Relation Mining in the Biomedical Domain using Entity-level Semantics

Kateryna Tymoshenko[1] **Swapna Somasundaran**[2] **Vinodkumar Prabhakaran**[3] **Vinay Shet**[4]

Abstract. This work explores the use of semantic information from background knowledge sources for the task of relation mining between medical entities such as diseases, drugs, and their functional effects/actions. We hypothesize that the semantics of medical entities, and the information about them in different knowledge sources play an important role in determining their interactions and can thus be exploited to infer relations between these entities. We capture entities' semantics using a number of resources such as Wikipedia, UMLS Semantic Network, MEDCIN, MeSH and SNOMED. Depending on coverage and specificity of the resources, and features of interest, different classifiers are learnt. An ensemble based approach is then used to fuse together individual predictions. Using a human-curated ontology as the gold standard, the proposed approach has been used to recognize ten medical relations of interest. We show that the proposed approach achieves substantial improvements in both coverage and performance over a distant supervision based baseline that uses sentence-level information. Finally, we also show that even a simple ensemble approach that combines all the semantic information is able to get the best coverage and performance.

1 INTRODUCTION

Relation mining in the biomedical domain attempts to find interactions between medical entities. This can enable Clinical Decision Support (CDS) systems in performing critical functions such as identifying potentially adverse drug interactions from patient health records. Adverse drug interactions may occur due to a wide variety of factors involving ingredients of the drugs, their mechanisms of action within the body, their physiological effects, contraindications with certain conditions, etc. It is therefore important to build relation mining systems that can recognize such interactions with good accuracy.

State of the art approaches to relation mining (e.g. [10, 18]) rely on human annotated corpora, where sentences containing entities of interest are annotated with their relation. This approach, however, is not feasible for our task due to the lack of human annotated corpora for all our clinical relations of interest.

In order to overcome this challenge, in this work, we exploit the hypotheses that *biomedical entities have certain inherent properties that are indicative of their interactions*, and *the way knowledge sources organize information regarding medical entities can be harnessed to infer their interactions*. Consequently, we exploit two different types of *entity-level semantics*. The first set of semantics correspond to the first hypothesis and is based on individual entity properties. For example, Aspirin, a drug, has a property of being anti-inflammatory, and anti-inflammatory drugs have the property of treating pain. Thus, by using this knowledge and the knowledge that Headache is a type of pain, we can infer that the entity Aspirin is likely to have a $treat$ relation with the entity Headache. The second set of semantics, corresponding to the second hypothesis, is based on the entity pair under consideration, and captures how information in standard knowledge sources links a given pair of entities. For example, a Wikipedia page for a drug typically mentions the diseases (or types of diseases) the drug treats in a "uses" section.

We test our hypotheses on the recognition of 10 different clinical relations from the National Drug File – Reference Terminology (NDF-RT)[5] using a number of knowledge sources such as Wikipedia and UMLS. We encode semantic features such as entity-category/taxonomy (derived from UMLS etc.) and entity-pair linkage information (derived from Wikipedia) into a machine learning algorithm. Based on the coverage and specificity of the resources and the features, we explore different feature combinations and construct different classifiers. Finally, we combine all the individual predictions using an ensemble approach.

Our investigations with entity-level semantic classifiers built using different knowledge source combinations reveal their strengths and weaknesses for large-scale biomedical relation mining. We compare our approach to distant supervision-based approaches that have been shown promising for relation mining between named entities (e.g. [14]). Experiments carried out over 97,000 entity pairs reveal that in the biomedical domain, distant supervision-based approaches that use sentence-level information face a number of challenges in terms of coverage and performance. Our approach that employs entity-level semantics from various knowledge sources is able to achieve substantial improvements in both: we get an average improvement of 44 percentage points in coverage and 39 percentage points in performance (Fmeasure). Finally, we show that even a simple ensemble approach that combines all the semantic information is able to get the best coverage and performance.

2 ENTITY-LEVEL SEMANTICS

Relation mining approaches for named entities such as Persons and Organizations have exploited human annotated corpora, such as ACE

[1] FBK-irst, via Sommarive 18, Povo (Trento), I-38123, Italy, email: tymoshenko@fbk.eu. This author contributed to the work presented in this paper while she was affiliated with Siemens Corporate Research

[2] Siemens Corporate Research, Princeton, NJ 08540, USA, email: swapna.somasundaran@siemens.com

[3] Columbia University, New York, NY 10027, USA email: vinod@cs.columbia.edu. This author contributed to the work presented in this paper while he was affiliated with Siemens Corporate Research

[4] Siemens Corporate Research, Princeton, NJ 08540, USA, email: vinay.shet@siemens.com

[5] `http://evs.nci.nih.gov/ftp1/NDF-RT/`

[3], to construct systems that leverage linguistic and contextual information within text surrounding a given pair of co-occurring entities. This approach is not feasible for our task due to the absence of an annotated corpus for our relations of interest. However, to our advantage, biomedical relations are characterized by the properties of the involved entities. Additionally, the clinical domain has a number of knowledge sources providing information about medical entities in an organized fashion. We call this entity-level semantics, and harness it to develop our relation mining system.

Our relation mining is motivated by the goal to assist Clinical Decision Support (CDS) Systems in identifying and flagging adverse drug interactions. Specifically, we focus on drugs and a subset of their interactions with other medical entities in the NDF-RT ontology. The medical entities of interest in this work are: Drugs, Diseases, Drug Pharmacology (Chemical) Class, Drug Physiological Effects, Drug Ingredients, Drug Mechanism of Action. Table 1 describes the relations of interest involving these entities. Notice that drugs can have different types of relations with the same types of medical entities (e.g. Mechanism of Action).

Table 1. Relations of interest from NDF-RT

Name	Description
may_treat	Drug A may treat Disease B
may_prevent	Drug A may prevent Disease B
may_diagnose	Drug A may diagnose Disease B
induces	Drug A induces Disease B
CI_with	Drug A is contraindicated (known to cause adverse reaction) with Disease B
has_Ingredient	Drug A has Ingredient B
has_PE	Drug A has Physiological Effect B
has_MoA	Drug A has Mechanism of Action B
CI_MoA	Drug A is contraindicated with Mechanism of Action of drug B
CI_ChemClass	Drug A is contraindicated with Chemical Class of drug B

As mentioned previously, entity-level semantics involves two different types of information. The first, *entity-specific semantics*, is based on the individual entity's properties and the second, *entity pair linkage*, is based on information on how the entities are linked in knowledge sources. For instance, the drug Aspirin is a type of analgesic (painkiller) drug that has the property of treating diseases (conditions) involving pain, such as Headache and Toothache. This is an example of the first type of entity-level semantics where the class and taxonomic information of the drug and the disease clue their interaction. As an example for the second type of entity-level semantics, let us consider the Wikipedia page for the drug Ibuprofen. The page mentions the disease Fever under "Medical Uses". Similarly, Wikipedia pages for drugs Paracetamol and Codeine also have "Medical Uses" sections where diseases that they cure are listed. Here, the manner in which a knowledge source such as Wikipedia links the two entities can clue to the type of relation between them.

We use Wikipedia[6], UMLS semantic network[7], and UMLS metathesaurus resources such as MEDCIN[8], SNOMED–CT[9] and MeSH[10] as our knowledge sources. All resources are used for extracting category and taxonomy information, while Wikipedia is used to capture linkage semantics.

3 SEMANTIC FEATURES

Our semantic features can be broadly categorized as entity-specific features and entity pair features. The former includes category/taxonomy-based features while the latter includes link-based features for entity pairs.

3.1 Entity-specific features

These features are based on the category of the entity and capture the class properties of the individual entities. Categories and taxonomy represent topical and semantic class information about the entities. Category features are extracted from all knowledge sources listed above. Some of the entity specific features are as follows.

- **wikiCategory**. This is a set of features that capture the category of the Wikipedia page corresponding to an entity e, and its ancestors in the Wikipedia category taxonomy up to two levels up. For instance, the page for `Aspirin` has categories `Acetate_esters` and `Antiplatelet_drugs`.
- **umlsPF**. These features capture the taxonomical path information in various UMLS knowledge sources. Path is calculated from an entity of interest e to the root of a specific UMLS source and represented as: [$root.node_n.node_{n-1}.<...>.node_0$], where $node_0$ is a direct parent of e in a source, and $node_{i+1}$ is a parent of $node_i$. *umlsPF* feature set also includes more generic subpaths of the full path shown above. For example, the following subpaths are also created as features: [$root$], [$root.node_n$], [$root.node_n.node_{(n-1)}$], ... , [$root.node_n.node_{(n-1)}.....node_1$]. Depending on the knowledge sources, there are different feature sets:
  - **umlsPF:::SNOMED**. This is umlsPF with SNOMED CT as the source. For instance, for *Aspirin*, umlsPF:::SNOMED would include [*Drug or medicament.Musculoskeletal system agent.Anti-rheumatic agent.Anti-inflammatory agent.Non-steroidal anti-inflammatory agent .Salicylate*].
  - **umlsPF:::MSH**. This is umlsPF with MeSH as the source. For instance, for *Aspirin*, it would include [*Chemicals and Drugs (MeSH Category).Organic Chemicals.Phenols.Hydroxybenzoic Acids.Salicylic Acids*].
  - **umlsPF:::MEDCIN**. This is umlsPF with MEDCIN as the source. For instance, for *Aspirin*, it would include [*therapy.medications and vaccines.analgesics.salicylates*].
- **umlsSemType**. This feature set captures the semantic types of an entity in the UMLS Semantic Network, and is similar to wikiCategory features. For example, *Aspirin* has UMLS semantic types *Organic Chemical* and *Pharmacologic Substance*.
- **umlsCUI**. This is the UMLS Concept Unique Identifier (CUI) of an entity, (e.g. *C0004057* for *Aspirin*) and captures the identity of the entity.

3.2 Entity pair linkage features

The entity pair linkage features capture how information about one entity refers to the other entity, or how both entities refer to other concepts that are common to them. We encode two different types of entity features using Wikipedia as the knowledge source. In this work, we only focus on the linking and sectioning information.

[6] `http://www.wikipedia.org/`
[7] `http://semanticnetwork.nlm.nih.gov/`
[8] `http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MEDCIN/`
[9] `http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html`
[10] `http://www.ncbi.nlm.nih.gov/mesh`

- **pairwiseLinkFeature**. These consider direct links between entities. There are two types of pairwise link features: (1) name of the section(s) in which the Wikipedia page corresponding to entity e_1 points to the Wikipedia page about entity e_2; (2) the same information in the opposite direction. For example, a link to the `Aspirin` page occurs in the *Prevention* section of the `Migraine` page, while the reverse link occurs in the *Medical uses* section of the `Aspirin` page.
- **sectLinkSectPath**. This feature set captures indirect links between the entities and includes the concatenated names of the sections of Wikipedia pages corresponding to e_1 and e_2 having common outgoing links. For example, `Aspirin` links to `Tension_headache` in its *Medical uses* section, and `Migraine` links to `Tension_headache` in its *Cause* section. Thus the sectLinkSectPath path constructed for the `Aspirin` – `Migraine` entity pair is *Medical_Uses:::Cause*

4 EXPERIMENTS

We perform experiments in two parts. In the first part (Section 4.3), we evaluate the utility of using entity-level semantics over a standard approach. The insights from the first part are then used to create an overall better relation recognizer in the second part (Section 4.4).

4.1 Data

We extracted the experimental dataset from the National Drug File–Reference Terminology (NDF-RT). NDF-RT is an extended formal ontological version of the National Drug File (NDF), a list of drugs and their properties released by U.S. Department of Veterans Affairs, Veterans Health Administration (VHA). It contains information about drugs and their relations with other biomedical entities, including interactions, physiological effects, methods of action, etc.

Entity pairs are extracted from the NDF-RT ontology, which provides the relation labels for each entity pair. Given an entity pair, we construct features based on entity-level semantics described above. This is then used to train a supervised relation classifier.

The dataset is a set of labeled examples. An example is a triple (e_1, R, e_2), where e_1 (subject) and e_2 (object) are UMLS entities corresponding to NDF-RT entities. R is either one of the NDF-RT relations listed in Table 1, or, if e_1 and e_2 are not related, $R = NOREL$ (and the entity pair is considered as a negative example).

We extracted positive examples by searching NDF-RT for all the entity pairs engaged in a given relation of interest. All entity pairs having more than one relation in NDF-RT were discarded to remove ambiguity during evaluation. Additionally, entities with symbols in their name (e.g. "%", ",", "/") were discarded, as these entities are likely to have no coverage in the knowledge sources (for our systems as well as the baseline). Negative examples were randomly generated following the closed world assumption. We randomly draw (e_1, e_2) and check whether NDF-RT contains information about relation between them. If it does not, then the entity pair is considered an example of a $NOREL$ relation.

The resulting dataset, *AUTONDF*, contains 48,519 positive and 48,519 negative examples. The number of entity pair examples per relation are as follows, CI_ChemClass: 1,113; CI_MoA: 318; CI_with: 13,819; has_Ingredient: 1,630; has_MoA: 6,509; has_PE: 10,449; induces: 271; may_diagnose: 386; may_prevent: 882; may_treat: 13,142; NOREL: 48,519.

For each (e_1, R, e_2) example we extract a set of features described in Section 3. In order to obtain features from UMLS Semantic Network and UMLS Metathesaurus features, we queried the off-line distribution of UMLS for CUIs of interest. Wikipedia-based features were extracted using JWPL Wikipedia API[26], from the Wikipedia version of December, 2011[11]. If there was more that one page retrieved for either e_1 or e_2, all the pages were exploited as feature sources.

Due to size limitations of knowledge sources, they may not have coverage over all instances. For example, one or both entities in a pair may not have a corresponding page in Wikipedia, making it impossible to extract Wikipedia-based features. When training a classifier, instances that do not find coverage in the knowledge sources it uses are skipped.

4.2 Baseline

Our baseline, *DS*, is a system using distant supervision and sentence-level features. This approach has been suggested to circumvent the lack of sufficiently large, labeled corpus for relation extraction [14]. In distant supervision, for each pair of entities that are in a particular relation, all sentences containing those two entities are extracted from a large unlabeled corpus and a relation classifier is trained using textual features of these sentences. The underlying hypothesis is that "if entities e_1 and e_2 are known to be in relation R, then any sentence containing a mention of both e_1 and e_2 is likely to express the relation R".

Table 2. Baseline system performance.

Relation	DS-covered			
	Count (Coverage)	P	R	F_1
CI_ChemClass	138 (12.40%)	61.90	9.42 (1.17)	16.35 (2.29)
CI_MoA	0 (0%)	0.00	0.00 (0.00)	0.00 (0.00)
CI_with	905 (6.55%)	83.63	20.88 (1.37)	33.42 (2.69)
has_Ingredient	64 (3.93%)	93.75	23.44 (0.92)	37.50 (1.82)
has_MoA	48 (0.74%)	77.78	29.17 (0.22)	42.42 (0.43)
has_PE	117 (1.12%)	0.00	0.00 (0.00)	0.00 (0.00)
induces	60 (22.14%)	0.00	0.00 (0.00)	0.00 (0.00)
may_diagnose	24 (6.22%)	0.00	0.00 (0.00)	0.00 (0.00)
may_prevent	183 (20.75%)	43.75	7.65 (1.59)	13.02 (3.06)
may_treat	2320 (17.65%)	59.22	98.66 (17.42)	74.02 (26.92)
NOREL	324 (0.67%)	66.67	0.62 (0.00)	1.22 (0.01)
Overall	4183 (4.31%)	44.25	17.26 (2.11)	19.81 (3.38)

We built *DS* using our *AUTONDF* dataset and PubMed as the source of sentences. We queried PubMed for abstracts and titles containing pairs of entities from our dataset using *NCBI Entrez Utilities Web Service*[12], and labeled sentences containing e_1 and e_2 with relation R. Overall we have extracted 122,466 sentences for the entity pairs from the *AUTONDF* dataset. These sentences were then used to train a system to predict relations between entities in the context of a sentence. We used features motivated by lexical features presented in [14]. Specifically, we used word lemmas and part of speech tags of three words to the left and right of both entities, word lemmas between the entities and a binary feature denoting which entity comes first in the sentence. In addition, we also used the distance between both entities in terms of words.

In testing phase, to predict the relation between an entity pair, we used the majority prediction by this system on the set of all sentences

[11] `http://dumps.wikimedia.org/enwiki/20111201/`

[12] `http://www.ncbi.nlm.nih.gov/entrez/query/static/esoap_help.html`

Table 3. Performance of best feature sets per relation on test instances covered by a specific feature set. R and F_1 for the same feature set on the full data set are reported in parentheses (P remains the same under both conditions).

Relation	Best Feature Set	Count (Coverage)	P	R	F_1
CI_ChemClass	umlsPF:::SNOMEDCT, umlsSemType, umlsCUI	939 (84.37%)	91.39	94.99 (80.14)	93.16 (85.4)
CI_MoA	wikiCategory, pairwiseLinkFeatures, umlsCUI	115 (36.16%)	95.65	95.65 (34.59)	95.65 (50.81)
CI_with	umlsPF:::SNOMEDCT, umlsCUI	10571 (76.50%)	93.41	94.76 (72.49)	94.08 (81.63)
has_Ingredient	umlsPF:::MEDCIN, wikiCategory, pairwiseLinkFeatures, sectLinkSectPath	134 (8.22%)	83.33	67.16 (5.52)	74.38 (10.36)
has_MoA	umlsSemType, umlsCUI	6509 (100.00%)	95.06	96.67 (96.67)	95.86 (95.86)
has_PE	umlsPF:::SNOMEDCT, wikiCategory, pairwiseLinkFeatures, sectLinkSectPath	26 (0.25%)	100	100 (0.25)	100 (0.5)
induces	umlsPF:::SNOMEDCT, umlsSemType, umlsCUI	194 (71.59%)	92.22	85.57 (61.25)	88.77 (73.61)
may_diagnose	umlsPF:::SNOMEDCT, umlsSemType, umlsCUI	132 (34.20%)	85.47	75.76 (25.91)	80.32 (39.76)
may_prevent	umlsPF:::SNOMEDCT, umlsCUI	598 (67.80%)	84.92	71.57 (48.53)	77.68 (61.76)
may_treat	umlsPF:::SNOMEDCT, umlsSemType, umlsCUI	10003 (76.11%)	88.6	93.97 (71.53)	91.2 (79.15)
NOREL	umlsPF:::MSH, umlsSemType, umlsCUI	12918 (26.62%)	97.78	95.7 (25.48)	96.73 (40.43)

extracted for that pair from PubMed. The baseline system is implemented using the multi-class linear kernel support vector machine (SVM)[7] classifier. More specifically, we used *libsvm* library[5].

4.3 Entity-level Semantics (ELS) Systems

Each individual ELS system is a linear SVM classifier operating upon a vector of a subset of features described in Section 3. The only difference between different individual systems is the feature set employed. We created 49 individual systems, based on different feature type combinations. The combinations with very small coverage are not considered. A feature set coverage is considered too small if the corresponding covered subset of *AUTONDF* did not contain enough instances to carry out a reliable evaluation.

4.3.1 Results

Performance is evaluated using 10-fold cross-validation on the *AUTONDF* dataset. Results over individual folds are averaged in order to obtain the results over the entire dataset. We report the performance of our systems in terms of precision (P), recall (R), and Fmeasure (F_1). Here we use standard formulas for P, R and F_1 in multi-class setting.

Table 2 reports the performance of the distant supervision baseline for each relation type. Here, precision, recall and Fmeasure are calculated over the instances for which the classifier is able to make a prediction (instances not covered by the classifier are skipped from evaluation). Fmeasure and recall over the full dataset are reported in parentheses, precision remains the same under both conditions. The column *Count (Coverage)* reports the number and percentage of entity pairs of a relation type for which the classifier is able to find sentences and create instances. First, we can see that the coverage of DS is rather poor. Due to this, the classifier is not able to learn reliable models in many cases (e.g. CI_MoA, and induces). There is only one relation, may_treat, for which the classifier finds adequate number of instances for training, resulting in a reasonable Fmeasure. We also evaluated the baseline on the entire dataset (the table is not shown due to space limitations). The precision remains the same (as the number of instances retrieved does not change with the evaluation set), but the recall numbers drop drastically, resulting in very poor Fmeasures. In spite of using a rich resource such as Pubmed, we found that this classifier faces coverage issues because first, not all relations of interest are commonly expressed in sentences, and second, not every entity pair, from our large entity pair dataset, always co-occurs in sentences.

Table 3 reports the performance of our ELS classifiers. Again, performance is calculated for the covered instances. For space reasons, only the best performing classifier (based on Fmeasure) is shown for each relation type. Note that Overall numbers are not shown in this table as these are different classifiers. The second column (Best FS) reports the feature set of the best-performing classifier and the third column reports its coverage. First, we notice that the coverage of these classifiers are much higher than DS for most relation types (except for has_PE). Specifically, there is a 44 percentage point improvement on average. Second, the precision, recall and Fmeasures obtained by using entity-level semantics are substantially higher than that obtained by DS. Specifically, all Fmeasures are greater than 75%, and for six relations, the Fmeasure achieved is greater than 90%. On an average, this is a 39 percentage point improvement. This indicates that, for the detection of our medical relations, features using entity-level semantics is better than sentence-level features.

Observe that the best performing classifier is different for different relation types. For example, recognition of CI_MoA is most benefited by Wikipedia and entity-pair features, while may_treat is best benefited by category features from UMLS. Interestingly, observe that the very simple feature set (umlsSemType, umlsCUI) is the best performer for has_MoA. Additionally, by virtue of being available for all entities in our dataset, we also observed that this is the only feature set that has 100% coverage.

4.4 Ensemble of entity-level semantics classifiers

The previous section showed that systems using entity-level semantics have better performance than a system using sentence-level information. We also saw that systems with complex features may suffer from coverage issues while systems with simple features may not be discriminative enough. However, due to the difference in coverage and performance for different relation types, it is difficult to select one universally best system. Further, in many cases, a new instance to be classified has coverage in more than one ELS system, and it is difficult to decide which system's prediction is to be considered.

In order to get the best in terms of performance as well as coverage, we combine all 49 ELS systems in a *ensemble* system which takes outputs of individual systems for an instance as input, and predicts a single relation class label.

The ensemble classifier has a feature corresponding to each ELS classifier. Given an entity pair, the feature value for an ELS classifier feature will be its relation prediction for that entity pair (or "notCovered" if there in no coverage for that classifier). This classifier is also implemented using libsvm.

4.4.1 Results

Table 4 reports the performance of the ensemble classifier on the entire *AUTONDF* data. Ensemble classifier for i-th test fold of cross validation was trained on the outputs obtained by the individual classifiers on $1, 2, i-1, i+1$, 10-th test folds. For comparison, we also report the performance of the best ELS system that has full coverage, STCUI. STCUI uses only the simple semantic features: Umls semantic type, and CUI. Here we see that in addition to full coverage, the ensemble is also able to achieve better performance than STCUI for all relation types. The improvement in Fmeasure is due to the improvement in both precision and recall. The improvements that are significant at $p < 0.01$ are shown in bold. Thus, by combining the individual ELS classifiers, it is possible to harness different types of entity-level semantics to achieve good coverage as well as performance for relation mining.

Table 4. Performance of ensemble and STCUI baseline systems. Overall is obtained by macro-averaging over results for individual relations.

Relation	Ensemble			STCUI		
	P	R	F1	P	R	F1
CI_ChemClass	**93.47**	92.54	**93**	91.56	93.62	92.58
CI_MoA	**93.25**	95.6	**94.41**	88.56	94.97	91.65
CI_with	**93.67**	**94.49**	**94.08**	92.45	93.58	93.01
has_Ingredient	**78.11**	**63.25**	**69.9**	73.11	53.87	62.03
has_MoA	**95.26**	**97.05**	**96.15**	95.06	96.67	95.86
has_PE	95.82	96.5	96.16	95.76	96.53	96.14
induces	91.67	81.18	**86.11**	88.26	74.91	81.04
may_diagnose	89.91	**76.17**	**82.47**	87.77	72.54	79.43
may_prevent	**81.45**	**68.71**	**74.54**	77.13	54.31	63.74
may_treat	**90.57**	92.48	**91.51**	87.51	92.2	89.79
NOREL	**96.49**	**96.38**	**96.43**	96.33	95.7	96.01
Overall	**90.88**	**86.76**	**88.61**	88.50	83.54	85.57

5 DISCUSSION

UMLS semantic type has been frequently used as one of the most useful semantic features[2]. However, we found that, these features can be too coarse to be discriminative for our task. For instance, consider the entity pair *Secretin* (e_1)- *Liver Diseases* (e_2). UMLS semantic types of an entity e_1 was found to be *Hormone*, *Pharmacologic_Substance*, *Amino_Acid,_Peptide,_or_Protein*, while e_2 has semantic type *Disease_or_Syndrome*. On the other hand SNOMED CT contains information that e_1 is *Gastrointestinal hormone* and *Peptide hormones and their metabolites and precursors*, while e_2 is a *Liver finding*, *Disorder of abdomen* and a *Disorder of digestive organ* . Intuitively such fine-grained information is more discriminative. Table 3 corroborates this intuition – most of the top classifiers that use entity category information infact make use of SNOMED CT features.

The semantic features we employ vary from simple identity-based features such as umlsCUI, to complex pair-based features such as pairwiseLinkFeatures. Entity pair features are complex, and relatively sparse, which makes learning them reliably a challenge. However, the information they capture can lead to creating more precise predictions. On inspecting instances that were incorrectly classified by classifiers using only simple category features such as umlsCUI, but correctly classified using pairwiseLinkFeatures features, we found that the classifier with only identity-based features predicted the most common relation that the given entities were involved in, while the classifier incorporating pairwiseLinkFeatures overcame this pitfall.

Finally, we experimented with extending the sentence-level baseline classifier, DS, with the simple semantic features. Here we augmented the existing feature vectors constructed using linguistic features with semantic information such as umlsSemType and umlsCUI. This approach does not change the coverage of DS, but allows us to inspect the impact on precision due to entity-level semantics. Table 5 reports the results, similar to Table 2. Here we see that, with the addition of even simple entity-level semantics, not only has the precision for most relations improved, but the recall of the relation types are improved as well, resulting in much higher Fmeasures. Addition of more complex entity-level semantics and combining the sentence-level system with semantics-based system are directions for our future explorations.

Table 5. Distance Supervision - Using STCUI

Relation	DS + STCUI		
	P	R	F_1
CI_ChemClass	72.96	84.06	78.11
CI_MoA	0.00	0.00	0.00
CI_with	89.77	60.11	72.01
has_Ingredient	82.76	37.50	51.61
has_MoA	78.18	89.58	83.50
has_PE	96.23	87.18	91.48
induces	87.10	45.00	59.34
may_diagnose	100.00	8.33	15.38
may_prevent	84.51	32.79	47.24
may_treat	79.51	97.67	87.66
NOREL	83.58	70.68	76.59
Overall	77.69	55.72	60.27

6 RELATED WORK

Biomedical relation extraction: Approaches to relation extraction in the Biomedical domain has employed pattern based approaches [1, 19, 17], machine learning approaches [18, 10, 21, 12, 13] or a combination of the two [2]. For example, [1] use a set of relation-specific patterns, [19] use a set of syntactic patterns, and [17] extract relations matching a set of manually designed rules using an enriched syntactic parse tree representation of sentences. Our focus in this work is on supervised methods.

Supervised statistical machine learning (ML) approaches automatically learn patterns in the labeled data. [18] recognize *disease*, *treatment* semantic role and seven semantic relations, and extract 7 binary and unary relations between them: *cure, only DIS, only TREAT, Prevent, Vague, Side Effect, NO Cure* using discriminative models. [10] distinguish between three relation classes, *cure*, *prevent* and *side-effect*, experimenting with various feature representations. The best results were achieved using rich feature sets (bag of words, noun phrases, verb phrases, UMLS semantic types). The authors mention that better results are achieved when ontological knowledge is employed. We too use a supervised setting, and some of our semantic features overlap with theirs. However, our work focuses on exploring an assortment of semantic features alone, as sentences and consequently sentence-based features have low coverage for our task. Our results corroborate that semantic features are important for relation extraction in this domain. However, we focus on a different set of biomedical relations from the above. Additionally, we show that using classifier ensembles can overcome the difficulties due to lack of coverage.

Relation extraction using semantic knowledge: Background knowledge sources (e.g. WordNet [9] and Wikipedia) have been widely exploited for relation extraction both in general and specialized domains. Wikipedia has been used to extract features indicating whether entities of interest are related, and whether they are in *parent-child* relation[4], and to obtain features indicating the semantic relatedness between nominals of interest [22]. In our work, we focus on Wikipedia category and linkage features that are important for biomedical relation extraction.

In the biomedical domain semantic knowledge is exploited previously by [18] who used MeSH IDs of the words occurring in a sentence being classified as features. UMLS features have been added to sentence-level features in relation mining with promising results in [10] and [2]. We found that sentence-based systems have poor coverage for our task, which we remedy using a variety of semantic information and then fusing them.

Distant supervision for relation mining: Distant supervision (DS) approaches for relation mining have used Freebase[13][14] and YAGO[14][15] to extract labeled sentences from Wikipedia. [25] use an undirected graphical model for both relation and entity type prediction and use Freebase as a source of seeds, and Wikipedia and New York Times corpus as source of sentences. In the biomedical domain, [23] investigated distant supervision for protein-protein interaction (PPI). The problems of DS approaches are the noise in the data and absence of knowledge about negative instances and their distribution. Moreover, in our task, the sentence retrieval lacks coverage.

Ensemble Learning: Ensemble learning methods have been applied to a variety of natural language processing applications such as those for text categorization [20], parsing [6], word-sense disambiguation [16, 8]. In relation mining, they have been used for ontology learning within a system called OntoLancs [11]. [24] use ensemble feature selection for biomolecular text mining. They show that their feature selector is able to discard a large fraction of machine-generated features, improving classification performance of state-of-the-art text mining algorithms. While we use an ensemble approach, the main focus of our work is on exploration of a variety of entity-level semantics for detecting different clinical relations.

7 CONCLUSION

This work explores the use of rich knowledge about biomedical entities obtained from various sources for relation mining. Our entity-level semantics includes taxonomic information about individual entities as well as linkage information between entity pairs. We built individual classifiers that harness entity semantics as well as a meta classifier to achieve advantages of performance and coverage. Our approach was tested on a large dataset obtained from a standard human-curated ontology.

Our experiments reveal that the distant supervision approach that uses sentence-level information does not perform well for our domain and relation types – it has issues with both coverage and performance. We discovered that different types of semantics are useful for different relation types, and that performance and coverage vary based on the scope and depth of the knowledge sources used. Our ensemble approach proved successful in solving the problem of coverage, while achieving good overall performance.

REFERENCES

[1] A.B. Abacha and P. Zweigenbaum, 'Automatic extraction of semantic relations between medical entities: a rule based approach', *Journal of Biomedical Semantics*, **2**(Suppl 5), S4, (2011).

[2] A.B. Abacha and P. Zweigenbaum, 'A hybrid approach for the extraction of semantic relations from medline abstracts', in *In Proceedings of CICLing-2011*, pp. 139–150. Springer-Verlag, (2011).

[3] ACE, 'Automatic Content Extraction', *http://www.ldc.upenn.edu/Projects/ACE/*, (2000-2005).

[4] Y.S. Chan and D. Roth, 'Exploiting background knowledge for relation extraction', in *In Proceedings of COLING-2010*, pp. 152–160. ACL, (2010).

[5] C.C. Chang and C.J. Lin, 'LIBSVM: A library for support vector machines', *ACM Transactions on Intelligent Systems and Technology*, **2**, 27:1–27:27, (2011). Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[6] M. Collins and T. Koo, 'Discriminative reranking for natural language parsing', *Computational Linguistics*, **31**(1), 25–69, (2005).

[7] N. Cristianini and J. Shawe-Taylor, *An introduction to Support Vector Machines*, Cambridge University Press, March 2000.

[8] G. Escudero, L. Màrquez, and G. Rigau, 'Boosting applied to word sense disambiguation', *Proceedings of ECML-00*, 129–141, (2000).

[9] *WordNet An Electronic Lexical Database*, ed., C. Fellbaum, The MIT Press, Cambridge, MA ; London, May 1998.

[10] O. Frunza and D. Inkpen, 'Extraction of disease-treatment semantic relations from biomedical sentences', in *Proceedings of BioNLP-2010*, pp. 91–98. ACL, (2010).

[11] R. Gacitua and P. Sawyer, 'Ensemble methods for ontology learning - an empirical experiment to evaluate combinations of concept acquisition techniques', in *Proceedings of ICIS-2008*, pp. 328 –333, (2008).

[12] C. Giuliano, A. Lavelli, and L. Romano, 'Exploiting shallow linguistic information for relation extraction from biomedical literature', in *Proceedings of EACL-2006*, pp. 401–408, (2006).

[13] J. Li, Z. Zhang, X. Li, and H. Chen, 'Kernel-based learning for biomedical relation extraction', *Journal of the American Society for Information Science and Technology*, **59**(5), 756–769, (2008).

[14] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, 'Distant supervision for relation extraction without labeled data', in *Proceedings of the ACL-IJCNLP 2009*, pp. 1003–1011. ACL, (2009).

[15] T.V.T. Nguyen and A. Moschitti, 'End-to-end relation extraction using distant supervision from external semantic repositories', in *Proceedings of ACL-HLT*, pp. 277–282. ACL, (2011).

[16] T. Pedersen, 'A simple approach to building ensembles of naive bayesian classifiers for word sense disambiguation', *NAACL 2000*, 63–69, (2000).

[17] C. Ramakrishnan, K. Kochut, and A. Sheth, 'A framework for schema-driven relationship discovery from unstructured text', *The Semantic Web-ISWC 2006*, 583–596, (2006).

[18] B. Rosario and M.A. Hearst, 'Classifying semantic relations in bioscience texts', in *ACL 2004*, pp. 430–437. ACL, (2004).

[19] S. Sahay, S. Mukherjea, E. Agichtein, E. V. Garcia, S. B. Navathe, and A. Ram, 'Discovering semantic biomedical relations utilizing the web', *ACM Trans. Knowl. Discov. Data*, **2**, 3:1–3:15, (April 2008).

[20] F. Sebastiani, 'Machine learning in automated text categorization', *ACM Comput. Surv.*, **34**, 1–47, (March 2002).

[21] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.

[22] G. Szarvas and I. Gurevych, 'Tud: semantic relatedness for relation classification', in *Proceedings of SemEval-2010*, pp. 210–213. ACL, (2010).

[23] P. Thomas, I. Solt, R. Klinger, and U. Leser, 'Learning protein protein interaction extraction using distant supervision', in *Proceedings of Workshop on Robust Unsupervised and Semisupervised Methods in NLP*, pp. 25–32, Hissar, Bulgaria, (sep 2011).

[24] S. Van Landeghem, T. Abeel, Y. Saeys, and Y. Van de Peer, 'Discriminative and informative features for biomolecular text mining with ensemble feature selection', *Bioinformatics*, **26**(18), i554–i560, (2010).

[25] L. Yao, S. Riedel, and A. McCallum, 'Collective cross-document relation extraction without labelled data', in *Proceedings of EMNLP 2010*, pp. 1013–1023. ACLs, (2010).

[26] T. Zesch, C. Müller, and I. Gurevych, 'Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary', in *Proceedings of LREC 2008*, (2008).

[13] `http://freebase.com/`

[14] `http://www.mpi-inf.mpg.de/yago-naga/yago/`