Leveraging Sublanguage Features for the Semantic Categorization of Clinical Terms

Leonie Grön
Quantitative Linguistics and
Lexical Variation (QLVL)
KU Leuven
leonie.gron@kuleuven.be

Ann Bertels
Leuven Language Institute (ILT)
Quantitative Linguistics and
Lexical Variation (QLVL)
KU Leuven
ann.bertels@kuleuven.be

Kris Heylen
Quantitative Linguistics and
Lexical Variation (QLVL)
KU Leuven
kris.heylen@kuleuven.be

Abstract

The automatic processing of clinical documents, such as Electronic Health Records (EHRs), could benefit substantially from the enrichment of medical terminologies with terms encountered in clinical practice. To integrate such terms into existing knowledge sources, they must be linked to corresponding concepts. We present a method for the semantic categorization of clinical terms based on their surface form. We find that features based on sublanguage properties can provide valuable cues for the classification of term variants.

1 Background

Structured terminologies and ontologies play a pivotal role in the automatic processing of health data, as they provide the framework for mapping unstructured information into a machine-readable format. Moreover, the term bases themselves can serve as input for the identification of medical entities in free text. Even though methods from machine learning are gaining popularity, many state-of-the-art systems rely strongly on pre-compiled terminologies (e.g. Savova et al. 2010). The performance of such applications thus relies crucially on the lexical coverage of the term base. However, the major biomedical terminologies, such as the Systematic Nomenclature of Medicine – Clinical Terms (SNOMED CT)[1] and the Unified Medical Language System (UMLS)[2] do not adequately reflect the range of term variants encountered in clinical practice. Especially in languages other than English, where the available terminologies are less comprehensive, this discrepancy can harm performance (Henriksson et al. 2014; Skeppstedt et al. 2014). One strategy to overcome this bottleneck is to enrich the available terminologies with additional variants acquired from domain corpora. Concretely, this involves the recognition of variants in text, and their association with the semantic classes or concepts provided by the respective terminology.

The focus of this paper is on the second task, i.e. the semantic categorization of term variants. In particular, we investigate whether the features of a given sublanguage can be leveraged to associate individual variants with semantic classes. According to sublanguage theory, specialized languages can be characterized by semantic constraints, as well as stylistic preferences and distinctive syntactic patterns (Friedman, Kra, and Rzhetsky 2002; Harris 1982, 1991; 2002). In the medical domain, such differences manifest themselves at fine-grained levels, e.g. between clinical specialties and different document types (Feldman, Hazekamp,

[1] https://browser.ihtsdotools.org/?

[2] https://uts.nlm.nih.gov/home.html

Proceedings of the BioNLP 2019 workshop, pages 211–216
Florence, Italy, August 1, 2019.

Section	Function	Stylistic properties
Anamnesis	Assess environmental and behavioral factors that could influence the patient's condition.	Narrative; high proportion of abbreviations
Comments	Inform colleagues about the current state and further course of treatment.	Telegraphic; high proportion of abbreviations and non-standard variants
Complaints	Summarize the current mental and physical state as experienced by the patient himself.	Narrative; high proportion of lay terms
Conclusion	Inform the patient's GP about the outcome of the consultation and the course of therapy.	Narrative; well-formed syntax; standard terms
Examination	Report on procedures carried out during the consultation.	Telegraphic; high proportion of abbreviations
History	Enumerate prior conditions and procedures that the patient underwent.	List-style; mostly nominal forms; standard terms
Medication	List the pharmaceutical substances administered to the patient.	List-style; mostly nominal forms
Therapy	Document further therapeutic measures.	List-style; mostly nominal forms

Table 1: Overview of the sections of the EHRs in the corpus, their communicative function and style.

and Chawla 2016). We capitalize on this phenomenon for the semantic classification of clinical terms: Drawing on the observation that, even within one clinical document, there are fundamental semantic and stylistic differences between the individual sections, we consider the languages found in different parts of the EHR sublanguages of their own. Based on the assumption that, within the context of a sublanguage, certain variation processes pattern with conceptual properties, we use properties of the surface form as predictors for the semantic classification of the term.

The remainder of this paper is structured as follows: In Section 2, we give an overview of related research. In Section 3, we describe our materials and methods. After the presentation of the results (Section 4) and their discussion (Section 5), we conclude in Section 6.

2 Related research

Especially in emerging domains and under-resourced languages, domain corpora are a valuable resource for terminology development. Automatic Term Recognition (ATR) from biomedical and clinical text is thus a well-studied field (cf. e.g. Spasić et al. 2013; Carroll, Koeling, and Puri 2012; Doing-Harris, Livnat, and Meystre 2015; Zhang et al. 2017 for state-of-the-art systems).

To leverage the acquired terms for NLP, they are typically organized according to their semantic properties. If the target categories are not yet defined, clustering can be used to group semantically related terms and infer taxonomical relations (Siklósi 2015). However, the more common scenario is that the newly acquired variants need to be integrated into an existing knowledge source. To associate terms with pre-defined semantic categories, both external and internal features of the terms have been used. Most approaches rely on external context. In particular, they draw on the core assumption of distributional semantics, which is that semantically similar words tend to occur in similar lexical contexts and syntactic constellations (Sibanda et al. 2006; Weeds et al. 2014). A number of studies showed, though, that term-internal properties can inform the task as well: Medical terms contain a high number of descriptive elements, such as neoclassical affixes or roots associated with a semantic type. Such features have been successfully employed to classify biological concept names and validate the assignment of semantic types in biomedical knowledge sources (Torii, Kamboj, and Vijay-Shanker 2004; Fan, Xu, and Friedman 2007). Morpho-semantic decomposition has also been employed for the semantic grouping of medical compounds in a multilingual setting (Namer and Baud 2007).

However, these approaches only work for a very confined group of terms, namely specialized terms that are based on neoclassical roots, spelled out in their full form, and adhere to grammatical and orthographic conventions. While these conditions might be met in the biomedical genre, they are unrealistic when dealing with input from the clinical domain: In clinical practice, medical staff use both specialized terms and lay variants, which do not contain neoclassical elements. Moreover, clinical

Feature	Criteria	Example term from corpus
REGISTER	Standard term as in SNOMED CT	hypotensie "hypotension"
REDUCTION	Abbreviation or acronym	asp "aspirine"
MORPHO-SYNTACTIC VARIANT	Derivation, paraphrase or compound	thoraxwand "thorax wall"

Table 2: Formal term features.

Semantic class	Example concepts (SNOMED CT term)
ANATOMY	Thoracic structure
CHEMICALS & DRUGS	Human insulin analog product
CONCEPTS & IDEAS	Chronic persistent
DISORDERS	Hypotension
PROCEDURES	Thyroid panel

Table 3: Semantic classes and example concepts.

records are composed in a hectic environment and primarily intended for peer-to-peer communication. They are thus known to contain a high proportion of irregular or intransparent forms, such as misspellings and abbreviations. Therefore, in this paper, we investigate whether the approach can be taken to a more abstract level. Instead of using the words themselves as predictors, we employ a set of non-lexical features reflecting formal properties of the surface form.

3 Materials and Methods

3.1 Corpus Characteristics

We evaluate the approach on a set of terms extracted from a clinical corpus written in Belgian Dutch. This corpus consists of 4,426 EHRs, which were provided by a Belgian hospital. All of them relate to patients diagnosed with diabetes, who visit the hospital in regular intervals for routine check-ups. The EHRs were exported from the clinical data warehouse and de-identified by the ICT team of the hospital. In particular, all personal information concerning the patients themselves, their families, or members of clinical staff was removed. In addition, all researchers that had insight into the data signed confidentiality agreements with the hospital.

All EHRs relate to individual clinical encounters. They were composed with a semi-structured template, which contains different sections relating to the individual stages of a consultation. These sections differ with regard to their thematic scope and communicative function, resulting in characteristic semantic structures and stylistic properties. They can thus be considered distinct sublanguages. For example, the section *complaints* serves to assess the current mental and physical condition. This section is composed in interaction with the patient, which manifests itself in the narrative style and a high proportion of lay terms. By contrast, the *comments* are used for the informal exchange among colleagues. This section is composed in a telegraphic style, containing a high proportion of ungrammatical constructions and jargon expressions. Table 1 gives an overview of the sections and their characteristics.

3.2 Semantic and Formal Annotation

In an earlier project, all EHRs in the corpus were manually annotated with concept codes from SNOMED CT. After manual validation of the term-concept association, a total of 15,025 unique terms, relating to 7,687 different concepts, remain. All concepts were mapped to the semantic groups of the UMLS (McCray, Burgun, and Bodenreider 2001). In a second pass, the terms obtained in the earlier stage were also annotated at the formal level. To this end, the unique terms were manually annotated with a set of binary features reflecting the term's register, morpho-syntactical alternations and reduction processes. Table 2 gives an overview of the formal term features. [3]

Each term was inspected individually. For those features that applied to the term, a positive value was assigned; for the remaining features, the values remained negative by default. For example, the term *hypotens* 'hypotensive' would be assigned the following features: REGISTER – *positive*; REDUCTION – *negative*; MORPHO-SYNTACTIC VARIANT – *positive*.

[3] While the original set of features was more extensive, we used a reduced version for the present study to create more realistic conditions. In a real-life scenario, it is unlikely that resources would be available for the manual coding of term features. Therefore, we only included those features that could be assigned automatically, e.g. by dictionary lookup or morphological analysis.

Section	Number of terms	Number of target classes	F1-score	REGISTER	REDUCTION	MORPHO-SYNTAX
Anamnesis	1081	5	0.73	0.08	0.23	**0.69**
Comments	3105	5	0.64	0.14	**0.54**	0.31
Complaints	1592	5	0.5	0.16	0.02	**0.82**
Conclusion	8214	5	0.48	**0.49**	0.03	0.48
Examination	804	3	0.85	0.25	0.31	**0.44**
History	4202	5	0.45	**0.6**	0.15	0.26
Medication	3529	3	0.99	0.24	0.11	**0.65**
Therapy	508	4	0.86	0.27	0.03	**0.7**
	23035		**0.69**	**0.28**	**0.18**	**0.55**

Table 4: Details of the terms and the results of the classification by section. The last three columns specify the mean importance of the different predictor types; for each section, the highest value is printed in bold. The last row provides the sum of the second column and the mean values of the last four columns.

3.3 Composition of the Concept and Term Sample

For the classification task, we focused on the five most frequently occurring semantic groups, namely DISORDERS, PROCEDURES, CONCEPTS & IDEAS, CHEMICALS & DRUGS and ANATOMY (cf. Table 3). For each group, the associated concepts were ranked by absolute frequency and the number of associated variants. Five concepts per group were chosen for the classification task. The final selection of concepts was based on the diversity of formal alternations observed in the associated variants. For instance, a concept whose terms showed variation in both morpho-syntax and reduction (e.g. a noun phrase and a paraphrase, and an abbreviation and a full form) would be preferred over a concept whose terms only vary at the morpho-syntactical level. Moreover, we aimed to compose the sample such that the full spectrum of the semantic class would be covered. For instance, for ANATOMY, we chose concepts relating to visible body parts (e.g. *leg*) as well as internal organs (e.g. *thyroid*) The final sample consisted of 25 concepts. For each concept, the annotated terms were retrieved from our corpus and sorted by the section of occurrence. Concepts occurring with a frequency of less than 500 within a section were excluded. Consequently, the number of semantic classes varies across sections.

3.4 Experimental Setup

We approached the categorization task as a multi-class classification problem with multiple predictors: *Given the observation of a term in a particular section, predict the semantic category based on the formal features*. Our hypothesis is that the sub-language features of each section influence the informativity of the formal predictors. For example, in a narrative section like the *complaints*, MORPHO-SYNTACTICAL features should be better predictors than in the *medication*, which contains few full sentences, but merely enumerates drugs and dosage instructions. On the other hand, the REDUCTION feature is likely more insightful in the *comments*, which are dominated by informal expressions, than in the *conclusion*, where well-formed expressions prevail.

For the classification experiment, we used a Python implementation[4] of the Random Forest Classifier (Breiman 2001). For each section, the list of annotated terms is split into a training and test set, containing 70% and 30% of all terms respectively. One model is trained and tested per section. To evaluate the results, we calculate the F1-score as well as the mean importance of the different predictor types.

4 Results

Overall, the best results were achieved in those sections that only contain a small number of target classes, namely the *medication*, *therapy* and *examination*. By contrast, the F1-values tend to be lower in those sections that are more diverse. On average, the MORPHO-SYNTACTIC features are the most important predictors, followed by the REGISTER feature. The REDUCTION feature, on the other hand, seems less informative overall.

At the same time, the relative contribution of the feature types varies considerably across the sections: In the *conclusion* and *history*, REGISTER is

[4] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

the strongest predictor; however, in the *conclusion*, the MORPHO-SYNTACTIC features are almost on par with REGISTER. While REDUCTION is most important in the *comments*, it also has a substantial effect in the *examination*. The MORPHO-SYNTACTIC features make the strongest contribution in the *complaints*, *therapy*, *anamnesis* and *medication*; they are also strongest, but not quite as dominant, in the *examination*. Table 4 provides the full results.

5 Discussion

The results show that the semantic complexity of the respective sublanguage influences classification performance. The best F1-scores were achieved in those sections devoted to a very confined topic, while the values were lower in the more heterogeneous ones. This tendency corroborates the findings of previous work studying the effect of sublanguage properties on NLP in the clinical domain (Doing-Harris et al. 2013).

However, we found striking differences in the relative importance of the predictor types. On the whole, the contribution of the predictors patterns with the stylistic properties of the respective sublanguages: For instance, MORPHO-SYNTACTIC features are most informative in those sections composed in a narrative style; REDUCTION is strongest in the informal parts of the document. This finding confirms our initial hypothesis. At a closer look, though, another effect emerges: In semantically homogeneous sections, infrequent features can serve to identify conceptual outliers. For instance, in the therapy-centered sections, which are dominated by nouns relating to pharmaceutical substances, the presence of non-nominal morphological properties, such as an adjective ending, is a strong predictor for a term belonging to another semantic class, such as a temporal modifier.

Our study has its limitations, as it only considers a very small sample of highly frequent concepts. Possibly, for low-frequency concepts, the formal features would not be informative enough to allow a reliable classification. Therefore, in future work, we plan to replicate the experiment at a larger scale, including a more diverse concept sample. Besides, in order to test the generalizability of the method, it would be interesting to evaluate the performance on data from different clinical specialties, and from multiple clinical institutions.

6 Conclusion

We presented a first attempt for the classification of clinical terms by formal features alone. While there is much variation in the results, our experiment demonstrates that sublanguage properties can be exploited to associate terms acquired from domain corpora with semantic categories. This approach could be integrated with other systems to support the enrichment of medical terminologies. In further research, we plan to replicate the study at a larger scale.

Acknowledgements

This work was supported by Internal Funds KU Leuven.

References

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45: 5–32. https://doi.org/10.1023/A:1010933404324.

Carroll, John, Rob Koeling, and Shivani Puri. 2012. "Lexical Acquisition for Clinical Text Mining Using Distributional Similarity." *Lexical Computational Linguistics and Intelligent Text Processing. CICLing 2012,* 232–4. https://doi.org/10.1007/978-3-642-28601-8_20.

Doing-Harris, Kristina, Olga Patterson, Sean Igo, and John Hurdle. 2013. "Document Sublanguage Clustering to Detect Medical Specialty in Cross-Institutional Clinical Texts." *Proc ACM Int Workshop Data Text Min Biomed Inform*, 9–12. https://doi.org/10.1145/2512089.2512101.

Doing-Harris, Kristina, Yarden Livnat, and Stephane Meystre. 2015. "Automated Concept and Relationship Extraction for the Semi-Automated Ontology Management (SEAM) System." *J Biomed Sem* 6: 15. https://doi.org/10.1186/s13326-015-0011-7.

Fan, Jung Wei, Hua Xu, and Carol Friedman. 2007. "Using Contextual and Lexical Features to Restructure and Validate the Classification of Biomedical Concepts." *BMC Bioinform* 8: 264. https://doi.org/10.1186/1471-2105-8-264.

Feldman, Keith, Nicholas Hazekamp, and Nitesh V. Chawla. 2016. "Mining the Clinical Narrative: All Text Are Not Equal." *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, 271–80. https://doi.org/10.1109/ICHI.2016.37.

Friedman, Carol, Pauline Kra, and Andrey Rzhetsky. 2002. "Two Biomedical Sublanguages: A Description Based on the Theories of Zellig Harris." *J Biomed Inform* 35 (4): 222–35. https://doi.org/10.1016/S1532-0464(03)00012-1.

Harris, Zellig. 1982. “Discourse and Sublanguage.” In *Sublanguage. Studies of Language in Restricted Semantic Domains*, edited by Richard Kittredge and John Lehrberger, 231–36. Berlin/New York: De Gruyter.

Harris, Zellig. 1991. *A Theory of Language and Information. A Mathematical Approach.* Oxford: Clarendon Press.

Harris, Zellig. 2002. “The Structure of Science Information.” *J Biomed Inform* 35 (4): 215–21. https://doi.org/10.1016/S1532-0464(03)00011-X.

Henriksson, Aron, Hans Moen, Maria Skeppstedt, Vidas Daudaravičius, and Martin Duneld. 2014. “Synonym Extraction and Abbreviation Expansion with Ensembles of Semantic Spaces.” *J Biomed Sem* 5: 6. https://doi.org/10.1186/2041-1480-5-6.

McCray, Alexa T., Anita Burgun, and Olivier Bodenreider. 2001. “Aggregating UMLS Semantic Types for Reducing Conceptual Complexity.” Stud Health Technol Inform 84 (0 1): 216–20. https://doi.org/10.3233/978-1-60750-928-8-216.

Namer, Fiammetta, and Robert Baud. 2007. “Defining and Relating Biomedical Terms: Towards a Cross-Language Morphosemantics-Based System.” Int J Med Inform 76 (2–3): 226–33. https://doi.org/10.1016/j.ijmedinf.2006.05.001.

Savova, Guergana K, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, Christopher G Chute. 2010. “Mayo Clinical Text Analysis and Knowledge Extraction System (CTAKES): Architecture, Component Evaluation and Applications.” *J Am Med Inform Assoc* 17 (5): 507–13. https://doi:10.1136/jamia.2009.001560 .

Sibanda, Tawanda, Tian He, Peter Szolovits, and Ozlem Uzuner. 2006. “Syntactically-Informed Semantic Category Recognizer for Discharge Summaries.” *AMIA Annu Symp Proc 2006*, 714–18.

Siklósi, Borbála. 2015. “Clustering Relevant Terms and Identifying Types of Statements in Clinical Records.” *Computational Linguistics and Intelligent Text Processing. CICLing 2015.* 619–30. https://doi.org/10.1007/978-3-319-18117-2_46.

Skeppstedt, Maria, Maria Kvist, Gunnar H. Nilsson, and Hercules Dalianis. 2014. “Automatic Recognition of Disorders, Findings, Pharmaceuticals and Body Structures from Clinical Text: An Annotation and Machine Learning Study.” *J Biomed Inform* 49: 148–58. https://doi.org/10.1016/j.jbi.2014.01.012.

Spasić, Irena, Mark Greenwood, Alun Preece, Nick Francis, and Glyn Elwyn. 2013. “FlexiTerm: A Flexible Term Recognition Method.” *J Biomed Sem* 4: 27. https://doi.org/10.1186/2041-1480-4-27.

Torii, Manabu, Sachin Kamboj, and K. Vijay-Shanker. 2004. “Using Name-Internal and Contextual Features to Classify Biological Terms.” *J Biomed Inform* 37: 498–511. https://doi.org/10.1016/j.jbi.2004.08.007.

Weeds, Julie, James Dowdall, Gerold Schneider, Bill Keller, and David J. Weir. 2014. “Using Distributional Similarity to Organise BioMedical Terminology.” *Terminology* 11 (1): 107–41. https://doi.org/10.1075/bct.2.07wee.

Zhang, Rui, Jialin Liu, Yong Huang, Miye Wang, Qingke Shi, Jun Chen, and Zhi Zeng. 2017. “Enriching the International Clinical Nomenclature with Chinese Daily Used Synonyms and Concept Recognition in Physician Notes.” *BMC Med Inform Decis Mak* 17: 54. https://doi.org/10.1186/s12911-017-0455-z.