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. In this paper we present an approach for the extraction of Snomed CT definitions from natural language text. We test and evaluate the approach using two types of texts.

1 Introduction

Snomed CT [6] 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.

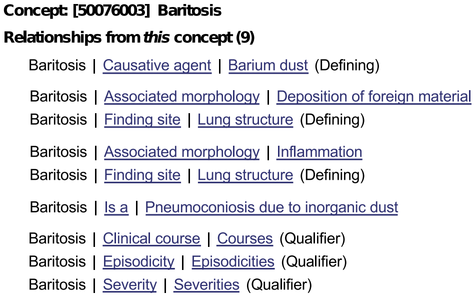
While setting Snomed CT apart from medical vocabularies such as MeSH the formal semantics also comes at a cost. Adding new concepts to a formal ontology is a tedious, costly and error-prone process, that needs to be performed manually by specially trained knowledge engineers [5]. Thus, researchers have developed services providing assistance in ontology design and maintenance, some of which can extract formal DL-based definitions from text [3, 7, 8, 2].

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.

[*] 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

Baritosis ≡
∃Causative_agent.Barium_dust
⊓ ∃Associated_morphology.
Deposition_of_foreign_material
⊓ ∃Finding_site.Lung_structure
⊓ ∃Associated_morphology.Inflammation
⊓ ∃Finding_site.Lung_structure
⊓ Pneumoconiosis_due_to_inorganic_dust
⊓ ∃Clinical_course.Courses
⊓ ∃Episodicity.Episodicities
⊓ ∃Severity.Severities

Existing approaches for ontology generation mostly focus on learning superclass or subclass relations [8] and therefore fail to make use of existential restrictions allowed by $\mathcal{EL}$++. 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 multiclass 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 by the tool Dog4Dag [8].

Task Description

Our approach is based on the observation that in Snomed CT 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 if 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 formal definitions for new concepts as in Table 2.

For example, for the target concept Baritosis we expect the Causative_agent relation to Barium_dust and the Finding_site relation to Lung_structure to be recognizable from the following two sentences: (1) “Baritosis is a benign type of pneumoconiosis, which is caused by long-term exposure to barium dust.” (2) “Baritosis is due to inorganic dust lies in the lungs.”

2 Related Work

Formal ontology generation is an important but non-trivial task [3]. It is particularly challenging for specific domains, such as Snomed CT. [7] describes some first approaches which apply syntactic transformation rules to generate OWL DL concept definitions for generic domains. When directly applying their approaches to SNOMED CT concept definition generation, we may encounter unresolved reference roles such as $\exists\mathsf{Of}$. Moreover, different formal expressions (e.g. $\exists\mathsf{Caused\_by}$, $\exists\mathsf{Due\_to}$, ...) will be generated from variant expressions (e.g. "caused by", "due to", ...), even though they all express the same relation $\exists\mathsf{Causative\_agent}$ in SNOMED CT. By contrast, our approach naturally avoids unresolved reference roles and the lexical variant problems by the prefixed set of SNOMED CT roles.

In addition, [7] does not specifically consider $\mathcal{EL}{++}$ constructors, while [2] is similar to the present work where $\mathcal{EL}{++}$ is the target language. However, [2] 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 [3]. We extend this idea to generate formal definitions of SNOMED CT concepts. Among the approaches for relation extraction, ours is similar to *distance supervision* [4] in that no manually labelled data is required. However, [4] is not proposed for formal concept definition purposes and works at the entity level. Moreover, we use features independently instead of feature conjunctions as in [4] 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 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 3, 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 aligned with Causative_agent, with the latter called an aligned role.

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 of 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. For the features, besides the standard lexical features (between-words of annotated phrases [4]), we use eleven semantic types, including *organism*, *finding*, and *disorder*, which are provided by Snomed CT for each concept. A flag denotes if the two concepts occur in the same order in the sentence as 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.

4 Evaluation and Conclusion

The two corpora chosen for experiments are named WIKI and D4D. WIKI is obtained by querying Wikipedia with one-word Snomed CT concept names, resulting in around 53,943 distinct sentences with 972,038 words. D4D contains textual definitions extracted by querying Dog4Dag[2] [8] over concepts that occur in the relationships of three well populated roles (i.e., Causative_agent, Associated_morphology, Finding_site) examined in this paper, obtaining 7,092 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

[2] Dog4Dag is a system that can retrieve and rank textual definitions from the web. However, it has query number restrictions.

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

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. The Stanford maximum entropy classifier[3] is used for the implementation and micro average F-measure is the evaluation metric. Different feature settings (with/without type information) are explored. Table 4 compares the results based on different training and test data. We can have the following conclusions:

- The semantic type information described in Section 3 significantly improved the results for all the experiments except for the MIX training data with the TD test data. This suggests that type is an important feature to be used in our system for predicating formal definitions of concepts.
- 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.

In the future, we will improve the system by using logic reasoning to avoid unreasonable predicted relationships. Text quality appears to be crucial with D4D outperforming WIKI. So we will consider high quality MeSH textual definitions.

References

1. Baader, F., Brandt, S., Lutz, C.: Pushing the $\mathcal{EL}$ envelope. In: Proceedings of IJCAI'05, Morgan Kaufmann (2005)
2. Chitsaz, M., Wang, K., Blumenstein, M., Qi, G.: Concept learning for EL++ by refinement and reinforcement. In: Proceedings of PRICAI'12. (2012) 15–26
3. Cimiano, P.: Ontology learning and population from text - algorithms, evaluation and applications. Springer (2006)
4. 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
5. Simperl, E.P.B., Tempich, C., Sure, Y.: A cost estimation model for ontology engineering. In: Proceedings of ISWC'06. (2006) 625–639
6. Snomed *Clinical Terms.* Northfield, IL: College of American Pathologists (2006)
7. Völker, J.: Learning expressive ontologies. PhD thesis, Universität Karlsruhe (2009)
8. Wächter, T., Fabian, G., Schroeder, M.: Dog4dag: semi-automated ontology generation in obo-edit and protégé. In: Proceedings of SWAT4LS'11. (2011) 119–120

[3] http://nlp.stanford.edu/software/classifier.shtml