Learning Formal Definitions for SNOMED CT from Text

Yue Ma* and Felix Distel

Institute of Theoretical Computer Science, Technische Universität Dresden, Dresden, Germany, {mayue,felix}@tcs.inf.tu-dresden.de

Abstract. SNOMED CT is a widely used medical ontology which is formally expressed in a fragment of the Description Logic $\mathcal{EL}$++. The underlying logics allow for expressive querying, yet make it costly to maintain and extend the ontology. Existing approaches for ontology generation mostly focus on learning superclass or subclass relations and therefore fail to be used to generate SNOMED CT definitions. In this paper, we present an approach for the extraction of SNOMED CT definitions from natural language texts, based on the distance relation extraction approach. By benefiting from a relatively large amount of textual data for the medical domain and the rich content of SNOMED CT, such an approach comes with the benefit that no manually labelled corpus is required. We also show that the type information for SNOMED CT concept is an important feature to be examined for such a system. We test and evaluate the approach using two types of texts. Experimental results show that the proposed approach is promising to assist SNOMED CT development.

1 Introduction

SNOMED CT [7] is a medical ontology and now a widely accepted international standard. It describes concepts such as anatomical structures, disorders, organisms among others. It has been adopted in many countries worldwide as a standard for electronic health records and is also used in clinical decision support systems. Users can access SNOMED CT through browsers such as NIH Browser (cf. Table 1).

Unlike simpler medical vocabularies SNOMED CT has a formal logic-based foundation, based on Description Logics (DL), more precisely the lightweight DL $\mathcal{EL}$++ [1], a fragment of the standard OWL2EL[1]. While this is hidden to most users, it is a key advantage of SNOMED CT compared to other systems. The formal semantics results in a computer processable knowledge base that can be extended, debugged and queried through reasoning services.

While setting SNOMED CT apart from medical vocabularies such as MeSH the formal semantics also comes at a cost. Adding new concepts to a formal ontology is a tedious, costly and error-prone process, that needs to be performed

* We acknowledge financial support by the DFG Research Unit FOR 1513, project B1.

[1] http://www.w3.org/TR/owl2-profiles

Table 1. The concept Baritosis as displayed by NIH SNOMED CT Browser

Concept: [50076003] Baritosis

Relationships from *this* concept (9)

Baritosis	Causative agent	Barium dust	(Defining)
Baritosis	Associated morphology	Deposition of foreign material	
Baritosis	Finding site	Lung structure	(Defining)
Baritosis	Associated morphology	Inflammation	
Baritosis	Finding site	Lung structure	(Defining)
Baritosis	Is a	Pneumoconiosis due to inorganic dust	
Baritosis	Clinical course	Courses	(Qualifier)
Baritosis	Episodicity	Episodicities	(Qualifier)
Baritosis	Severity	Severities	(Qualifier)

Table 2. The concept description of Baritosis in $\mathcal{EL}$-syntax

$$
\begin{aligned}
&\mathsf{Baritosis} \equiv \\
&\exists \mathsf{Causative\_agent.Barium\_dust} \\
&\quad \sqcap \exists \mathsf{Associated\_morphology.} \\
&\qquad \mathsf{Deposition\_of\_foreign\_material} \\
&\quad \sqcap \exists \mathsf{Finding\_site.Lung\_structure} \\
&\quad \sqcap \exists \mathsf{Associated\_morphology.Inflammation} \\
&\quad \sqcap \exists \mathsf{Finding\_site.Lung\_structure} \\
&\quad \sqcap \mathsf{Pneumoconiosis\_due\_to\_inorganic\_dust} \\
&\quad \sqcap \exists \mathsf{Clinical\_course.Courses} \\
&\quad \sqcap \exists \mathsf{Episodicity.Episodicities} \\
&\quad \sqcap \exists \mathsf{Severity.Severities}
\end{aligned}
$$

manually by specially trained knowledge engineers [6]. Thus, researchers have developed services providing assistance in ontology design and maintenance, some of which can extract formal DL-based definitions from text [4, 8, 9, 3].

DL vocabulary consists of concept names such as Baritosis, Lung_structure, etc. and relationships, typically called roles, such as Causative_agent, Finding_Site. Roles link concepts to one another. Using concept constructors, new concepts can be defined using existing ones. $\mathcal{EL}$++ provides the constructors conjunction ($\sqcap$) and existential restrictions ($\exists$) among others. Table 2 shows how the relationships from Table 1 can be expressed in the DL syntax.

Existing approaches for ontology generation mostly focus on learning superclass or subclass relations [9] and therefore fail to make use of the full expressivity of $\mathcal{EL}$++. To overcome this, we propose an approach, named *Snomed-supervised relation extraction*, for automatically extracting relationships for concepts (or existential restrictions in DL lingo) from natural language texts. A key advantage of our approach is that no manually labeled training data is required by profiting from the large amount of existing formal knowledge in SNOMED CT. It uses a maximum entropy classifier to classify sentences according to the relationships they describe (if any). To test the approach we use text data from Wikipedia, as well as textual definitions found on the web using the tool DOG4DAG [9].

2 Related Work

Formal ontology generation is an important but non-trivial task [4]. It is particularly challenging for specific domains, such as SNOMED CT. [8] describes some first approaches by applying syntactic transformation rules to generate OWL DL concept definitions for generic domains. When directly applying their approaches on SNOMED CT concept definition generation, we may encounter unresolved reference roles such as $\exists$Of. Moreover, different formal expressions (e.g. $\exists$Caused_by, $\exists$Due_to, $\exists$Result_from) will be generated from variant expressions (e.g. "caused by", "due to", "result from"), even though they all express the same relation $\exists$Causative_agent in SNOMED CT. By contrast, our approach naturally

avoids unresolved reference roles and the lexical variant problems by fixing the set of relationships in advance.

In addition, [8] does not specifically consider $\mathcal{EL}$++ constructors, while [3] is similar to the present work where $\mathcal{EL}$++ is the target language. However, [3] is based on the inductive logic programming technique and requires a large amount of facts about individual entities (called ABox in DL lingo) instead of merely conceptual descriptions of concepts as in the case of SNOMED CT.

Relation extraction is often used to construct ABoxes from ontologies [4]. We extend this idea to generate formal definitions of SNOMED CT concepts. Among the approaches for relation extraction, ours is similar to *distance supervision* [5] in that no manually labelled data is required. However, [5] is not proposed for formal concept definition purposes and works an entity level. Moreover, we use features independently instead of feature conjunctions as in [5] because of the limited data available for the medicine domain. And we show that medicine domain specific features (SNOMED CT types) are important for the system.

3 Task Description

Since ontology construction is costly [6] we provide assistance in this process by automatically extracting relationships for a target concept from text. Our approach is based on the assumption that the set of roles remains relatively stable while the set of concepts constantly increases. To facilitate adding new concept descriptions, we create a system that for a given input sentence annotated with two SNOMED CT concepts is able to decide whether the sentence describes a relationship between the two concepts and which relationship. Since systems for learning subclass and superclass relations already exist, this will eventually enable us to obtain complete $\mathcal{EL}$++ descriptions for new concepts as in Table 2.

4 Architecture

Textual information from the medical domain is widely available from publicly accessible resources, such as the web or textbooks. The methodology used in our system makes use of both textual data and existing SNOMED CT definitions. In the following we describe the three steps used in our method.

Automatic Data Preparation During data preparation SNOMED CT roles and SNOMED CT concept labels are aligned to textual sentences. We achieve this automatically as follows.

Relationship extraction: Through DL reasoning we generate the set of all relationships $A|R|B$ that logically follow from SNOMED CT: $\mathcal{RB} = \{A|R|B : \text{SNOMED} \models A \sqsubseteq \exists R.B\}$. Reasoning provides a way to use implicit information encoded in SNOMED CT. For example, for Finding_site 630,547 relation pairs are obtained through reasoning compared to only 43,079 explicitly given ones.

Annotation: Using the tool *Metamap* developed at the U.S. National Library of Medicine we annotate the textual sentences with SNOMED CT concepts to identify all concepts occurring in a sentence.

Relationship Alignment: Annotated sentences are aligned with a relationship if they contain two concepts that are in a relationship in SNOMED CT. This is illustrated in Table 4, where "Baritosis" and "barium dust" in the sentence are annotated with concepts Baritosis_(disorder) and Barium_Dust_(substance), respectively, by Metamap. The inferred role base $\mathcal{RB}$ contains the relationship Baritosis_(disorder) | Causative_agent | Barium_dust_(substance). The sentence is thus annotated with Causative_agent.

Table 3. Text Alignment and Features

Annotated Sentence	"*Baritosis*/Baritosis_(disorder) is pneumoconiosis caused by *barium dust*/Barium_Dust_(substance)."		
SNOMED CT relationship	Baritosis_(disorder) | Causative_agent | Barium_Dust_(substance)		
Features	left type	between-words	right type
	disorder	"is pneumoconiosis caused by"	*substance*

Training Phase Once the relationship alignment is done, features will be extracted from the corresponding sentences. The assumption here is that such sentences likely represent role relationships via the aligned role. Since several sentences can be aligned to the same role, weights for different features extracted from different sentences will be learned by a multi-class classifier. Feature extraction from text corpora is an important step in Natural Language Processing applications [2]. For the features, besides the standard lexical features (between-words of annotated phrases [5]), we use eleven types of concepts, including *organism*, *finding*, and *disorder*, which are among the content of SNOMED CT. A flag is used to denote if the occurrence order of two concepts in a sentence is the same as it is in SNOMED CT.

Test Phase Test data consists of textual sentences that are candidates for describing a relation. Such sentences are first annotated with SNOMED CT concepts by Metamap, and then features are extracted. Based on these a multi-class classifier can predict role relationships between the target concept and other concepts appearing in the sentences. Note that the roles considered in the current system are disjoint, i.e. no pair of concepts can be related via two different roles. However, for one target concept different roles can be predicted for the same successor concept, which conflicts the above fact. For aggregation we select the role which maximizes the predicted weight according to the classifier as following:

$$R_{C_{tbd},A} = \arg\max_{R \in \mathcal{R}} w(R, C_{tbd}, A),$$

where $\mathcal{R}$ is the space of all considered roles, and $w(R, C_{tbd}, A)$ is the confidence value that C_{tbd} and A are in relationship R according to the classifier.

Table 4. Evaluation over training datasets WIKI, D4D, and MIX and test datasets TW, and TD with and without the type features

	TW			TD		
	WIKI	D4D	MIX	WIKI	D4D	MIX
Without Type	0.00	0.40	0.20	0.27	0.45	0.59
With Type	0.40	**0.80**	0.60	0.50	**0.64**	0.59

5 Evaluation

The corpora chosen for experiments are two texts: named WIKI and D4D. WIKI is obtained by querying Wikipedia with one-word SNOMED CT concept names, resulting in a document consisting of around 53,943 distinct sentences with 972,038 words. D4D contains textual definitions extracted by querying DOG4DAG[2] [9] over concepts that occur in the relationships of three roles[3], obtaining 7,092 distinct sentences with 112,886 words. MIX is a combination of WIKI and D4D.

The SNOMED CT relationship set is divided for testing and training: only relationships not concerning a target concept can be considered for training. Negative examples are generated for the classifier to recognize sentences which do not describe any relationship. We test the approach on two test datasets: TW and TD. TW contains sentences from Wikipedia about the concepts to be defined and TD is TW combined with sentences from DOG4DAG for the same concepts. As the evaluation metric, micro average F-measure (each test item counts equally) is used due to the multi-class classification. Table 5 compares the results based on different training and test data. We can conclude the following.

- Except for the MIX case, type information significantly improved the results for all training and test data. This suggests that type is an important feature to be used in our system.
- D4D training data outperformed WIKI and MIX on both of the test data. This shows that precomputed textual definitions by DOG4DAG are helpful for generating formal definitions of concepts of SNOMED CT.

For higher quality text (D4D) the results appear promising. For illustration, for Baritosis as the target concept, the system correctly recognizes the Causative_agent relation to Barium_dust and the Finding_site relation to Lung_structure.

[2] DOG4DAG is a system capable of retrieving and ranking textual definitions from the web. However, it has query number restrictions so that we cannot query as many as SNOMED CT concepts.

[3] As a preliminary experiment, we focus on three well populated roles of SNOMED CT Causative_agent, Associated_morphology, Finding_site for defining concepts.

6 Conclusion and Future Work

In this paper, we have designed a system that can extract $\mathcal{EL}$ definitions from texts according to the SNOMED CT format. Having examined on different textual data and three well popluated SNOMED CT roles with different parameter settings, we conclude that such an approach can serve as a good start for generating SNOMED CT definitions of new concepts.

In the future, we will improve the system for more SNOMED CT roles and using logic reasoning to avoid unreasonable predicted relationships. Text quality appears to be crucial with D4D outperforming WIKI. Hence, we plan to consider high quality textual definitions from the MeSH vocabulary.

References

1. Baader, F., Brandt, S., Lutz, C.: Pushing the $\mathcal{EL}$ envelope. In: Proceedings of IJCAI'05, Morgan Kaufmann (2005)
2. Broda, B., Kędzia, P., Marcińczuk, M., Radziszewski, A., Ramocki, R.: Fextor: a feature extraction framework for natural language processing. a case study in word sense disambiguation, relation recognition and anaphora resolution. In Przepiorkowski, A., Piasecki, M., Jassem, K., Fuglewicz, P., eds.: Computational Linguistics. Springer Berlin Heidelberg (2013) To appear
3. Chitsaz, M., Wang, K., Blumenstein, M., Qi, G.: Concept learning for EL++ by refinement and reinforcement. In: Proceedings of PRICAI'12. (2012) 15–26
4. Cimiano, P.: Ontology learning and population from text - algorithms, evaluation and applications. Springer (2006)
5. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of ACL/AFNLP'09. (2009) 1003–1011
6. Simperl, E.P.B., Tempich, C., Sure, Y.: A cost estimation model for ontology engineering. In: Proceedings of ISWC'06. (2006) 625–639
7. SNOMED *Clinical Terms*. Northfield, IL: College of American Pathologists (2006)
8. Völker, J.: Learning expressive ontologies. PhD thesis, Universität Karlsruhe (2009)
9. Wächter, T., Fabian, G., Schroeder, M.: Dog4dag: semi-automated ontology generation in obo-edit and protégé. In: Proceedings of SWAT4LS'11. (2011) 119–120