Automated Learning of Domain Taxonomies from Text using Background Knowledge

Julia Hoxha[a], Guoqian Jiang[b], Chunhua Weng[a,c]

[a]*Department of Biomedical Informatics, Columbia University, New York, NY, USA*
[b]*Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA*
[c]*Corresponding Author, Department of Biomedical Informatics, Columbia University, 622 W 168th Street, PH-20, New York, NY 10032, USA, Email: chunhua@columbia.edu,Tel: 646-734-9159*

Appendix A. Concept Coverage Experiment

Experimental Protocol: The goal of this experiment is to evaluate the coverage of the automatically-extracted domain concepts. This step is performed using available dictionary lookup tools (i.e. ELIXR and MedEx), with no additional contribution from Ontofier. So the purpose of this experiment is to assess whether by choosing this approach for concept extraction, rather than relying on lexical rules or any semi-manual pattern mining methods, we make a good decision and have better results. We conduct this experiment for both evaluation cases: clinical trial descriptions (dataset D_1) and MEDLINE abstracts (dataset D_4).

- **Case 1**: Clinical trials studying a particular disease often employ common concepts (variables) to determine patient eligibility, e.g. *"hemoglobin A1c"*. American Heart Association (AHA) has published 95 key cardiovascular disease Common Data Elements (CDEs) defined by human experts [1], which is used as gold standard for this experiment. CDEs are data elements defined upon agreement of among medical experts that are considered to be commonly used across different studies and forms of documents for a particular disease.

  Recently, Luo et al. [2] proposed a semi-automated approach to discover CDEs from text, and report their mined concepts with respect to the AHA standard. We

use this dataset to evaluate and compare the percentage of the concepts automatically identified by Ontofier in dataset D_1.

- **Case 2**: we assess the coverage of the concepts related to pharmacological substances (drugs) that were automatically identified in the MEDLINE abstracts by Ontofier (applying MedEx). As gold standard, we use the manual annotation of drugs specified in the MEDLINE corpus.

  We compare to a state-of-the-art taxonomy learning method called Ontolearn [3], using the taxonomy[1] (and respective terminology[2]) extracted with this approach from the same MEDLINE corpus. We also compare to the ADTCT framework [**?**], which is another framework for taxonomy learning from text. We use an implementation provided by the authors of this work.

Experimental Results: The results of this experiment are illustrated in Table A.1. For the case of clinical trial descriptions in dataset D1, we have identified with Ontofier 95.75% of the Common Data Elements defined for the cardiovascular disease in the American Heart Association (AHA) standard. Ontofier clearly outperforms the semi-automated approach of Luo et al. [2]. Partial mapping refers to a learned concept that partially misses the semantics of standard element, e.g. for the term *"syncope lower 3 months"* we identified the concept named only *"syncope"*, assessing it as partial match to the AHA element *"Date of syncope"*.

In the experiment with MEDLINE corpus, as component for term extraction we used MedEx [24] that identified 444 unique concepts of type drug (mapped to UMLS CUIs) from a total of 834 sentences. The number of automatically identified concepts that match MEDLINE manual annotations is 206 from the set of 246 available (i.e. concept

[1] `http://lcl.uniroma1.it/ontolearn_reloaded/files/DDI/DDI_TREE.tsv`

[2] `http://lcl.uniroma1.it/ontolearn_reloaded/files/DDI/DDI_terminology.txt`

		Mapped		Unmapped
Case 1. Clinical trials	Ontofier	full: 88.3%	partial: 7.45%	**4.25%**
		95.75%		
	Luo et al.	78.72%		21.28%
Case 2. MEDLINE abstracts	Ontofier	**83.7%**		**16.3%**
	Ontolearn	0.06%		99.04%
	ADTCT	11.38%		88.62%

Table A.1: Coverage of concepts: case 1) concepts extracted from the clinical trials description of dataset D_1, mapped to the standardized CDEs of Cardiovascular Disease; case 2) concepts (type drug) extracted from the MEDLINE dataset D_4.

coverage is 83.7%). There was an additional set of 238 concepts identified and mapped to UMLS, but they do not occur in the provided MEDLINE annotations.

Results show that our approach is superior to the other methods. Concept coverage in ADTCT framework is only 11.38%. The other method Ontolearn generates a terminology in which only 14 concepts match to the manual annotations (i.e. 0.06%). We observe that the terminology identified by Ontolearn contains highly generic concepts, such as *"compound"*, *"various complex phenolic substance"*, etc. that are extracted more often from sentences containing definitions of terms. It also contains variations of terms extracted from these definitions, e.g. *"absence of glucose"*, *"valproic acid diminished binding"*, etc. Meanwhile, our approach is able to provide a domain-specific terminology, also demonstrated by the high coverage of concepts that are mapped to the manual annotations provided in the MEDLINE corpus.

In Table A.2, we illustrate examples of extracted concepts that are fully- and partially-mapped to the standardized Common Data Elements.

Standardized CDE	**Found concept (CUI)**	**Coverage (full/ partial/ empty)**	**Explanation**
Myocardial infarction	myocardial infarction(C0027051)	full	
Hematocrit	hematocrit(C0518014)	full	
Tobacco use	smoke(C0037366); tobacco(C0040329)	full	
Heart failure	heart failure congestive(C0018802);heart failure(C0018801)	full	
Systolic blood pressure	systolic blood pressure (C0871470)	full	
Mitral valve area	mitral valve area (C0428818)	full	
Date of Syncope	syncope(C0039070)	partial	*extracted as procedure not as date (temporal criterion)*
Direct rennin inhibitors	rennin(C1150116)	partial	*phrase not occuring in text: we find 'rennin' instead of 'rennin inhibitors'*
Date of cardiac arrest	cardiac arrest (C0018790)	partial	*detected as concept, not as date (temporal criterion)*
P2Y12 blocker		empty	
Left ventricle size (Quality)		empty	

Table A.2: Examples of extracted concepts by Ontofier that are fully-mapped, partially-mapped, or not mapped to the Common Data Elements (CDEs) standardized by the American Heart Association (AHA) for the cardiovascular disease.

References

[1] W. C. Members, Accf/aha 2011 key data elements and definitions of a base cardiovascular vocabulary for electronic health records, in: Circulation, 2011.

[2] Z. Luo, R. Miotto, C. Weng, A human-computer collaborative approach to identifying common data elements in clinical trial eligibility criteria, Journal of Biomedical Informatics 46 (1) (2013) 33–39. doi:10.1016/j.jbi.2012.07.006. URL http://dx.doi.org/10.1016/j.jbi.2012.07.006

[3] P. Velardi, S. Faralli, R. Navigli, Ontolearn reloaded: A graph-based algorithm for taxonomy induction, Computational Linguistics 39 (3) (2013) 665–707.