Learning Formal Definitions for Biomedical Concepts

George Tsatsaronis[1], Alina Petrova[1], Maria Kissa[1], Yue Ma[2], Felix Distel[2], Franz Baader[2], and Michael Schroeder[1]

[1] Biotechnology Center, Technische Universität Dresden
{george.tsatsaronis,alina.petrova,maria.kissa,ms}@biotec.tu-dresden.de
[2] Institute of Theoretical Computer Science, Technische Universität Dresden
{mayue,felix,baader}@tcs.inf.tu-dresden.de

Abstract. Ontologies such as the *SNOMED Clinical Terms* (SNOMED CT), and the *Medical Subject Headings* (*MeSH*) play a major role in life sciences. Modeling formally the concepts and the roles in this domain is a crucial process to allow for the integration of biomedical knowledge across applications. In this direction we propose a novel methodology to learn formal definitions for biomedical concepts from unstructured text. We evaluate experimentally the suggested methodology in learning formal definitions of *SNOMED CT* concepts, using their text definitions from *MeSH*. The evaluation is focused on the learning of three roles which are among the most populated roles in *SNOMED CT*: *Associated Morphology*, *Finding Site* and *Causative Agent*. Results show that our methodology may provide an *Accuracy* of up to 75%. For the representation of the instances three main approaches are suggested, namely, *Bag of Words*, *word n-grams* and *character n-grams*.

1 Introduction

The biomedical domain is characterized by an exponential growth in the produced data volumes, primarily scientific published articles, knowledge and databases, nucleotide sequences and protein structures. To handle such amounts of data and information, the notion of organizing the biomedical knowledge using ontologies has been the focus point of many initiatives and activities in the biomedical domain [6]. The basic motivation is that since ontologies represent a conceptualization of how things are organized in reality in the underlying domain, this formal representation may provide an actual language for the community, with which they can talk about entities and concepts, and exchange data in the same representation. Moreover, sharing the same conceptualization of entities in the biomedical domain allows researchers to communicate new facts and knowledge referring to the same concepts that may be found with different labels across several different data sources.

More formally, an ontology is a set of logical axioms which model the reality of the domain. With the advent of *description logics* (*DL*) [2] and *OWL*'s *description logic* flavor *OWL DL* [www.w3.org/TR/owl-guide/], the task of designing and implementing formally ontologies has become easier, as the ontology engineers may express the ontology concepts and their relations without losing computational completeness, and in parallel retain decidability of reasoning systems. In practice *DL* has become the leading formalism for representing ontologies, a trend which nowadays is also supported by many popular ontology editors such as *Protégé* [protege.stanford.edu/] and *OBO-Edit*[oboedit.org/]. Notably, many large biomedical ontologies have adopted this formalism, such as *GALEN* [17], which was also the first biomedical ontology to be developed in *DL* and the *NCI Thesaurus* [ncit.nci.nih.gov/].

In particular, *SNOMED CT* [www.ihtsdo.org/snomed-ct/] has adopted the lightweight description logic $\mathcal{EL}++$, which allows for tractable reasoning. For several years now, research on how other biomedical ontologies may be translated in *DL* has been conducted [9, 21, 12]. Strickingly, the application of formal ontologies in the biomedical domain has produced interesting results, e.g., the works of Rubin *et al.* [19] and King *et al.* [13] to name a few. In the former work, the authors used the *Foundational Model of Anatomy* (*FMA*) ontology to develop a methodology through which they can automate reasoning about penetrating injuries. In the latter work the authors presented *Adam*, a laboratory robot that can perform independent experiments to test hypotheses and interpret findings without human guidance.

It is, thus, evident that coherent formalization of biomedical ontologies has valuable applications in the biomedical domain. In this work we present a novel methodology to learn formal definitions of biomedical concepts from their textual definitions, which can be considered as a first step towards the automated process of creating formal biomedical ontologies. We approach the problem from three different perspectives: (i) learning the *Bag of Words* (*BOW*) representation that participate in the expression of each role[3] within the textual definitions of concepts, (ii) learning the word *n-grams* that participate in the expression of each role, and, (iii) learning the character *n-grams* that participate in the expression of each role. The first approach is a standard representation methodology in text mining, while the main difference between the other two approaches is that the former considers the order of words in the definitions and the fact that the words may form composite terms, i.e., *word n-grams*, while the latter considers the order of characters, i.e., *character n-grams*. Finally, we merge the three representations into one, by combining their features, in an effort to get the best of all worlds.

The rest of the paper is organized as follows. Section 2 refers to related work pertaining to the generation of formal definitions from unstructured text. Section 3 introduces formally the problem that we are addressing. Section 4 introduces the methodologies used, and describes how the textual definitions we are using for our analysis were obtained and annotated. Section 5 presents the results of our experimental evaluation and discusses the findings, and Section 6 concludes and provides pointers to possible extensions, applications and future work.

2 Related Work

As argued above, formal ontologies are useful, but their creation is a labor intensive task. Hence, it is desirable to automate aspects of it. Towards the direction of processing automatically text descriptions from biomedical ontologies, there exists much related work (e.g., [3, 5]); however, most of these approaches assume the existence of an

[3] For the remaining of the paper, the words *property* and *role* might be used interchangeably to refer to the properties of the concepts.

ontology to be enriched, while our method is mostly related to approaches that can create an ontology from scratch. Under this scope, in the following we refer to approaches that aim to generate axioms from unstructured text.

2.1 Domain Agnostic

Several approaches use a deep syntactic analysis of natural language definitions [24], and others apply lingustic patterns [20]. Fuzzy logic components have been also developed towards the automated generation of logical axioms [16], whilst semantic analysis and word sense disambiguation have also contributed [1] towards the completion of the task of generating formal descriptions of entities and relations. Finally, there exist also running systems which are designed to generate ontologies from text, an example being *Text-To-Onto* [15], which, however, in their majority do not take into account axioms and instances.

2.2 Biomedical Domain

In the biomedical domain little work exists concerning the automated axiom generation and learning of formal representations [14]. More specifically, regarding *MeSH*, to the best of our knowledge there is only one work that attempts its representation in *OWL*, but in an indirect way that makes use of *CISMeF* (Catalogue and Index of French-speaking Medical Sites) which encapsulates the French version of *MeSH* [21]. Other works that target the expansion of *MeSH*, attempt to automatically suggest its expansion with synonyms and provide alternative definitions for the concepts, but not in a formalized way [25].

A non-exclusive list of other recent examples of works that attempt to model formally concepts in the biomedical domain are the works by: Boelling *et al.* [7], who attempt to model biochemical processes; Chepelev and Dumontier [8], who define the *Chemical Entity Semantic Specification* (*CHESS*) for the representation of polyatomic chemical entities, their substructures, bonds, atoms, and reactions using Semantic Web technologies; Stenzhorn *et al.* [23], who attempt to map clinical documentation to the formal representation of *SNOMED CT*; Jupp *et al.* [11], who introduce *Populous*, a tool that may populate ontologies from the analysis of spreadsheets; and, finally, the work by Hastings *et al.* [10], who use *OWL* and description graphs to represent classes of chemical entities, such as molecules, ions and groups.

Evidently, there is significant research work towards representing the biomedical knowledge using a formal representation, yet there is a gap regarding the generation of formal definitions from unstructured biomedical text. Our work addresses this open challenge by presenting a novel methodology for extracting formal definitions from unstructured text. We argue that the proposed methodology may aid the time-consuming and demanding process of ontology generation, evolution and maintenance, by constituting the first step to transit successfully from unstructured text to the extraction of formal biomedical concept definitions.

3 The Problem of Learning Formal Definitions

3.1 Formal Definitions in *SNOMED CT*

SNOMED CT ([22]) is a medical ontology describing concepts such as anatomical structures, disorders, organisms, and it is becoming adopted by a growing number of countries worldwide as a reference vocabulary in clinical research [18]. Its underlying structure is based on formal logics, more specifically on the *lightweight Description Logic* $\mathcal{EL}++$. *Description Logics* (*DLs*) can express a rich network of different types of relationships between concepts. For example, using the SNOMED CT vocabulary one can express that the concept Baritosis is caused by Barium_dust by writing:

$$\mathsf{Baritosis} \sqsubseteq \exists \mathsf{Causative\_agent.Barium\_dust} \tag{1}$$

linking the two concepts with the relationship Causative_agent. Other well populated roles include Finding_site, and Associated_morphology. Another example from SNOMED CT is the formal definition of the *Fox-Fordyce Disease*, in which we can find that its *finding sites* are *Apocrine Glands* and *Intraepidermal Apocrine Ducts*, its *causative agents* are *Obstruction* and *Rupture* and its *Associative Morphology* is *Papular Eruptions*. The aforementioned example may be formally written as follows:

$$\mathsf{Fox} - \mathsf{Fordyce\_Disease} \sqsubseteq \exists \mathsf{Finding\_site.Apocrine\_glands}, \tag{2}$$

$$\exists \mathsf{Finding\_site.Intraepidermal\_apocrine\_ducts}, \tag{3}$$

$$\exists \mathsf{Causative\_agent.Obstructure}, \tag{4}$$

$$\exists \mathsf{Causative\_agent.Rupture}, \tag{5}$$

$$\exists \mathsf{Associative\_morphology.Papular\_eruptions} \tag{6}$$

3.2 Problem Formulation

The formal semantics of SNOMED CT are a key advantage, however, they come at a cost. Adding new concepts to a formal ontology is a tedious, costly and error-prone process, that needs to be performed manually by specially trained knowledge engineers. The suggested methodology can provide assistance in this process by automatically extracting the relations between concepts from text. The approach is based on the assumption that the set of roles (relations) remains relatively stable while the set of concepts constantly increases. To facilitate the addition of new concept description, we formulate the following problem: for a given input sentence in natural language that is annotated with two SNOMED CT concepts decide whether the sentence describes a role between the two concepts and which role precisely.

More formally, we can express this problem as a classification problem. Let C be the class label based on which the training of the classifiers takes place. C can be the label of any role R contained in SNOMED CT. Each example (instance) is a sentence, denoted with I, which is annotated with SNOMED CT concepts, and for which a set of features X has been computed, which are explained in Section 4. Thus, $I = [X_1, ..., X_N]$. If I is a sentence which describes a role R_i between two SNOMED CT concepts, where $R_i \in R$, then I is a positive example for this role, and, hence $C = R_i$ in this case. Therefore, the problem of role extraction from unstructured text can be seen as a multi-class classification problem, following the aforementioned representation. In the following section we explain in detail the methodology of building a data set for learning three SNOMED CT roles, which are among the most widely populated in SNOMED CT.

4 Methodology

The methodology comprises three steps: (i) creating a data set with labelled instances from which the roles can be learned based on a set of features, (ii) representing formally the instances, for which we explore three different representations, and, (iii) using a machine learning methodology to train a model in the produced data set following the suggested instance representation that may recognize any of the labelled roles in an unseen input sentence. For the purposes of our evaluation this latter step is conducted on the created data set, using $10-$fold cross validation to measure the performance of the tested methods. Steps (i) and (ii) are explained respectively in Sections 4.1 and 4.2 that follow.

4.1 Generating a Data Set to Learn SNOMED CT Roles

In order to obtain high quality sentences that describe the role between two SNOMED CT concepts we need to obtain sentences that primarily contain both concepts, and in turn, describe a relation between the two. For that purpose, we chose to use *MeSH* definitions that contain SNOMED CT concepts, since *MeSH* definitions are produced manually by medical experts, and, thus, constitute precise, scientifically valid, and high quality sentences.

The first step for the aforementioned transition is to obtain a mapping between *MeSH* and SNOMED CT concepts. Such a mapping exists via the *Unified Medical Language System* (*UMLS*). *UMLS* defines a *Concept Unique Identifier* (*CUI*) for each of the *UMLS* concepts. Each *CUI* may be associated with one or more concepts from external libraries or thesauri, such as *MeSH* and SNOMED CT. Analyzing this association, we extracted the *CUI*s that are associated with both a *MeSH* and a SNOMED CT concept, which is interpreted as a mapping between the two concepts. Using the latest *UMLS* version (*2012AB*), we obtained in this manner a total of $21,461$ mappings.

Next, we used the produced mappings to create a high quality data set for learning roles between SNOMED CT concepts. For the purposes of our data set creation we focused into three widely populated roles in SNOMED CT, namely *Associated Morphology* (*AM*), *Causative Agent* (*CA*) and *Finding Site* (*FS*). To explain in detail the process of the data set creation we define the following notation. Let R_i be a SNOMED CT role, where $R_i \in AM, CA, FS$. Let A and B be two SNOMED CT concepts that populate R_i such that: $\mathsf{A} \sqsubseteq \exists \mathsf{R_i}.\mathsf{B}$. From all of the *MeSH* definitions we retained only those which define any A involved in R_i and we further filtered with the definitions that contain B. After filtering there were 424 *MeSH* definitions remaining. For the purposes of filtering we identified the definitions that contain B through annotating

Role	Associated Morphology	Causative Agent	Finding Site
Number Of Instances	121	95	208
Word Occurrences	938	723	$1,550$
Avg. # of Words	7.75	7.61	7.45
# of Distinct Words	433	218	547

Table 1. Description of the produced data set. The data set contains in total 424 instances from three SNOMED CT roles: *Associated Morphology*, *Causative Agent* and *Finding Site*.

Annotated Sentence	"*Baritosis*/Baritosis_(disorder) is pneumoconiosis caused by *barium dust*/Barium_Dust_(substance)."		
SNOMED CT relationship	Baritosis_(disorder) — Causative_agent — Barium_Dust_(substance)		
Alignment	left type	between-words	right type
	$disorder$	"is pneumoconiosis caused by"	$substance$
BoW	{is,pneumoconiosis,caused,by}		
Word n-grams	{is,pneumoconiosis,caused,by,is pneumoconiosis,pneumoconiosis caused,caused by}		
Char. n-grams	{i,s, ,p,n,e,u,m,o,c,a,d,b,y,is,s , p,pn,ne,eu,um,mo,oc,co,on,ni,io,os,si, c,ca,au,us,se,ed,d , b,by}		

Table 2. Text alignment and example of an instance representation using boolean feature values. For the *n-gram* representations a value of $n = 2$ is used.

the sentences with SNOMED CT concepts. For the annotation we used two different tools: (a) *Metamap* [metamap.nlm.nih.gov/], which may annotate any text with *UMLS* concepts, and, (b) *SnomedAnnotator* developed in house, which may annotate any text with SNOMED CT concepts. The two annotators were used sequentially to provide a broader coverage of annotations; hence, we considered the union of the provided annotations from the two tools. Following the aforedescribed steps, the produced dataset contains 424 instances in total for the three roles (*AM*, *CA* and *FS*). Its details are summarized in Table 1.

4.2 Instance Representation for Learning SNOMED CT Roles

Using the aforedescribed dataset, we can now proceed with providing a description of how the features can be generated, with which the instances may be represented for the learning process. For the feature engineering, we use three approaches: (i) *Bag of Words*, (ii) *Word n-grams*, and, (iii) *Character n-grams*. The three approaches are described next, and are summarized with an example in Table 2. The example is drawn from the formal description presented in Equation 1. In all three approaches, the annotated sentences are split in a way such that the words that occur between A and B, may be isolated and processed. The basic assumption behind this alignment lies in the hypothesis that each role R_i has a characteristic way of being expressed in natural language text, which may be captured by the analysis of the words that occur between concepts A and B. All three representations have a default feature weight equal to the value of 1 if they occur in this text, or 0 otherwise. We also expand these representations to their weighted versions, i.e., instead of boolean representation of the features, a real value is used.

Bag of Words (*BOW*) Representation: The representation of text following the *Bag of Words* model has been used traditionally both in the fields of *information retrieval* and *text mining* [4]. According to this representation, each distinct term constitutes a dimension of the collection. More formally, in our case let T be the text between A and B in an annotated sentence (instance). T is naturally a series of ordered words, e.g., $T = [w_1 w_2 w_3 ... w_k]$. The *BoW* representation of this instance will be the unordered set of all unique words $w_i \in T$. Thus, the feature space according to *BoW* comprises the union of all unique terms appearing in all text definitions T. Each instance can then be represented as a set of features $X_i : w_i$. In its simple (unweighed) version, as a value of each feature we use 0 or 1 (boolean representation), depending on whether $X_i = w_i$ occurs in T (1) or not (0).

Word *n-grams* Representation: Given T and an assigned value to a parameter n, we can expand the *BoW* representation in order to represent each instance with all the possible *word n-grams* occurring in T. For the extraction of the *word n-grams* we are using a sliding window of search in the ordered words of T. Note also that this representation includes at least all features of the *BoW* representation; in fact, if $n = 1$, the *word 1-gram* representation is reduced to the *BoW* representation. Regarding the weight of each feature, in the simple (unweighed) version, we use a boolean representation, as previously.

Character *n-grams* Representation In an analogy to representing instances at a *word n-gram* level, we can also represent instances at the *character n-gram* level. Given again T and a value for the parameter n, we now examine T as an ordered series of characters, instead of words. For the extraction of the *character n-grams*, as in the case of *word n-grams*, we are using a sliding window of search in the ordered characters of T, and we do not exclude space characters. Again the weight of each feature in the simple (unweighed) version, follows a boolean representation, as previously.

Weighted Feature Representations: In all of the three aforedescribed feature representations of instances, we have assumed a boolean representation for the feature values. Ideally, we would like to have a real value v_i for each feature X_i acting as a weight that would discriminate their flat contribution of the boolean representation. For this purpose, we utilize the dataset and define a global weight for each feature, which can be computed always on the part of the dataset kept for training. A local weight is not a realistic option, as the *MeSH* definition sentences are usually short, significantly shorter than text passages or documents. Hence, for each feature X_i of any of the aforedescribed representations, we define a weight: $v_i = P(X_i)$, where $P(X_i)$ is the probability of occurrence of feature X_i in the training corpus. However, since the training corpus contains instances from several roles R_i (class labels), it is important to discriminate the probabilities of the features' occurrences per role (class). Hence, to create the weighted representations of instances following any of the aforementioned schemes, we create for each feature X_i, k features, where k is the number of all roles (labels): $X_{i1}...X_{ik}$, with the real value of each feature being: $v_{im} = PX_i|R_m$, and $m \in [1..k]$. Thus, the weight of each feature is the probability of its occurrence in the respective role, and each instance in the weighted version may be now represented with $k * X$ features.

Combined Feature Representations: A final consideration is the representation of the instances using the union of all the features that were described in each case. This potentiality can show whether the synergy of *word n-grams* and *character n-grams* may provide better predicting power for the extraction of roles from unstructured text. Naturally, this combined representation can be utilized both for the weighted and the unweighed versions of the aforedescribed instance representations.

5 Experimental Evaluation

The experimental evaluation was conducted on the dataset that was created as explained in Section 4.1 and summarized in Table 1. As machine learning methodologies we compared four different state of the art supervised algorithms, namely: (1) *Logistic Regression*, (2) *Support Vector Machines*, (3) *Multinomial Naive Bayes*, and, (4) *Random Forests*[4]. For the evaluation we apply $10-$fold cross validation and for performance measuring we report on the *overall accuracy*, *precision*, *recall* and *F-Measure* per role (*AM*, *CA*, and *FS*), and *macro-averaged precision*, *recall* and *F-Measure* over all roles. The results are reported in Tables 3 and 4 for the unweighed and weighted versions of the instance representations respectively.

Analyzing the results from the perspective of how difficult it is to learn a model than can recognize each of the roles, the reported numbers examining both tables suggest that the easiest role to identify is *CA*, with an *F-Measure* that can reach up to 82.9%, the second easiest is *FS*, with an *F-Measure* that can reach up to 80.3%, and the hardest role is *AM*, with an *F-Measure* that can reach up to 65.8%. A second finding when examining the results from the point of view of the underlying feature representation is that the character *n-grams* tend to report better results than the rest representations. More precisely, the character *3-grams* report the top performance in terms of *accuracy* and *macro-averaged F-Measure* (75.71% and 74.91% respectively). In addition, the *combined* representation does not seem to boost the performance of the character *n-grams*, which is probably due to the fact that the word *n-grams* cannot perform individually equally well as the character *n-grams*. However, further ensemble or combination, or feature selection methodologies will have to be explored in the future in order to assess whether there are feature subsets that can boost the overall performance with the concurrent reduction of the feature space complexity. Considering the tested classifiers, though the performance differences in absolute numbers are subtle, *SVM* tends to produce the better *accuracy* and *F-Measure* in the majority of the cases. Furthermore, in

[4] The standard *Weka* v3.6 platform implementations were used

Method	ML	Acc.	AM			CA			FS			All		
			P	R	F	P	R	F	P	R	F	P	R	F
BoW	LR	68.63	54	72.7	62	70.9	82.1	76.1	82.8	60.1	69.6	69.23	71.63	69.23
	SVM	69.1	63.6	34.7	44.9	83.3	78.9	81.1	65.7	84.6	73.9	70.87	66.07	66.63
	NB	66.27	57.5	38	45.8	64.8	85.3	73.6	70.3	74	72.1	64.2	65.77	63.83
	RF	66.5	50.8	52.1	51.4	72.9	82.1	77.2	73.1	67.8	70.3	65.6	67.33	66.3
Word 2-grams	LR	70.04	53.8	63.6	58.3	84.3	73.7	78.7	75.8	72.1	73.9	71.3	69.8	70.3
	SVM	66.03	54.5	44.6	49.1	73.2	74.7	74	68	74.5	71.1	65.23	64.6	64.73
	NB	63.91	52.7	40.5	45.8	59.3	90.5	71.7	73.1	65.4	69	61.7	65.46	62.16
	RF	65.56	51.7	50.4	51	82.1	67.4	74	67.1	73.6	70.2	66.96	63.8	65.06
Word 3-grams	LR	68.86	51.2	72.7	60.1	90.3	68.4	77.8	77.2	66.8	71.6	72.9	69.3	69.83
	SVM	66.03	54.9	51.2	53	71.6	76.8	74.1	69.4	69.7	69.5	65.3	65.9	65.53
	NB	63.44	53.6	43	47.7	56.4	92.6	70.1	75.4	62	68.1	61.8	65.86	61.96
	RF	65.09	50.8	51.2	51	79	67.4	72.7	67.9	72.1	69.9	65.9	63.56	64.53
Word 4-grams	LR	64.62	51.3	32.2	39.6	78.2	71.6	74.7	64	80.3	71.2	64.5	61.36	61.83
	SVM	66.74	53.8	52.9	53.3	73.7	73.7	73.7	71	71.6	71.3	66.16	66.06	66.1
	NB	62.5	54.2	43	47.9	54	91.6	68	75.4	60.6	67.2	61.2	65.06	61.03
	RF	64.22	49.6	48.8	49.2	87.5	66.3	75.4	65.2	73.1	68.9	67.43	62.73	64.5
Character 2-grams	LR	67.68	52.6	57.9	55.1	77.5	65.3	70.9	73.5	74.5	**80.3**	67.86	65.9	68.76
	SVM	74.29	61.3	60.3	60.8	87.2	78.9	**82.9**	76.3	80.3	78.2	74.93	73.16	73.96
	NB	62.97	46.7	35.5	40.4	69.4	71.6	70.5	66.7	75	70.6	60.93	60.7	60.5
	RF	67.68	59.6	48.8	53.6	84	66.3	74.1	66	79.3	72.1	69.86	64.8	66.6
Character 3-grams	LR	69.33	55	58.7	56.8	75.8	78.9	77.3	75.5	71.2	73.3	68.76	69.6	69.13
	SVM	**75.23**	69.8	61.2	65.2	83.7	75.8	79.6	74.6	83.2	78.6	76.03	73.4	**74.46**
	NB	67.92	58.9	46.3	51.9	68.5	80	73.8	71.6	75	73.2	66.33	67.1	66.3
	RF	68.39	64.3	44.6	52.7	73.3	66.3	69.6	68.1	83.2	74.9	68.56	64.7	65.73
Character 4-grams	LR	68.39	58.9	52.1	55.3	75.5	74.7	75.1	70	75	72.4	68.13	67.26	67.6
	SVM	75	69.1	62.8	**65.8**	78.4	80	79.2	76.5	79.8	78.1	74.66	74.2	74.36
	NB	67.68	58.9	43.8	50.2	66.7	80	72.7	71.8	76	73.8	65.8	66.6	65.56
	RF	63.91	58.5	45.5	51.2	69.5	60	64.4	64.1	76.4	69.7	64.03	60.63	61.76
Combined 2-grams	LR	67.45	52.5	43.8	47.7	80.7	74.7	77.6	68.9	77.9	73.1	67.36	65.46	66.13
	SVM	75	66.3	57	61.3	84.3	78.9	81.5	75.3	83.7	79.3	75.3	73.2	74.03
	NB	65.09	55	27.3	36.5	75.6	71.6	73.5	63.9	84.1	79.3	64.83	61	63.1
	RF	66.27	55.6	45.5	50	82.6	60	69.5	66	81.3	72.8	68.06	62.26	64.1
Combined 3-grams	LR	71.46	62.7	52.9	57.4	79.3	76.8	78.1	72.2	79.8	75.8	71.4	69.83	70.43
	SVM	74.52	65.3	63.6	64.4	83.9	76.8	80.2	75.8	79.8	77.8	75	73.4	74.13
	NB	67.68	58.2	38	46	71.2	77.9	74.4	69.3	80.3	74.4	66.23	65.4	64.93
	RF	69.1	58.8	47.1	52.3	82.1	67.4	74	69.1	82.7	75.3	70	65.73	67.2
Combined 4-grams	LR	66.98	57.1	49.6	53.1	74.4	70.5	72.4	68.6	75.5	71.9	66.7	65.2	65.8
	SVM	72.87	64.7	62	63.3	76.3	77.9	77.1	75.8	76.9	76.4	72.26	72.26	72.26
	NB	66.27	57.5	41.3	48.1	66.1	80	72.4	69.8	74.5	72.1	64.46	65.26	64.2
	RF	67.68	57.7	52.9	55.2	82.9	66.3	73.7	67.5	76.9	71.9	69.36	65.36	66.93

Table 3. Overall Accuracy (*Acc.*), *Precision* (*P*), *Recall* (*R*) and *F-Measure* (*F*) per role and over all roles for the unweighed representations. *Logistic Regression* (*LR*), *Support Vector Machines* (*SVM*), *Multinomial Naive Bayes* (*NB*) and *Random Forests* (*RF*) are compared.

Method	**ML**	**Acc.**	**AM**			**CA**			**FS**			**All**		
			P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
BoW	**LR**	66.74	52.5	34.71	41.79	84.33	73.68	78.65	65.51	82.21	72.92	67.45	63.53	64.45
	SVM	69.1	59.22	50.41	54.46	70.47	77.89	74	73.14	75.96	74.52	67.61	68.09	67.66
	NB	49.05	0	0	0	0	0	0	49.05	100	65.81	16.35	33.33	21.94
	RF	63.67	47.41	45.45	46.41	74.19	72.63	73.4	67.97	70.19	69.03	63.19	62.75	62.94
Word 2-grams	**LR**	68.16	60.49	40.49	48.51	77.41	75.78	76.59	67.2	80.76	73.36	68.36	65.67	66.15
	SVM	70.99	64.44	47.93	54.97	73.58	82.1	77.61	72.36	79.32	75.68	70.12	69.78	69.42
	NB	49.05	0	0	0	0	0	0	49.05	100	65.82	16.35	33.33	21.94
	RF	66.5	53.26	40.49	46	75.25	76.84	76.04	68.08	76.92	72.23	65.53	64.75	64.75
Word 3-grams	**LR**	69.81	64.7	45.45	53.39	77.89	77.89	77.89	68.44	80.28	73.89	70.34	67.87	68.39
	SVM	70.28	59.22	50.41	54.46	75.72	82.1	78.78	72.93	76.44	74.64	69.29	69.65	69.29
	NB	49.05	0	0	0	0	0	0	49.05	100	65.82	16.35	33.33	21.94
	RF	66.98	52.94	44.62	48.43	84.52	74.73	79.32	66.8	76.44	71.3	68.08	65.26	66.35
Word 4-grams	**LR**	67.21	59.3	42.14	49.27	76.92	73.68	75.26	66.39	78.84	72.08	67.53	64.88	65.53
	SVM	69.1	59.22	50.41	54.46	72.64	81.05	76.61	72.09	74.51	73.28	67.98	68.65	68.11
	NB	49.05	0	0	0	0	0	0	49.05	100	65.82	16.35	33.33	21.94
	RF	69.1	59.59	48.76	53.63	77.52	72.36	75	69.91	79.32	74.32	69	66.81	67.65
Character 2-grams	**LR**	66.04	53.96	61.98	57.69	65.59	64.21	64.89	75	69.23	72	64.85	65.14	64.86
	SVM	73.82	63.39	58.68	60.94	80.9	75.79	78.26	76.23	81.73	78.89	73.5	72.06	72.69
	NB	58.25	42.22	47.11	44.53	67.05	62.11	64.48	65.17	62.98	64.06	58.14	57.4	57.69
	RF	64.38	56.44	47.11	51.35	68.67	60	64.04	66.25	76.44	70.98	63.78	61.18	62.12
Character 3-grams	**LR**	72.16	60.74	53.71	57.01	80.43	77.89	79.14	74.22	80.28	77.13	71.79	70.62	71.09
	SVM	**75.71**	68.47	62.81	**65.52**	82.95	76.84	79.78	76.44	82.69	**79.45**	75.95	74.11	**74.91**
	NB	62.03	48.03	50.41	49.19	65.31	67.37	66.32	69.35	66.35	67.81	60.89	61.37	61.10
	RF	68.87	58.49	51.24	54.63	85.71	63.16	72.73	68.55	81.73	74.56	70.91	65.37	67.3
Character 4-grams	**LR**	71.46	59.59	48.76	53.63	82.95	76.84	79.78	72.15	82.21	76.85	71.56	69.27	70.08
	SVM	73.11	62.16	57.02	59.48	82.42	78.95	80.65	74.77	79.81	77.21	73.11	71.92	72.44
	NB	61.08	47.15	47.93	47.54	64.65	67.37	65.98	67.82	65.87	66.83	59.87	60.39	60.11
	RF	65.57	58.42	48.76	53.15	72.22	68.42	70.27	66.09	74.04	69.84	65.57	63.74	64.42
Combined 2-grams	**LR**	70.05	56.31	47.93	51.79	80.85	80	80.42	71.81	78.37	74.94	69.65	68.76	69.05
	SVM	74.59	66.67	57.38	61.67	82.95	76.84	79.78	75	83.65	79.09	70.83	72.62	73.51
	NB	58.71	44.62	47.93	46.22	62.37	64.44	63.39	66.33	62.5	64.36	57.77	58.29	57.99
	RF	65.48	56.94	33.88	42.49	79.73	62.77	70.24	63.9	85.1	72.99	66.85	60.58	61.9
Combined 3-grams	**LR**	70.75	60.61	49.59	54.55	77.17	74.74	75.94	72.53	81.25	76.64	70.1	68.52	69.04
	SVM	75.47	66.97	60.33	63.48	83.33	78.95	**81.08**	76.44	82.69	**79.45**	75.58	73.99	74.67
	NB	63.21	49.59	50.41	50	67.37	67.37	67.37	69.42	68.75	69.08	62.12	62.17	62.15
	RF	65.8	57.14	42.98	49.06	73.33	57.89	64.71	66.67	82.69	73.82	65.71	61.18	62.53
Combined 4-grams	**LR**	69.81	59.3	42.15	49.28	80.22	76.84	78.49	69.64	82.69	75.6	69.72	67.22	67.79
	SVM	74.06	65.18	60.33	62.66	81.52	78.95	80.21	75.45	79.81	77.57	74.05	73.03	73.48
	NB	61.08	47.11	47.11	47.11	64.65	67.37	65.98	67.65	66.35	66.99	59.8	60.27	60.02
	RF	66.51	61.05	47.93	53.7	75.61	65.26	70.06	65.59	77.88	71.21	67.41	63.69	64.99

Table 4. Overall Accuracy (*Acc.*), *Precision* (*P*), *Recall* (*R*) and *F-Measure* (*F*) per role and over all roles for the weighted representations. *Logistic Regression* (*LR*), *Support Vector Machines* (*SVM*), *Multinomial Naive Bayes* (*NB*) and *Random Forests* (*RF*) are compared.

some cases *Naive Bayes* tends to select the majority class (cf., Table 4), which means that further training examples would be needed in these representations for the specific setup.

Finally, with regards to the contribution of the weighted representations the reported numbers in Table 4 suggest that this is minor, and, in fact there are several cases where the weighting of the features drops the performance, compared to the numbers shown in Table 3. It seems that the selected global weighting is not enough to differentiate the predictive power of the features. Perhaps, a local weighting of the features would be able to make a fine-grained differentiation, taking into account the way *TF-IDF* works for the term features in typical text mining tasks. However, such an application is not feasible in our case since the instances are short and the representation is extremely sparse.

6 Conclusions and Future Work

In this paper we introduced a methodology for learning formal definitions from unstructured text, formulating the problem from the point of view of learning roles between concepts. From this perspective the results are encouraging and we showed that for three widely populated SNOMED CT roles, namely *Associated Morphology*, *Causative Agent* and *Finding Site*, the task can be achieved with accuracy reaching up to 75%. On the other hand side, the suggested methodology has limitations and the experimental results showed that there is definitely room for improvement regarding the underlying representations. In this direction, the next step is to analyze the cases that the method fails, and bring into the surface the reasons, as well as the nature of additional features that might be used to correct these cases. In the same direction, no syntactic or semantic information was taken into account for the feature engineering, which we plan to integrate in future work. In addition, the text pre-processing for the preparation of the dataset can be enriched with more elaborate steps such as: better alignment of sentences which contain roles inside concept names, or nested roles. Finally, another direction that we will look into in our future work is the extraction of roles from *MeSH* definitions in comparison to employing the existing SNOMED CT roles for learning formal definitions, motivated by the different nature of *MeSH* and SNOMED CT.

References

1. I. Augenstein, S. Padó, and S. Rudolph. Lodifier: Generating linked data from unstructured text. In *ESWC*, pages 210–224, 2012.
2. F. Baader, I. Horrocks, and U. Sattler. Description logics. In *Handbook on Ontologies*. 2004.
3. M. Bada and L. Hunter. Enrichment of obo ontologies. *J. of Biomedical Informatics*, 40(3):300–315, 2007.
4. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.
5. W. A. Baumgartner, H. L. Johnson, H. L. Johnson, K. B. Cohen, K. B. Cohen, W. A, Z. Lu, Z. Lu, M. Bada, M. Bada, T. Kester, T. Kester, H. Kim, H. Kim, L. Hunter, and L. Hunter. Evaluation of lexical methods for detecting relationships between concepts from multiple ontologies. pac symp biocomput. In *Pac Symp Biocomput*, pages 28–39, 2006.

6. O. Bodenreider and R. Stevens. Bio-ontologies: current trends and future directions. *Briefings in Bioinformatics*, 7(3):256–274, 2006.
7. C. Boelling, M. Dumontier, M. Weidlich, and H.-G. Holzhütter. Role-based representation and inference of biochemical processes. In *ICBO*, 2012.
8. L. L. Chepelev and M. Dumontier. Chemical entity semantic specification: Knowledge representation for efficient semantic cheminformatics and facile data integration. *J. Cheminformatics*, 3:20, 2011.
9. U. Hahn and S. Schulz. Towards a broad-coverage biomedical ontology based on description logics. In *Pacific Symposium on Biocomputing*, pages 577–588, 2003.
10. J. Hastings, M. Dumontier, D. Hull, M. Horridge, C. Steinbeck, R. Stevens, U. Sattler, T. Hörne, and K. Britz. Representing chemicals using owl, description graphs and rules. In *OWLED*, 2010.
11. S. Jupp, M. Horridge, L. Iannone, J. Klein, S. Owen, J. Schanstra, R. Stevens, and K. Wolstencroft. Populous: A Tool for Populating Templates for OWL Ontologies. In *SWAT4LS*, 2010.
12. S. Jupp, R. Stevens, and R. Hoehndorf. Logical Gene Ontology Annotations (GOAL): exploring gene ontology annotations with OWL. *J Biomed Semantics*, 3 Suppl 1:S3, 2012.
13. R. D. King, J. J. Rowland, W. Aubrey, M. Liakata, M. Markham, L. N. Soldatova, K. E. Whelan, A. Clare, M. Young, A. Sparkes, S. G. Oliver, and P. Pir. The robot scientist Adam. *IEEE Computer*, 42(8):46–54, 2009.
14. K. Liu, W. R. Hogan, and R. S. Crowley. Natural language processing methods and systems for biomedical ontology learning. *Journal of Biomedical Informatics*, 44(1):163–179, 2011.
15. A. Maedche, E. Maedche, and S. Staab. The text-to-onto ontology learning environment. In *Software Demonstration at ICCS-2000 - Eight International Conference on Conceptual Structures*, 2000.
16. T. T. Quan, S. C. Hui, A. C. M. Fong, and T. H. Cao. Automatic generation of ontology for scholarly semantic web. In *International Semantic Web Conference*, 2004.
17. A. L. Rector and J. Rogers. Ontological and practical issues in using a description logic to represent medical concept systems: Experience from galen. In *Reasoning Web*, pages 197–231, 2006.
18. R. L. Richesson, J. E. Andrews, and J. P. Krischer. Use of SNOMED CT to represent clinical research data: A semantic characterization of data items on case report forms in vasculitis research. *Journal of the American Medical Informatics Association*, 13(5):536–546, 2006.
19. D. L. Rubin, O. Dameron, Y. Bashir, D. Grossman, P. Dev, and M. A. Musen. Using ontologies linked with geometric models to reason about penetrating injuries. *Artificial Intelligence in Medicine*, 37(3):167–176, 2006.
20. D. Sánchez, A. Moreno, and L. D. V. Terrientes. Learning relation axioms from text: An automatic web-based approach. *Expert Syst. Appl.*, 39(5):5792–5805, 2012.
21. L. F. Soualmia, C. Golbreich, and S. J. Darmoni. Representing the MeSH in OWL: Towards a Semi-Automatic Migration. In *Proceedings of the KR Workshop on Formal Biomedical Knowledge Representation*, pages 81–87, 2004.
22. M. Q. Stearns, C. Price, K. A. Spackman, and A. Y. Wang. SNOMED clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, 2001.
23. H. Stenzhorn, E. J. Pacheco, P. Nohama, and S. Schulz. Automatic mapping of clinical documentation to SNOMED CT. In *MIE*, pages 228–232, 2009.
24. J. Völker, P. Hitzler, and P. Cimiano. Acquisition of OWL DL axioms from lexical resources. In *ESWC*, pages 670–685, 2007.
25. T. Wächter and M. Schroeder. Semi-automated ontology generation within obo-edit. *Bioinformatics [ISMB]*, 26(12):88–96, 2010.