Learning Inter-Sentence, Disorder-Centric, Biomedical Relationships from Medical Literature.

[1]Anton H. van der Vegt, BE, BSc [1]Guido Zuccon, PhD [2]Bevan Koopman, PhD
[1]The University of Queensland, St Lucia, Qld, Australia; [2]CSIRO, Brisbane, QLD, Australia

Abstract

Relationships between disorders and their associated tests, treatments and symptoms underpin essential information needs of clinicians and can support biomedical knowledge bases, information retrieval and ultimately clinical decision support. These relationships exist in the biomedical literature, however they are not directly available and have to be extracted from the text. Existing, automated biomedical relationship extraction methods tend to be narrow in scope, e.g., protein-protein interactions, and pertain to intra-sentence relationships. The proposed approach targets intra and inter-sentence, disorder-centric relationship extraction. It employs an LSTM deep learning model that utilises a novel, sequential feature set, including medical concept embeddings. The LSTM model outperforms rule based and co-occurrence models by at least +78% in F1 score, suggesting that inter-sentence relationships are an important subset of all disorder-centric relations and that our approach shows promise for inter-sentence relationship extraction in this and possibly other domains.

Introduction

Deriving the relationships between a disorder and its related symptoms, tests and treatments is a critical part of medicine. Consider the relationships that exist between symptoms and their associated disorders in Figure 1. Such disorder-centric relationships are necessary for a myriad of medical tasks and research including automated problem lists[1], clinical decision support, medical information retrieval, data mining and knowledge base construction. However, in medical thesauri such as the UMLS and SNOMED CT, such relationships either do not exist, or at least are sporadic. While hand-coding such relationships is prohibitively laborious, it is possible to automatically derive these from free-text using biomedical relationship extraction (BRE) systems. Most systems, however, consider one specific, highly focused relationship set; for example, locations of bacteria[2] or protein-protein interactions[3].Those that are broader in coverage, and can be used for disorder-centric relationship extraction, exhibit poor performance (in terms of precision and recall).

Figure 1: The title (in bold) and first 2 sentences of Medline article PMID=19707415. Medical concepts have been extracted with MetaMap[4] and disorders have been underlined, symptoms have been *italicised*. Medical terms mapped as both disorders and symptoms are both *italicised and underlined* .

> **The old and new therapeutic approaches to the treatment of giardiasis: where are we?**
> Giardia lamblia is the causative agent of giardiasis, one of the most common parasitic infections of the human intestinal tract. This disease most frequently affects children causing *abdominal pain*, *nausea*, *vomiting*, acute or chronic *diarrhea*, and malabsorption syndrome.

Figure 1 also highlights another important issue: relationships span across multiple sentences. Without the capability to detect inter-sentence relationships, all of the symptoms (*nausea*, *vomiting*, etc.) in the second sentence would not be related to the disorder giardiasis in the title and first sentence. Current state-of-the-art relationship extraction systems such as SemRep[5] and PASMED[6] only extract intra-sentence relationships. New methods for inter-sentence relationship extraction are needed.

Extracting inter-sentence, disorder-centric relationships involves a number of challenges[7]. Grammar rules reset between sentences so features, such as shortest dependency path (SDP) analysis, cannot be relied upon. Patterns are harder to identify for inter-sentence, rule-based approaches and the number of possible patterns and/or relationships expands dramatically. To overcome these issues, we recast the inter-sentence relationship extraction task as a sequential labelling problem; a Long Short-Term Memory (LSTM) model is proposed to process a multi-sentence input for a given relationship label occurring across sentence borders. The main contributions of this research are:

1. A deep learning model to perform inter-sentence, disorder-centric biomedical relationship extraction.
2. An empirical evaluation of existing relationship extraction techniques when applied to inter-sentence relationship extraction. In addition, we make available an inter-sentence, disorder-centric, human labelled dataset for training and evaluation purposes, as well as UMLS concept embeddings based on 23 million MEDLINE citations.

Background and Related Work

Extracting medical concepts (or entities) from biomedical free-text has had considerable attention[8]. There are well established tools for mapping free-text to UMLS concepts such as MetaMap[4] and QuickUMLS[9]. While these tools identify concepts (including disorders), they do not identify the important relationships that exist between them.

Clinicians tend to ask a common set of medical questions, with the following 4 accounting for 40% of all questions[10]: (1) How should I treat condition x?; (2) What is the drug of choice to treat condition x?; (3) What is the cause of condition x?; and (4) What test is indicated in situation x? Based on these findings we focus on extracting relationships between: Disorder and treatments (DT); Disorders and symptoms (DS); Disorders and medical tests (DE) and; Disorders and other disorders (DD), as these are often viewed as symptoms for disorders, e.g., hypertension.

We now review previous relationship extraction approaches; these fall into four categories: co-occurrence based, rule based, machine learning based, and more specialised deep learning based.

Co-occurrence based: A common approach is to simply infer a relationship between two concepts if they *co-occurr* within some window of text (e.g., a sentence). Ding et al.[11] used a variety of window sizes — phrase, sentence, sentence pair and entire abstract — to assess the impact on precision and recall. A window size set at the sentence or abstract level provided the best combination of recall and precision; adjacent sentences produced the worst results. However, interestingly, adjacent sentences nearly doubled the number of distinct relationships found and generated high recall, but also a considerable drop in precision. These insights suggested to us that if the target relationships were much broader than biochemical-noun interactions (relationships) and the selection model more discriminating than the simple co-occur model, then inter-sentence BRE could dramatically improve the recall of biomedical relationships, without a corresponding drop in precision.

Rule based methods: incorporate NLP to parse biomedical text and identify specific relationships with either hand-crafted, or automated rules. Rules are created on the basis of syntactic or semantic patterns which describe specific relationship cases; for example, using POS tagging, noun-phrase chunking and dependency parsing to define the rule set[12]. Dependency parse trees enable relationships to be identified where the medical concepts do not appear close together. Other approaches use verbs in the sentence by analysing the phrase-level conjunction to extract relationships[13]. The main disadvantage of rule-based approaches is that they are usually applied to a narrow set of relationships.

We further detail the two benchmark ruled-based approaches — SemRep & PASMED, which are much broader in their relationship extraction. Semrep[5] is a sentence-based relationship extractor identifying predicates from Medline articles; a predicate consists of two UMLS concepts and one of 58 relationships. Applying SemRep to the whole of Medline, the authors released SemMedDB[14]: 93 million relationships extracted from 17.4 million Medline citations. PASMED[6] also constructs a wide-relationship set from Medline, but in addition targets higher recall of *all* relationships. PASMED incorporates predicate-argument-structure (PAS) patterns to detect candidates for relationships. Rather than capture explicit relationship types, PASMED simply retains the actual verb that relates the two concepts. PASMED generated 137 million relationships from MEDLINE. An evaluation on 500 sentences showed PASMED had better recall but worse precision compared to SemRep. SemRep and PASMED are both effective and include a range of disorder-centric relationships; thus both are included as benchmarks in our experimental evaluation.

Machine learning methods are commonly employed to perform relationship extraction with supervised classification solutions, such as support vector machines (SVM) and conditional random fields (CRF). Most machine learning methods require careful feature engineering, with a mix of both domain knowledge and NLP experience. Rink et al.[15] trained an SVM on eight relationship types between concepts within a sentence. While 21 features were used, context features proved to be very important and in particular the concept and concept types. In contrast, Uzuner et al.[16] found that lexical features, in particular the tokens occurring between candidate concepts, were most informative in their SVM approach. Quirk et al.[17]employed a distant supervision ML model to extract relationships across sentence

Table 1: Classification of UMLS concepts into categories for disorder-centric relationship capture.

Category	Defining UMLS Semantic Types (abbreviation)
Disorders	Acquired abnormality (acab), anatomical abnormality (anab), cell or molecular dysfunction (comd), congenital abnormality (cgab), disease or syndrome (dsyn), experimental model or disease (emod), injury or poisoning (inpo), mental or behavioral dysfunction (mobd), neoplastic process (neop), pathalogic function (patf)
Symptoms	Sign or symptom (sosy), Finding (fndg)
Treatments	Health care activity (hlca), therapeutic or preventative procedure (topp), pharmacologic substance (phsu)
Tests	Lab procedure (lbpr), diagnostic procedure (diap)

boundaries, and this model was used with success for extracting drug-gene interactions from biomedical literature. However this model did not incorporate UMLS concept-to-concept relationships and the relationships detected were very niche. Like most machine learning approaches, these methods suffer from complex feature engineering.

Deep learning methods have proven effective across a range of biomedical tasks, including relationship extraction[18]. Deep learning approaches can avoid the laborious feature engineering of the aforementioned machine learning methods. For relationship extraction, a deep auto encoder was used to produce word embeddings of the medical concept word features[18]. These were then fed into a CRF classifier to identify relationships. Results showed only subtle improvements over the standard feature input set. However, we posit that perhaps using a much broader medical concept vector, across the whole of Medline, could provide greater benefit. Outside the medical domain, long short term memory (LSTM) recurrent neural network (RNN) models have proved highly effective on other tasks[19]. This approach incorporated shortest dependency paths (SDP) and linguistic information into a multi-channel RNN. Each SDP for each sentence is sequenced from either direction of the SDP using LSTM units and this is done for 4 different channels of information (word representation, POS tag, grammatical relationship and WordNet hypernyms). It was insightful to note that the word embedding had by far and away the greatest impact on performance, and each other channel added at most 1% to the score or combined, only 1.6%. Li et al.[20] used a similar, bi-directional LSTM model, in the biomedical domain, for the extraction of adverse drug events and bacteria biotopes. Their model also employed SDP information as well as word, POS and character embeddings of the words in the sentence. Gupta et al.[21] also made use of SDPs, however within a deep learning model incorporating bidirectional RNNs to identify relationships across sentences in biomedical text. This model evaluated with the Bacteria Biotope relationship extraction task, showed very promising performance, however the task incorporated very specific relationships (habitats of bacteria), and unlike our proposed approach, medical text rather than medical concepts were the source of relationships. This is an important distinction because concept space disrupts the natural grammar and syntax of sentences, needed for SDP analysis. The use of SDP within LSTM deep learning models is popular, however for inter-sentence relationship extraction, it is not really possible[22]. Kim et al.[7] raised this as a general challenge for participants in the BioNLP shared task in 2009. For this reason, we investigate alternative designs for using the LSTM RNNs to employ for inter-sentence, UMLS concept relationship extraction.

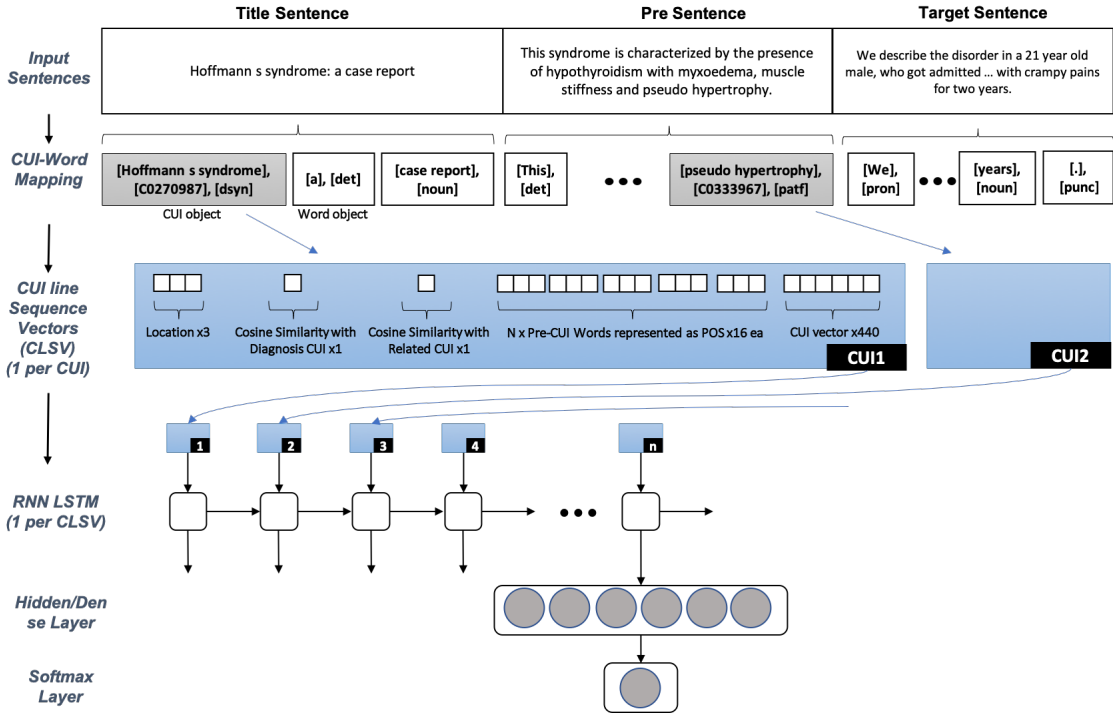
Proposed Disorder-Centric Relation Extraction Model

Typically, deep learning relationship extraction models operate on an intra-sentence basis and therefore the input sequence is limited to, at-most, a sentence of words and often much less if the two entities are closer together. The objective of our approach is to extract relationships within 1 to N sentences, resulting in a significantly larger potential sequence of words between the related entities. To reduce the complexity of this problem, we decided to divide the relationship extraction task into two components: relationship detection and relationship classification. In this way, the deep learning model is employed to detect the presence of a relationship between a disorder and one of treatment, test or symptom concepts, i.e., a binary classification task.

Relationships classification is then done on the basis of the concept types involved in the relationship; for example, if a relationship exists between a symptom and a disorder concept, it is assumed to be a DS relationship. The concept types are based on the UMLS semantic types (semType). SemTypes provide a broad, yet consistent, classification of every UMLS concept. Table 1 defines the semType sets used for classification of concepts into the four types.

Disorders are derived directly from the *Disorders* semGroup excluding symptoms and findings, which are used in the Symptoms concept category. Tests and treatments semTypes are selected using the semMed DB[14] predication table. For test semTypes, we counted the predications found for each semTypes involved in a *diagnoses* relationship and

Figure 2: Pipeline diagram showing both the data preprocessing and basic LSTM network architecture.



selected the top 2. For treatments we counted the predications for semTypes involved in a *TREATS* relationship and selected the top 3 excluding the Amino Acid, Peptide or Protein semType.

Figure 2 describes the information flow and network architecture of the proposed disorder-centric, relationship detection model. Unlike Xu et al.[19] and Mehryary et al.[22], the sequence components in our model are not divided into feature channels, such as POS, words or grammatical relations. Instead, we use UMLS concepts as the basis of each sequence line and can therefore capitalize on UMLS concept embeddings that span the Medline data set. Concepts have a number of advantages over words for use in embeddings. First, the same medical entity expressed in different ways is aggregated into a single concept making for a denser set of relationships between entities. Second, n-grams are also encapsulated within a single concept — especially important in biomedicine where multi-term entities are common and used as a means of distinguishing them; e.g., over 100 are disorders entitled "[something] hypertrophy" (e.g., *pseudo hypertrophy*). We posit that embeddings of real biomedical entities (rather than the words 'pseudo' and 'hypertrophy') are more discriminative. The individual elements of Figure 2 are detailed next.

Input Sentences consist of zero or more title sentences, a target sentence and the sentence before the target (if it exists), called the pre-sentence. The target sentence must contain at least the related concept. The disorder concept may exist in any of the sentences.

CUI-Word Mapping. The input sentences are converted to UMLS concepts using MetaMap (with word sense disambiguation) to extract the best match candidates from each sentence, as well as the POS of each word, and punctuation. Sentences are then converted into a *CUI-Word Mapping* format, which expresses each word or punctuation mark in the sentence as either part of a CUI, together with its semType, or a word, together with its POS.

CUI Line Sequence Vectors (CLSV) are then generated. The lines are constructed starting from the beginning of the title sentence and moving through to the pre-sentence, finishing at the end of the target sentence. Each CUI that is found along the way, forms a CLSV. A CLSV consists of the words leading up to a single CUI and the CUI itself. Therefore, the sequence will consist of n lines where there are n CUIs found across the input sentences. Each CLSV is a concatenation of feature vectors which include:

1. CUI2VEC embedding for the CUI in the CLSV: Two CUI embeddings are tested. The first is the 500 dimensional CUI embedding recently generated by Beam et al.[23] containing 108,477 CUIs, based on 20M clinical notes and 1.7M full text biomedical journal articles. We developed the second using the DL4J word2vec model, which uses Skipgram, CBOW or DBOW feature extraction. The best mapping CUI candidates were extracted from over 23M Medline citations by parsing the 2015 MetaMapped Medline collection. These CUIs were then input to the word2vec model and by applying a window size of 10 and a minimum word frequency of 1 a set of 440-dimensional embeddings were generated for 558,764 CUIs. (500 or 440 dimensions)
2. Location: A one-hot vector representation of the location of the CUI in this CLSV. The locations are within either the title, pre-sentence or target sentence. (3 dimensions.)
3. COSine similarity: Calculated using the CUI embedding vectors above, the cosine similarities between the CUI in the CLSV and the assessed (i) diagnosis CUI and; (ii) related CUI. Note, the CUIs assessed for relationship presence will appear in at least one CLSV each across the sequence of CLSVs. (2 dimensions.)
4. SemType: A one-hot vector representation of the semType of the CUI in this CLSV. (135 dimensions.)
5. POS tag of up to N words preceding the CUI: A one-hot representation of the 11 POS tags plus a separate tag for each of full stops, commas, other punctuation and invalid words. In addition a present/not present flag dimension was used to identify all zero POS embeddings for no-word situations. This occurs frequently where less than N words or punctuation marks appear before the CUI. (N x 16 dimensions)

RNN LSTM model: An LSTM unit was applied to each sequence line; i.e., each CLSV. For the LSTM, the 'no peep-hole' variant[24] was instanced through DL4J. Standard settings include weight initialization, $tanh$ activation function for the LSTM layer and a cross-entropy loss function combined with SoftMax activation for the RNN output layer[25].

Deep Learning Model Tuning & Testing: During this testing phase a number of model and feature set changes were trialled in order to develop the most effective overall disorder-centric relationship detection system. Important model/feature combinations are reported later in the results section (Table 4). The layer size was set to 300 throughout; batch size to 32 and learning rate to 0.005. From Table 4, the extra LSTM layer (Id=S2) was added in series immediately after the first layer. The dense layer (Id=S3) is a fully connected feed forward layer, connected to the output of the second LSTM layer and feeds into the final RNN output layer.

Training and Evaluation Methodology

A label data set for training the deep learning model and evaluating the BRE systems was developed. Existing label collections were too specific to utilise; for example, labels derived for protein-protein interactions[3] or adverse drug events[26]. The closest suitable label data set, the i2b2 challenge[4], was developed for clinical text, not biomedical literature, which is very different, rendering it unsuitable. The high cost of constructing a *gold* label collection — one constructed from expert annotations, of sufficient size — made the gold option not possible. However, based on the reduced requirement for relationship identification, rather than classification, non-expert labellers (via crowdsourcing, with appropriate quality controls) were employed to detect pre-identified concepts within 754 Medline citations. The final label values were assigned through a voting system across multiple labellers (n=5) and multiple tests of the same relationship, found in different parts of the text, with a positive label awarded when the proportion of actual to possible votes for the relationship exceeded a threshold of 0.4. Details of this *silver*, disorder-centric relationship, label set are provided in Table 2. Note the collection is heavily weighted towards negative labels, which is as expected, because across one or more sentence boundaries, the majority of concepts are not directly related.

Table 2: Final label counts for the disorder-centric silver label set. Numbers summed by relationship type. DS=Disorder-Symptom DE=Disorder-Test, DT=Disorder-Treatment, DD=Disorder-Disorder.

Relationship Type:	**DS**	**DE**	**DT**	**DD**	**Totals(% of total)**
Negative label	1,927	1,080	1,959	4,064	9,030 (70.6%)
Positive label	559	658	1,194	1,359	3,770 (29.4%)
Totals	2,486	1,738	3,153	5,423	12,800 (100%)

The evaluation metrics were the standard precision, recall and F1-score. A detected relationship was counted as true positive if the silver label set and model agreed on the existence of the relationship.

Table 3: Definition of comparative baselines; specific settings and relationship collection statistics included.

Setting/Statistic	Baselines				
	CO-OAS-BT	**CO-OAS-BF**	**SR-OAS**	**SR-MED**	**PAS-MED**
Corpus Filter	OA-Subset2014	OA-Subset2014	OA-Subset2014	None	None
Concept Extraction Method	QuickUMLS*	QuickUMLS	SemRep**	SemRep	MetaMap***
Concept Filter:					
- Best Match Setting	true	false	N/A	N/A	N/A
- Similarity filter	>0.7	>0.7	N/A	N/A	N/A
Relation extraction method	co-occur	co-occur	SemRep	SemRep	PasMed
Test Set Statistics:					
- # PMID used	675,052	675,052	675,052	17,470,394	23,343,329
- # distinct disorders	29,298	52,434	10,826	26,354	45,374
- # distinct relationships	1,539,931	6,934,505	158,674	1,333,218	5,489,990
- # distinct Symp-Dis rels	353,638	1,733,461	5,802	83,950	738,801
- # distinct Treat-Dis rels	333,493	1,848,135	108,822	616,372	2,136,403
- # distinct Test-Dis rels	133,199	776,312	11,436	161,104	500,091
- # distinct Dis-Dis rels	719,602	2,576,597	35,614	471,792	2,114,695

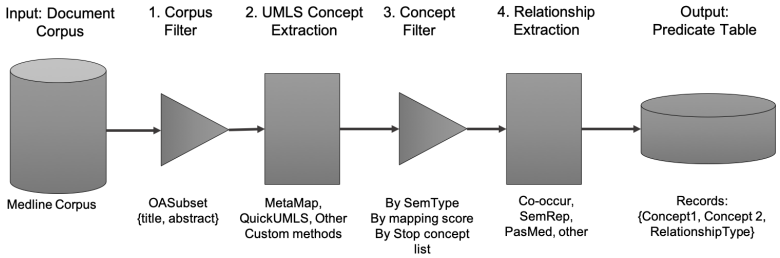
*QuickUMLS[9] version 1.2 built on UMLS dataset 2018AA; ** SemMed v3.1[14] using SemRep[5] v1.7. Predication table containing the relationships processed Medline articles up to December 31, 2017; *** Metamap[4] v2016v2 with UMLS dataset 2018AA.

For the LSTM model comparison with the existing models, five fold cross validation was employed where 80% of the label data was used for training and 20% was set aside for testing. The best average F1 score across the five folds for a single epoch was then used for comparison. Cross-fold validation was not employed for trialling and comparing the LSTM models and features, however training and test data was divided using the same 80/20 split.

Instantiation of Existing Models For Detecting Disorder-Centric Biomedical Relationships

The proposed LSTM model is compared with a number of existing models; these baselines were selected according to the following criteria: (a) Relationships are extracted from the Medline corpus and/or the TREC 2014[27] document collection, consisting of the Open Access Subset of Pubmed Central, taken on January 21, 2014, herein called OA-Subset2014. This subset of Medline was selected because it is well defined, commonly used and easily accessible to researchers; (b) The extracted relationship types must include at least the four disorder-centric relationships targeted in this study; (c) The extracted biomedical relationships must be expressed as a relationship between two UMLS concepts so that evaluation between extraction methods can take place in the UMLS concept space. Table 3 defines each baseline model, the settings employed for each and the statistics of the resulting test collection. Figure 3 describes the generic pipeline employed for preparing each model to generate the test collections.

Figure 3: Pipeline Diagram with the text along the top describing the steps used to extract relationships and the text along the bottom providing examples of each step



For all models, a semType filter was applied, matching the disorder-centric concept classifications specified in Table 1. In addition, because SemRep also included a fixed set of predicates, we limited the predicates to those appropriate for each relation (e.g., disorder-treatment (DT) relationships were limited to *treats, neg_treats* predicates). (For the sake of brevity, these are not listed here, however are available upon request.) Generic, high-use concepts, such as *treatment, patient, disease* were removed prior to labelling, and therefore not tested in our evaluations.

In addition to the settings identified in Table 3, the following pre-processing steps relate to specific models. For the co-occurrence models, a text window size of 5, within sentence was used to capture two concepts within a relationship. The BestMatch setting relates to the CUI(s) selected within the co-occur window. When bestMatch=false, all possible concepts extracted from the window of text are included within the window, i.e. often more than one concept for the same text and concept borders can overlap, whereas when BestMatch=true, concept borders can not overlap and only the best concepts are selected. PASMED relationships were extracted by firstly filtering for the correct semType combinations, then MetaMapping both of the medical terms in each relationship to identify the CUIs and filtering across only those CUIs within the label data set. Because an actual relationship type was not provided, but rather the verb relating the two CUIS, *all* relationships were considered valid if the CUIs formed a valid disorder-centric semType pair, irrespective of the order of the CUIs. This provided PASMED results with optimistic recall.

Results and Discussion

LSTM Model Selection Results

Table 4 lists the model and feature definition changes and their impact on the extraction system performance. The changes made at each step are cumulative, such that each model builds upon the model and feature set of the previous model. The cumulative impact on F1-score is graphed in Figure 4.

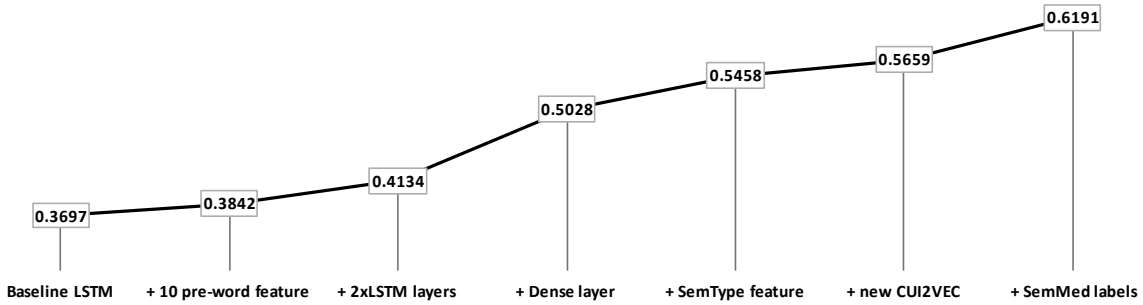
The baseline model consists of a single LSTM layer and output RNN with input features including the location, cosine similarities, POS of a single pre-word to the CUI and CUI embeddings[23] (500 dimensions). Improving the input features had a smaller overall impact (16.1%) on F1-scores, when compared with improving the model (38.6%). Incorporating a dense layer, i.e., a fully connected feed forward layer, at the output of the second LSTM, was the most important model improvement, adding 21.6% to F1 with all of the benefit arising from a 58.5% improvement in recall. The intuition behind the addition of the dense layer was to introduce a learning layer at a higher level of abstraction that might identify patterns across the output of the LSTM units. In particular the authors speculate that the because of the wide variation in input sequences, the dense layer can take this variability into account to identify enable the identification of more positive relationships. The most important feature addition was the CUI semType embedding, improving recall, over the previous model, by almost 15%, resulting in the best overall, recall model (R=70.3%). No available learned semType embeddings from Medline were available, and so one-hot vectors were employed. A learned embedding of semTypes is likely to result in further gains, which was left for future work.

Table 4: Deep learning, disorder-centric models and performance. The table depicts the incremental modifications made to either the model or the feature set used to train the model. Performance is measured by precision, recall and F1-score.* The (%) is a measure of the percentage change over the previous model

Id	Model Definition	Feature Definition	Precision (%*)	Recall (%*)	F1-score (%*)
BL	M1=Baseline LSTM	F1=Baseline features	0.4467 (0.0)	0.3154 (0.0)	0.3697 (0.0)
S1	M1	F2=F1+10 pre-word POS	0.4347 (-2.7)	0.3442 (+9.1)	0.3842 (+3.9)
S2	M2=M1+extra LSTM layer	F2	0.4457 (+2.5)	0.3856 (+12.0)	0.4134 (+7.6)
S3	M3=M2+dense layer	F2	0.4271 (-4.2)	0.6113 (+58.5)	0.5028 (+21.6)
S4	M3	F3=F2+semType	0.4463 (+4.5)	**0.7024** (+14.9)	0.5458 (+8.5)
S5	M3	F4=F3+new CUI2VEC	0.4799 (+7.5)	0.6895 (-1.8)	0.5659 (+3.7)
S6	M4=M3+semMed training data	F4	**0.6185** (+28.9)	0.6395 (-7.2)	**0.6191** (+9.4)

Introducing positive semMed[14] training data had the second highest impact (9.4%) on model performance and the greatest positive impact on precision across all changes. It is widely recognised that training deep learning models on imbalanced label sets can adversely impact model performance. The silver label data set contains 2.4 times more negative than positive labels (Table 2). To balance the training data set, disorder-centric semMed relationships were selected from source documents outside of the silver label set but within the OA-Subset2014 collection, until the positive/negative label ratio was equal. The balanced and larger training set improved precision by 28.9%, however it is unclear what proportion of this change is due to the larger training set or better label balance. Training the model with a balanced silver data set would reduce it's size by 41%, thereby rendering the results equally unclear and semMed only provides positive labels, so balance cannot be achieved by adding equal positive/negative labels. Better understanding the causal elements for this improvement has been left for further investigation.

Figure 4: Line graph showing the F1-scores for each model identified in Table 4



Although the input features had a lesser impact on performance, each feature added at least 3.7% to the F1 score. The most important feature, CUI embeddings, were incorporated into the baseline model. Switching to our custom embedding, improved the F1 score by 3.7%, however this may have been a result of the expanded training base. The custom embedding supported more than 5 times the number of CUIs enabling a 39% increase in the training set size. Like the addition of semMed training data, the performance increase was the net result of improved precision and decreased recall (+7.5% vs -1.8%). It is unknown whether the improvement was due to the increased training set or the change in CUI embedding used. Understanding the relative benefit would be helpful and was left for future work. Also left for future work was replacing the CUI embedding with a deep, pre-trained model, such as those used in language model or text classification (E.g., ELMo, ULMFiT) which incorporate contextual information beyond that of the semantic embeddings used in this work.

Comparative Results

Table 5 lists the results by test system evaluated on the silver label data set. The label collection was derived from 754 documents within the OA-Subset2014 citation collection so that all systems tested on this corpus have equal access to the source documents.

The LSTM, inter-sentence relationship detection model achieved the best performance, compared with all systems, whether they utilised the whole of Medline or the OA-Subset2014 collection to derive their relationship sets. It achieved improvements of +82% in precision, of +57% in recall and of +78% in F1-score, when compared with the best corresponding result for any other system. Three limitations to the evaluation warrant mention here. First, the label collection quality is classed as 'silver': the detection of relationships is done by non-experts assessing each sentence from an English grammar standpoint, rather than a medical one. The use of multiple labellers and multiple validation paths for each label is designed to minimise poor labels, however until the data set is validated by a domain expert, errors may exist. Explication and validation of this silver dataset is the subject of future work. Second, at this point it is not possible to confirm that the superior performance of the LSTM model is due to the ability to detect more relationships, via inter-sentence relationship capture, or other factors. Analysis of the label collection reveals that 58% of the labels are inter-sentence, which might favour this hypothesis, however confirming this is left to future research. Third, the LSTM model was trained via 5-fold cross validation on the label collection. While over-fitting may occur for such a method on a smaller dataset, the standard deviation of F1 scores across five folds was 0.0307, which was 5.0% of the F1 average, suggesting minimal over-fitting.

Table 5: Performance results for each system based on the data collection.

Test System	Data Collection	Precision	Recall	F1-Score
CO-OAS-BF	OA-Subset2014	0.2966	0.4146	0.3458
CO-OAS-BT	OA-Subset2014	0.2933	0.3493	0.3189
SR-OAS	OA-Subset2014	0.3400	0.1037	0.1589
SR-MED	MEDLINE	0.3218	0.2735	0.2957
PAS-MED	MEDLINE	0.3103	0.3976	0.3486
LSTM-DL	NA	0.6162	0.6487	0.6189

Between the test systems that were constructed with the OA-Subset2014 citation collection, the co-occurrence model with bestMatch=false produced the best F1 score (0.3458) with the highest recall (0.4146). It was hypothesised that this model would generate higher recall than the alternative bestMatch=true model because of the increased number

of CUIs and therefore relationships that are identified. However, it was not expected to produce similar, or even higher precision. This result is possibly explained by the quality of CUI mapping, which despite the setting name, *bestMatch*, struggles to select the correct CUI for a set of medical text. By using the BestMatch=false, more CUI options are created within the selection window, generating more correct relationships and because of the disorder-centric filter, less of the invalid relationships are kept. Although semRep generated particularly low recall results (0.1037), the relationships that were identified were more likely to be correct than in the co-occurrence models, reflected by a +14.6% improvement in precision over the best co-occurrence model precision.

Expanding the source of relationships from the OA-Subset2014 citation collection to the whole of Medline improved the SemRep system F1 performance by +86%, solely through a +164% increase in recall. PASMED, which demonstrated a +71% increase in recall over SemMed in Nguyen et al.'s work[6], showed a +45.5% increase in this comparison. The lower increase could be a result of the specific disorder-centric nature of the relationships extracted in this work. Overall, the PASMED system produced the best F1 score of all the existing baselines.

All of the existing baselines utilise within-sentence relationship extraction methods and the maximum recall achieved by any system was 0.4146, which would indicate that many of the labelled relationships were indeed inter-sentence. This helps to explain why the LSTM model could achieve a step-wise improvement in relationship-detection, over these existing baselines. This hypothesis could be tested in future work by extending the co-occurrence window to citation level co-occurrence, which should capture all relationships. If many relationships are cross-sentence, as indicated, then the disorder-centric LSTM model may have broader application in other BRE domains where relationships also occur between sentences.

Conclusion

In this paper we introduce a deep learning model to detect disorder-centric relationships between medical concepts across sentence boundaries. Relationships between disorders and their associated tests, treatments and symptoms underpin essential information needs of clinicians and can support biomedical knowledge bases, information retrieval and ultimately clinical decision support. The task was re-cast as a sequential labelling problem that could be tackled by a deep learning LSTM model approach. An inter-sentence, disorder-centric *silver* quality label collection was created by non-expert humans for training and evaluation purposes.

The cumulative contributions of word POS, punctuation and semType embeddings are compared together with important model changes. The resulting solution is then tested against other existing broad, biomedical relationship extraction methods, including co-occurrence, semRep and PASMED. The LSTM model outperforms these other models by at least +78% in F1 score suggesting that (a) inter-sentence relationships form an important component of disorder-centric relationships in biomedical literature and (b) the proposed LSTM deep-learning model and input sequence definition is a suitable approach for extracting these relationships. Although the LSTM model tested here was a suitable initial investigation, the next step is to consider other more recent deep learning models, for example transformer and end-to-end learning models, to compare the impact on BRE.

We also explore the impact of concept extraction methods (bestMatch = true/false) and collection size (Medline versus a 700K subset of Medline) on relationship extraction performance. Corpus size improves recall with a reduction in precision and using a broad concept selection policy (bestMatch=false) improves recall without deterioring precision.

In conclusion, we provide a promising, new approach to inter-sentence, disorder-centric biomedical concept relationship extraction with noted limitations and numerous avenues to expand the technique to other biomedical domains.

References

1. Murthy Devarakonda and Ching-Huei Tsou. Automated problem list generation from electronic medical records in IBM Watson. In *Twenty-Seventh IAAI Conf.*, 2015.
2. Louise Deleger, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferre, Philippe Bessieres, and Claire Nedellec. Overview of the bacteria Biotope task at BioNLP shared task. In *Proc. BioNLP.*, pages 12–22, 2016.
3. Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(1):50, 2007.

4. Alan R Aronson and François-Michel Lang. An overview of MetaMap: historical perspective and recent advances. *J. Am. Med. Informatics Assoc.*, 17(3):229–236, 2010.
5. T. Rindflesch and M. Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J. Biomed. Inform.*, 36(6):462–477, 2003.
6. Nhung T H Nguyen, Makoto Miwa, Yoshimasa Tsuruoka, Takashi Chikayama, and Satoshi Tojo. Wide-coverage relation extraction from MEDLINE using deep syntax. *BMC Bioinformatics*, 16(1):107, 2015.
7. Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. Extracting bio-molecular events from literature-the bioNLP'09 shared task. *Comput. Intell.*, 27(4):513–540, 2011.
8. S. M Meystre, G. Savova, K. Kipper-Schuler, and J Hurdle. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb. Med. Inform.*, 17(01):128–144, 2008.
9. Luca Soldaini and Nazli Goharian. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR Work. SIGIR*, 2016.
10. J W Ely, J a Osheroff, P N Gorman, M H Ebell, M L Chambliss, E a Pifer, and P Z Stavri. A taxonomy of generic clinical questions: classification study. *BMJ*, 321(7258):429–432, 2000.
11. Jing Ding, Daniel Berleant, Dan Nettleton, and Eve Wurtele. Mining MEDLINE: abstracts, sentences, or phrases? In *Biocomput. 2002*, pages 326–337. World Scientific, 2001.
12. Katrin Fundel, Robert Küffner, and Ralf Zimmer. RelEx—Relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, 2006.
13. Abhishek Sharma, Rajesh Swaminathan, and Hui Yang. A verb-centric approach for relationship extraction in biomedical text. In *Semant. Comput. (ICSC), 2010 IEEE Fourth Int. Conf.*, pages 377–385. IEEE, 2010.
14. 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.
15. Bryan Rink, Sanda Harabagiu, and Kirk Roberts. Automatic extraction of relations between medical concepts in clinical texts. *J. Am. Med. Informatics Assoc.*, 18(5):594–600, 2011.
16. Ozlem Uzuner, Jonathan Mailoa, Russell Ryan, and Tawanda Sibanda. Semantic relations for problem-oriented medical records. *Artif. Intell. Med.*, 50(2):63–73, 2010.
17. Chris Quirk and Hoifung Poon. Distant supervision for relation extraction beyond the sentence boundary. *arXiv Prepr. arXiv1609.04873*, 2016.
18. X. Lv, Y. Guan, J. Yang, and J. Wu. Clinical relation extraction with deep learning. *IJHIT*, 9(7):237–248, 2016.
19. Xu Yan, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. Classifying Relations via Long Short Term Memory Networks along Shortest Dependency Path. In *Proc. EMNLP.*, pages 1785–1794, 2015.
20. Fei Li, Meishan Zhang, Guohong Fu, and Donghong Ji. A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinformatics*, 18(1):198, 2017.
21. Pankaj Gupta, Subburam Rajaram, Hinrich Schütze, Bernt Andrassy, and Thomas Runkler. Neural relation extraction within and across sentence boundaries. *arXiv Prepr. arXiv1810.05102*, 2018.
22. Farrokh Mehryary, Jari Björne, Sampo Pyysalo, Tapio Salakoski, and Filip Ginter. Deep learning with minimal training data: TurkuNLP entry in the BioNLP shared task 2016. In *Proc. BioNLP*, pages 73–81, 2016.
23. Andrew L Beam, Benjamin Kompa, Inbar Fried, Nathan P Palmer, Xu Shi, Tianxi Cai, and Isaac S Kohane. Clinical concept embeddings learned from massive sources of medical data. *arXiv Prepr. arXiv1804.01486*, 2018.
24. Klaus Greff, Rupesh K Srivastava, Jan Koutnik, Bas R Steunebrink, and Jürgen Schmidhuber. LSTM: A search space odyssey. *IEEE Trans. neural networks Learn. Syst.*, 28(10):2222–2232, 2017.
25. Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proc. Thirteen. Int. Conf. Artif. Intell. Stat.*, pages 249–256, 2010.
26. Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *J. Biomed. Inform.*, 45(5):885–892, 2012.
27. Matthew S Simpson, Ellen Voorhees, and William Hersh. Overview of the TREC 2014 Clinical Decision Support Track. In *TREC 2014*, volume 2, pages 1–7, 2014.