e-Health – For Continuity of Care
C. Lovis et al. (Eds.)

doi:10.3233/978-1-61499-432-9-565

Sublanguage Analysis of Medical Weblogs

Kerstin DENECKE
ICCAS Innovation Centre Computer Assisted Surgery

Abstract. Analysing medical social media data gains in importance given an increased availability of such data. In this paper, we analyse the language of medical blogs by means of a sublanguage analysis. More specifically, verb usage, semantic categories of used words as well as co-occurrence patterns are determined by means of natural language processing tools. The results show that in this text type, many concepts refer to the semantic categories Living Beings and Chemicals and Drugs. In contrast to clinical documents, the spectrum of verbs in blogs is very broad creating semantic relations of different types. From these language characteristics, we conclude for automatic processing tools for medical blogs that methods for reference resolution and for relation extraction where the relation type does not need to be specified in advance are required.

Keywords. Information extraction, Natural language processing, Healthwebscience

Introduction

The advances in internet and mobile technologies changed the way people access, use and share information. A process of ”revolutionizing healthcare” by improving healthcare and quality of life has begun [1]. Data and experiences on diseases, symptoms and medical treatments are exchanged via instant messaging, blogs, social networking (e.g. Facebook) or video sharing (e.g. YouTube). These tools opened new ways of communicating and enabled for timeless and location-independent information exchange. Patients increasingly rely on the Internet when looking for medical information and advice that extends their ability to share personal experiences and opinions on health concerns [7].

The huge amount of health information available online requires automatic methods for supporting its analysis and interpretation. Among others automatic processing tools that are able to identify relevant pieces of information in medical social media texts are required. Linguistic peculiarities of that particular data need to be considered in the development of appropriate processing tools. Existing language processing tools are especially designed for processing clinical documents and are optimised for that particular sublanguage (i.e., short sentences or even telegraphic style of writing, verbless clauses, use of formal language). It is still an open question, what the differences between the language and word usage in clinical and medical blog are and how existing clinical document analysis tools perform on medical blogs. Clearly, when providing health information or discussing health issues in the Web, the language used is domain-specific, i.e. medical terms and concepts are exploited. Medical social-media data or blogs are written for other purposes than clinical texts and biomedical literature, even though authors can be healthcare professionals, clinical researchers as well as non-professionals. Thus, the term usage and content are expected to be different.

Whereas the linguistic characteristics of clinical and biomedical texts have been analysed comprehensively by other researchers [6] [10] [4], the literary composition of medical social-media data has unfortunately not yet been analysed with the same degree of precision. In this paper, we study the language and content characteristics of medical blogs by means of a sublanguage analysis. We want to find out whether there are characteristic linguistic or semantic patterns in medical blogs that can support or need to be considered in automated processing which is necessary for information retrieval tasks in the medical domain or other applications using medical social media.

1. Methods

1.1. Sublanguage analysis

The sublanguage theory was proposed by Zellig Harris [5]. He claimed that languages of technical domains have a certain structure and regularity which can be specified in computer-processable manner. His theory incorporates domain-specific semantic information and syntactic peculiarities of a domain language. Thus, a sublanguage is in this way a subset of a natural language characterized by the fact that only a subset of the vocabulary and certain grammatical rules are used.

Sublanguage analysis is a technique for discovering units of information and their relationships in narrative text. Semantic categorization of terms, co-occurrence patterns or constraints, usage of terminologies and controlled vocabularies within the sublanguage and other characteristics are assessed within sublanguage analysis [4]. The clinical sublanguage was comprehensively analysed within the context of developing natural language processing tools for clinical documents [12] and compared to the biomedical sublanguage [4], to general English [2] and to newspaper language. One of the most often used mean to analyse the (clinical) sublanguage, is a study of word usage. In this work, we study the sublanguage in medical social media which was so far not considered by existing sublanguage analyses, but can provide useful insights for developing language processing tools for such data. NLP work on social media focused so far mainly on sentiment analysis. Domain-specific usage of terms and language patterns has not yet been analyzed before.

1.2. Study design

We characterise the medical blog sublanguage by studying verb usage, semantic categories of words, and co-occurrence patterns of semantic categories with verbs. For this purpose, language processing tools are applied to the data, namely (1) DragonToolkit [11] for identifying medical concepts in free text and their categories, (2) ReVerb [3] for identifying relations in sentences, (3) openNLP for part of speech tagging and chunking.

As underlying terminology, we are using the Metathesaurus of the Unified Medical Language System (UMLS). Each UMLS concept is assigned to at least one of the 134 semantic types that in turn have been aggregated into a set of 15 semantic groups to reduce complexity. We chose the open information extraction tool ReVerb for extracting relations in form of two arguments connected by a relation type. ReVerb is designed for Web-scale information extraction, where the target relations cannot be specified in advance and speed is important [3]. We selected ReVerb for our study

since it provides correlation patterns without the need of specifying relation types of interest in advance. ReVerb takes as input a sentence tagged with part of speech information and NP chunks and returns relation triples consisting of two arguments and a relation phrase.

A co-occurrence pattern as used in the sublanguage analysis consists of one or two sets of main categories connected by a verb. It is received by first applying ReVerb to a sentence. The arguments extracted by ReVerb are processed by the DragonToolkit to map to UMLS concepts and to get semantic categories of the words in the argument string. An example of a co-occurrence pattern is: *[Procedures + show + Disorders]*. The left argument was mapped to a concept of the semantic category *Procedures* and the left argument to a concept of the semantic category *Disorders*.

The data set underlying our study comprises a random sample of around 12800 texts collected over a period of four years starting in 2009 from RSS feeds provided through the blogs of WebMD (http://www.webmd.com) and EverydayHealth (http://www.everydayhealth.com), from the Medicinenet website (http://www.medicinenet.com), as well as blogs from several sources. WebMD provides health information provided in a blog style by physicians. EverydayHealth.com is a provider of online health information designed for non-health professionals. MedicineNet.com provides authoritative medical information for consumers. The RSS feeds were collected, the URL to the complete text identified and the text extracted from the HTML page. From WebMD 1000 texts were considered; from the other sources 100 texts per source. We chose only physician-written texts to be able to compare with clinical texts that are also written by doctors.

2. Results

In the medical social-media data words that belong to the category *Living Beings* are used often among all studied blog data sets. In addition, words falling into the categories *Disorders, Chemicals and Drugs, Concept and Ideas* are exploited frequently.

The most frequently used co-occurrence patterns determined are:

- Patterns referring to having or developing diseases or symptoms (e.g. [have + Disorders], [develop + Disorders]),
- Patterns on drug consumption (e.g., [take + Chemicals and Drugs], [receiv + Chemicals and Drugs]),
- Patterns representing statements of persons on medical issues (e.g. [Living Beings + say + Concepts and Ideas, Phenomena], [say + Living Beings], and
- Patterns referring to the use of devices for confirming a diagnosis or performing procedures (e.g. [Devices + discov + Disorders, Procedures]).

It can be seen that frequently occurring patterns are referring to having or developing diseases or symptoms and performing actions / medical procedures including drug consumption. Verbs that are used frequently in medical blogs fall into the following categories. (Note that verb forms of *be, have* and *do* were not considered even though they had a high frequency as well.)

- Verbs describing effects of medical procedures, drugs and proteins or genes: affect, regulate, reduce, prevent, improve, help, treat, increase, enhance, enable, benefit, strengthen

- Verbs describing observations: appear
- Verbs referring to education and research: learn, discuss, study, empower, educate
- Verbs describing disease development and transfer: contribute, depend, infect, require, acquire.

Table 1. Linguistic characteristics of medical blogs

Sentence structure	Rather long sentences
Word usage	Adjectives, descriptive and narrative words
Language	Clinical terminology, medical terms are very frequent, Consumer health vocabulary, common language,
Spelling	Abbreviations, misspellings
Semantic categories of words	Living Beings, Disorders, Chemicals and Drugs, Concept and Ideas

3. Discussion

This paper studied the characteristics of medical blogs by means of a sublanguage analysis. Content-wise information on treatments, diseases and drugs are given rather than information on an individual in social media. In contrast to clinical documents, medical blogs can refer to an individual person, but contain seldom measured (clinical) values and in our data set authors often refer to general health statistics or health information. Friedman et al. studied the sublanguage in clinical and biomedical texts [4]. They found out that the clinical sublanguage is characterised by a usage of words belonging to the semantic categories Behaviour, Findings, Medication, Device, Body function, Labtest, Procedure [4]. This is in contrast to our results for medical blogs where concepts referring to *Living Beings, Disorders, Chemicals and Drug* are used most frequently. Also verb usage is different. The biggest differences to language used in clinical documents were found to be the usage of a broad spectrum of verbs in medical blogs, which is in contrast to clinical documents and the reference to terms referring to living beings. This results in challenges for automatic tools that include reference resolution, processing of verbs and detecting relations. The broader spectrum of verbs used in medical blogs results in a need for tools that are able to collect these verbs and analyze their meanings when it comes to automatically analyze social media text. A broad range of verbs is used in medical blogs, which makes it difficult to specify in advance verb patterns or extraction rules that contain verb roots. In contrast a limited set of verbs are used throughout the discharge summaries – a characteristic that is used in natural language processing from those texts. An extended linguistic analysis of medical blogs should therefore concentrate on verb usage, which was studied for medical corpora already [15]. Wang et al. studied operative notes and their content and language by analyzing and categorizing verbs [14]. An alternative approach to study differences between language and content in clinical notes of different types and specialities was suggested by Patterson and Hurdle [13]. They proposed a document clustering approach to study in more detail characteristics of clinical sublanguages. To determine patterns of semantic types, we applied open information extraction, which was so far not considered for sublanguage analysis. We decided to use that approach to be able to extract the relation type. This goes beyond co-occurrence analysis of semantic types and is interesting, since verb meanings would get lost during the

mapping process to UMLS. A limitation of this approach is that only relations with two arguments are extracted by the chosen implementation.

Corpus analysis methods [9] would be an additional option for analyzing medical social media. We chose sublanguage analysis to be able to compare with results from sublanguage analysis from clinical texts.

It is clear that an assessment as described in this paper can only consider a subset of medical social-media data. The data set underlying our study mainly comprised blogs written by health professionals. The content is thus more informational; information on medical treatments is described. Content and language could be different when only patient-written blogs are considered. A generalization of the results to patient-written texts is difficult. A limitation of the study is the mapping to UMLS concepts. The DragonToolkit does not use the latest UMLS version. The tool has already been evaluated on biomedical abstracts collected from MEDLINE and achieved a precision rate of 71.6% and a recall of 75% [11]. However, quality might be different for mapping text of medical social-media to UMLS concepts. Our study focused on identifying linguistic characteristics that might influence the future development of natural language processing tools for medical social-media data. This quality assessment is still an open research questions.

References

[1] L. Aase, G. D., M. Gould, J. Noseworthy, and F. Timimi. Bringing the Social-media Revolution to Health Care. Mayo Clinic Center for Social-media, 2012.

[2] D. A. Campbell and S. B. Johnson. Comparing syntactic complexity in medical and non-medical corpora. Proc AMIA Symp, p.90–94, 2001.

[3] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, p.1535–1545, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[4] C. Friedman, P. Kra, and A. Rzhetsky. Two biomedical sublanguages: a description based on the theories of zellig harris. Journal of Biomedical Informatics, 35:222–235, 2002.

[5] Z. Harris. A theory of language and information: a mathematical approach. Clarendon Press, Oxford, 1991.

[6] I. Kovic, I. Lulic, and G. Brumini. Examining the Medical Blogosphere: An Online Survey of Medical Bloggers. Journal of Medical Internet Research, 10(3), 2008.

[7] A. Lau, K. Siek, L. Fernandez-Luque, and et al. The role of social media for patients and consumer health. Yearb Med Inform., 6:131–8, 2011.

[8] A. McCray. An upper-level ontology for the biomedical domain. Comp Funct Genom, 4:80–84, 2003.

[9] E. Miscin. Use of Corpus Analysis Tools in Medical Corpus Processing. INFuture2013: "Information Governance", 187-196: 2013 .

[10] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle. Extracting information from textual documents in the electronic health record: A review of recent research. p.128–144, 2008.

[11] X. Zhou, X. Zhang, and X. Hu. Dragon toolkit: Incorporating autolearned semantic knowledge into large-scale text retrieval and mining. In Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), October 29-31, 2007, Patras, Greece, 2007.

[12] N. Sager and N. T. Nhan. The computability of strings, transformations, and sublanguage. In B. E. Nevin and S. M. Johnson, editors, The legacy of Zellig Harris: Language and information into the 21st century, p.79–120. John Benjamins, 2002.

[13] O. Patterson, J.F. Hurdle. Document Clustering of Clinical Narratives: a Systematic Study of Clinical Sublanguages, AMIA Annu Symp Proc., p.1099-1107, 2011.

[14] Y. Wang , S. Pakhomov, N.E. Burkart et al. A study of actions in operative notes. Annu Symp Proc., p.1431-40, 2012.

[15] O. W. Tchami, M-C. L'Homme, N.Grabar. Discovering Semantic Frames for a Contrastive Study of Verbs in Medical Corpora.Terminologie Intelligence Artificielle (TIA) 2013. Villetaneuse, France.