# Inferring Implicit Causal Relationships in Biomedical Literature

**Halil Kilicoglu**
Lister Hill National Center for Biomedical Communications
National Library of Medicine
Bethesda, MD, 20894, USA
kilicogluh@mail.nih.gov

## Abstract

Biomedical relations are often expressed between entities occurring within the same sentence through syntactic means. However, a significant portion of such relations (in particular, causal relations) are expressed implicitly across sentence boundaries. Inferring these discourse-level relations can be challenging in the absence of syntactic clues. In this paper, we present a study of textual characteristics that contribute to expression of implicit causal relations across sentence boundaries. Focusing on a chemical-disease relationship corpus, we identify and investigate the contribution of various features that can assist in identifying such inter-sentential relations. Using these features for supervised learning, we were able to improve previously reported best results by more than 13%. Our results demonstrate the usefulness of the proposed features and the importance of using a balanced dataset for this task.

## 1 Introduction

Causal associations between entities, events, and processes are central to biomedical knowledge (Mihăilă et al., 2013). Such associations extend from physical causation, such as gene-disease relationships and adverse drug reactions, to rhetorical causation between claims and their justifications. Detecting causal associations in biomedical literature can assist in biocuration of pathways and databases, such as the Comparative Toxicogenomics Database (CTD)[1], and support tasks such as drug discovery and pharmacovigilance. Recognizing this need, the recent BioCreative V challenge included a task (CID) on extraction of chemical-induced disease relationships from Medline abstracts (Wei et al., 2016).

Chemical-disease relationships that the CID task focuses on are causal relationships in which a chemical acts as the *cause* and a disease or an adverse effect acts as the *effect*. In the simplest case, these relationships can be expressed intra-sententially through syntactic means. For example, in the sentence below (taken from the CDR corpus used in the CID task), the causal relationship between the drug *tacrolimus* and the disease *myocardial hypertrophy* is expressed explicitly with the causal trigger *induce*, which has the drug mention as its subject and the disease mention as its direct object.

(1) *Thus, we conclude that tacrolimus* ***induces*** *reversible myocardial hypertrophy.*

Assuming that the named entities have been successfully recognized by a named entity recognition (NER) system, lexical clues (the causal trigger *induce*) and syntactic dependency path between the entities and the trigger can be used to establish a causal link. However, not all causal relationships are expressed intra-sententially, and crucial information may be missed if the implicit, discourse-level relationships are simply ignored. For illustration, consider the discourse fragment below.

(2) *We investigated the efficacy and toxicity of a 3-hour paclitaxel infusion in a phase II trial in patients with inoperable stage IIIB or IV NSCLC. . . . Hematologic toxicities were mild: only one patient (2%) developed grade 3 or 4 neutropenia, while 29% had grade 1 or 2. Grade 1 or 2 polyneuropathy affected 56% of patients while only one (2%) experienced severe polyneuropathy. Similarly, grade 1 or 2 myalgia/arthralgia was observed in*

[1] http://ctdbase.org/



*Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 46–55,
Berlin, Germany, August 12, 2016. 

*63.2% of patients, but only 14.3% experienced grade 3 or 4. Nausea and vomiting were infrequent, ...*

Limiting relation extraction to sentence level, we would miss the causal relationships between the drug *paclitaxel* and the adverse effects (*neutropenia, polyneuropathy, myalgia, arthralgia, nausea,* and *vomiting*). The difficulty of extracting implicit, discourse-level relationships is due to several factors. First, the role of syntax in expressing relationships is limited; no syntactic dependency exists between the entities. Secondly, discourse-level phenomena, such as coreference, implicit argumentation, and rhetorical relations between sentences play a larger role. Resolving such phenomena could aid in identifying implicit relationships; however, these are all challenging NLP tasks in their own right. Thirdly, potential relationships between all entities occurring in the document may need to be considered, which can lead to a data sparsity/imbalance problem due to the smaller number of relations expressed across sentence boundaries.

In the biomedical domain, to our knowledge, there is little research specifically focusing on implicit, inter-sentential relations. In the GENIA event corpus (Kim et al., 2008), one of the major corpora for biomedical relation extraction, 7.8% of all events cross sentence boundaries and the majority of these events (4.8%) are causal. In contrast to the text-bound and linguistically-motivated annotation in the GENIA event corpus, the CDR corpus annotation is not concerned with explicit event triggers and implicit causal inferences are annotated much more frequently, as illustrated in the example above. 27.2% of all relations in the corpus are expressed only at the discourse level; that is, their arguments never co-occur within the same sentence. Therefore, the CDR corpus provides a good opportunity to study implicit causal relationships. While systems participating in the CID task have addressed discourse-level relations to some extent, only a few have explicitly reported results on discourse-level relations. Among these, the top-ranked system (CD-REST) (Xu et al., 2016) incorporated a document-level classifier, which uses entity and context-based features as well as knowledge-based features. Knowledge-based features, particularly those extracted from the CTD database, proved to be the difference, since this database provides manually curated relationships between chemical and diseases.

In this paper, we aim to elucidate the textual characteristics that play a role in implicit, discourse-level relations. While the CD-REST system (Xu et al., 2016) demonstrates that curated knowledge about chemical-disease relationships in structured resources can be used to great advantage, we approach the problem purely as a natural language processing task and specifically focus on characteristics that can be derived from the text, since presence of curated relationships cannot be assumed for all relation extraction tasks and therefore such an approach may not be generalizable. Based on the characteristics that are discussed, we propose specific features that can play a role in recognizing implicit relations, use these features for supervised learning and investigate their effect. To address the imbalance of the data, we also experiment with different training sizes. Our results show that the features we propose aided by a balanced training set can provide state-of-the-art performance in recovering implicit causal relationships and indicate that named entity recognition has a significant impact on the performance.

## 2 Related Work

In the general domain, Swampillai and Stevenson (2011) used an SVM-based approach to address inter-sentential relations in the MUC6 dataset. Adapting structural features used for intra-sentential relation extraction (e.g., parse trees) to the inter-sentential case and addressing the data sparsity problem by hyperplane adjustment, they were able to obtain comparable performance to intra-sentential relation extraction. A relevant research thread in semantic role labeling (SRL) is concerned with implicit arguments of predicates. Gerber and Chai (2010) studied implicit arguments of a small number of nominal predicates, such as *price* and *shipping*. Their model used a variety of features such as VerbNet classes and semantic roles for predicates and arguments, sentence distance, predicate frequency, and pointwise mutual information between arguments to identify implicit arguments. The SemEval-2010 Task 10: Linking Events and their Participants in Discourse (Ruppenhofer et al., 2010) addressed the same problem on a larger set of event predicates. The participating systems performed very poorly; however, more recent studies

were able to improve results, by casting the problem as an anaphora resolution task (Silberer and Frank, 2012) and by using the previously identified explicit arguments of a given predicate in linking (Laparra and Rigau, 2013). Causal relations have also been studied in the general domain from a wide range of perspectives. For example, Girju (2003) learned patterns indicating causal relationships between noun phrases to improve question answering. Other research focused on causal relations between discourse segments (rather than individual entities) and generally reported poorer results on causal relations than other types of discourse relations (Subba and Di Eugenio, 2009). It should be noted that most research on implicit arguments and causal relations assume the presence of explicit triggers (e.g., *produce*, *as a result*).

In the biomedical domain, there is little work that specifically addresses implicit arguments. Focusing on consumer health questions, Kilicoglu et al. (2013) incorporated resolution of anaphora and ellipsis to their question frame extraction pipeline and reported an 18 point improvement in $F_1$ score due to implicit argument resolution. Coreference resolution has been studied as a strategy to recover implicit arguments and improve event extraction and varying degrees of improvement due to coreference resolution have been reported (Yoshikawa et al., 2011; Miwa et al., 2012; Kilicoglu and Bergler, 2012; Lavergne et al., 2015; Kilicoglu et al., 2016).

Regardless of whether they are expressed implicitly, a wide range of causal relations have also been addressed in biomedical text. GENIA event corpus (Kim et al., 2008) and BioInfer corpus (Pyysalo et al., 2007) contain causal relationships between genes/proteins (e.g., REGULATION, POSITIVE_REGULATION, and NEGATIVE_REGULATION), in addition to other relation types. Causal relations in these corpora were often found to be more challenging to identify than other relation types (Kim et al., 2012). In the BioCause corpus (Mihăilă et al., 2013), causality was addressed as a discourse coherence relation and 850 causal discourse relations from full-text journal articles on infectious diseases (94% of which have explicit causal triggers) were annotated. In the BioDRB corpus (Prasad et al., 2011), a larger number of discourse relation types were annotated, one of which is causality. Mihăilă and Ananiadou (2014) focused on discourse causality in BioCause and used a semi-supervised method to recognize causal triggers and their arguments in biomedical discourse. They did not address implicit discourse causality.

BioCreative V CID task involved chemical-disease relationships at the discourse level, even though they were often not specifically addressed. The top-ranked system (CD-REST) (Xu et al., 2016) incorporated a discourse-level classifier, which interestingly performed better than the sentence-level classifier; however, most of the performance gain was due to features extracted from curated resources, particularly CTD. Similarly, the next best system (Pons et al., 2016) used domain knowledge from various databases, and one of better performing systems, UET-CAM (Le et al., 2015), incorporated features from coreference resolution into an intra-sentential relation classifier. The present study diverges from these studies by specifically addressing implicit, discourse-level causality and focusing on textual characteristics.

## 3 Methods

In this section, we first describe the corpus we used for analysis and experiments. Next, we discuss the linguistic characteristics of inter-sentential, implicit causal relationships. In the following subsection, we describe our supervised learning approach and features that we developed. Finally, we discuss our evaluation.

### 3.1 CDR Corpus

For our analysis and experiments, we used the CDR corpus that was used in the BioCreative V CID task (Wei et al., 2016). This corpus consists of 1,500 Medline abstracts, annotated with chemical and disease mentions, normalized to MeSH identifiers, and the abstract-level chemical-disease causal relationships between the normalized entities. The corpus is split into three, one-third is used for training, one-third for development, and the rest for testing. Causal triggers have not been annotated in the corpus. The distribution of chemical/disease entities as well as that of the relations are given in Table 1. For our experiments, we focused on relations that are solely expressed across sentences (i.e., entity pairs co-occurring in the same sentences are ignored). The statistics for these relations are also given in Table 1. We did not perform any named entity recognition or nor-

| Dataset | # Diseases | # Chemicals | # Relations | # Discourse-level Relations |
|---|---|---|---|---|
| Training | 4,182 (1,965) | 5,203 (1,467) | 1,038 | 283 |
| Development | 4,244 (1,865) | 5,347 (1,507) | 1,012 | 246 |
| Testing | 4,244 (1,988) | 5,385 (1,435) | 1,066 | 320 |
| TOTAL | 12,670 (5,818) | 15,935 (4,409) | 3,116 | 849 |

Table 1: CDR corpus characteristics

malization and conducted our analysis and experiments using the gold entities. For comparison, we also used DNorm (Leaman et al., 2013) for disease and tmChem (Leaman et al., 2015) for chemical name recognition/normalization. On the test portion of the corpus, DNorm achieves 81% $F_1$ score and tmChem achieves 91% $F_1$ score.

### 3.2 Characteristics of implicit causal relations

Focusing only on inter-sentential relations in the training set, we examined the linguistic characteristics that play a role in expressing them. We examine and exemplify some of the important characteristics below.

#### 3.2.1 Causal ordering of events

A significant portion of the implicit chemical-disease relationships can be seen as inferences, rather than explicit assertions. One minimal condition for such causal inference is temporality: if a chemical causes a disease in a patient, then the chemical administration has to occur before the manifestation of the disease. In biomedical abstracts, language describing such event ordering is present, particularly in descriptions of experiments. An example, shortened from the original text, is shown below, with relevant chemical and disease mentions underlined.

(3) *We report on a combination of everolimus and tacrolimus in 24 patients ... with either myelodysplastic syndrome ... or acute myeloid leukemia .... All patients engrafted, and only 1 patient experienced grade IV mucositis. ... Transplantation-associated microangiopathy ... occurred in 7 patients ..., with 2 cases of acute renal failure.*

Similarly, case studies often involve language describing a sequence of events that lead to a medical problem. An example is given below.

(4) *We present a case of a 5-year-old child with cerebral palsy and seizure disorder, receiving clonidine for restlessness, who presented for placement of a baclofen pump. Without the knowledge of the medical personnel, the patient's mother administered three doses of clonidine during the evening before and morning of surgery to reduce anxiety. During induction of anesthesia, the patient developed bradycardia and hypotension ...*

#### 3.2.2 Coreference

The role of coreference in expressing implicit arguments has been acknowledged (Silberer and Frank, 2012). Anaphora relations can create explicit links between sentences and assist in resolving implicit arguments. In the following example, the definite noun phrase *this regimen* and the personal pronoun *it* corefer with *combination therapy with pegylated interferon and ribavirin* in the previous sentence. If these anaphora relations are resolved, the anaphoric expressions can simply be substituted with the antecedent, simplifying the problem to sentence-bound relation extraction.

(5) *The current best treatment for HCV infection is combination therapy with pegylated interferon and ribavirin. Although **this regimen** produces sustained virologic responses (SVRs) in approximately 50% of patients, **it** can be associated with a potentially dose-limiting hemolytic anemia.*

Bridging (or associative) anaphora (Poesio et al., 1997), a type of indirect coreference that is distinguished by relations such as hypernymy (is-a) or meronymy (part-of), is also used considerably to indicate implicit causal relations. In the following example, the causal relation between *ventricular fibrillation* and the chemicals *sodium citrate* and *disodium edetate* can be identified, if we can recognize that there is a meronymic relationship between these chemicals and *Renografin*.

(6) *Renografin contains the chelating agents sodium citrate and disodium edetate,*

*while Hypaque contains calcium disodium edetate and no sodium citrate. Ventricular fibrillation occurred significantly more often with* ***Renografin****.*

#### 3.2.3 Document Topic as Implicit Argument

Since abstracts are relatively short, it is common to have the main focus of the article mentioned only once and referred to implicitly throughout the abstract. For example, in an article investigating the side effects of a drug, the drug name is often mentioned early on (in some cases, only in the title), and the side effects of the drug are revealed later in the abstract. In the following example, the logical object argument of the predicate *treatment* is uninstantiated, and this implicit argument refers to the document topic, the drug *CCNU (lomustine)*.

(7) *CCNU (lomustine) toxicity in dogs: a retrospective study (2002-07) ... CCNU was used most commonly in the treatment of lymphoma, mast cell tumour, .... Throughout* ***treatment****, 56.9% of dogs experienced neutropenia, 34.2% experienced anaemia and 14.2% experienced thrombocytopenia.*

#### 3.2.4 Document Structure

The title and the abstract of an article need to convey the gist of the study in a small, often predetermined, number of words. To ensure that the content of the abstract is representative of the study, some journals require the abstracts to conform to a formal structure (structured abstracts), with sections such as Objective, Methods, and Results. Important findings are more likely to be reported in the Results section, and implicit causal relationships between entities in the Results section and the main topics of the articles are frequent. In the following example, *desipramine* is one of the main topics of the article and the only mention of *ventricular arrhythmias* is in the Results section.

(8) *Effect of calcium chloride and 4-aminopyridine therapy on desipramine toxicity in rats ... The incidence of ventricular arrhythmias (p = 0.004) and seizures (p = 0.03) in the CaCl2 group was higher than the other groups.*

### 3.3 Supervised learning of implicit causal relationships

We formulate implicit causal relation extraction as a binary classification task, where examples consist of chemical-disease mention pairs whose corresponding normalized entities do not co-occur intra-sententially in the abstract. Positive examples are mention pairs that are causally related, and negative examples are those that are not. We used linear SVM (Fan et al., 2008) to train the binary classifier and empirically set the regularization parameter $C$ to 0.1. To address the imbalance of the dataset (approximately 85% of all examples are negative), we trained the classifier with varying number of negative examples (undersampling). We selected negative examples from the documents in proportion to the number of all examples extracted from the document.

The classifier uses features developed based on the analysis presented in the previous section as well as standard n-gram (unigram and bigram) features. Features that proved predictive are provided in Table 2 and illustrated on the *desipramine:ventricular arrhythmias* pair from Example (8). In Table 2, we also indicate whether the feature or an approximation was used by the top-performing system (Xu et al., 2016) in the CID task. We distinguish between lexical, semantic, and discourse features.

*Lexical features* are simple n-gram features extracted from the sentences of the target mentions. We use unigrams and bigrams of the mentions as well as those of sentences that the mentions appear in.

*Semantic features* include conceptual knowledge about the entities (their MeSH identifiers and the MeSH identifiers of their ancestors in the MeSH hierarchy) as well as other semantic information that occur in the sentence context. For this purpose, we use an existing dictionary of causal predicates, previously compiled from several corpora. The list consists of 201 predicates and mainly includes triggers for regulatory events (e.g., *induce*, *effect*, *develop*) as well as discourse connectives that describe causal (e.g., *as a result*) or temporal relations (e.g., *before*, *after*). We also use a feature that indicates whether an experiencer (e.g., *patient*, *rats*) is mentioned in the sentence context. Finally, a binary feature indicates whether any mention belonging to the opposing semantic class occurs in the sentence (i.e., if the classified example includes a chemical mention in the current sentence, this feature is true if the current sentence contains a disease mention).

*Discourse features* are mainly features based on

| Feature | Description | CD-REST |
|---|---|---|
| *Lexical features* | | |
| $F_1$ | Uncased unigrams of the mentions | ✓ |
| $F_2$ | Uncased bigrams of the mentions | ✓ |
| $F_3$ | Uncased unigrams of the mention sentence(s) | |
| $F_4$ | Uncased bigrams of the mention sentence(s) | |
| *Semantic features* | | |
| $F_5$ | Uncased causal predicate lemmas preceding the chemical mention ($\{$*effect*$\}$) | S |
| $F_6$ | Uncased causal predicate lemmas following the chemical mention ($\emptyset$) | S |
| $F_7 - F_8$ | Same as $F_5 - F_6$, for the disease mention ($\{\emptyset,\emptyset\}$) | S |
| $F_9$ | Whether the opposing semantic class in the mention pair exists in the sentence (*true*) | |
| $F_{10}$ | Whether an experiencer trigger exists in either mention sentence (*true*) | ✓ |
| $F_{11}$ | Disease MeSH identifier (*D001145*) | ✓ |
| $F_{12}$ | chemical MeSH identifier (*D003891*) | ✓ |
| $F_{13}$ | disease MeSH hypernyms ($\{$*D002318, D006331, D010335, D013568*$\}$) | ✓ |
| $F_{14}$ | chemical MeSH hypernyms ($\{$*D003984, D006571, D006575*$\}$) | ✓ |
| *Discourse features* | | |
| $F_{15}$ | chemical in focus (*true*) | S |
| $F_{16}$ | disease in focus (*false*) | |
| $F_{17}$ | normalized section name of the chemical (*TITLE*) | |
| $F_{18}$ | normalized section name of the disease (*RESULTS*) | |
| $F_{19}$ | main verb POS sequence in target and intervening sentences (*NONE*) | |
| $F_{20}$ | whether the sentences of the mentions are adjacent (*false*) | ✓ |
| $F_{21}$ | the document contains sortal anaphors (*true*) | |
| $F_{22}$ | MeSH descendant of the disease occurs in the document (*true*) | ✓ |
| $F_{23}$ | MeSH ancestor of the disease occurs in the document (*true*) | ✓ |

Table 2: The features used by the binary classifier (S: a similar feature is used)

our analysis. To address causal ordering of events by capturing tense information, we include a feature that concatenates the part-of-speech tags of the main verbs of the mention sentences and those of the sentences intervening between them (ignoring title sentences). Adjacent sentences are often implicitly related, and therefore, we include a binary feature that indicates whether the mention sentences are adjacent. To address anaphora, we include a binary feature that indicates whether the abstract contains any sortal anaphors that can refer to chemical or disease mentions (e.g., *this drug*, *the condition*). We extracted this information using the Bio-SCoRes tool (Kilicoglu and Demner-Fushman, 2016). With regards to bridging anaphora, we use binary features that indicate whether a MeSH ancestor or descendant of one of the entities in the pair appear in the abstract, addressing hypernymy. Whether the chemical entity and the disease entity may be document topics are also included as features. We simply included all entities that appear in the title of the article as document topics. To capture document structure, we normalized the section names in structured abstracts using the mappings curated at the NLM[2]. If the abstracts are not structured, we simply used TITLE or ABSTRACT as the normalized section name.

Feature extraction presupposes a standard linguistic processing pipeline (i.e., tokenization, part-of-speech tagging, syntactic parsing). We performed this processing using the Stanford CoreNLP toolkit (Manning et al., 2014).

[2]http://structuredabstracts.nlm.nih.gov/Downloads/Structured-Abstracts-Labels-110613.txt.

### 3.4 Evaluation

In separate experiments, we used gold standard entities and those recognized by DNorm (Leaman et al., 2013) and tmChem (Leaman et al., 2015) as the basis for relation extraction. Following the CID baseline, we took simple abstract-level entity co-occurrence as the baseline method. We also compared our results to those reported with CD-REST (Xu et al., 2016). This comparison is made somewhat difficult by the fact that their discourse-level classifier considers entities co-occurring in the same sentences as candidates, as well. Another complicating factor in comparison is their classifier's use of curated knowledge-base features. In particular, two features from CTD provide more than 18% improvement in their overall $F_1$ score using gold standard entities (from 56.7% to 67.1%). For a fair comparison, we implemented these CTD features and incorporated them into our best model.

- CTD relation between the chemical and the disease: *null, inferred-association, therapeutic,* or *marker/mechanism*
- Whether the disease has a marker/mechanism association with any chemical in CTD

In addition, we performed an ablation study to better understand the contribution of various feature sets. We used the standard evaluation metrics, precision, recall, and $F_1$ score, to assess relation extraction performance.

## 4 Results and Discussion

The results of implicit causal relation extraction on the test set using the gold standard entities and DNorm/tmChem entities are provided in Table 3. The effect of CTD features on classification performance is also shown. We obtained the highest $F_1$ score and recall when we undersampled negative examples to yield a 1:1 positive/negative sample ratio (balanced training)[3]. The highest precision was obtained in both cases when all available data are used for training.

The improvement due to CTD features was less dramatic than that found by Xu et al. (2016) but still significant (more than 11% improvement with the gold entities, from 66.1% to 73.7%). However, we believe that the results obtained without CTD features are a better representation of the state-of-the-art for implicit causal relation extraction from a purely NLP perspective. In this setting, we obtained 66.1% $F_1$ score with gold entities and 48.6% $F_1$ score with DNorm/tmChem entities (in italics).

In comparing our performance to that of CD-REST, we find that our approach overall outperforms CD-REST. Using gold entities, CTD features, and a balanced training set, we outperformed their system by more than 9% (67.3% vs. 73.7%). They have not used a balanced training set, so the difference with their reported system is even wider (56.7% to 73.7%). Without CTD features, we slightly outperformed their reported results (66.1% vs. 64.9%), indicating that our approach in some sense compensates for the CTD knowledge and suggesting that it could support biocuration of these relationships. The performance they reported with gold entities is somewhat higher than what we obtained with our implementation of their features (64.9% vs. 56.7%); however, it is worth pointing out that their classifier takes into account mention pairs that co-occur in the same sentences, as well, which can explain the difference to some extent. The small differences in our implementation of their features could also account for some of the difference. CD-REST uses its own named entity recognition tool, which outperforms the DNorm/tmChem combination, and this is partly reflected in the performance difference between using DNorm/tmChem entities with their features and their reported end-to-end performance (50.2% vs. 56.8%).

To better understand the contribution of features, we performed an ablation study in which we removed a set of features, retrained our classifier, and assessed the performance. The results of this evaluation are shown in Table 4. In these experiments, we used gold entities and a balanced training set and did not include CTD features. The results show that lexical and discourse features contribute similarly to implicit causal relation extraction, while the contribution of semantic features is much smaller. We observe that the effect of lexical features is to improve precision, whereas discourse features contribute significantly to recall, with a minor degradation in precision.

While the discourse features we used were overall successful, our attempts at using more sophisticated discourse features have often resulted in performance loss. For example, coref-

[3]Not all the ratios we experimented with are shown.

| Experiment | Precision | Recall | $F_1$ |
|---|---|---|---|
| **Using DNorm/tmChem entities** | | | |
| Baseline | 20.3 | 67.3 | 31.2 |
| *Balanced training* | *46.9* | *50.5* | *48.6* |
| Balanced + CTD features | 56.4 | 54.9 | 55.6 |
| Unbalanced | 56.4 | 36.2 | 44.1 |
| **Using gold entities** | | | |
| *Balanced* | *59.7* | *74.0* | *66.1* |
| Balanced + CTD | 68.0 | 80.3 | 73.7 |
| Unbalanced | 67.6 | 52.4 | 59.0 |
| **Our CD-REST implementation** | | | |
| Balanced + CTD w/ gold entities | 70.1 | 64.8 | 67.3 |
| Unbalanced + CTD w/ gold entities | 79.4 | 44.1 | 56.7 |
| Balanced + CTD w/ tmChem/DNorm entities | 60.5 | 42.9 | 50.2 |
| **Reported CD-REST performance (Xu et al., 2016)** | | | |
| Using gold entities | 68.4 | 61.8 | 64.9 |
| End-to-end results | 64.1 | 50.5 | 56.8 |

Table 3: Evaluation results

| Experiment | Precision | Recall | $F_1$ |
|---|---|---|---|
| All | 59.7 | 74.0 | 66.1 |
| -Lexical features | 42.8 | 91.4 | 58.3 |
| -Semantic features | 58.6 | 73.7 | 65.3 |
| -Discourse features | 60.7 | 55.9 | 58.2 |

Table 4: Feature ablation results

erence emerged as an important aspect of implicit causal relations, and it seemed that fully resolving disease/chemical coreference in the abstract could improve the performance. We adapted the Bio-SCoRes framework (Kilicoglu and Demner-Fushman, 2016) to extract anaphora relations and incorporated more sophisticated features based on these relations into our classifier, such as whether a mention corefers with an anaphor in the sentence of the other mention in the pair (Example 5). While this improved precision (59.7% to 66.2%), the recall loss was more significant (74.0% to 62.9%), leading to a lower $F_1$ score (64.5%). Similar, unsuccessful features include a binary feature indicating whether there is a potential bridging anaphora that involves the chemical or the disease mention itself. On the other hand, a simplistic discourse feature that indicates whether the document contains any sortal anaphor at all improved the $F_1$ score from 65.1% to 66.1%. Along the same lines, using normalized structured abstract section labels improved the classification performance. However, most abstracts are not structured, and our attempts to automatically assign section labels using the sentence position in the abstract in such cases did not improve results.

The named entity recognition tools we used have reported relatively high performance on the test set (81% and 91% $F_1$ scores for DNorm and tmChem, respectively). However, the performance difference when using these tools in comparison to using gold entities is relatively large; gold entities yield more than 30% higher $F_1$ on average. This indicates that the relation extraction performance is highly sensitive to entity recognition and normalization, and that even small performance drop in this task can cause a major performance drop in relation extraction.

Data sparsity is a well-known problem for inter-sentential relation extraction (Swampillai and Stevenson, 2011). To deal with this problem, we experimented with training various positive/negative sample ratios, and found that a balanced training set led to superior overall performance, at the expense of loss of precision. This result is similar to that of Swampillai and Stevenson (2011), which they achieved with hyperplane adjustment.

There are several limitations to the study presented. First, we have not investigated the generalizability of the approach to other relation types expressed implicitly. The GENIA event corpus, with

its text-bound event triggers, presents an opportunity to study implicit argumentation more widely from a semantic role labeling perspective, even though the number of relevant events in the corpus is relatively small. Secondly, whether the method can be extended to extracting relations from full-text articles remains to be seen. Thirdly, there are NLP methods that can provide more predictive features that we have not attempted to incorporate into our models. For example, temporal ordering of events have been the subject of much research recently, in both general (Chambers et al., 2014) and clinical (Bethard et al., 2015) domains, and tools based on these methods can provide useful features to detect causal ordering of events. Similarly, while our simple sentence position-based heuristics to assign sections to unstructured abstract sentences did not yield predictive features, more advanced methods to classify sentences into rhetorical categories (Agarwal and Yu, 2009) could be beneficial.

## 5 Conclusion

We presented a method to extract implicit, intersentential causal relationships from Medline abstracts. The method incorporates lexical, semantic, and discourse features and a simple undersampling approach for data sparsity to achieve state-of-the-art results. In this study, we specifically focused on implicit relationships across sentences, since they are more challenging from an NLP perspective, and future work involves combining the proposed method with methods that extracts sentence-bound, mostly explicit relationships. Improving feature extraction and named entity recognition/normalization are likely to be beneficial in further improving the state-of-the-art in causal relationship extraction. Joint learning of named entities and causal relationships could further improve performance by preventing, to some extent, the propagation of named entity recognition errors to relation extraction step.

## Acknowledgments

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

## References

Shashank Agarwal and Hong Yu. 2009. Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion. *Bioinformatics*, 25(23):3174–3180.

Steven Bethard, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. SemEval-2015 Task 6: Clinical TempEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Matthew Gerber and Joyce Chai. 2010. Beyond NomBank: A Study of Implicit Arguments for Nominal Predicates. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1583–1592, Uppsala, Sweden.

Roxana Girju. 2003. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*, pages 76–83.

Halil Kilicoglu and Sabine Bergler. 2012. Biological Event Composition. *BMC Bioinformatics*, 13 (Suppl 11):S7.

Halil Kilicoglu and Dina Demner-Fushman. 2016. Bio-SCoRes: A Smorgasbord Architecture for Coreference Resolution in Biomedical Text. *PLoS ONE*, 11(3):e0148538.

Halil Kilicoglu, Marcelo Fiszman, and Dina Demner-Fushman. 2013. Interpreting consumer health questions: The role of anaphora and ellipsis. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, pages 54–62.

Halil Kilicoglu, Graciela Rosemblat, Marcelo Fiszman, and Thomas C. Rindflesch. 2016. Sortal anaphora resolution to enhance relation extraction from biomedical literature. *BMC Bioinformatics*, 17:163.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9:10.

Jin-Dong Kim, Ngan Nguyen, Yue Wang, Jun'ichi Tsujii, Toshihisa Takagi, and Akinori Yonezawa. 2012. The Genia Event and Protein Coreference tasks of the BioNLP Shared Task 2011. *BMC Bioinformatics*, 13 Suppl 11:S1.

Egoitz Laparra and German Rigau. 2013. ImpAr: A Deterministic Algorithm for Implicit Semantic Role Labelling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1180–1189.

Thomas Lavergne, Cyril Grouin, and Pierre Zweigenbaum. 2015. The contribution of co-reference resolution to supervised relation detection between bacteria and biotopes entities. *BMC Bioinformatics*, 16 (Suppl 10):S6.

Hoang-Quynh Le, Mai-Vu Tran, Thanh Hai Dang, and Nigel Collier. 2015. The UET-CAM System in the BioCreAtIvE V CDR Task. In *Fifth BioCreative challenge evaluation workshop*, pages 208–213.

Robert Leaman, Rezarta Islamaj Dogan, and Zhiyong Lu. 2013. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917.

Robert Leaman, Chih-Hsuan Wei, and Zhiyong Lu. 2015. tmChem: a high performance approach for chemical named entity recognition and normalization. *Journal of Cheminformatics*, 7(S-1):S3.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Claudiu Mihăilă and Sophia Ananiadou. 2014. Semi-supervised learning of causal relations in biomedical scientific discourse. *BioMedical Engineering OnLine*, 13(2):1–24.

Claudiu Mihăilă, Tomoko Ohta, Sampo Pyysalo, and Sophia Ananiadou. 2013. BioCause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics*, 14(1).

Makoto Miwa, Paul Thompson, and Sophia Ananiadou. 2012. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. *Bioinformatics*, 28(13):1759–1765.

Massimo Poesio, Renata Vieira, and Simone Teufel. 1997. Resolving bridging references in unrestricted text. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 1–6.

Ewoud Pons, Benedikt F.H. Becker, Saber A. Akhondi, Zubair Afzal, Erik M. van Mulligen, and Jan A. Kors. 2016. Extraction of chemical-induced diseases using prior knowledge and textual information. *Database*, 2016.

Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi, and Hong Yu. 2011. The biomedical discourse relation bank. *BMC Bioinformatics*, 12:188+.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007. BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8:50.

Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 2010. SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 45–50.

Carina Silberer and Anette Frank. 2012. Casting implicit role linking as an anaphora resolution task. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics*, pages 1–10.

Rajen Subba and Barbara Di Eugenio. 2009. An effective discourse parser that uses rich linguistic information. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 566–574.

Kumutha Swampillai and Mark Stevenson. 2011. Extracting Relations Within and Across Sentences. In *Proceedings of RANLP*, pages 25–32.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Jiao Li, Thomas C. Wiegers, and Zhiyong Lu. 2016. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. *Database*, 2016.

Jun Xu, Yonghui Wu, Yaoyun Zhang, Jingqi Wang, Hee-Jin Lee, and Hua Xu. 2016. CD-REST: a system for extracting chemical-induced disease relation in literature. *Database*, 2016.

Katsumasa Yoshikawa, Sebastian Riedel, Tsutomu Hirao, Masayuki Asahara, and Yuji Matsumoto. 2011. Coreference Based Event-Argument Relation Extraction on Biomedical Text. *Journal of Biomedical Semantics*, 2 (Suppl 5):S6.