Title: A Supervised Machine Learning Framework for the Extraction of Drug-Drug Interactions from Structured Product Labels

Authors and affiliations:

Johann Stan, PhD[1]

Dina Demner-Fushman, MD, PhD[1]

Kin Wah Fung, MD, MS[1]

Sonya E. Shooshan, MLS[1]

Laritza Rodriguez MD, PhD[1]

Olivier Bodenreider, MD, PhD[1]

[1]U.S. National Library of Medicine

Lister Hill National Center for Biomedical Communications

8600 Rockville Pike, Bethesda, MD 20894, USA

Corresponding Author

Olivier Bodenreider, MD, PhD[1]

[1]U.S. National Library of Medicine

Lister Hill National Center for Biomedical Communications

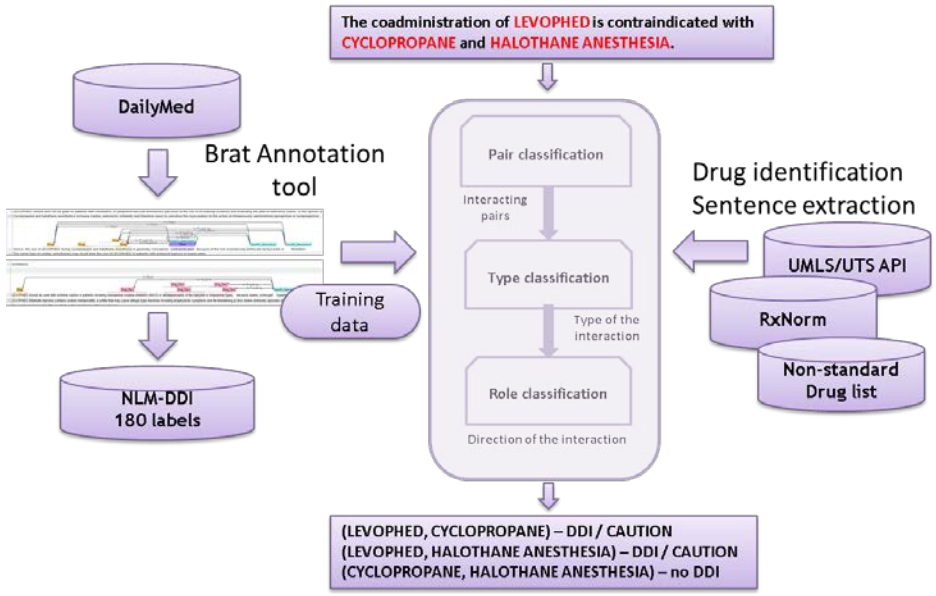
8600 Rockville Pike, Bethesda, MD 20894, USA

olivier.bodenreider@nih.gov

Keywords

Drug–drug interaction, Natural Language Processing, Support Vector Machines, Pharmacovigilance, DailyMed Structured Product Labels

Graphical Abstract



Highlights

- We design a complete machine learning framework for drug-drug interaction extraction using support vector machines and a shallow linguistic features space.
- We report that drug-drug interactions can be extracted from product labels published on DailyMed with high statistical accuracy scores.
- Our feature space and model generalize well for different types of annotation schemas and datasets (DailyMed, MEDLINE, DrugBank).
- Our annotated corpus is available at http://lhce-brat.nlm.nih.gov/NLMDDICorpus.htm.

Abstract

Background: Information about drug-drug interactions (DDIs) is found in the medical literature and in drug package inserts published on DailyMed in addition to commercial drug databases.

Objectives: To develop a machine learning framework for the extraction of DDIs from structured product labels (SPLs).

Methods: We develop a supervised machine learning framework (support vector machine classifier with shallow linguistic features) that extracts sentences containing a drug interaction relation, classifies interaction types and identifies the object and precipitant drug. We evaluate the framework performance against three document sets: the annotated corpus of 180 cardiovascular drug SPLs we created for framework development, the corpus of the SemEval DDIExtraction 2013 challenge, and a reference list of DDIs.

Results: The performance on our SPL corpus (F-score = 0.84) is competitive with the best performance in the SemEval DDIExtraction 2013 challenge. We demonstrate the portability of our approach to another corpus (SemEval DDI corpus).

Conclusions: Our work is the first attempt to extract drug-drug interactions from structured product labels. Our annotated corpus of 180 SPLs is available for download at http://lhce-brat.nlm.nih.gov/NLMDDICorpus.htm. Future work includes resolution of drugs referred to anaphorically and through drug classes.

1. Introduction

A drug-drug interaction (DDI) is defined as a modification in the effect of a drug when administered with another drug. It can be an increase or a decrease in the action of either drug, or an adverse effect not normally associated with the drugs when administered on their own [1, 2]. Exposure to a DDI occurs when a patient is prescribed or administered two or more drugs that are known to interact [3].

The incidence of DDIs in patients is estimated to range from 4.7% to 8.8% [4]. While exposure to a DDI does not always result in an adverse drug event [5], such events are a significant source of preventable drug-related harm. An analysis of sixteen cohort and case-control studies revealed an elevated risk of hospitalization in patients who were exposed to DDIs [6]. Clinically important events attributable to DDI exposure are estimated to occur in 5.3% to 14.3% of inpatients, and are responsible for 0.02% to 0.17% of the nearly 130 million emergency department visits that occur each year in the United States.

The effective use of clinical decision support in electronic health records has been shown to reduce medication errors, including adverse events related to DDIs [7]. Availability of comprehensive, up-to-date, and machine-readable knowledge about DDI is a prerequisite to more widespread implementation of clinical decision support systems. However, many knowledge bases fail to include important drug interactions and contain outdated, irrelevant, or even incorrect information [4]. The medical literature and drug package inserts (drug labels) are two major and commonly used sources of DDI information. In these sources, however, DDIs are expressed in natural language and do not have a predefined format [8]. In other words, DDI knowledge from these sources is not machine-readable. Examples of DDIs expressed in textual form include *"Do not coadminister aliskiren with Diovan in patients with diabetes."* and *"Limit*

the dose of simvastatin in patients on amlodipine to 20 mg daily." The lack of publicly available sources of machine-readable DDI information motivates our work on extracting this information from the Food and Drug Administration (FDA) structured product labels provided by DailyMed.

This variability observed in the expression of DDIs suggests that machine learning techniques will likely be more successful than symbolic methods (e.g. regular expressions) for extracting DDIs. Supervised machine learning techniques require an annotated corpus for their development and evaluation. Therefore, the first step in our work is to generate a corpus of DailyMed structured product labels annotated with DDI information.

The main objective of this work is to develop a supervised machine learning framework for the extraction of DDIs from DailyMed structured product labels. A secondary objective is to share our annotated corpus with the community in order to promote DDI extraction research. This work, done in collaboration with colleagues at the FDA, is the first attempt to extract DDI information from DailyMed structured product labels. Our current pilot focuses on cardiovascular drugs.

2. Background

This section introduces our source of DDI information, the DailyMed structured product labels, discusses related work and highlights the specific contribution of our work. It also presents two additional resources used in our evaluation: the SemEval DDI corpus and the ONC list of DDIs (high-priority DDIs curated by an expert panel [18] for the Office of the National Coordinator for Health Information Technology as part of the *Meaningful Use* incentive program).

2.1. DailyMed structured product labels

We use the DailyMed structured product labels[1] as our source of the drug-drug interactions. This is the most comprehensive, current and authoritative source of drug information available to the public in the form of drug package inserts (structured product labels). DailyMed is developed collaboratively by the National Library of Medicine and the FDA (Food and Drug Administration) and provides high quality information about some 67,876 marketed drugs. The Structured Product Labeling (SPL) document markup standard defines the structure of the human readable documents that contain the information provided in drug package inserts. A structured product label mostly contains textual information and is organized in several sections. The sections (described with their names and corresponding LOINC[2] codes) where DDI information may be present include: Boxed Warning section (34066-1), Contraindications section (34070-3), Dosage and Administration (34068-7), Drug and/or laboratory test interaction section (34074-5), Drug Interactions section (34073-7), Precautions section (42232-9), Warnings and Precautions section (43685-7), and Warnings section (34071-1). The DailyMed structured product labels used in this study were downloaded from the NLM DailyMed[1] website on August 10 2013.

2.2. Related Work

We review approaches to extracting information from the DailyMed structured product labels (SPLs), as well as approaches to extracting DDIs from textual sources, before highlighting the specific contribution of our work.

---

[1] DailyMed Structured Product Labels - http://dailymed.nlm.nih.gov/dailymed/index.cfm - accessed 05/05/2014

[2] LOINC codes - http://loinc.org/ - accessed 10/30/2014

2.1.1. Information Extraction from DailyMed Structured Product Labels

The SPLs have been used by several research groups to extract various kinds of information, including drug indications and adverse drug events.

Indications. Fung et al. [8] extracted focus drug indications from the SPLs. Their framework parses the indication section of the labels and analyses the corresponding text segments with the MetaMap program, restricting the extraction to disorder and finding semantic types and ignoring high-level concepts. The overall recall, precision, and F score were 0.95, 0.77, and 0.85, respectively, demonstrating that natural language processing (NLP) approaches are effective for extracting drug indications from SPLs. Li et al. [9] describe a system, called AutoMExtractor, that extracts medical conditions from SPLs. Instead of MetaMap, they employ conditional random fields, a statistical machine learning approach, for the extraction. The system was trained on a corpus containing 6611 manually annotated medical conditions. The authors report 0.92 precision, 0.73 recall, and 0.82 F-score, similar to the performance reported by Fung et al. [8].

Adverse drug events. Duke et al. [10] developed a tool, called SPLICER, to extract and codify adverse drug reaction information from SPLs. Tagging of adverse drug events (ADEs) is accomplished by a set of specific rules tailored to the different text sections and formatted structures (e.g., tables, lists) of the SPL. SPLICER demonstrated high accuracy in ADE extraction. High statistical scores (recall, precision and F-measure of 0.96, 0.97 and 0.97 respectively) show that rules-based systems perform well for specific information extraction goals. Kuhn et al. developed SIDER[3], a publicly available ADE knowledge base, which was obtained by extracting ADE information from SPLs [11]. Specific sections of the SPLs were

[3] SIDER - http://sideeffects.embl.de/ - accessed 10/20/2014

analyzed by text mining tools using a dictionary of side effects derived from the UMLS Metathesaurus. SIDER contains 62,269 drug–adverse event pairs and covers a total of 888 unique drugs and 1450 distinct side effects. The performance of the extraction tools is not reported. Smith et al. [12] reflect on the challenges in identifying pharmacovigilance information from multiple sources, including the SPLs. They processed SPLs using the KnowledgeMap Concept Identifier [13], an NLP tool developed at Vanderbilt University. The authors highlight complex logical and temporal sentence structures in SPLs, which standard NLP approaches currently fail to handle properly.

2.2.1. Extraction of drug-drug interactions from biomedical text

As illustrated by the success of recent challenges (SemEval 2011, SemEval 2013), drug-drug interaction extraction from textual sources is an active field of research in medical informatics. The statistical scores reported in these challenges are much lower than those reported for drug indication extraction, underscoring the inherent difficulty to accurately classify relations between drugs. Most participants in the SemEval challenges have used machine learning approaches to extracting DDI information, particularly support vector machine classifiers. (An overview of the existing methods for DDI extraction can be found in [14] [15].) The sources of DDI information used in the SemEval challenges have included the biomedical literature (MEDLINE®) and DrugBank [16]. Surprisingly, the DailyMed structured product labels, used for extracting drug indications and adverse drug events, have not yet been used as a source for the extraction of DDI information.

2.2.2. Specific contribution

As we have seen, the SPLs have been used as a source for other elements of drug information (indications, ADEs), but not DDI. Similarly, DDI extraction has been applied to different sources

of biomedical text, but not the SPLs. The main contribution of this work is to bridge this gap, i.e., to propose an approach to extracting DDI information from the DailyMed structured product labels. An additional contribution is the refinement of DDI extraction through role classification within a drug interaction (i.e. the identification of object and precipitant drugs). Finally, we share with the community a manually annotated DDI corpus of 180 SPLs developed in the course of this work.

2.3. Resources used for evaluation purposes

We use several resources for the evaluation of our DDI extraction framework. In addition to our own corpus, we use two external references, namely the corpora used in the SemEval challenges and a list of high-priority DDIs identified for clinical decision support purposes.

2.3.1. SemEval Corpus for Drug and DDI Extraction

The annotated corpus developed as the gold standard for the SemEval 2011 and 2013 challenges (drug entity recognition and drug-drug interaction classification) [17] has been published recently. This corpus contains 792 texts selected from the DrugBank database and 233 MEDLINE abstracts. It was annotated with a total of 18,502 instances of pharmacologic substances and 5028 DDIs, including both pharmacokinetic (PK) and pharmacodynamic (PD) interactions. The corpus enumerates all drug pairs in a sentence. The following DDI information is recorded wherever applicable: “mechanism” (how the interaction occurs), “effect” (the consequence of the interaction), “advice” (recommendation or advice) and “int” (when no further information is mentioned). The specific spans that identify the textual evidence for an interaction are not mentioned in the corpus. We use this resource in our evaluation in order to assess whether our approach to extracting DDIs from SPLs can be applied to different corpora.

2.3.2. ONC High-Priority DDIs

Given the multitude of overlapping drug interaction resources, a set of high-severity, clinically significant drug-drug interaction resource was needed to serve as a reference for interactions that every clinical decision system should contain. An expert panel under the supervision of Bates et al. was convened to curate existing DDI lists and identify high-priority DDIs for the Office of the National Coordinator for Health Information Technology (ONC) as part of the *Meaningful Use* incentive program [18]. Candidate DDIs were assessed by the panel based on the consequence of the interaction, severity levels assigned to them across various medication knowledge bases, availability of therapeutic alternatives, monitoring/management options, predisposing factors, and the probability of the interaction based on the strength of evidence available in the literature. The list contains 360 interacting pairs of individual drugs corresponding to 88 distinct drugs. We use this resource in our evaluation in order to assess the degree to which DDIs from these lists can be extracted from SPLs.

3. Materials

Our corpus consists of 180 cardiovascular DailyMed structured product labels independently annotated and reviewed by four experts. We present the annotated corpus we created to support our supervised machine learning approach, from the perspective of the annotation schema, the annotation process, its characteristics and inter-annotator agreement.

3.1. NLM DDI Annotation Schema

DDIs are analyzed at the sentence level without performing anaphora resolution. In other words, both drug entities involved must be present in the sentence, not referred to through an anaphor.

Drug entities. We annotate *Pharmacologic substances*, including *drugs*, *drug classes and other substances (e.g. food)*. Unlike SemEval, we do not distinguish between generic names and brand names (they are both annotated as *Drug*), because the distinction is generally not significant for DDIs. We annotate both standard drug classes (i.e., found in standard drug classification systems, e.g., *anti-hypertensives*) and non-standard drug classes (e.g., *drugs that may be indicated for the treatment of the cirrhosis of the liver). Substances* refer to any material entities that are not drugs or drug classes. These include foods, nutritional supplements and other things that can be found in the environment (e.g. grapefruit juice, alcohol etc.).

DDI roles. For the roles of drugs in the interaction, we reuse the schema from [19] (i.e., object and precipitant for the role of interacting drugs or substances). The object of interaction is a drug, drug class, or substance whose effect is altered by the precipitating entity. The precipitant drug, drug class, or substance is the entity that alters the pharmacologic and/or other action of the object entity. In our study, we derived a new schema to represent DDI knowledge. The most general mention of a DDI is the caution interaction. This is roughly equivalent to "advice" and "int" in the SemEval dataset. (We decided not to distinguish these, as they both imply some degree of caution without mentioning specifics.) We categorize interactions into *increase* and *decrease* interactions, according to the polarity of the effect of the precipitant on the object drug. This distinction is critical for clinical decision support, because an action may be required to maintain the therapeutic effect or minimize the toxicity of the object drug. *Increase/decrease* interactions are generally due to some pharmacokinetic mechanism, so the closest match with the SemEval schema would be "mechanism", which encompasses both types of interactions. We use *specific interaction* to capture any specific reaction resulting from a DDI. This is similar to "effect" in the SemEval schema.

3.2. Annotation Process

Two expert annotators (a medical librarian, SES, and a medical doctor, LR, both trained in medical informatics) carried out the annotation task. The Brat rapid annotation tool [20] was used to support the annotation process. The two experts independently annotated all the sentences extracted from the relevant sections of the 180 SPLs. Two experts (KWF and DDF) with both medical and informatics expertise then reviewed the annotations from the two annotators and reconciled them as necessary.

3.3. Corpus characteristics

The characteristics of our DDI corpus are listed in Table 4. For the 180 SPLs, a total of 8444 drug entities and 5059 interactions were annotated. With 54.2%, individual drugs have the highest frequency of occurrence in the corpus. Of note, drug classes are also encountered very frequently (33.3%). The most frequent interactions are specific interactions (52.2%), followed by caution interactions (25.7%).

Our corpus is available in a format similar to the SemEval corpus. Our format follows the standoff annotation principle in which the original sentence text is preserved and all entities are stored as offsets. Our corpus also contains negative examples, corresponding to sentences containing two drug entities but no DDI annotation. Our annotated corpus can be downloaded at http://lhce-brat.nlm.nih.gov/NLMDDICorpus.htm.

3.4. Inter-Annotator Agreement

In order to assess the quality of our gold standard, we computed an inter-annotator agreement score for each type of drug entity and DDI role. Our inter-annotator agreement score is measured by the F-measure, when considering the annotations of the first annotator as the reference. Detailed inter-annotator agreement scores are listed in Table 1. Overall, the scores range between

.72 and .90, reflecting good agreement between the two annotators, with the exception of the exact span of text used as evidence, for which the agreement is lower (.56), as can be expected for something that specific.

Table 1. Characteristics of our DDI corpus and Inter-annotator agreement scores.

	Entities	**Total number**	**%**	**Inter-Annotator Agreement**
Drug entities	Drug	4584 (592 distinct)	54.2%	0.81
	Drug Class	2816 (670 distinct)	33.3%	0.84
	Substance	221 (33 distinct)	2.7%	0.82
	Span	823 (290 distinct)	9.8%	0.56
	Total	**8444**	**100%**	
DDI roles	Specific Interaction	2595 (560 distinct triggers)	52.2%	0.79
	Caution Interaction	1308 (204 distinct triggers)	25.7%	0.72
	Increase Interaction	894 (289 distinct triggers)	17 %	0.84
	Decrease Interaction	262 (128 distinct triggers)	5%	0.9
	Total	**5059**	**100%**	

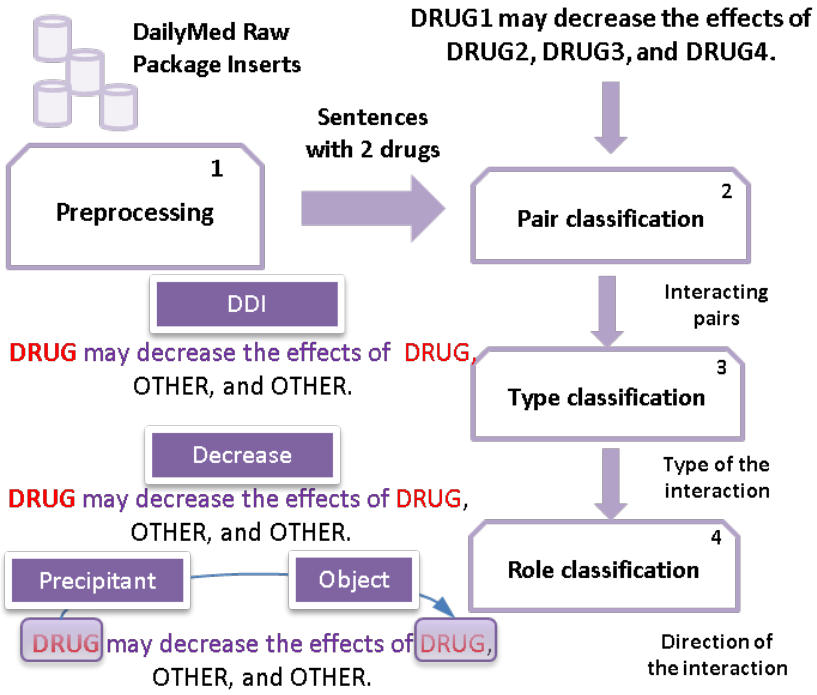
4. Methods and Results

In this section, we present the details of our machine learning approach, as well as our evaluation strategy (against our own corpus, the SemEval corpus, and the ONC high-priority list of DDIs).

4.1. Machine Learning Approach for DDI Extraction

DDI extraction can be thought of as relation classification (i.e. the identification of the type of relation between two entities in text). As mentioned earlier, DDIs can be expressed in a variety of ways, making machine learning techniques the method of choice for the automatic extraction of DDIs. Our DDI extraction classifiers perform the following steps: (i) recognize if two drugs within a sentence are in a DDI relation or not; (ii) recognize the specific type of the DDI relation and (iii) recognize the direction of the DDI relation (i.e., which drug is the object and which is the precipitant). Using LIBSVM [21], an open-source Java implementation of support vector machines and inspired by the jSRE system [22], we trained a classifier for each subtask. Every classifier was trained with shallow features, such as stems, n-grams of stems, part-of-speech (POS) tags, n-grams of POS tags and specific orthographic information about tokens. The feature space is described in more details below. An overview of our three-step classifier is illustrated in Figure 1.

Figure 1. Multi-step Classification Framework



4.1.1. Preprocessing

In the preprocessing step, we transform the original sentences containing more than 2 drug entities by creating one sentence for each pair of drug entities, allowing the classifier to focus on one pair per sentence. Consider the sentence *"Warfarin is contraindicated with NSAIDs and hypertensives."* In this case, the interaction (or its absence) between two drug entities has to be classified for three pairs: *(Warfarin, NSAIDs)*, *(Warfarin, hypertensives)* and *(NSAIDs, hypertensives)*. At the same time, we also abstract away from the specific names of drug entities, replacing them by the label DRUG (for the two drug entities in a pair of interest) and OTHER (for other drug entities). In practice, for the pair *(Warfarin, hypertensives)*, we create the following instance of the original sentence *"DRUG is contraindicated with OTHER and DRUG."*. One additional sentence is created for each of the other pairs.

4.1.2. Multi-Step Classification

The classification process is composed of three sub-classifiers, each one providing the input for the next. The first classifier performs Pair classification (1). It takes pre-processed sentences for each candidate pair as input and labels each pair as *interacting* or *not interacting*. The second classifier performs Type Classification (2). It takes each interacting pair as input and assigns it a specific interaction type. The final classifier performs Role Classification (3). It classifies the drug entities within a pair and labels their roles (object, precipitant). Role labelling is especially important for the *increase* and *decrease* interactions. All classifiers use the linear kernel setting of the LIBSVM library. Grid search with a 10-fold cross-validation is used for estimating the best cost parameter for our linear kernel.

4.1.3. Features Spaces for the Multi-Step Classifier

We distinguish between global and local context features. The global context feature space uses information from the whole sentence, while the local context features represent information from the neighborhood of the entities.

Global context feature space. The global context feature space divides the sentence into three fragments relative to the position of the two drug entities. It leverages the observation that a relation is often expressed using the words before and between the entities (“fore-between”), only between them (“between”), or between and after them (“between-after”). Features extracted from each sub-space are the stems of words, n-grams of the stems of words, part-of-speech (POS) tags of words and n-grams of POS tags of words. We introduce sparse stems, which is a bigram of stems, meaning that we connect the token at position i with the token at position $i+2$. An illustration of the global context feature space is in Figure 2. A sparse binary vector is used to store each feature space. If a feature occurs, it is represented in this vector as the value 1 at a

given position in the feature space. We also use subtrees of the parse tree of the sentence (i.e., all nodes of the parse tree along with all their descendants) that correspond to a given region (fore-between, between, between-after). In order to use them with a linear kernel, we transform these subtrees into sequences of nodes using the pre-order traversal. Since only the structure of the tree is important here, terminal nodes (stems) are removed. The list of global context features for the example sentence ("between" space) is shown in Table 2. We normalize all feature spaces by dividing each value by the Euclidean norm of the vector. The implementation of our feature space model was inspired by [22], where more detail is provided.

Figure 2. Global Context Feature Spaces

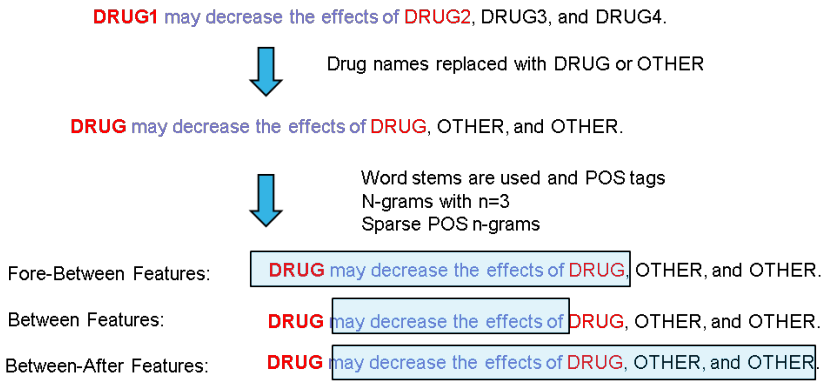


Table 2. Global Context Features for the example sentence (“between” space)

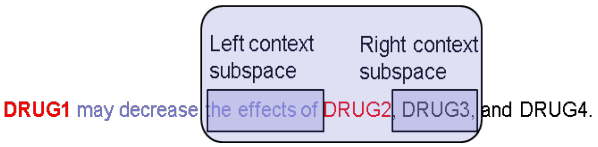
Feature type	**Features**
Unigram	Stems: may, decrease, the, effect, of POS tags: MD, VB, DT, NNS, IN
Bigram	Stems: may_decrease, decrease_the, the_effect, effect_of POS tags: MD_VB, VB_DT, DT_NNS, NNS_IN
Trigram	Stems: may_decrease_the, the_effect_of POS: MD_VB_DT, VB_DT_NNS, DT_NNS_IN
Subtrees	Between subspace: “may decrease the effects of”. Parse tree: (ROOT (SINV (VP (MD) (VP (VB))) (NP (NP (DT) (NNS)) (PP (IN))))) Features examples: DT_NNS, PP_IN, NP_NP_DT_NNS_PP_IN, VP_MD_VB, NP_DT _NNS.
Sparse bigrams of stems	may_the, decrease_effect, the_of

Local context features space. The local context features represent information from the neighborhood of the drug entities. These features describe the context of the drug entities within a window of several tokens on the left and right side of the entity (Figure 3). We empirically determined the optimal window size as described in Section 4.2.1.1. For each token, the original token, its stem, POS tag and orthographic class are extracted and included in the feature space. The orthographic class feature can cover a wide range of orthographic categories, such as

whether the token begins with uppercase or lowercase, is capitalized, has punctuation, or is a numeric value.

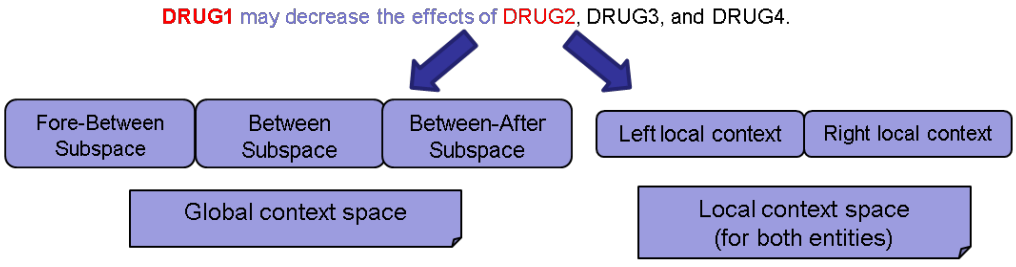
The original token, its stem and POS tag together with the corresponding role label are first added, along with position information (relative to that of the drug entity). The orthographic categories of entities relative to a given position are also added (e.g. is it a punctuation mark, a word, a number, lowercase or uppercase?). The UMLS semantic types for the entities are also included in the feature space. The local context space is generated for each of the two drug entities in a candidate pair.

Figure 3. Local context of an entity



Combined feature space. The global feature and the individual local context feature spaces for each of the two drugs are combined. This combined feature vector is then used by the linear kernel. The final feature vector is depicted in Figure 4.

Figure 4. Feature vector composition



4.2.Performance of the Machine Learning Framework

In order to assess the performance of our machine learning framework for extracting DDIs, we perform three experiments. First, we evaluate the performance on our corpus for the 3 main

tasks: pair categorization, type categorization and role categorization. Then we apply our DDI extraction framework to the SemEval DDI corpus, in order to assess the framework portability. Finally, we assess the ability of our framework to extract DDIs for drugs other than the cardiovascular drugs in our annotated corpus. To this end, we attempt to identify in structured product labels the DDIs from the ONC high-priority list.

4.2.1. Evaluation on the NLM CD (Cardiovascular) Corpus

We used LIBSVM with a linear kernel to perform all classification tasks. The DailyMed dataset was divided into 70% training and 30% test set, using statistical sampling for each extraction task. A 10-fold cross validation on the training subset was performed in order to compute the cost parameter C, part of the linear kernel. The cost parameter value that yielded the highest accuracy was 2. This parameter is used in all the experiments below. Results for each sub-task are shown in separate sections.

4.2.1.1. Pair Classification

Feature Space Parameter Selection. In order to find the best configuration of our feature space, we designed a set of experiments to study the contribution of the n-gram and window sizes of our feature spaces. The best performance (precision = 0.818; recall = 0.869; F-score = 0.842) was obtained with an n-gram size of 3 for the global context features, and with a window-size of 3 for the local context features. The results of the calibration experiments are summarized in Appendix 1. Unless mentioned otherwise, these parameters are used in the following experiments.

Contribution of Individual Feature Spaces. The best performance was obtained with the combined feature space, integrating both the global and local features. In particular, the combined feature space significantly improves precision. In terms of F-measure, the combined

feature space outperforms the global features spaces by 3% and the local feature space by 6%. The contribution of individual feature spaces is summarized in Appendix 2.

Contribution of global context feature categories. The combination of global features resulting in the best performance includes n-grams of stems, n-grams of POS tags and sparse stems. The contribution of the various combinations studied is presented in Appendix 3.

4.2.1.2.Type Classification

The overall classification accuracy was 88.068% (716/813), highest for *specific* and *caution* interactions (> .90) and lowest for *decrease* interactions (.71). The details of the performance for each DDI type are summarized in Table 3.

Table 3. Results for type classification

Type of interaction	**Total**	**Precision**	**Recall**	**F-measure**
Specific	813	0.893	0.926	0.909
Caution	813	0.918	0.888	0.903
Increase	813	0.843	0.897	0.869
Decrease	813	0.833	0.625	0.714

The overall effect of combining the two feature spaces (global and local context) seems beneficial for type classification (Appendix 4). The recall significantly improves for *decrease* interaction, while there is small decline in precision for *specific* interaction and recall for *caution* interaction.

4.2.1.3 Role Classification

The overall classification accuracy was 100% for role classification, i.e. for determining which of the two drug entities is the object vs. precipitant drug. The local context kernel does not seem to have any effect on this classification task in our dataset. Our results are summarized in Appendix 5.

4.2.2. Evaluation on the SemEval Corpus

We evaluate our machine learning approach on the SemEval Drug interaction Task 9 (2013) corpus in order to see the robustness of our feature space when applied to text other than structured product labels. The SemEval corpus is composed of two types of texts, from DrugBank and MEDLINE, which we treat as a combined corpus. The classifier was retrained on this combined corpus. As with our own corpus, a 10-fold cross-validation was performed for cost parameter estimation. The best cross-validation accuracy, 91.59%, was obtained for C value of 3.. Our performance is reported in Table 4 together with highest scores obtained in the competition. Overall, our framework systematically outperforms the best method in the SemEval competition, demonstrating that our method is able to extract DDI relations not only from DailyMed, but also from different types of text (MEDLINE, DrugBank).

Table 4. Results for the SemEval corpus

Medline + DrugBank Accuracy = 77.41585233441911% (713/921) (classification) Mean squared error = 0.4223669923995657 (regression) Squared correlation coefficient = 0.5642510733129408 (regression)	Type	Tp	Fp	fn	Total	Precision	Recall	F-score	Best F-scores
	effect	175	49	43	921	**0.781**	**0.803**	**0.792**	**0.662**
	mechanism	229	41	58	921	**0.848**	**0.798**	**0.822**	**0.679**
	advice	268	113	54	921	**0.703**	**0.832**	**0.762**	**0.692**
	int	41	5	53	921	**0.891**	**0.436**	**0.586**	**0.547**

4.2.3. Finding evidence for the ONC High-Priority DDIs in DailyMed Structured Product Labels

The ONC high-priority drug list contains 360 interacting pairs, involving 88 distinct ingredients from a variety of pharmacologic classes [18]. 84 of these drugs are covered by DailyMed. We hypothesize that these DDIs should be described in the DailyMed Structured Product Labels, and we use this list to evaluate the capacity of our tool to find evidence of drug-drug interactions in the DailyMed labels. Our goal with this experiment is to assess the performance of our DDI extraction framework beyond the corpus of cardiovascular drugs on which it was trained.

We selected one DailyMed Structured Product Label for each of the 84 drugs not already covered by our corpus. Priority was given to labels for injectable forms when available, to oral forms otherwise (as opposed to topical forms, for which DDIs may not be systematically mentioned). We ran the extraction process on this dataset (2554 sentences) and identified 620 sentences containing a drug interaction. Of these, 104 sentences contained the two drugs corresponding to a DDI from the ONC high-priority list, 310 sentences contained only one of the two drugs corresponding to a DDI from the ONC high-priority list, and 206 sentences did not contain any of the two drugs corresponding to a DDI from the ONC high-priority list. Overall, of the 360 DDIs from the ONC high-priority list, only 59 pairs could be found within the 104 sentences in which the two drugs were found.

The performance on this dataset seems mediocre, since we could find evidence for only 59 of the 360 DDIs. However, the main problem here is not really that our system fails to extract DDI relations expressed between two specific drugs in a sentence, but rather that a majority of these DDIs are expressed without mentioning the two specific drugs. In the discussion section, we present a specific analysis of these sentences and make recommendations for future work.

5. Discussion

In this section we evaluate the significance of our work, present a failure analysis (false positive and false negative cases in DDI extraction), and analyze specifically the cases of missed evidence for the ONC High-Priority DDIs. We also briefly discuss the limitations and possible applications of this work.

5.1. Significance

Overall. This investigation is the first to consider the DailyMed structured product labels as a source of DDIs for automatic extraction. The performance of our DDI extraction framework is competitive with that of state-of-the-art machine learning systems reported in the SemEval DDI challenge. Moreover, we achieve good performance not only on our DailyMed corpus, but also on the corpus used in the SemEval DDI challenge itself, demonstrating that our DDI extraction framework is effective beyond the corpus on which it was originally developed. Of note, lower recall is observed for *decrease* interactions, for which there are few annotations. The performance observed on the task of finding evidence in DailyMed for the ONC high-priority DDIs is both surprising and disappointing. However, it turns out that the expression of the DDIs through classes (rather than individual drugs) is responsible for it. In other words, our system did not extract these DDIs, mostly because the two specific drugs were not mentioned together in a sentence. A detailed analysis of these cases is provided later.

Parameter selection and cross-validation accuracy. Among the trained models, we selected the model maximizing both F-measure and precision (n-gram = 3, window-size = 3). These parameters provide the best and most robust configuration for the automatic extraction of drug-drug interactions from DailyMed, as they minimize the number of false positives to be reviewed by curators subsequently. As expected, a small n-gram size generally favored recall, while a

larger n-gram favored precision. We analyze the generalization potential by comparing 10-fold cross-validation accuracy scores with those on the test sets. On our DailyMed annotated DDI corpus, we report an average accuracy of 97.20% (training set) and 95.18% (test set) for pair classification, 99% and 100% (respectively) for role classification, and 91% and 88.19% (respectively) for type classification. As expected, 10-fold cross-validation accuracy is lower on the combined SemEval DDI corpus because of its diversity (DrugBank and MEDLINE).

5.2. Failure Analysis of DDI extraction

Here are some examples of misclassified sentences, along with typical reasons for misclassification. False positives often occur in long sentences with complex structures. Several errors are also due to inaccurate concept identification by MetaMap. Finally, cases in which the two drug entities identified are a drug class and one of its instances could be eliminated with additional filtering. Examples of false positives are presented in Table 5. False negatives are generally associated with very short sentences or complex sentences containing multiple interactions. Examples of false negatives are shown in Table 6.

Table 5. False Positives

Sentence in DailyMed	False Positive Pair	Possible Explanation
Hypokalemia may develop with LASIX, especially with brisk diuresis, inadequate oral electrolyte intake, when cirrhosis is present, or during concomitant use of corticosteroids, ACTH, licorice in large amounts, or prolonged use of laxatives.	corticosteroids-laxatives	Long sentence with several drugs.
In patients with an activated renin-angiotensin-aldosterone system, such as volume - or salt-depleted patients receiving high doses of diuretics, symptomatic hypotension may occur in patients receiving renin-angiotensin-aldosterone system (RAAS) blockers.	diuretics-renin-angiotensin-aldosterone	MetaMap wrongly identifies an entity renin-angiotensin-aldosterone is not a drug
As with other beta blockers, when discontinuation of TENORMIN is planned, the patients should be carefully observed and advised to limit physical activity to a minimum.	beta blockers-TENORMIN	Drug instance of a drug class, not drug-drug class interaction. Such cases can be captured using standard terminologies.

Table 6. Examples of False Negatives

Sentence in DailyMed	False Negative Pair	Possible Explanation
Phenytoin decreases serum amiodarone levels.	Phenytoin-amiodarone	Short sentences may lead to false negatives due to lack of context.
Use amiodarone with caution in patients receiving - receptor blocking agents (e.g. , propranolol, a CYP3A inhibitor) or calcium channel antagonists (e.g., verapamil, a CYP3A substrate, and diltiazem, a CYP3A inhibitor) because of the possible potentiation of bradycardia, sinus arrest, and AV block, if necessary, amiodarone can continue to be used after insertion of a pacemaker in patients with severe bradycardia or sinus arrest.	amiodarone-calcium channel antagonists	Long sentences describing multiple interactions, composed of drug classes with examples of drug class members. Such sentences need additional processing, first decomposed and simplified. This example would then be transformed into several short sentences that are much simpler for the classifier.

5.3. Analysis of missed evidence for the ONC High-Priority DDIs

We conducted a specific analysis of the missed evidence for the ONC High-Priority DDIs. As mentioned earlier, in many cases the sentences found in DailyMed for these drugs contained the mention of only one of the two individual drugs from the ONC high-priority DDIs. Causes for the missing mention of the other specific drug include reference to it through a drug class and anaphoric reference (e.g., through a pronoun). While our system is designed to extract DDIs between a specific drug and a drug entity, it does not resolve drug classes to their individual drug members. It does not perform anaphora resolution either.

We performed further analysis of 310 DDI sentences in which only one specific drug was found, in order to determine which of these two causes was primarily responsible for missed evidence for the ONC High-Priority DDIs. In all but one case (309), the second drug entity was referred to through a drug class. The remaining case corresponds to an anaphoric reference.

We further investigated the mentions of drug classes within the 309 sentences. We attempted to resolve drug class names against standard terminologies, such as ATC, MeSH and NDF-RT, using the RxClass API[4]. Only 7 standard classes could be found. Using the RxClass API, we retrieved the members of these drug classes and, in all cases, were able to find the second drug of the ONC high-priority DDI among the members of the class we identified. Most of the classes that were not present in standard terminologies corresponded to classes defined in relation to the Cytochrome P450 (CYP). The majority of them (250) were inducers or inhibitors of CYP enzymes (e.g., *"Coadministration of a CYP3A4 and UGT1A1 inhibitor has the potential to increase systemic exposure to SN-38, the active metabolite of irinotecan."*), which we extracted

[4] RxClass - http://mor.nlm.nih.gov/RxClass/ - accessed 10/29/2014

using simple string matching techniques. Members of these classes were resolved against the Cytochrome P450 Drug Interaction Table (www.drug-interactions.com). After retrieving the drugs corresponding to these CYP classes, we were able to find the second drug of the ONC high-priority DDI among the members of the class we identified in 240 of the 250 cases.

This specific analysis demonstrates the prevalence of drug classes, including non-standard drug classes, in the expression of DDIs in DailyMed structured product labels. It also emphasizes the need for including resolution of these classes as part of DDI extraction systems, in particular when reference DDIs are expressed at the level of individual drugs.

5.4. Limitations and future work

Although our system shows good performance on several DDI corpora, we also noted several limitations, namely lack of anaphora resolution (for drugs mentioned through a reference in the DDI sentence) and lack of resolution of drug classes. We are currently addressing anaphora resolution, which is made possible by the fact that anaphoric references were annotated in our corpus (but originally not taken into account by our DDI extraction system). The analysis presented earlier demonstrates the possibility of resolving most drug classes, standard or not. However, specific processing of these drug classes is required, because non-standard drug classes are generally not recognized by MetaMap.

5.5. Application Scenarios

The main application of this research is to support the curation of DDIs systematically extracted from DailyMed structured product labels. This use case was suggested to us by our partner, the FDA. Beyond the pilot phase reported on in this paper, systematic extraction would require processing all the DailyMed structured product labels. The collection of DDIs obtained after curation is expected to be an important resource for clinical decision support.

6. Conclusion

Our work is novel as the first attempt to extract drug-drug interactions from the DailyMed structured product labels, piloted on cardiovascular drugs. We provide the first publicly available toolkit that transforms a product label into a structured list of DDIs. We also make the annotated corpus of 180 structured product labels we used to develop our extraction process available for download at http://lhce-brat.nlm.nih.gov/NLMDDICorpus.htm. Our feature space, which is an extension of the shallow linguistic kernel proposed by [27] shows good generalization capability, both to other corpora and to other drug classes. Our future work includes the extraction of DDIs from the entire DailyMed dataset, as well resolution of drugs referred to anaphorically and through drug classes. We therefore consider our work as a significant step towards the design of efficient, scalable and robust methods for DDI extraction.

7. Acknowledgments

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine. The authors would like to thank Bill Hess and Dr. Randy Levin from the U.S. Food and Drug Administration for providing the motivation for this work. Our thanks also go to Ms. Josephine O'Para for technical support with the annotation tool.

References

[1] S. Phansalkar, A. Wright, G. J. Kuperman, A. J. Vaida, A. M. Bobb, R. A. Jenders, T. H. Payne, J. Halamka, M. Bloomrosen, and D. W. Bates, "Towards Meaningful Medication-Related Clinical Decision Support: Recommendations for an Initial Implementation," *Applied Clinical Informatics,* vol. 2, pp. 50-62, 2011.

[2] E. M. van Mulligen, A. Fourrier-Reglat, D. Gurwitz, M. Molokhia, A. Nieto, G. Trifiro, J. A. Kors, and L. I. Furlong, "The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships," *J Biomed Inform,* vol. 45, pp. 879-84, Oct 2012.

[3] L. E. Hines, D. C. Malone, and J. E. Murphy, "Recommendations for generating, evaluating, and implementing drug-drug interaction evidence," *Pharmacotherapy,* vol. 32, pp. 304-13, Apr 2012.

[4] S. Duda, C. Aliferis, R. Miller, A. Statnikov, and K. Johnson, "Extracting drug-drug interaction articles from MEDLINE to improve the content of drug databases," *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium,* pp. 216-20, 2005.

[5] J. R. Nebeker, P. Barach, and M. H. Samore, "Clarifying adverse drug events: a clinician's guide to terminology, documentation, and reporting," *Annals of internal medicine,* vol. 140, pp. 795-801, May 18 2004.

[6] L. E. Hines and J. E. Murphy, "Potentially harmful drug-drug interactions in the elderly: a review," *The American journal of geriatric pharmacotherapy,* vol. 9, pp. 364-77, Dec 2011.

[7] D. W. Bates, J. M. Teich, J. Lee, D. Seger, G. J. Kuperman, N. Ma'Luf, D. Boyle, and L. Leape, "The impact of computerized physician order entry on medication error prevention," *J Am Med Inform Assoc,* vol. 6, pp. 313-21, Jul-Aug 1999.

[8] K. W. Fung, C. S. Jao, and D. Demner-Fushman, "Extracting drug indication information from structured product labels using natural language processing," *J Am Med Inform Assoc,* vol. 20, pp. 482-8, May 1 2013.

[9] Q. Li, L. Deleger, T. Lingren, H. Zhai, M. Kaiser, L. Stoutenborough, A. G. Jegga, K. B. Cohen, and I. Solti, "Mining FDA drug labels for medical conditions," *BMC medical informatics and decision making,* vol. 13, p. 53, 2013.

[10] J. D. Duke and J. Friedlin, "ADESSA: A Real-Time Decision Support Service for Delivery of Semantically Coded Adverse Drug Event Data," *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium,* vol. 2010, pp. 177-81, 2010.

[11] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork, "A side effect resource to capture phenotypic effects of drugs," *Molecular systems biology,* vol. 6, p. 343, 2010.

[12] **J. C. Smith, J. C. Denny, Q. Chen, H. Nian, A. Spickard, 3rd, S. T. Rosenbloom, and R. A. Miller, "Lessons learned from developing a drug evidence base to support pharmacovigilance," *Applied Clinical Informatics,* vol. 4, pp. 596-617, 2013.**

[13] **J. C. Denny, J. D. Smithers, R. A. Miller, and A. Spickard, 3rd, ""Understanding" medical school curriculum content using KnowledgeMap," *Journal of the American Medical Informatics Association : JAMIA,* vol. 10, pp. 351-62, Jul-Aug 2003.**

[14] **I. Segura-Bedmar, P. Martinez, and M. Herrero-Zazo, "SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013)," in *In Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013).* 2013.**

[15] **I. Segura-Bedmar, P. Martinez, and M. Herrero-Zazo, "Lessons learnt from the DDIExtraction-2013 Shared Task," *Journal of biomedical informatics,* vol. 51, pp. 152-64, Oct 2014.**

[16] **M. Herrero-Zazo, I. Segura-Bedmar, P. Martinez, and T. Declerck, "The DDI corpus: an annotated corpus with pharmacological substances and drug-drug interactions," *Journal of biomedical informatics,* vol. 46, pp. 914-20, Oct 2013.**

[17] **M. Herrero-Zazo, I. Segura-Bedmar, P. Martinez, and T. Declerck, "The DDI corpus: an annotated corpus with pharmacological substances and drug-drug interactions," *J Biomed Inform,* vol. 46, pp. 914-20, Oct 2013.**

[18] **D. A. Phansalkar S, Bell D, Yoshida E, Doole J, Czochanski M, Middleton B, Bates DW. and S.-O.-d. E. A. 26., "High-priority drug-drug interactions for use in electronic health records.," *J Am Med Inform Assoc. ,* 2012 Apr 26 2012.**

[19] **R. Boyce, G. Gardner, and H. Harkema, "Using natural language processing to identify pharmacokinetic drug-drug interactions described in drug package inserts," presented at the Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, Montreal, Canada, 2012.**

[20] **S. P. Pontus Stenetorp, Goran Topic, Tomoko Ohta, Sophia Ananiadou and Jun'ichi Tsujii, " brat: a Web-based Tool for NLP-Assisted Text Annotation," *In Proceedings of the Demonstrations Session at EACL 2012,* 2012.**

[21] **C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.,* vol. 2, pp. 1-27, 2011.**

[22] **A. L. Claudio Giuliano, Lorenza Romano, "Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature," *In 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL '06),* pp. 401-408 2006.**

Appendices

1. Feature Space Parameter Selection

Feature set	Configuration N=n-gram size W=window size	Precision	Recall	F-measure
Bag of words Features (baseline)	N=1 N=2 N=3	0.573 0.689 0.689	0.902 0.894 0.888	0.701 0.778 0.776
Global Context Features (n-grams of stems)	N=1 N=2 N=3	0.687 0.738 0.728	0.885 0.861 0.855	0.774 0.795 0.786
Global Context Features (n-grams of stems, POS tags and sparse stems)	N=1 N=2 N=3	0.676 0.744 0.793	0.887 0.853 0.846	0.767 0.795 0.819
Shallow Linguistic (n-grams of stems and POS tags, sparse stems and local context features)	N=1, W=3 N=2, W=3	0.797 0.808	0.872 0.869	0.833 0.837
Shallow Linguistic (95.52 CV accuracy) **(same features as above)**	**N=3, W=3**	**0.818**	**0.869**	**0.842**

2. Contribution of Individual Feature Spaces

Type of feature space	Precision	Recall	F-measure
Local context only (LC) Window size 3	0.74	0.826	0.781
Global context only (GC) n-gram size 3	0.778	0.854	0.814
SL (LC + GC) Window, n-gram 3	**0.818**	**0.869**	**0.842**

3. **Contribution of global context feature categories**

Feature category	Precision	Recall	F-measure
1. N-grams of stems only	0.766	0.864	0.812
2. + N-grams of POS tags	0.816	0.869	0.841
3. 1+2 + Sparse stems	**0.818**	**0.869**	**0.842**
4. 1+2 + Sparse POS Tags	0.797	0.861	0.828
5. 1+ 2+ 3 +4	0.805	0.854	0.829

4. **Contribution of features spaces for type classification**

Type of feature space	Type	Precision	Recall	F-measure
Local context only (LC) Window size 3	Specific	0.837	0.889	0.862
	Caution	0.865	0.851	0.858
	Increase	0.794	0.832	0.813
	Decrease	0.795	0.484	0.602
Global context only (GC) n-gram size 3	Specific	0.907	0.910	0.909
	Caution	0.898	0.913	0.905
	Increase	0.842	0.892	0.866
	Decrease	0.812	0.609	0.696
SL (LC + GC) Window, n-gram 3	**Specific**	**0.893**	**0.926**	**0.909**
	Caution	**0.918**	**0.888**	**0.903**
	Increase	**0.843**	**0.897**	**0.869**
	Decrease	**0.812**	**0.625**	**0.714**

5. Contribution of features spaces for Role classification

Type of feature space	Type	Precision	Recall	F-measure
Local context only (LC) Window size 3	1	0.769	0.938	0.845
	2	0.983	0.930	0.956
Global context only (GC) n-gram size 3	1	1	1	1
	2	1	1	1
SL (LC + GC) Window, n-gram 3	1	1	1	1
	2	1	1	1