Extracting Disease-Symptom Relationships by Learning Syntactic Patterns from Dependency Graphs

Mohsen Hassan, Olfa Makkaoui, Adrien Coulet and Yannick Toussaint
LORIA (CNRS, Inria, Université de Lorraine),
Campus scientifique, Vandoeuvre-lès-Nancy, F-54506, France
{mohsen.sayed,olfa.makkaoui,adrien.coulet,yannick.toussaint}@loria.fr

Abstract

Disease-symptom relationships are of primary importance for biomedical informatics, but databases that catalog them are incomplete in comparison with the state of the art available in the scientific literature. We propose in this paper a novel method for automatically extracting disease-symptom relationships from text, called SPARE (standing for Syntactic PAttern for Relationship Extraction). This method is composed of 3 successive steps: first, we learn patterns from the dependency graphs; second, we select best patterns based on their respective *quality* and *specificity* (their ability to identify only disease-symptom relationships); finally, the patterns are used on new texts for extracting disease-symptom relationships. We experimented SPARE on a corpus of 121,796 abstracts of PubMed related to 457 rare diseases. The quality of the extraction has been evaluated depending on the pattern *quality* and *specificity*. The best F-measure obtained is 55.65% (for $specificity \geq 0.5$ and $quality \geq 0.5$). To provide an insight on the novelty of disease-symptom relationship extracted, we compare our results to the content of phenotype databases (OrphaData and OMIM). Our results show the feasibility of automatically extracting disease-symptom relationships, including true relationships that were not already referenced in phenotype databases and may involve complex symptom descriptions.

1 Introduction

Disease-Symptom (D-S) relationships are of major importance for biomedical informatics since they provide a fine-grained description of disease that could be used to guide medical diagnosis in clinical care. However, biomedical databases that catalog D-S relationships such as OrphaData and OMIM are incomplete in comparison with the state of the art available in the scientific literature (Köhler et al., 2014). In addition, extracting this information manually from the literature by experts requires a lot of time and effort, which motivates the need for developing automatic methods.

Our study focuses on extracting symptoms in relation with rare diseases (RDs). These are diseases that affect a small percentage of the population, ranging from 1/1,000 to 1/200,000. As their number is relatively important (between 6,000 and 8,000 (Mazzucato et al., 2014)), RDs have received a particular attention in the medical domain.

In this context, we propose an automatic method, called SPARE (Syntactic PAttern for Relationship Extraction), for D-S relationship extraction based on shortest path patterns generated from the dependency graphs (DGs) of texts. We applied SPARE to the extraction of D-S relationships associated with rare diseases. Because symptoms associated with rare diseases may be uncommon and complex (*i.e.,* they can not be expressed with one word or a simple expression), we particularly focus on enabling the recognition of symptoms that are not listed in phenotype databases or ontologies.

As a result, objectives of this work are threefold: (1) learning patterns specific for diseases-symptom relationships extraction; (2) identifying symptom description that is pointed by specific pattern; and (3) extracting D-S relationships.

This article is organized as follow: we introduce D-S relationship relative issues in section 2. Section 3 presents main methods for relationship extraction. In section 4, we detail the SPARE method. Section 5 describes experiments and re-

Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015), pages 71–80,
Beijing, China, July 30, 2015. ©2015 Association for Computational Linguistics

sults. Finally, we discuss and conclude about the results described in the article.

2 Disease-Symptom Relationships

OrphaData and OMIM are two examples of databases that catalog D-S relationships. OrphaData[1] is the database accessible from Orphanet, the portal for rare diseases and orphan drugs. It includes description of symptoms (clinical signs) of rare disease. OMIM[2] (Online Mendelian Inheritance in Man) is a database for genetic diseases. It contains disease descriptions that include a list of symptoms named "clinical synopsis".

Due to the fact that their content is manually curated by experts, OrphaData and OMIM are high quality resources. However, these resources do not contain a complete list of relationships between diseases and symptoms that exist in the biomedical literature. As shown in Table 1, among the 8,644 diseases listed by OrphaData only 2,689 diseases (31.11%) are associated with clinical signs and symptoms. Indeed, one can use cross references between OrphaData and OMIM[3] to associate OrphaData diseases to symptoms described in OMIM. Nevertheless, even when considering these additional symptoms, only 4,856 (56.18%) OrphaData diseases have symptoms. The rest, 3,788 OrphaData diseases, is not related to any symptom. This motivates us to extract these relations from the literature.

	#Diseases	#Diseases associated with symptoms	#Symptoms	#D-S Relations
OrphaData	8,644	2,689	1,273	52,503
OMIM	23,929	23,910	46,369	432,760

Table 1: Information about OrphaData and OMIM databases

Recognizing diseases and symptoms in texts is a preliminary step for D-S relationships extraction. Previous work on disease recognition achieved good results (Leaman and Lu (2014) obtained 78.25% F-Measure, 76.3% recall and 80.3% precision). Less works aimed at recognizing symptoms. Their performances are low in comparison with those of disease recognition. For example, Martin *et al.* (2014) used HPO[4] (Köhler et al., 2014) for symptom extraction and obtained 36.8% F-Measure, 23.7% recall and 82.2% precision.

Extracting D-S relationships automatically is a challenging task mainly due to the following two reasons: first, there is no complete dictionary of symptoms to guide their recognition; second, symptoms are complex entities that are hard to recognize in text. Indeed, HPO, which contains 11,021 phenotypes terms, covers only symptoms related to genetic diseases. Thus, a simple "exact match" approach to recognize HPO symptoms in text would give a low recall: "serositis" in example 2.1 is not known as a symptom in HPO.

In addition, Named Entity Recognition (NER) tools recognize symptoms with low recall. This is the case of MetaMap (Aronson, 2001), a tool that annotates texts with concepts from UMLS (Bodenreider, 2004). In example 2.2 MetaMap annotates "Familial Mediterranean Fever" as disease but does not annotate "fever" or "attacks of fever" as a symptom.

Ex 2.1. *"<disease>Familial Mediterranean Fever</disease> is characterized by serositis"*

Ex 2.2. *"<disease>Familial Mediterranean Fever</disease> (FMF) is an autosomal recessive disorder characterized by attacks of fever"*

Recognizing the full description of symptoms is another challenge for symptom recognition, in particular with rare diseases where symptom description can be complex phrases. Some cases of partial annotations occur when HPO or MetaMap annotates only a part of the entity. For instance, example 2.3 shows that "pure spasticity of the lower limbs" is a symptom but MetaMap annotates only "spasticity".

Ex 2.3. *"One patient with <disease>Krabbe disease</disease> presented with pure <symptom>spasticity</symptom> of the lower limbs"*

The ambiguity between diseases and symptoms is another factor of complexity as diseases play, in some situations, the role of symptoms. For instance, example 2.4 shows that "muscle wasting" is recognized by MetaMap as a disease. However, it can be considered as a symptom for "Duchenne muscular dystrophy".

Ex 2.4. *"<disease>Duchenne muscular dystrophy</disease> is characterized by <disease>muscle wasting</disease>"*

[1] OrphaData website: http://www.orphadata.org/

[2] OMIM website: http://www.omim.org/

[3] 4,162 OrphaData diseases have cross references to OMIM diseases.

[4] HPO (The Human Phenotype Ontology) provides a structured and controlled vocabulary for the phenotypic features of diseases.

3 Related Works

Various works have proposed methods to extract relationships from text. They are based on different approaches such as statistics, pattern-based or rule-based, and machine learning.

A co-occurrence method is a simple method to identify relationships between two entities that co-occur in the same sentence (Bunescu et al., 2006). It is based on the hypothesis that if two entities are mentioned frequently together, they are likely to be in a relation. Approaches based on co-occurrences of entities do not employ NER techniques. The type and the direction of relationships are not captured by these methods. Various statistical measures are used to decide whether the two entities co-cited together are in relation or not (Lee et al., 2007; Ramani et al., 2005). Examples of these measures are Pointwise Mutual Information, Chi-Square or Log-Likelihood Ratio (Manning and Schütze, 1999), which use the co-occurrence statistics of the two entities to hypothesize about the existence of a relationship between them. Ramani *et al.* (2005) use random co-citation model based on the hypergeometric distribution. Co-occurrence methods have been successfully applied to the automated construction of networks of biomolecules such as gene-protein and gene regulatory networks (Šarić et al., 2006; Friedman et al., 2001).

Pattern- and rule-based methods generate symbolic patterns or rules to extract relationships, with advantage that they are easy to interpret (Agichtein and Gravano, 2000). These patterns or rules can be generated manually (Divoli and Attwood, 2005) or automatically by learning from annotated corpus (Hakenberg et al., 2005). They are based on different levels of linguistic information like lexical, syntactic or dependency information and different levels of structures like sequences, trees and graphs. These methods tend to have a high precision but a low recall (Cellier et al., 2010; Béchet et al., 2012; Liu et al., 2013; Martin et al., 2014; Hassan et al., 2014).

Liu *et al.* (Liu et al., 2013) proposed a graph-based approach to learn rules for event extraction (that can be compared to relationship extraction). The rules are represented by the information on the shortest path between entities in an undirected DG. Béchet *et al.* (2012) and Cellier *et al.* (2010) proposed a method based on sequential pattern mining to extract disease-gene and gene-gene relationships. As the number of their patterns is very large, they introduced constraints for patterns filtration to reduce them. Close to our objectives, Martin *et al.* (2014) used sequential patterns for recognizing unidentified symptoms. Also, Hassan *et al.* (2014) proposed a pattern-based method for D-S relationship extraction, where diseases and symptoms are previously recognized and annotated by a NER tool. The patterns are learned from shortest paths between diseases and symptoms in directed DGs.

Machine Learning (ML) methods consider a relationship extraction task as a classification problem. Two ML techniques are mainly employed: feature-based and kernel-based methods. Feature-based methods such as support vector machines or conditional random fields have been employed by (Krallinger et al., 2008; Bundschus et al., 2008) for relationship extraction. Kernel methods use a kernel function to measure the similarity between a large amount of features *e.g.,* sub-sequences, trees, graphs (Zelenko et al., 2003; Zhang et al., 2008; Airola et al., 2008).

Bunescu and Mooney (2005) proposed a shortest path kernel method that uses the shortest path between two entities in an undirected DG for relationship extraction. This work is based on the hypothesis that the relationship between two entities in the same sentence is typically captured by the shortest path between them in the DG. Chowdhury *et al.* (2012) proposed a hybrid kernel that uses different types of information (*e.g.,* syntactic, contextual, semantic) and their different representations (*i.e.,* flat features, tree structures and graphs). This hybrid kernel helps improving the results of relationship extraction.

4 Method

We describe in this section the SPARE method for D-S relationship extraction. This method is composed of three steps: first, learning patterns out of DGs that include both a disease and a symptom; second, selecting patterns in regard to their quality (*i.e.,* their capacity to identify true relationships) and their specificity (*i.e.,* their capacity to identify only D-S relationships); third, using selected patterns to extract D-S relationships from text.

The originality of the SPARE method relies on measuring how syntactic patterns between diseases and symptoms are specific to D-S relationships. Using highly specific patterns allow us to consider the case where symptoms are not recognized by NER tools, which consequently offers the opportunity to discover new symptom descriptions that can be potentially rare and complex.

SPARE is inspired from various previous works such as using the shortest path between entities of a DG as described by Bunescu et Mooney (2005), then applied by Chowdhury *et al.* (2012) and Liu *et al.* (2013). Similarly to Liu *et al.* (2013), we extract patterns represented by the whole subgraph (*i.e.,* all nodes and edges in the shortest path), but unlike them, we keep edge directions. Hassan *et al.* (Hassan et al., 2014) proposed a pattern-based method for D-S relationship extraction. They assume that diseases and symptoms are initially recognized by a NER tool. Here we relax patterns, similarly to Blohm *et al.* (2011), and use specificity to consider cases of unrecognized symptoms.

The following subsections detail the three steps of SPARE.

4.1 Learning Syntactic Patterns from DG

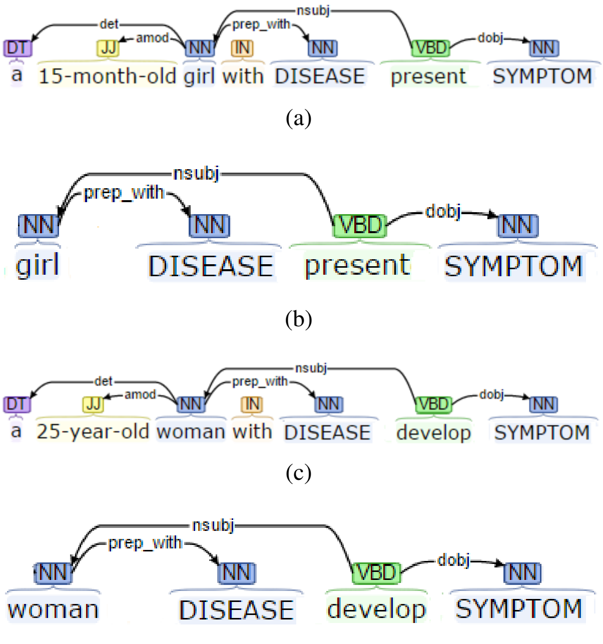
For pattern learning, only DGs of sentences that contain at least one disease and one symptom are considered as we are interested in extracting D-S relationships. DGs are explored to find the shortest paths between diseases and symptoms. Because one sentence can mention several diseases and symptoms, several shortest paths may be found.

Ex 4.1. *"A 15-month-old girl with <disease>propionic acidemia</disease> presented <symptom>muscular hypotonia </symptom>"*

Ex 4.2. *"A 25-year-old woman with <disease>cystic fibrosis</disease> developed <symptom>hemoptysis</symptom>"*

Figures[5] 1(a) and 1(c) show the DGs generated from sentences of examples 4.1 and 4.2 after the replacement of the annotated entities (*i.e.,* diseases and symptoms) by generic words (*i.e.,* DISEASE and SYMPTOM) and other words by their lemmas. Figures 1(b) and 1(d) show the shortest paths extracted from associated DGs. The whole shortest path is kept, including all nodes, edges and directions.

Next, patterns are generated on the basis of shortest paths, using a generalization process. In this process, two shortest paths (or more) can be merged and represented in one generalized pattern. Different shortest paths are aggregated to a pattern if those share the same edges and directions. Figure 2 illustrates this generalization process considering the shortest paths obtained from examples 4.1 and 4.2. If the values of the nodes in the pattern are different, then they are replaced by "*" (*i.e.,* matching any token). A list of values observed for each node is kept but for pattern documentation purpose only. The frequency of patterns is measured by their *support*, *i.e.,* how many sentences in our learning corpus match this pattern.


det
amod
prep_with
nsubj
dobj
DT
JJ
NN
IN
NN
VBD
NN
a 15-month-old girl with DISEASE present SYMPTOM

(a)

nsubj
prep_with
dobj
NN
NN
VBD
NN
girl DISEASE present SYMPTOM

(b)

det
amod
prep_with
nsubj
dobj
DT
JJ
NN
IN
NN
VBD
NN
a 25-year-old woman with DISEASE develop SYMPTOM

(c)

nsubj
prep_with
dobj
NN
NN
VBD
NN
woman DISEASE develop SYMPTOM

(d)

Figure 1: (a,c) DGs and (b,d) Shortest paths between disease and symptom respectively extracted from sentences of examples 4.1 and 4.2

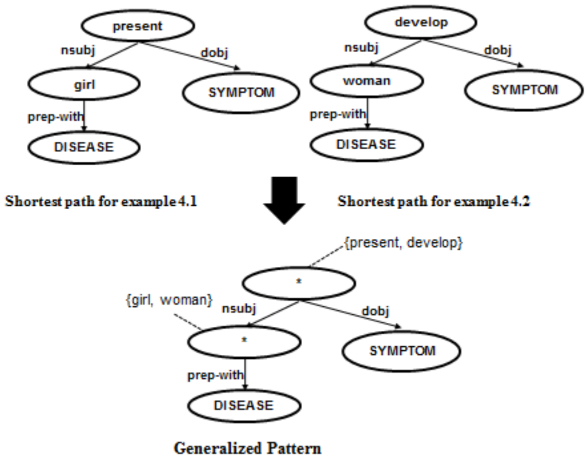


Figure 2: Example of pattern generation from two shortest paths

[5] DGs are processed by the Stanford Parser and drawn with the Brat tool at http://nlp.stanford.edu:8080/corenlp/

This generalization affects the precision and the recall of the patterns. Replacing the node value in the shortest path by using "*" (*i.e.,* any token) makes the pattern more generic, and has the consequence of increasing the recall of the patterns. On the other side, we assume that edges (*i.e.,* dependency types of DGs) and directions of the pattern guarantee its precision.

4.2 Pattern Selection

4.2.1 Quality-Based Selection

We classify patterns into two classes: positive and negative patterns. This classification relies both on the frequency and on the quality of patterns. The *quality* of patterns requires an evaluation procedure, on the basis of an annotated corpus, to be computed. The *quality* of a pattern is defined as:

$$quality = \frac{|T|}{|A|} \tag{1}$$

where T is the set of all true relationships and A is the set of all (true and false) relationships that are identified by the pattern. A relationship is qualified as true if it is annotated in the corpus, *i.e.,* if the sentence is actually mentioning the relationship. A pattern is considered positive if its *support* is greater than or equal to a minimum support denoted *min_support* and its *quality* is greater than or equal to a minimum quality denoted *min_quality*.

4.2.2 Specificity-Based Selection

In order to measure how much a pattern is specific to D-S relationships and not to other relationships, a *specificity* measure of the pattern is defined. To measure this specificity, we performed a new evaluation task for which we consider: *(i)* a novel set of annotated sentences, not including one disease and one symptom but including one disease and another entity (*e.g.,* a symptom, a gene, a treatment or a living being); *(ii)* patterns from which we removed the constraint on the symptom node (*i.e.,* SYMPTOM is replaced by "*"). The pattern specificity is computed by the following formula:

$$specificity = \frac{|DS|}{|A|} \tag{2}$$

where DS is the set of true D-S relationships extracted by the pattern and A is the set of all (true and false) relationships that are extracted by the pattern (including D-S, disease-gene, disease-treatment and disease-living being relationships). For example, if the pattern extracts 23 true D-S relationships and 7 disease-any entity relationships, then pattern specificity is 23/30. The specificity measure is used to select the patterns that are the most specific to D-S relationships by selecting those that have a specificity greater than or equal to a minimum specificity denoted *min_specificity*.

Both quality and specificity are associated with the precision of patterns (the ratio of true relationships on all extracted relationships, see formula 3) but are used in different contexts. The quality of a pattern is calculated based on the extracted relationships (D-S relationships only) of the training corpus. In order to keep the patterns that are able to extract true relationships with high precision, we restrict the pattern on disease and symptom constraints. In contrast, the specificity of a pattern is calculated using the relationships (D-S or disease-any entity relationships) in the whole corpus when the pattern is relaxed on the symptom constrain. Specificity is used to keep the patterns that are specific to D-S relationships only.

4.3 Relationship Extraction

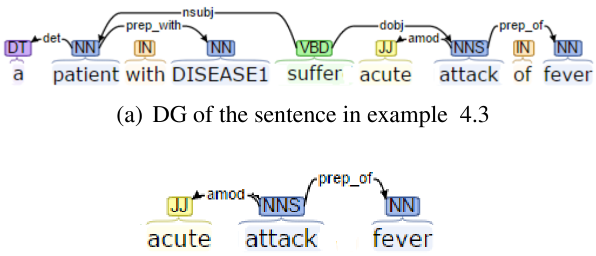
4.3.1 Pattern Relaxation for Unknown Symptoms

Patterns with $specificity \geq min\_specificity$ are relaxed on the symptom constraint, meaning that one entity must be annotated as a disease, but there is no requirement for the second entity to be annotated as a symptom. This enables us to identify symptoms that are not recognized by NER tools. Similarly to the learning phase, DGs are generated from the text to explore for D-S relationships. Then, a pattern matching between DGs and the pattern set is applied to extract D-S relationships.

4.3.2 Extraction of Complex Symptoms

During pattern matching, the word that matches with the node of the second entity (not constrained) is considered to be a symptom. Indeed, it is considered to be a symptom if this word is a leaf of the DG, but is considered as the "head" of a more complex symptom description if it is not a leaf. To extract the complete description of the symptom, we explore the subtree that has, as a head, the node that matched as a symptom. For example, if a pattern matching is applied to the sentence provided in example 4.3, we obtain a match with the pattern presented in Figure 2. In this case, the word that is considered to extract the symptom description is "attack". Exploring the subtree represented in Figure 3(b) enables us to reconstruct the full symptom that is involved in the relationship. This reconstruction uses every word of the subtree, dependency types plus the initial order of words to reconstruct the symptom description, "acute attack of fever" in our example. This example illustrates the usefulness of DGs in identifying and representing complex entities like symptoms.

Ex 4.3. *"A patient with <disease>Familial Mediterranean Fever</disease> suffered acute attacks of fever"*



(a) DG of the sentence in example 4.3

(b) The subtree of a complex symptom description

Figure 3: An example of complex symptom extraction

5 Experiments

5.1 Data Preparation

5.1.1 Rare Disease Corpus

Our rare disease corpus is composed of 121,796 PubMed abstracts obtained by querying PubMed with 457 rare diseases of OrphaData.[6] These diseases are selected because they fulfill following criteria: (1) they are associated with symptoms (namely "clinical signs") in OrphaData; (2) they can be mapped to an OMIM disease throught UMLS CUI; (3) their corresponding OMIM reference is annotated with symptoms (namely "clinical synopsis") in OMIM. This enables having a corpus of a reasonable size and guarantees that the selected diseases are associated with symptoms in both OrphaData and OMIM. This set of diseases and associated symptoms are used in subsection 5.5 to compare our relationships with the content of OrphaData and OMIM.

5.1.2 Preprocessing

The 121,796 abstracts are first split into 907,088 sentences using LingPipe[7]. These sentences are then annotated by MetaMap in order to label diseases, symptoms, genes, treatments and living beings with UMLS CUI. Finally, sentences that do not contain diseases are filtered out. Therefore, we obtained 301,599 sentences with at least one disease.

5.2 Pattern Learning

To learn patterns for D-S relationship extraction, 2,341 sentences with at least one disease and one symptom are kept. These sentences are split into a *learning corpus* made of 90% of sentences (randomly selected) and a *testing corpus* made of 10% of sentences. Both corpora are manually annotated by only one person to identify true and false relationships: the annotation task mainly requires linguistics and NLP skills. A true relationship is counted when a pair D-S is found and a relationship between them is actually mentioned in the text; whereas a false relationship is listed when the pair is found but no relationship is mentioned. The sentence in example 5.1 shows instances of both true and false relationships. "Schwartz Jampel syndrome"-"blepharospasm" is a true relationship, while "rare neuromuscular disorder"-"blepharospasm" is false. Table 2 shows the size of the learning and testing corpora in term of number of sentences, and of true and false relationships in each corpus. We use the Stanford parser to generate a DG for each sentence (de Marneffe et al., 2006). Shortest paths are computed from the 2,107 prepared DGs to generate 1,049 patterns. Figure 4 presents 7 examples of patterns generated.

Ex 5.1. *"<disease>Schwartz Jampel syndrome</disease> is a <disease>rare neuromuscular disorder</disease> characterized by <symptom>blepharospasm</symptom>"*

Corpus	#Sentences	#True Relations	#False Relations
learning	2,107	2,680	2,294
testing	234	330	326

Table 2: Size and content of the learning and testing corpora used for pattern learning and selection.

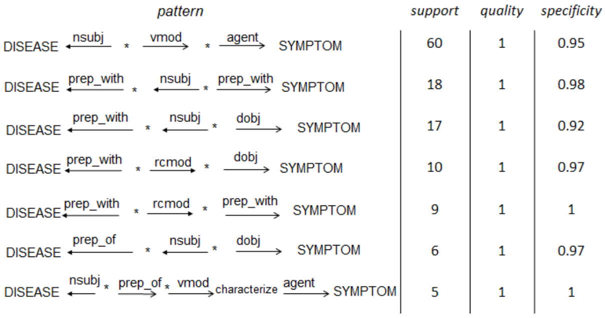
5.3 Pattern Selection

5.3.1 Quality-based Selection

Increasing the *min_support* value from 1, to 2, then to 3, reduces the number of patterns from 1,049, to 257, then to 118. To avoid rare patterns that can result from parser errors or complex sentences, we fixed *min_support* = 2.

[6] The list of 457 rare diseases is available at https://sourceforge.net/projects/spare2015/files/457-diseases

[7] LingPipe website: http://alias-i.com/lingpipe/



pattern	*support*	*quality*	*specificity*
$\text{DISEASE} \xleftarrow{nsubj} * \xrightarrow{vmod} * \xrightarrow{agent} \text{SYMPTOM}$	60	1	0.95
$\text{DISEASE} \xleftarrow{prep\_with} * \xleftarrow{nsubj} * \xrightarrow{prep\_with} \text{SYMPTOM}$	18	1	0.98
$\text{DISEASE} \xleftarrow{prep\_with} * \xleftarrow{nsubj} * \xrightarrow{dobj} \text{SYMPTOM}$	17	1	0.92
$\text{DISEASE} \xleftarrow{prep\_with} * \xrightarrow{rcmod} * \xrightarrow{dobj} \text{SYMPTOM}$	10	1	0.97
$\text{DISEASE} \xleftarrow{prep\_with} * \xrightarrow{rcmod} * \xrightarrow{prep\_with} \text{SYMPTOM}$	9	1	1
$\text{DISEASE} \xleftarrow{prep\_of} * \xleftarrow{nsubj} * \xrightarrow{dobj} \text{SYMPTOM}$	6	1	0.97
$\text{DISEASE} \xleftarrow{nsubj} * \xrightarrow{prep\_of} * \xrightarrow{vmod} \text{characterize} \xrightarrow{agent} \text{SYMPTOM}$	5	1	1

Figure 4: 7 examples of patterns from our pattern set and their *support*, *quality* and *specificity*.

We fixed *min_quality* $= 0.5$ to reduce our selected patterns to 235. This choice is guided by the *F-measure* that we computed for each *quality* threshold, as presented in Figure 5. This optimal *F-measure* is 56.97% (*precision* 87.97%, *recall* 42.12%) on the testing corpus. If used on the testing corpus, these 235 patterns extract 139 true relationships and 19 false relationships on a total of 330 relationships. Formulas for precision, recall and F-measure are recalled hereafter:

$$precision = \frac{all\ true\ extracted\ relations}{all\ extracted\ relations} \tag{3}$$

$$recall = \frac{all\ true\ extracted\ relations}{all\ relevant\ relations} \tag{4}$$

$$F-measure = 2 * \frac{precision * recall}{precision + recall} \tag{5}$$

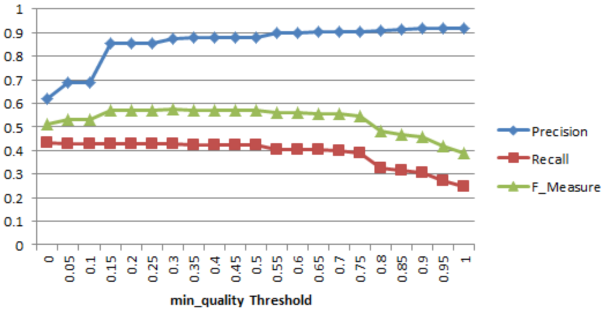


Figure 5: The effect of quality threshold on precision, recall and F-Measure values

5.3.2 Specificity-based Selection

For computing the pattern specificity, all sentences that contain at least one disease and another UMLS entity are selected. This produces 9,233 sentences. Then, all the 235 previously selected patterns are applied to the DGs of these sentences, resulting in the extraction of 5,197 D-S relationships and 391 disease-non symptom relationships (182 disease-gene, 182 disease-treatment, 27 disease-living being relationships). Finally, the specificity of each pattern is computed (see formula 2). Figure 6 shows that using *min_specificity* $= 0.5$ achieves the best *F-measure*, 55.65% (*precision* $= 89.87\%$ and *recall* $= 40.3\%$), on the testing corpus. Finally we keep 220 patterns with *quality* ≥ 0.5 and $specificity \geq 0.5$.[8]

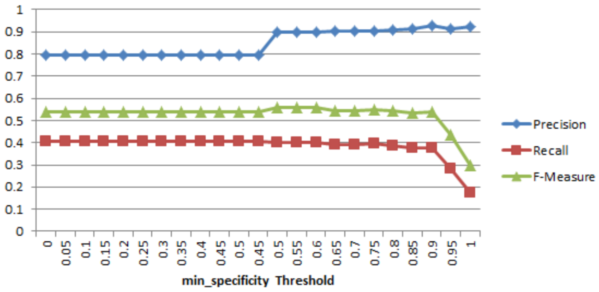


Figure 6: The effect of specificity threshold on precision, recall and F-Measure values

5.4 Application of Relationships Extraction

We applied selected patterns to the whole corpus[9] (301,599 sentences with at least one disease). The extracted relationships are divided into two groups. The first group contains 4,886 D-S relationships where symptoms were previously recognized by MetaMap. The second group contains 6,572 D-S relationships where symptoms were not recognized by MetaMap. After manual checking[10], these extractions achieved respectively 90.69% and 83.13% *precision*. The number of distinct symptoms in the second group is 3,849.

5.5 Comparison with Phenotype Databases

5.5.1 Comparison Approach

The novelty of extracted relationships is evaluated based on the comparison with D-S relationships available in OrphaData, and in OMIM. Results of the comparison are categorized into 3 groups: matched, partial matched and new relationships. To realize this comparison, it is required to map

[8] The list of 220 patterns is available at http://sourceforge.net/projects/spare2015/files/220Patterns

[9] The whole corpus is used (including the training and testing corpus) because the purpose of this task is to extract as much as possible D-S relationships and then, to compare them to the content of phenotype databases.

[10] The manual checking is done by only one person.

diseases of extracted relationships to OMIM diseases. Indeed, MetaMap provides, for each extracted disease, a UMLS CUI that may be mapped to OMIM.

For symptom mapping, we implemented a similarity measure to evaluate the similarity between the extracted symptom and those referenced in OMIM clinical synopsis and HPO. Our similarity value is based on the Jaccard index and is computed following formula:

$$Jaccard\ Index = \frac{text\_words \cap symp\_words}{text\_words \cup symp\_words} \quad (6)$$

where $text\_words$ are the words of the extracted symptom string and $symp\_words$ are the words that are describing a symptom defined either in OMIM or HPO.

Before computing the Jaccard index, each word in the extracted symptom and HPO (or OMIM) symptom is replaced by its lemma, stop words[11] are removed and a list of synonyms from WordNet (Fellbaum, 1998) is associated with each word. The synonym list of a word is used in case of the word does not match with any other word. The similarity value is then computed by the Jaccard index. For each symptom, the first three closest symptoms found in OMIM and HPO (six in total) are manually checked to select the best match if exists. A label "exact", "partial" or "new" is assigned to express if the match is exact or partial, or if the symptom is not listed in OMIM and HPO, thus considered as new.

5.5.2 Comparison Results

The relationships in the first group are compared automatically to Orphadata and OMIM relationships (because both their disease and symptom are associated with a UMLS CUI). The number of true D-S relationships is 4,431, including 803 relationships available in OrphaData and 646 available in OMIM. The union of these 2 sets counts up to 1,074 distinct D-S relationships already listed in OrphaData, OMIM or in both. Consequently, about 3,357 D-S relationships are potentially new and must be added to phenotype databases.

Regarding the relationships in the second group, the extracted symptoms are mapped to symptoms in HPO and OMIM. In this step, 3,236 symptoms (from 3,849 distinct symptoms in the relationships) are mapped to HPO and OMIM symptoms. The extracted relationship pairs are then compared to relationships in HPO and OMIM, which results in 1,422 matched relationships. As a result, we identified 613 $(3,849-3,236)$ new symptoms descriptions that may be of interest in rare disease studies and 4,041 $(5,463-1,422)$ potentially new D-S relationships[12][13].

6 Discussion

In SPARE, the choice of $min\_quality$ and $min\_specificity$ have important consequences on the results of the relationship extraction. Figures 5 and 6 show how the quality of the extraction changes when these two values are changed. In both cases, we observe relatively few evolution of the F-Measure. In Figure 5, $min\_quality$ between 0.35 and 0.5 achieve the best *F-measure* of 56.97%. They give the same result because the number of extracted patterns with $min\_quality$ between 0.35 and 0.5 is the same (235 patterns). Consequently, we chose arbitrarily $min\_quality = 0.5$. As shown in Figure 6, we chose $min\_specificity = 0.5$ because it achieves the best F-Measure. The result of F-Measure is constant when $min\_specificity$ between 0 and 0.45 because the number of patterns in this interval is the same.

We obtain a relatively good precision but a low recall. We Consider that a larger corpus for learning patterns could enable us to increase the recall. Our learning corpus is annotated manually with true and false relationships and increasing its size would require annotating additional relationships.

The corpus used in the learning task is relatively small, subsequently it is not enough to train ML methods. We propose to increase the size of the annotated corpus in order to apply ML methods on this corpus and compare with the results of our SPARE method.

Studying the novelty of our extracted relationships requires the comparison with the relationships of phenotype databases. For now, this comparison is semi-automatic and partial matching relies on a rather naive similarity measure. We

[11] We considered stop words listed in http://xpo6.com/list-of-english-stop-words/

[12] A list of extracted D-S relationship examples is available at https://sourceforge.net/projects/spare2015/files/D-SRelationsExamples.csv

[13] A list of extracted symptom examples is available at https://sourceforge.net/projects/spare2015/files/symptom-examples

would like to develop a more systematic approach by enabling a fine-grained comparison of phenotype descriptions. This could be achieved by normalizing then, comparing DGs of symptom descriptions.

SPARE method is a supervised classification process, in which threshold is selected manually. This selection can be computed automatically by considering the best F-Measure value.

7 Conclusion

In this paper, we proposed a pattern-based method that we call SPARE for extracting D-S relationships. The patterns are learned from shortest paths observed between the entities of interest (diseases and symptoms) within DGs. Using only the shortest path is simple and it captures the most important features required to describe the relationship between two entities. For extracting relationships involving rare or complex symptoms, we selected a subset of patterns that are specific to D-S relationships. In turn, a DG is helpful to extract and define complex symptoms, that are not recognized by other tools such as MetaMap. The novelty of relationship extracted has been compared with relationships listed in OrphaData and OMIM. This shows the ability of the SPARE to discover existing and potentially new relationships and the ability to identify new and complex symptom as well.

Acknowledgments

This work is supported by the ANR (French National Research Agency), project Hybride ANR-11-BS02-002.

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, DL '00, pages 85–94, New York, NY, USA. ACM.

A. Airola, S. Pyysalo, J. Björne, T. Pahikkala, F. Ginter, and T. Salakoski. 2008. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC bioinformatics*, 9 Suppl 11.

A. R. Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. *Proc AMIA Symp*, pages 17–21.

Nicolas Béchet, Peggy Cellier, Thierry Charnois, Bruno Crémilleux, and Marie-Christine Jaulent. 2012. Sequential pattern mining to discover relations between genes and rare diseases. In *CBMS*, pages 1–6.

Sebastian Blohm, Krisztian Buza, Philipp Cimiano, and Lars Schmidt-Thieme, 2011. *Relation Extraction for the Semantic Web with Taxonomic Sequential Patterns*, pages 185–209. Applied Semantic Web Technologies. Taylor and Francis Group.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database-Issue):267–270.

Markus Bundschus, Mathaeus Dejori, Martin Stetter, Volker Tresp, and Hans-Peter Kriegel. 2008. Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics*, 9(1):207.

Razvan C. Bunescu and Raymond J. Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 724–731, Stroudsburg, PA, USA. Association for Computational Linguistics.

Razvan Bunescu, Raymond Mooney, Arun Ramani, and Edward Marcotte. 2006. Integrating co-occurrence statistics with information extraction for robust retrieval of protein interactions from medline. In *Proceedings of the HLT-NAACL Workshop on Linking Natural Language Processing and Biology (BioNLP'06)*, pages 49–56, New York, NY, June.

Peggy Cellier, Thierry Charnois, and Marc Plantevit. 2010. Sequential patterns to discover and characterise biological relations. In A. F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing (CICLing)*, LNCS 6008, pages 537–548. Springer.

Md. Faisal Mahbub Chowdhury and Alberto Lavelli. 2012. Combining tree structures, flat features and patterns for biomedical relation extraction. In *EACL*, pages 420–429.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *IN PROC. INT'L CONF. ON LANGUAGE RESOURCES AND EVALUATION (LREC*, pages 449–454.

A. Divoli and T. K. Attwood. 2005. BioIE: extracting informative sentences from the biomedical literature. *Bioinformatics*, 21(9):2138–9.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.

Carol Friedman, Pauline Kra, Hong Yu, Michael Krauthammer, and Andrey Rzhetsky. 2001. Genies: a natural-language processing system for the extraction of molecular pathways from journal articles. *Comput. Appl. Biosci.*, 17(suppl_1):S74–82, June.

Jörg Hakenberg, Conrad Plake, Ulf Leser, Harald Kirsch, and Dietrich Rebholz-schuhmann. 2005. Lll'05 challenge: genic interaction extraction – identification . . . with alignments and finite state automata. In *IN PROC LEARNING LANGUAGE IN LOGIC WORKSHOP (LLL05) AT THE 22ND INT CONF ON MACHINE LEARNING*, pages 38–45.

Mohsen Hassan, Adrien Coulet, and Yannick Toussaint. 2014. Learning subgraph patterns from text for extracting disease - symptom relationships. In *Proceedings of the 1st International Workshop on Interactions between Data Mining and Natural Language Processing co-located with The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, DMNLP@PKDD/ECML 2014, Nancy, France, September 15, 2014.*, pages 81–96.

Sebastian Köhler, Uwe Schoeneberg, Johanna C. Czeschik, Sandra C. Doelken, Jayne Y. Hehir-Kwa, Jonas Ibn-Salem, Christopher J. Mungall, Damian Smedley, Melissa A. Haendel, and Peter N. Robinson. 2014. Clinical interpretation of CNVs with cross-species phenotype data. *Journal of Medical Genetics*, pages jmedgenet–2014–102633+, October.

Martin Krallinger, Florian Leitner, Carlos Rodriguez-Penagos, and Alfonso Valencia. 2008. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome biology*, 9 Suppl 2(Suppl 2):S4+.

Robert Leaman and Zhiyong Lu. 2014. Disease named entity recognition and normalization with dnorm. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '14, pages 587–587, New York, NY, USA. ACM.

Insuk Lee, Zhihua Li, and Edward M. Marcotte. 2007. An improved, bias-reduced probabilistic functional gene network of baker's yeast, saccharomyces cerevisiae. *PLoS ONE*, 2(10).

Haibin Liu, Karin Verspoor, Donald C. Comeau, Andrew MacKinlay, and W John Wilbur. 2013. Generalizing an approximate subgraph matching-based system to extract events in molecular biology and cancer genetics. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 76–85, Sofia, Bulgaria, August. Association for Computational Linguistics.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

Laure Martin, Delphine Battistelli, and Thierry Charnois. 2014. Symptom extraction issue. In *Proceedings of BioNLP 2014*, pages 107–111, Baltimore, Maryland, June. Association for Computational Linguistics.

Monica Mazzucato, Laura Visonà Dalla Pozza, Silvia Manea, Cinzia Minichiello, and Paola Facchin. 2014. A population-based registry as a source of health indicators for rare diseases: the ten-year experience of the Veneto Region's rare diseases registry. *Orphanet Journal of Rare Diseases*, $item.volume:37+, March.

A.K. Ramani, R.C. Bunescu, Raymond J. Mooney, and E.M. Marcotte. 2005. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biology*, 6(5):r40.

Jasmin Šarić, Lars Juhl Jensen, Rossitza Ouzounova, Isabel Rojas, and Peer Bork. 2006. Extraction of regulatory gene/protein networks from medline. *Bioinformatics*, 22(6):645–650, March.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *J. Mach. Learn. Res.*, 3:1083–1106, March.

Min Zhang, GuoDong Zhou, and Aiti Aw. 2008. Exploring syntactic structured features over parse trees for relation extraction using kernel methods. *Inf. Process. Manage.*, 44(2):687–701, March.