A Hybrid Approach to Learn Description Logic based Biomedical Ontology from Texts*

Yue Ma[1] and Alifah Syamsiyah[1,2]

[1]Institute of Theoretical Computer Science, Technische Universität Dresden, Germany,
[2] Free University of Bozen-Bolzano
mayue@tu-dresden.de,alifah.syamsiyah@stud-inf.unibz.it

Abstract. Augmenting formal medical knowledge is neither manually nor automatically straightforward. However, this process can benefit from rich information in narrative texts, such as scientific publications. *Snomed-supervised relation extraction* has been proposed as an approach for mining knowledge from texts in an unsupervised way. It can catch not only superclass/subclass relations but also existential restrictions; hence produce more precise concept definitions. Based on this approach, the present work aims to develop a system that takes biomedical texts as input and outputs the corresponding $\mathcal{EL}$++ concept definitions. Several extra features are introduced in the system, such as generating general class inclusions (GCIs) and negative concept names. Moreover, the system allows users to trace textual causes for a generated definition, and also give feedback (i.e. correction of the definition) to the system to retrain its inner model, a mechanism for ameliorating the system via interaction with domain experts.

1 Introduction

Biomedicine is a discipline that involves a large number of terminologies, concepts, and complex definitions that need to be modeled in a comprehensive knowledge base to be shared and processed distributively and automatically. The National Library of Medicine (NLM) has maintained the world's largest biomedical library since 1836 [5]. One of the medical terminologies preserved by NLM is Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT). It is a comprehensive clinical vocabulary structured in a well-defined form that has the lightweight Description Logic $\mathcal{EL}$++ [2] as the underlying logic, which can support automatic checking of modeling consistency.

However, creating, maintaining, and extending formal ontology is an expensive process [6]. In contrast, narrative texts, such as medical records, health news, and scientific publications, contain rich information that is useful to augment a medical knowledge base. In this paper, we propose a hybrid system that can generate $\mathcal{EL}$ TBoxes from texts. It extends the formal definition candidates learned by the *Snomed-supervised relation extraction* process [4, 3] with linguistic patterns to give a finer-grained translation of the learned candidates. Besides generating concept name hierarchy that has been widely studied, the system can also generate definitions with existential restrictions to exploit the expressivity of $\mathcal{EL}$. Moreover, the implemented Graphical User Interface helps a user to visualize the flow of this framework, tracks textual sentences from which a formal definition is generated, and gives feedback to enhance the system interactively. The implementation of the system can be found from the link `https://github.com/alifahsyamsiyah/learningDL`.

* We acknowledge financial support by the DFG Research Unit FOR 1513, project B1. Alifah Syamsiyah was supported by the European Master's Program in Computational Logic (EMCL)

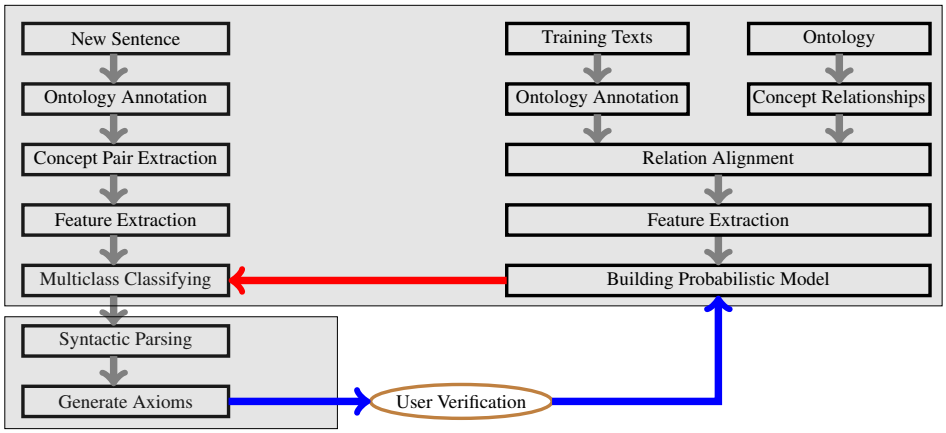

Fig. 1. Hybrid approach overview: the upper darked block is for machine learning phase to extract definition candidates, and the lower left darked block is for the pattern based transformation of definition candidates, and the brown ellipse is for interaction with users.

2 Task and Our Approach

Our task is to generate $\mathcal{EL}$ definitions from textual sentences. For example, from the sentence "Baritosis is a pneumoconiosis caused by barium dust", it is desired to have an automatic way to generate the formal $\mathcal{EL}$ axiom (together with some confidence value), as shown in the red frame of Figure 2. Moreover, to help users understand the origin of a generated definition and/or give their feedbacks, the system should be able to trace the textual sources from which a definition is generated (implemented with the question mark in our system), and allow users to correct automatically learned definitions (the "V" mark in Figure 2).

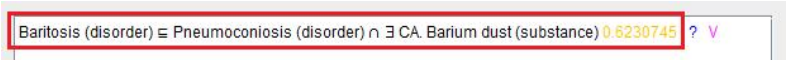

Fig. 2. An illustrative example for the functionality of the system (CA is shortened form for the SNOMED CT relation Causative_Agent)

Below we describe our hybrid system that has two components, as shown in Figure 1: one for extracting definition candidates by machine learning techniques, and the other for formulating final definitions from the definition candidates by linguistic patterns.

2.1 Extracting Formal Definition Candidates

The first part of the system is to generate definition candidates via the steps given in the upper block of Figure 1. It again contains two components: learning a model from the training data (training texts and ontology) and generating candidates from new texts. Each steps in the two components are described below.

Common steps in processing training and test texts. One is to recognize SNOMED CT concept names from a given textual sentence, called **ontology annotation** in Figure 1. In our implementation, this is done by invoking the tool Metamap [1]. Since we are only interested in the most specific and precise concepts, we filter the Metamap annotations by keeping merely those that refer to head of a phrase but not a verb. The other common processing step for training and test sentence is to extract textual features for a pair of concepts occurring in a sentence, called **feature extraction**. Currently, the system uses classical lexical features of n-grams over both characters and words as in [4].

A special processing on training texts is concerned to generate labelled training data and to learn multi-class classification model for each predefined relation [4]:

- Automatic generation of training data is realized by the step named **Relationship Alignment** that matches an annotated sentence by Metamap with relationships between concept names from ontology: If one sentence containing a pair of concepts that has a relation R according to the ontology, this sentence is considered as a textual representation of R, thus being labelled with R. Furthermore, we also consider the inverse roles that often appear in texts via active and passive sentences. Hence, if there are n predefined relations, there will be $2n$ possible labels for a sentence.
- **Building probabilistic model** is to learn a probabilistic multi-class classification model based on the textual features of labelled sentences from the previous step. For this, the current system uses the maximum entropy Stanford Classifier[1].

A special processing on test texts is to extract definition candidates from a new test sentence. A definition candidate is a triple (A, R, B) where A, B are concept names and R is a relation, meaning that A and B have a relation R according to a test sentence.

- **Concept pair extraction** is to get pairs of concepts from an annotated test sentence.
- **Multiclass classification** is to answer whether a pair of two concept names has a relation, and if yes, which relation it is. This part can be achieved by the model learned from training data by Stanford Classifier. A positive answer returned by the classifier gives a definition candidate (A, R, B). Slightly abusing of the notation, we also call $\exists r.B$ a definition candidate for A.

2.2 Pattern based Transformation of Definition Candidates

Once we get definition candidates, we first change the order of inverse role so it always appears as an active role. Next, different from [4], we distinguish two ways to formalize it: (1) into a subsumption $(A \sqsubseteq R.B)$ or (2) into a conjunction $(A \sqcap R.B)$. For example, the sentence "Baritosis is caused by barium dust" stands for the subsumption $Baritosis(disorder) \sqsubseteq Causative\_agent.Barium\_dust$; whilst "Chest pain from anxiety ..." corresponds to a conjunction $Chestpain(disorder) \sqcap \exists.Causative\_agent.Anxiety(disorder)$. To decide which transformation of a definition candidate, we follow the intuition observable from the above examples:

- A subsumption $A \sqsubseteq \exists R.B$ should be generated from a candidate (A, R, B) if A and B are connected in the sentence in a subject-object relation, called *S-form*.
- A conjunction $A \sqcap \exists R.B$ should be formed if A and B appearing in a noun phrase structure, called *NP-form*.

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

To implement this linguistic pattern based strategy, we use the Stanford Parser[2] to get syntactical parsing tree of a test sentence. The S-form and NP-form are detected in the following way: First, the phrases corresponding to A and B are recognized from the sentences, and then the least common node of these two phrases is searched from the syntactic parsing tree of the whole sentence. If the least common node has type S (resp. NP)[3], then A and B is in S-form (resp. NP-form). Otherwise, a parsing error is returned.

Negation Concept Names In natural language, sometimes we use negative way to define the opposite meaning. For example, the sentence "The disease from foot is not relative to heart attack" will be translated to $Disease\ (disorder) \sqcap FS.\ Foot(body\ structure) \sqsubseteq \neg Heart\_disease\ (disorder)$. This is achieved in the system based on negated atomic concept names detectable by Metamap version 2013.

2.3 Tracing Source Sentence and Classifier Model Retraining

There are two extra functions provided by the system, namely tracing to sentence and classifier model retraining. As given in Figure 2, if a user clicks the "?" mark, system will provide sentences from which the formal definition extracted. Note that the system uses machine learning approach to acquire definition candidates which may get wrong. Therefore, we provide a mechanism for user to validate the answer by clicking "V" symbol and then give the correct relation to link two concept names. As shown in Figure 3, the user changes the role relation from inverse of Finding Site (FS-1) to Causative Agent (CA).

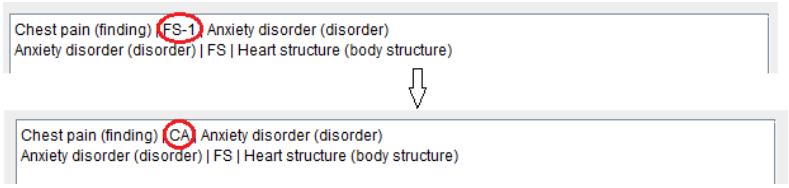

Fig. 3. Interaction with users: change the predicated relation in a definition candidate (FS is the shortened form for the SNOMED CT relation Finding_site, and FS-1 is the inverse role of FS)

References

1. Aronson, A.R., Lang, F.M.: An overview of metamap: historical perspective and recent advances. JAMIA **17**(3) (2010) 229–236
2. Baader, F., Brandt, S., Lutz, C.: Pushing the $\mathcal{EL}$ envelope. In: Proceedings of IJCAI'05. (2005)
3. Ma, Y., Distel, F.: Concept adjustment for description logics. In: Proceedings of K-Cap'13. (2013) 65-72.
4. Ma, Y., Distel, F.: Learning formal definitions for Snomed CT from text. In: Proceedings of AIME'13. (2013) 73–77
5. National Library of Medicine: NLM overview. http://www.nlm.nih.gov/about/index.html (2014)
6. Simperl, E., Bürger, T., Hangl, S., Wörgl, S., Popov, I.: Ontocom: A reliable cost estimation method for ontology development projects. Web Semantics: Science, Services and Agents on the World Wide Web **16**(5) (2012)

[2] http://nlp.stanford.edu/software/lex-parser.shtml

[3] "S" is for sentence, and"NP" for noun phase.