drug Therapy* subheading, the system would thus extract triples (**minocycline**, *treats*, textbfacne vulgaris) and (**minocycline**, *treats*, **drug hypersensitivity syndrome**), the latter of which is not only a negative example of *treats* but in fact an example for the opposing *causes* relation (as hinted by the "etiology" subheading). Given MSDS is bounded on both the *Drug Therapy* and *Therapeutic* Usage subheading, we would correctly ignore (**minocycline**, *treats*, **drug hypersensitivity syndrome**) as a positive example for *treats*. Consequently, we reduce the possibility of introducing training examples to the supervised model that are not only incorrect but actually contradictive.

The remainder of this section is organized as follows. Section 2.1 provides a background on the CrowdTruth method and corresponding crowd-sourced dataset. Section 2.2 serves as an overview of deep learning architectures while Section 2.3 discusses relation extraction techniques suited for the biomedical domain. In Section 2.4, we discuss an advanced approach for learning on noisy labels.

2.1. CrowdTruth

Dumitrache et al. [7] showed that, at least in the medical domain, crowd-sourced annotations are of similar or better quality when compared with expert annotations. A method was proposed, referred to as CrowdTruth, to obtain a *sentence-relation* score in $[0, 1]$ by measuring disagreement between multiple crowd-sourced annotations; this score, when thresholded, can be used to determine whether an example (that is, a subject/object pair and its textual context) is positive or negative with respect to a particular type of relation. The dataset used in experiments consisted of 3,984 sentences from PubMed that were originally collected by Wang and Fan [41] and re-annotated via the CrowdTruth method. Herein, we refer to

Annotation	Positives	Negatives	Total
TDS	683	2695	3378
CROWD	1127	2251	3378
GOLD	291	315	606

Table 1: Counts of positive/negative examples for each set of annotations for *treats*

the aforementioned dataset as the *CrowdTruth dataset.* Dumitrache et al. [7] demonstrated that, with enough crowd-sourced annotations for a particular example (specifically 15), the quality is on-par or better compared with using a single expert annotator — at least for the *treats* and *causes* predicates. This is accomplished by comparing the performance of models trained on different methods of annotation and evaluated on a common held-out adjudicated subset amounting to 975 sentences with "gold standard" annotations.

While the CrowdTruth dataset covers *causes* and *treats*, we focus specifically on *treats* for which there are 606 sentences with adjudicated "gold standard" annotations (simply referred to as *GOLD* labels) that can be used as a basis for direct model comparison. The remaining 3378 sentences are annotated with *TDS* and *CROWD* labels; the former refers to labels obtained via TDS while the latter refers to the crowd-sourced annotations obtained via CrowdTruth. The *GOLD* labels are well-balanced such that there is approximately a 1:1 positive-negative ratio, while *TDS* and *CROWD* exhibit more imbalanced ratios of 1:4 and 1:2 respectively. The exact distributions are recorded in Table 1. It is noted that while the same dataset is used, our experimental results are not directly comparable to those in the original study [7] given that the authors conducted experiments using 5-fold cross-validation over the test partition; that is, examples not in a test fold are used for training with the corresponding *TDS* or *CROWD* labels. In this study, we reserve the 606 sentences with gold annotations strictly for testing while opting for a *bootstrapped model averaging* setup (more in Section 4.1) as in a prior work [17] to obtain mean model performance.

2.2. Deep Learning and Bi-directional LSTMs

The recent state-of-the-art performance in natural language tasks such as text classification, relation extraction, named entity recognition, and machine translation are typically achieved with deep learning approaches such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs) [4, 5, 15, 17, 38]. These *deep* learning models are neural networks designed with many hidden layers that compose meaningful intermediate representations. High performance is additionally owed to use of neural word embeddings [19, 38, 40]. RNNs are particularly adept at modeling sequences which makes them suitable for natural language tasks. Long Short-Term Memory (LSTM) [11, 14] networks in particular are a type of RNN that feature a complex mechanism for memory management such that it is able to overcome issues such as the *vanishing gradients* [33] problem. We encourage readers to refer to Graves [13, Chapter 4] and Goldberg [12, Section 11] for thorough details of LSTMs and the corresponding derivations of gradients. Regular LSTMs model sequences accumulating at the last element, while bi-directional LSTMs (BiLSTMs) model a sequence jointly from both directions. The latter architecture has been shown to perform competitively especially for relation extraction tasks. With this model, words are fed as input to the network in the form of word embedding vectors. These word vectors are processed by a bi-directional LSTM, the output of which is max-pooled over the time-step dimension to produce a final feature vector. The feature vector is fully-connected to a softmax output layer with two units corresponding to a binary Yes/No output indicating whether or not there is a *treats* relation being conveyed. The additional use of *position vectors* may further enhance the performance of relation extraction models in our experience. These are learnable embedding vectors that represent the offset of a word to either entity.

2.3. Relation Extraction in the Biomedical Domain

Relation extraction approaches in the biomedical domain typically operate by exploiting the shortest dependency path between candidate entities according to a preprocessed dependency parse tree [2, 10, 20, 21, 37]. The concept of network centrality has also been explored [32] while other studies, including Frunza et al. [9], rely on more traditional linear methods that focus on syntactic and lexical features. More recent studies on relation extraction approaches are based on exploring meaningful deep learning architectures [17, 21, 36], including Segment-CNNs [22] and Graph-LSTMs [35]. Meanwhile, there is ongoing research effort to explore *end-to-end* relation extraction by jointly modeling entity recognition and relation detection to exploit inter-task correlations [16, 24, 42].

2.4. Dealing with Noisy Labels

A short-coming of utilizing annotations from distant supervision is noise arising from erroneously including examples as positive cases even when there is no relationship conveyed. Conversely, there will be examples that are included as negative cases even when there is evidence to the contrary. These sources of noise are a primary contributing factor to the quality of the training data and derived models. Supervised learning on data with noisy labels has been studied extensively [8]. One popular technique is to modify the loss function such that it is more robust to noisy labels [25, 34]. Natarajan et al. [25] established that it is possible to modify the original loss ℓ to a noise-resistant loss $\hat{\ell}$ such that training with $\hat{\ell}$ on noisy data is equivalent to training with ℓ on clean data — provided the noise rates are known *a priori*. Alternatively, it is possible to directly correct test predictions [39] without changing the architecture. Generally, the former is known as backward correction while the latter is known as forward correction. Patrini et al. [34] applied these ideas to the deep neural network setting and formalized an end-to-end, architecture-independent procedure to effectively train on data with noisy labels. Moreover, they propose a method for approximating the noise rates in case it is not known *a priori*. We discuss the modification of the loss function to deal with noisy labels (referred to as noisy-label loss) in our experiments in Section 4.1.

3. Methodology

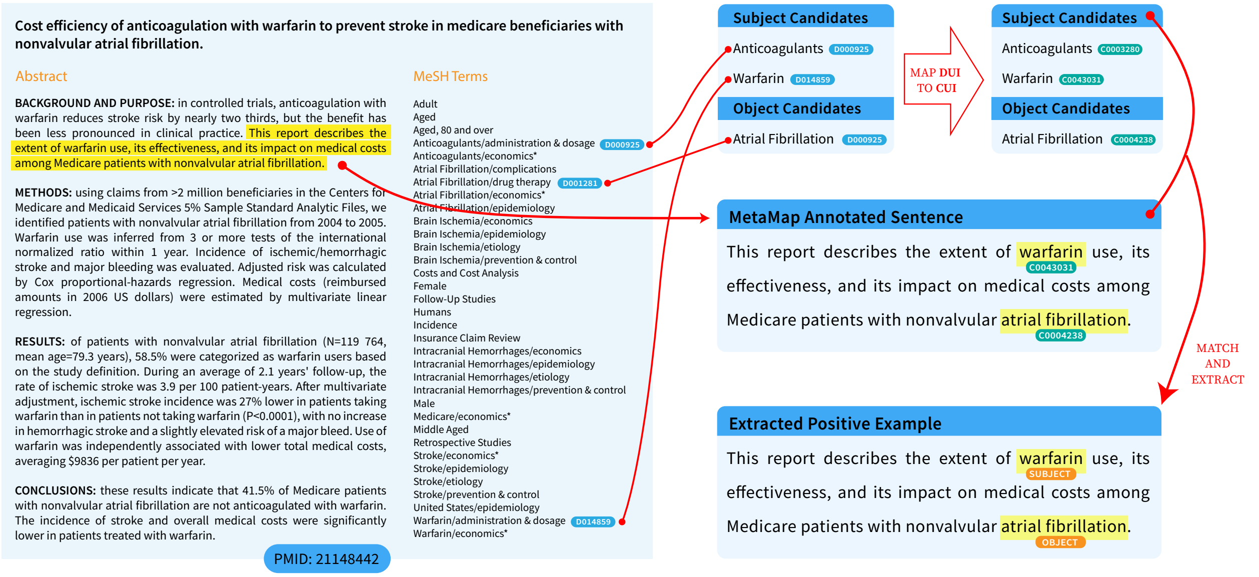
In this section, we formalize MSDS as a method for generating distantly-supervised examples for treatment relation extraction. Section 3.1 describes the article pruning process; not unlike document triage, the goal is to identify articles that contain expressions of treatment relationships and prune articles that do not. In Sections 3.2 and 3.3, we describe the formal process for generating positive and negative examples respectively.

3.1. Article Pruning

MSDS processes articles according to a list of PubMed Identifiers (PMIDs) and generates examples in the sequential order that they appear. We specifically extract examples from the title and the abstract (in that order) of the articles associated with each PMID. Clearly, it is prudent to avoid processing the entire PubMed database as this would be very time consuming with little return because the vast majority of articles do not pertain to medical treatment. It is ideal to target only the subset of articles that are very likely discussing illnesses and related therapeutic agents or procedures. To that end, a list of candidate PMIDs is generated by performing a Boolean search on PubMed with the following query: "therapy[sh] AND (therapeutic use[sh:noexp] OR administration and dosage[sh])". This query returns a list of approximately 2.18 million articles that contain both the *Therapy* and *Therapeutic Use* subheadings. As an aside, MeSH subheadings take upon a tree structure such that a parent subheading is only applied if the article does not fit into one of the more specific child subheadings. Searching for a particular subheading by default also includes articles with its child subheadings. In the case of the *Therapy* subheading, by excluding the *noexp* (no expansion) option, we consequently allow for all articles with child subheadings to be included; specifically, we allow for *Diet Therapy*, *Drug Therapy*, *Nursing*, *Prevention & Control*, *Radiotherapy*, *Rehabilitation*, *Surgery*, and *Transplantation*. In the case of *Therapeutic Use*, there are three child subheadings to consider: *Administration & Dosage*, *Adverse Effects*, and *Poisoning*. Here, we allow for *Administration & Dosage* while disallowing *Adverse Effects* and *Poisoning* which would otherwise be counterproductive with respect to the original objective. Henceforth, when mentioning *Therapy* or *Therapeutic Use*, we implicitly refer to the subheading itself and all child subheadings except the ones deemed "counterproductive."

3.2. Positive Examples

Before processing each article, we randomly shuffle the list of PMIDs according to a seed value. This allows us to extract examples uniformly such that the distribution of the resulting data is not biased toward any subheadings or publication period. We re-use the same seed value to naturally generate negative examples via the procedure described later



Cost efficiency of anticoagulation with warfarin to prevent stroke in medicare beneficiaries with nonvalvular atrial fibrillation.

Abstract

BACKGROUND AND PURPOSE: in controlled trials, anticoagulation with warfarin reduces stroke risk by nearly two thirds, but the benefit has been less pronounced in clinical practice. This report describes the extent of warfarin use, its effectiveness, and its impact on medical costs among Medicare patients with nonvalvular atrial fibrillation.

METHODS: using claims from >2 million beneficiaries in the Centers for Medicare and Medicaid Services 5% Sample Standard Analytic Files, we identified patients with nonvalvular atrial fibrillation from 2004 to 2005. Warfarin use was inferred from 3 or more tests of the international normalized ratio within 1 year. Incidence of ischemic/hemorrhagic stroke and major bleeding was evaluated. Adjusted risk was calculated by Cox proportional-hazards regression. Medical costs (reimbursed amounts in 2006 US dollars) were estimated by multivariate linear regression.

RESULTS: of patients with nonvalvular atrial fibrillation (N=119 764, mean age=79.3 years), 58.5% were categorized as warfarin users based on the study definition. During an average of 2.1 years' follow-up, the rate of ischemic stroke was 3.9 per 100 patient-years. After multivariate adjustment, ischemic stroke incidence was 27% lower in patients taking warfarin than in patients not taking warfarin (P<0.0001), with no increase in hemorrhagic stroke and a slightly elevated risk of a major bleed. Use of warfarin was independently associated with lower total medical costs, averaging $9836 per patient per year.

CONCLUSIONS: these results indicate that 41.5% of Medicare patients with nonvalvular atrial fibrillation are not anticoagulated with warfarin. The incidence of stroke and overall medical costs were significantly lower in patients treated with warfarin.

PMID: 21148442

MeSH Terms

Adult
Aged
Aged, 80 and over
Anticoagulants/administration & dosage D000925
Anticoagulants/economics*
Atrial Fibrillation/complications
Atrial Fibrillation/drug therapy D001281
Atrial Fibrillation/economics*
Atrial Fibrillation/epidemiology
Brain Ischemia/economics
Brain Ischemia/epidemiology
Brain Ischemia/etiology
Brain Ischemia/prevention & control
Costs and Cost Analysis
Female
Follow-Up Studies
Humans
Incidence
Insurance Claim Review
Intracranial Hemorrhages/economics
Intracranial Hemorrhages/epidemiology
Intracranial Hemorrhages/etiology
Intracranial Hemorrhages/prevention & control
Male
Medicare/economics*
Middle Aged
Retrospective Studies
Stroke/economics*
Stroke/epidemiology
Stroke/etiology
Stroke/prevention & control
United States/epidemiology
Warfarin/administration & dosage D014859
Warfarin/economics*

This report describes the extent of warfarin use, its effectiveness, and its impact on medical costs among Medicare patients with nonvalvular atrial fibrillation.

This report describes the extent of warfarin use, its effectiveness, and its impact on medical costs among Medicare patients with nonvalvular atrial fibrillation.

Figure 1: A simple example illustrating how a positive instance for the *treats* relation is extracted from a sentence appearing in the PubMed abstract with PMID 21148442

in Section 3.3. In order to generate positive examples, we apply the following procedure to each article in the shuffled list denoted as a sequence $x^1, \ldots, x^{\mathbf{n}}$.

Allowable subject and object entities for x^i are identified by gleaming the document's MeSH terms. Recall from Section 1 that a heading/descriptor is associated with a DUI and a subheading/qualifier is associated with a QUI. Generally, the former identifies a concept and the latter describes a qualifier with respect to the concept. Let $\bar{\mathcal{X}}^i$ be the set of concepts (DUIs mapped to CUIs via the UMLS [29] database) associated with the *Therapeutic Use* (Q000627) and *Administration & Dosage* (Q000008) subheadings; $\bar{\mathcal{X}}^i$ intuitively serves as the set of allowable *subject* concepts for the treatment relation. For candidate *object* concepts, we denote $\bar{\mathcal{Y}}^i$ as the set of all concepts (mapped to CUIs) associated with the following subheadings: *Therapy* (Q000628), *Diet Therapy* (Q000178), *Drug Therapy* (Q000188), *Nursing* (Q000451), *Radiotherapy* (Q000532), *Rehabilitation* (Q000534), *Surgery* (Q000601), and *Transplantation* (Q000637). *Prevention & Control* are excluded as they would be more appropriate for the *prevents* predicate despite the potential for overlap.

Next, we perform sentence splitting on x^i to obtain a list of $\mathbf{m}^i$ sentences $s^i_1, \ldots, s^i_{\mathbf{m}^i}$. This step is important as we are only concerned with intra-sentence relations. Let $\mathcal{C}$ be the set of all possible CUIs. We apply the following procedure for each sentence s^i_j:

1. The MetaMap tool is used to annotate s_j^i with a set of concepts $\mathcal{M}_j^i$; here, each concept is defined as a triple (α, β, c) where $\alpha \in \mathbb{N}$ is the beginning character offset, $\beta \in \mathbb{N}$ is the ending character offset, and $c \in \mathcal{C}$ is the CUI of the concept. When utilizing MetaMap, we disable the Word Sense Disambiguation (WSD) option[4] and we ignore concepts with mentions that are non-contiguous[5].

2. The set of candidate *subject* concepts $\mathcal{X}_j^i$ for sentence s_j^i is obtained by computing the intersection of concepts identified by MetaMap and allowable *subject* concepts as informed by the MeSH indexing for document i. Formally, we define this set as

   $$\mathcal{X}_j^i = \{(\alpha, \beta, c) : (\alpha, \beta, c) \in \mathcal{M}_j^i \wedge c \in \bar{\mathcal{X}}^i\}\,.$$

   Similarly, $\mathcal{Y}_j^i$ is the list of candidate *object* concepts, defined as

   $$\mathcal{Y}_j^i = \{(\alpha, \beta, c) : (\alpha, \beta, c) \in \mathcal{M}_j^i \wedge c \in \bar{\mathcal{Y}}^i\}\,.$$

3. The set of positive examples $[\mathcal{Z}_+]_j^i$ to be extracted from sentence j of article i is obtained by considering all pairs of candidate *subject* and *object* concepts with non-overlapping mentions. Concretely,

   $$\begin{aligned}[\mathcal{Z}_+]_j^i = \{((\alpha, \beta, c), (\alpha', \beta', c')) :& \\ &(\alpha, \beta, c) \in \mathcal{X}_j^i \quad \wedge \\ &(\alpha', \beta', c') \in \mathcal{Y}_j^i \quad \wedge \\ &((\alpha < \beta < \alpha' < \beta') \vee (\alpha > \beta > \alpha' > \beta'))\}\,.\end{aligned}$$

   Here, each extracted example is a pair of concepts along with their mention offsets.

[4] The WSD option determines the best concept given context if there are multiple potentially valid CUIs for a particular mention. Since we are matching identified concepts directly to MeSH headings, irrelevant CUIs will naturally be ignored and enabling WSD as a premature filtering step will only hurt recall.

[5] To clarify, we ignore mentions that have multiple pairs of starting and ending offsets each corresponding to a different segment of the full mention. In cases where there are multiple contiguous mentions of the same concept, we treat each mention as a separate entity.

The sentence and character offsets of each entity mention are necessarily recorded if they are to serve as training examples. Since it is possible for a single CUI to map to multiple mentions (corresponding to multiple start/end offset pairs), we consider each mention to be a distinct entity so that the examples are consistent with respect to linguistic considerations. Retaining this nuance allows more flexibility in future work which may involve identifying and purging "positive" examples that are not *semantically sound*. For example, consider the sentence, "Blood sugar levels are regulated by the hormone insulin$_1$; man-made insulin$_2$ is used to treat diabetes." Here, both (insulin$_1$,diabetes) and (insulin$_2$,diabetes) will be extracted as positive examples, but we only consider (insulin$_2$,diabetes) to be semantically sound given the linguistic context. Once each sentence is processed and all predications are extracted from the article abstract, we proceed to the next article in the sequence. An example illustrating the procedure is shown in Figure 1.

3.3. Negative Examples

Discriminative models require negative examples in addition to positive examples. Herein, we describe a complementary method to generate negative examples. The process is similar to positive example generation; however, external knowledge is leveraged to ensure that we only extract non-trivial examples and that we only extract examples *likely* to be negative. That is, we wish to avoid extracting an excessive number of false negatives while retaining the more nuanced or borderline cases. The proposed method heavily relies on the UMLS Semantic Network [26] (SemNet) which categorizes concepts and relationships in a hierarchical taxonomy. SemNet assigns broad categories to concepts (that is, CUIs) in the form of Semantic Types [28] (SemTypes). Moreover, the so called Semantic Relations [27] (such as *affects*, *causes*, *uses*, and *treats*) are associated with a set of SemType pairs serving as plausible entity types for that particular relationship. Intuitively, the SemNet constraints for *treats* provides a basis for choosing negative examples that are likewise plausible and therefore more nuanced. A training example in which a drug "treats" another drug is clearly a negative case, but its utility is limited if the goal is to train a robust classifier. Hence, as a rule, only negative examples with subject/object concepts that are consistent with SemNet constraints for the *treats* relation in SemNet are extracted. Let $\mathcal{T}$ denote the set of all

SemTypes defined by UMLS. As a preliminary step, we compute $\mathcal{R} \subseteq \mathcal{T} \times \mathcal{T}$ as the set of subject-object SemType pairs valid for the *treats* relation according to SemNet.

SemMedDB [18, 31], a repository of semantic predications extracted by the rule-based relation extraction tool SemRep [30], serves as another knowledge source that is used to filter out potential false negatives. Specifically, if a semantic predication appears in SemMedDB with the *treats* relation, then it is disregarded it as a candidate negative example. Essentially, recall is sacrificed in order to limit the introduction of false negatives (and therefore noise). With that in mind, we define $\mathcal{S} \subseteq \mathcal{C} \times \mathcal{C}$ as the set of subject-object CUI pairs that appears in SemMedDB at least once for the *treats* relation.

A feature of this approach is that it is capable of naturally generating negative examples alongside positive examples from the same list of abstracts. Consequently, the positive-negative imbalance naturally reflects the imbalance observed in a real-world setting. Here, positive examples are generated first via the method in Section 3.2 and then used as a filter when generating negative examples. Intuitively, a unique predication that has already been extracted as a positive example should not be extracted again as a negative example.

Using the same notations established in Section 3.2, the following procedure is proposed for the extraction of negative examples from sentence s_j^i:

1. As in generating positive examples, s_j^i is annotated with a set of concepts $\hat{\mathcal{M}}_j^i$ using the MetaMap tool. However, WSD is enabled such that each mention is mapped to only one of potentially many concepts. This is necessary as it is no longer possible to rely on MeSH subheadings to inform us of allowable *subject* and *object* concepts.

2. Next, we compute the set of predications based on the following filtering criteria. Each subject-object pair must follow the SemType constraints for *treats*, must not exist as a predication in SemMedDB, and must not exist as a previously-extracted positive example. Moreover, as when extracting positive examples, the entity mentions should likewise not overlap. Formally, the set of negative examples $[\mathcal{Z}_{-}]_j^i$ extracted from article

i, sentence j is defined as

$$\begin{aligned}[\mathcal{Z}_-]^i_j = \{((\alpha, \beta, c), (\alpha', \beta', c')) :& \\ &(\alpha, \beta, c) \in \hat{\mathcal{M}}^i_j \quad \wedge \\ &(\alpha', \beta', c') \in \hat{\mathcal{M}}^i_j \quad \wedge \\ &(\text{type}(c), \text{type}(c')) \in \mathcal{R} \quad \wedge \\ &(c, c') \notin \mathcal{S} \quad \wedge \\ &((\alpha, \beta, c), (\alpha', \beta', c')) \notin [\mathcal{Z}_+]^i_j \quad \wedge \\ &((\alpha < \beta < \alpha' < \beta') \vee (\alpha > \beta > \alpha' > \beta'))\}\end{aligned}$$

where type(c) is the SemType of concept c according to UMLS.

4. Experiments and Results

In order to evaluate the quality of our dataset, we trained supervised models using *TDS* and *CROWD* labels as well as data generated by our distant supervision method (simply referred to as *MSDS*). We observe the performance of trained models as a proxy for data quality as in past work [7]. Of the 3984 examples in the CrowdTruth dataset, 606 held-out sentences with *GOLD* labels are used exclusively for testing. The remaining 3378 examples with *TDS* and *CROWD* labels are used to train models that are evaluated to assess data quality. We provide more detail about our experimental design in Section 4.1 and discuss the corresponding results in Section 4.2.

4.1. Experimental Setup

As deep neural networks are not guaranteed to outperform traditional linear models for this particular task, we consider both traditional and deep neural models in our experiments; hence, we include both a traditional machine learning model (namely, logistic regression) and a deep learning model (namely, the BiLSTM as described in Section 2.2). The BiLSTM model is implemented as described in Kavuluru et al. [17, Section 5B]; however, the number of labels is fixed to 2 as the target task is strictly binary classification. This particular model is suitable as it is designed specifically for the task of relation extraction; with this in mind, the hyper-parameters are mirrored from [17] and fixed across all experiments to ensure a fair comparison. We additionally include a variant of the BiLSTM model with a modified noise-resistant loss function as describe in Section 2.4, referred to as BiLSTM-NLL. The implementation of the noisy-label loss is based on the *backward correction procedure* described in Patrini et al. [34, Section 4.1]. The matrix representing the *approximated* noise rates used in the cited method is computed by following the procedure for noise rate estimation [34, Section 4.3], where the set of "testing" instances for noise approximation is a sample of MSDS-generated data with 3000 examples. Note that regardless of the model, we perform an entity-binding step, wherein mentions of the subjects and objects are replaced with generic SUBJECT and OBJECT tokens respectively, as in prior work [17]. This implies that we are effectively evaluating models based entirely on its ability to learn the linguistic context without regard for subject-object pair correlations.

Bootstrapped Ensembling. The performance of each variant is measured based on the bootstrapped model averaging [17] technique wherein average behavior is studied through building and evaluating large numbers of ensembles. This is motivated by the fact that deep neural networks are trained using stochastic gradient descent; the result is that we will often find parameters corresponding to some "good enough" local minimum as opposed to an optimal global minimum. Different random parameter initializations of the network will converge to different solutions corresponding to these local minima. In order to arrive at a more stable model, it is typical to train a number of such models (each with a different parameter initialization) as part of an ensemble in an effort to improve both stability and accuracy. In a prior work [17], we proposed an experimental setup in which 20 deep neural models were trained for each architecture as part of a sampling pool. 10000 ensembles are assembled and evaluated by randomly sampling 10 models from the pool for each ensemble. With this setup, it is possible to assess mean performance and corresponding confidence intervals such that conclusions are drawn based on statistical significance. We apply the same methodology to assess the average behavior of each variant in this study. Although intended for deep neural networks, we apply bootstrapped model averaging to all models uniformly, including logistic regression, to ensure a fair comparison.

Method	TDS	MSDS	CROWD
Logistic Regression	64.57 ± 0.00646	82.86 ± 0.00889	90.46 ± 0.00187
BiLSTM	63.59 ± 0.02015	81.18 ± 0.01334	90.82 ± 0.00504
BiLSTM-NLL	64.33 ± 0.01708	81.38 ± 0.01371	90.57 ± 0.00441

Table 2: Results comparing quality between traditional distant supervision (*TDS*), Mesh Subheadings based Distant Supervision (*MSDS*), and crowd-sourced (*CROWD*) labels. We report the 95% confidence interval around mean PR-AUC over 10,000 ensembles across linear and deep neural methods.

4.1.1. Assessing MSDS Data Quality

We measure performance as a proxy for label quality using the Precision-Recall Area Under the Curve [6] (PR-AUC) metric instead of the more popular F1 metric. Our rationale for this decision is as follows. The difference in label distribution (even when binary) can serve as a misleading factor when comparing the quality of datasets. A model trained on a training dataset having a similar label distribution to that of the test set is at a significant advantage regardless of the quality of individual examples; this is especially the case when the F1 metric is considered given performance is dependent on predictions made at some probability estimate threshold (typically 50%). A *model-wide* evaluation method such as PR-AUC is more suitable when evaluating with an imbalanced dataset since it is not anchored at a specific threshold. Recall that the *GOLD* labels have a positive-negative ratio of 1:1 while *TDS* and *CROWD* labels have a ratio of 1:4 and 1:2 respectively. Given this imbalance, we measure data quality using PR-AUC as the primary evaluation metric so that imbalance-insensitive comparisons can be made between *CROWD*/*MSDS* and *TDS* labels. We report the results of this experiment in Table 2 for all three methods.

4.1.2. Augmenting Crowd-Sourced Labels with MSDS

In addition to assessing data quality of *MSDS* labels, we also designed an experiment to assess performance gains from augmenting the crowd-sourced examples (expensive to produce) with examples generated via MSDS (free and abundant). Here, we use mean F1 as the primary evaluation metric as we can overcome any imbalance issues by simply

	N	CROWD	1:1 CROWD/MSDS	1:2 CROWD/MSDS	1:3 CROWD/MSDS
Logistic Regression	3000	78.47 ± 0.00779	73.33 ± 0.01306	70.25 ± 0.01469	69.54 ± 0.01790
	6000	-	77.81 ± 0.00971	75.99 ± 0.01326	74.81 ± 0.01559
	9000	-	-	**78.60** ± 0.00781	76.30 ± 0.00905
	12000	-	-	-	77.32 ± 0.00570
BiLSTM	3000	80.84 ± 0.01499	79.22 ± 0.01917	78.08 ± 0.01681	77.08 ± 0.01839
	6000	-	**81.79** ± 0.01361	80.12 ± 0.01269	79.32 ± 0.01529
	9000	-	-	80.56 ± 0.01117	79.56 ± 0.01259
	12000	-	-	-	79.34 ± 0.01263
BiLSTM-NLL	3000	80.86 ± 0.01324	80.82 ± 0.01808	78.64 ± 0.01798	76.32 ± 0.02145
	6000	-	**82.02** ± 0.01572	80.11 ± 0.01109	79.40 ± 0.01447
	9000	-	-	80.83 ± 0.01188	79.75 ± 0.01259
	12000	-	-	-	79.17 ± 0.01181

Table 3: Results showing change in mean F1 with respect to varying training set size and proportion of *CROWD* to *MSDS* examples. We report the 95% confidence interval around mean F1 over 10,000 ensembles across linear and deep neural methods.

generating an *MSDS* based dataset such that there is positive-negative label ratio of 1:2 to match that of the *CROWD* labels. Moreover, we contend that F1 is more important for user-end applications since it is based on evaluating concrete label predictions.

Henceforth, we refer to models trained on data having a shared annotation method (or some combination thereof, more later) as being in the same "class" of models. For example, we refer to models trained on *TDS/CROWD* labels as simply being in the *TDS/CROWD* class of models. For the *MSDS* class of models, we generated three times as many examples as available in the crowd-sourced training set amounting to a total of 10,134 examples with the same label distribution. For the experiment, the exact number of examples generated by MSDS is immaterial as long as it is at least 9000.

To assess gains from augmenting crowd-sourced examples with MSDS, we include the following additional classes of models: *1:1 CROWD/MSDS*, *1:2 CROWD/MSDS*, and *1:3 CROWD/MSDS*. Each of these classes are named based on the ratio of crowd-sourced examples to MSDS-based examples used to train the model. We evaluated each class of models at fixed dataset sizes [6] of $N \in \{3000, 6000, 9000, 12000\}$.The purpose of evaluating at large values of N is to observe the scalability of model performance where crowd-sourced data is augmented with MSDS data. Herein, we refer to a particular N and class of model combination as a "variant". Note that when N is smaller than the total number of examples we have for a particular class of models, we simply sample N random examples from the pool of data available. For example, the models in the pool for *CROWD* at $N = 3000$ will each be trained using a different random sample of the 3378 available. Note that if we have less data than available for some N and some class of models, we ignore that corresponding variant. For example, we can evaluate *CROWD* at 3000 but not 6000, 9000, and 12000 since we only have 3378 examples total. Moreover, we can evaluate *1:1 CROWD/MSDS* at 3000 and 6000 but not 9000 and 12000 for similar reasons. We report these results in Table 3.

4.2. Results and Discussion

Table 2 displays results from our experiments to assess the quality of examples. Here, we observe that *CROWD* and *TDS* achieve roughly 90% and 65% PR-AUC respectively, while *MSDS* achieves a "middle-ground" of 80% PR-AUC across the three methods. Clearly, automatically-curated examples are incapable of competing against human annotations with respect to raw quality. However, we argue that there is value in being able to achieve approximately 80% in mean PR-AUC with *MSDS* when crowd-curated annotations achieve approximately 90% mean PR-AUC. This is especially the case when we consider that *CROWD* sentences and the test set sentences used to evaluate both *CROWD* and *MSDS* examples were collectively obtained via the same curation method (using the same seed articles and relations). Therefore, *CROWD* examples have a natural distributional advantage over *MSDS* examples within our evaluation framework. *MSDS* is therefore closer to *CROWD* in terms of performance compared with *TDS*. These results show that *MSDS*-generated examples are higher in quality compared to those obtained via traditional distant supervision examples and are actually closer in quality to crowd-sourced annotations.

Next, we examine the potential for using MSDS to augment crowd-sourced labels. As

[6] Each example in the dataset constitutes a pair of concepts, the sentence in which the pair is observed, and a Yes/No label indicating whether or not the pair is positive for the *treats* relation.

observed in Table 3, logistic regression achieves 78.47% mean F1 compared to 80.84% mean F1 by BiLSTM on *CROWD* at $N = 3000$. As the 95% confidence intervals do not overlap, the improvements are statistically significant. This indicates that deep learning may be more suitable than logistic regression for this particular task. We note that the deep neural model with noisy-label loss also exhibits better performance than without in most cases when MSDS labels are added. More importantly, we observe that regardless of method – but more so for deep learning models – there is an advantage in augmenting *CROWD* examples with instances generated by MSDS. And we can also observe that more data is not necessarily better, as performance peaks at certain proportions and tend to decrease when more noisy data is added. With BiLSTM, there is an approximate gain of one F1 point arriving at 81.79% when augmenting the 3000 *CROWD* examples with an additional 3000 *MSDS* examples. We see a slightly higher increase at 82.02% for the BiLSTM-NLL model. These improvements are statistically significant at the 95% level based on comparing confidence intervals.

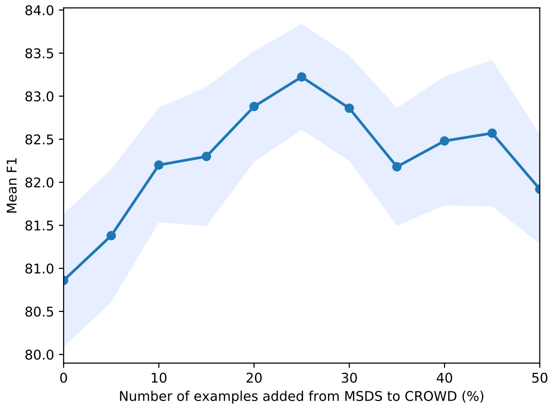


Figure 2: Change in mean F1 w.r.t. amount of examples added from *MSDS* to *CROWD*, indicated as a proportion of *CROWD* dataset size, for the BiLSTM-NLL model.

Based on results from Table 3, augmenting crowd-based examples with high quantities of MSDS is not necessarily better. As such, we perform an auxiliary experiment to determine an optimal balance between *CROWD* and *MSDS* examples, focusing on the case where there are *fewer* *MSDS* examples than *CROWD* examples. We find that by augmenting 3000 *CROWD* examples with only an additional 1500 (+50%) *MSDS* examples, the resulting model achieves 81.92% mean F1 with BiLSTM-NLL. The result is as high as 83.22% mean F1 when evaluating on 3000 *CROWD* examples augmented with 750 (+25%) *MSDS* examples. Based on these results, we can see that there is an optimal ratio of 4:1 between *CROWD* and *MSDS* examples where peak performance is observed. This trend is visualized in Figure 2; we note that the shaded area in this case represents standard deviation.

4.3. Error Analysis

In this section, we perform error analysis by assessing performance based on pairs of subject/object semantic types to identify cases that are difficult for a model trained on MSDS generated examples. In Table 4, we examine evaluation results on the test set by BiLSTM-NLL trained on *MSDS* at $N = 3000$ partitioned by pairs of subject/object SemTypes. We only include results for cases where there are at least 10 examples. An interesting aspect of the test set (with *GOLD* labels) is that there are many SemType pairs for which there are no positive examples; these pairs can be identified by rows where both true positives (TPs) and false negatives (FNs) rates are zero. Of course, this leads to an F1 of 0 which can be misleading at first glance; accuracy is more informative in this case. An example of this phenomenon exists when "Disease or Syndrome" occur as the SemType for both the subject and object — there are 50 such examples in the test set. These are all negative cases for *treats* which is consistent with reality given it is atypical for a disease to treat another disease. This phenomenon is problematic for models trained on MSDS as the corresponding examples are bounded by Semantic Network constraints; hence, there are no MSDS examples generated where both the subject and object are diseases. In other words, MSDS-based models are not trained on trivially negative examples (e.g., a disease treating another disease) and may have difficulties dealing with trivially negative cases at test time. It is possible to overcome this issue by introducing a filtering step in which we predict as negative all test examples that fail to adhere to SemNet constraints. However, this may adversely impact recall given there exists examples of treatment relations (in the wild) where the subject/object do not necessarily adhere to SemNet constraints.

One interesting case stems from the SemType pair "Neoplastic Process" (subject) and "Disease or Syndrome" (object) in which there is exactly one positive case in the groundtruth

Subject SemType	Object SemType	TP	TN	FP	FN	Total	P (%)	R (%)	F (%)	Acc (%)
Organic Chemical	Disease or Syndrome	74	11	10	25	120	88.10	74.75	80.87	70.83
Disease or Syndrome	Disease or Syndrome	0	40	10	0	50	0.00	0.00	0.00	78.43
Amino Acid, Peptide, or Protein	Disease or Syndrome	20	11	2	6	39	90.91	76.92	83.33	79.49
Organic Chemical	Sign or Symptom	9	4	0	9	22	100.00	50.00	66.67	59.09
Neoplastic Process	Finding	0	18	3	0	21	0.00	0.00	0.00	85.71
Neoplastic Process	Disease or Syndrome	0	14	4	1	19	0.00	0.00	0.00	73.68
Organic Chemical	Mental or Behavioral Dysfunction	13	1	0	3	17	100.00	81.25	89.66	82.35
Organic Chemical	Pathologic Function	5	3	0	8	16	100.00	38.46	55.56	50.00
Organic Chemical	Neoplastic Process	11	0	0	3	14	100.00	78.57	88.00	78.57
Neoplastic Process	Pathologic Function	0	8	4	0	12	0.00	0.00	0.00	66.67
Neoplastic Process	Sign or Symptom	0	10	2	0	12	0.00	0.00	0.00	83.33
Bacterium	Disease or Syndrome	0	6	5	0	11	0.00	0.00	0.00	54.55
Pharmacologic Substance	Disease or Syndrome	2	2	0	6	10	100.00	25.00	40.00	40.00

Table 4: Results at N=3000 for BiLSTM-NLL partitioned by subject/object semantic type

corresponding to a single false negative by the MSDS-based model. The example is as follows, with the subject and object underlined: "In patients with rapidly advancing disease characterized by B symptoms, massive lymphadenopathy and hepatosplenomegaly, consider CLL transformation (see disease specific drug treatment in patients with transformed CLL)." CLL as the subject refers to chronic lymphocytic leukaemia. From inspection, the linguistic phrasing in this case is understandably difficult for a machine learning system; the connecting word here is "consider", which is not as strong of an indicator as "treats" or "cures". There are subtle, logical inferences to be made that makes this and similar examples difficult for machine learning models. Moreover, CLL as a neoplastic process is more likely to be the object of a treats relation and the fact that it occurs as the subject in this case could be a puzzling factor. This is a stronger positive example and more semantically consistent if we consider the full mention "CLL transformation" to be the subject; that is, we suspect a minor error in the entity annotation that makes this example particularly difficult.

Another issue that leads to false negatives is the way in which entities are annotated when there are multiple mentions of a unique concept entity in the same sentence. First, we note that relation classification is performed between *mentions* of entities wherein the subject and object entities are bound to generic SUBJECT/OBJECT tokens. Hence, predictions are highly dependent on context. Consider the following sentence as an example, "In the trial based analysis, fondaparinux$_1$ was estimated to prevent 15.1 thromboebolic events per

1000 patients at three months compared with enoxaparin; fondaparinux$_2$ produced cost savings per patient at 30 days, 3 months, and 5 years postdischarge." Mentions of the relevant concepts are underlined. The concept *fondaparinux* has two mentions in the sentence which are discerned via numbered subscripts. The issue, in this case, stems from the fact that the gold treatment relation annotated for this sentence include (fondaparinux$_2$, *treats*, thromboebolic events) instead of (fondaparinux$_1$, *treats*, thromboebolic events). Based on manual examination of the linguistic context, the first mention of the subject (fondaparinux$_1$) is directly involved in a semantic relationship with the object (thromboebolic events), while the second mention of the subject (fondaparinux$_2$) is only involved in the relationship by association with the first mention. Hence, there is a logical inference aspect to the problem arising from the way entities are annotated that is not handled well by the model.

5. Extending to Other Relation Types

Despite its potential, it is important to stress that MSDS as presented is limited to treatment predications while TDS is more readily generalizable to other types of relations. However, we contend that it is possible to extend MSDS to other highly-important, functional relation types in the medical domain. For example, the *Prevention & Control* subheading may be straightforwardly used in place of the *Therapy* subheading, with minimal changes to the proposed method, to extract *prevents* instead of *treats* relations. Moreover, we may consider utilizing the *Etiology* subheading in combination with subheadings including *Methods* and *Complications* to identify the candidate entities for a *causes* relation. As an example, consider an abstract containing the following sentence: "This review provides an up-to-date insight into the aetiology of posterior shoulder dislocations; our results showed that seizures were most commonly implicated." The associated MeSH terms "Seizures/complications" and "Shoulder Dislocation/etiology" can be leveraged to extract an example for the relation triple (*seizures*, *causes*, *shoulder dislocation*). Likewise, the MeSH subheading *Diagnosis* in conjunction with *Methods* may be used to identify examples for the *diagnose* relation type; e.g., the co-occurence of "Behcet Syndrome/diagnosis" and "X-Ray Computed Tomography/methods" may indicate that Behcet's disease is diagnosable by X-Ray. Not only that, it is possible to move beyond *binary* treatment relations given MeSH indexing may include multiple terms with the *Therapeutic Use* subheading for a single article. That is, we can leverage MeSH terms to identify instances of *combination therapies*, wherein the treatment relation involves two or more drugs. However promising, these research avenues require further evaluation and analysis which is left for future work.

6. Conclusion

In this study, we introduced a distant supervision approach for relation extraction of medical treatment predications by exploiting MeSH subheadings. We demonstrated that our distant supervision method is a desirable compromise between traditional distant supervision and crowd-sourced annotations with the advantage that it is of reasonable quality and can be obtained without the costs associated with human involvement. We also showed that it is possible to use data obtained via our proposed method to augment existing crowd-sourced data for performance gains and this can be further improved by using a noise-resistant loss. In future efforts, we anticipate using the proposed distant supervision method to facilitate production of a large, high-quality human-annotated dataset solely for medical treatment relations.

Acknowledgment

We are grateful to the U.S. National Library of Medicine for offering the primary support for this work through grant R21LM012274. We are also thankful for additional support by the National Center for Advancing Translational Sciences through grant UL1TR001998. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

[1] Asma Ben Abacha and Pierre Zweigenbaum. Automatic extraction of semantic relations between medical entities: a rule based approach. *Journal of Biomedical Semantics*, 2(S5):1–11, 11 2011. ISSN 2041-1480.

[2] Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics*, 9(11):S2, 2008.

[3] Alan R. Aronson and François-Michel Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of 3th International Conference on Learning Representations (ICLR)*, 2015.

[5] Jason PC Chiu and Eric Nichols. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370, 2016.

[6] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.

[7] Anca Dumitrache, Lora Aroyo, and Chris Welty. Crowdsourcing ground truth for medical relation extraction. *ACM Transactions on Interactive Intelligent Systems (TiiS) Special Issue on Human-Centered Machine Learning*, 2017.

[8] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2014.

[9] Oana Frunza, Diana Inkpen, and Thomas Tran. A machine learning approach for identifying disease-treatment relations in short texts. *IEEE Transactions on Knowledge and Data Engineering*, 23(6): 801–814, 2011.

[10] Katrin Fundel, Robert Küffner, and Ralf Zimmer. RelEx - relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, 2007.

[11] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with LSTM. *Neural computation*, 12(10):2451–2471, 2000.

[12] Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.

[13] Alex Graves. *Supervised Sequence Labelling with Recurrent Neural Networks*, volume 385 of *Studies in Computational Intelligence*. Springer, 2012. ISBN 978-3-642-24796-5. doi: 10.1007/978-3-642-24797-2.

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[15] Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1700–1709, 2013.

[16] Arzoo Katiyar and Claire Cardie. Going out on a limb: Joint extraction of entity mentions and relations without dependency trees. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 917–928, 2017.

[17] Ramakanth Kavuluru, Anthony Rios, and Tung Tran. Extracting drug-drug interactions with word and character-level recurrent neural networks. In *Fifth IEEE International Conference on Healthcare Informatics (ICHI)*, pages 5–12. IEEE, 2017.

[18] Halil Kilicoglu, Dongwook Shin, Marcelo Fiszman, Graciela Rosemblat, and Thomas C Rindflesch. Semmeddb: a pubmed-scale repository of biomedical semantic predications. *Bioinformatics*, 28(23): 3158–3160, 2012.

[19] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D14-1181`.

[20] Jiexun Li, Zhu Zhang, Xin Li, and Hsinchun Chen. Kernel-based learning for biomedical relation extraction. *Journal of the Association for Information Science and Technology*, 59(5):756–769, 2008.

[21] Shengyu Liu, Kai Chen, Qingcai Chen, and Buzhou Tang. Dependency-based convolutional neural network for drug-drug interaction extraction. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1074–1080. IEEE, 2016.

[22] Yuan Luo, Yu Cheng, Özlem Uzuner, Peter Szolovits, and Justin Starren. Segment convolutional neural networks (Seg-CNNs) for classifying relations in clinical notes. *Journal of the American Medical Informatics Association*, 25(1):93–98, 2017.

[23] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011. Association for Computational Linguistics, 2009.

[24] Makoto Miwa and Mohit Bansal. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1105–1116, 2016.

[25] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in neural information processing systems*, pages 1196–1204, 2013.

[26] National Library of Medicine. The UMLS Semantic Network. https://semanticnetwork.nlm.nih.gov/, 2003.

[27] National Library of Medicine. Current Hierarchy of UMLS Predicates. http://www.nlm.nih.gov/research/umls/META3_current_relations.html, 2003.

[28] National Library of Medicine. Current Hierarchy of UMLS Semantic Types. http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html, 2003.

[29] National Library of Medicine. Unified Medical Language System Reference Manual. http://www.ncbi.nlm.nih.gov/books/NBK9676/, 2009.

[30] National Library of Medicine. SemRep - NLM's Semantic Predication Extraction Program. http://semrep.nlm.nih.gov, 2013.

[31] National Library of Medicine. Semantic MEDLINE Database. http://skr3.nlm.nih.gov/SemMedDB/, 2016.

[32] Arzucan Özgür, Thuy Vu, Güneş Erkan, and Dragomir R Radev. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, 24(13):i277–i285, 2008.

[33] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. *Proceedings of the 30th International Conference on Machine Learning*, 28:1310–1318, 2013.

[34] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1944–1952, July 2017.

[35] Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. Cross-sentence N-ary relation extraction with graph LSTMs. *Transactions of the Association for Computational Linguistics*, 5:101–115, 2017.

[36] Desh Raj, SUNIL SAHU, and Ashish Anand. Learning local and global contexts using a convolutional recurrent network model for relation classification in biomedical text. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 311–321, 2017.

[37] Bryan Rink, Sanda Harabagiu, and Kirk Roberts. Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association*, 18(5):594–600, 2011.

[38] Anthony Rios and Ramakanth Kavuluru. Convolutional neural networks for biomedical text classification: application in indexing biomedical articles. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 258–267. ACM, 2015.

[39] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. In *International Conference on Learning Representations (ICLR) Workshops*, 2015.

[40] Tung Tran and Ramakanth Kavuluru. Predicting mental conditions based on "history of present illness" in psychiatric notes with deep neural networks. *Journal of Biomedical Informatics*, pages S138–S148, 2017.

[41] Chang Wang and James Fan. Medical relation extraction with manifold models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 828–838, 2014.

[42] Suncong Zheng, Yuexing Hao, Dongyuan Lu, Hongyun Bao, Jiaming Xu, Hongwei Hao, and Bo Xu. Joint entity and relation extraction based on a hybrid neural network. *Neurocomputing*, 257:59–66, 2017.