Chapter 2
The Burgeoning of Medical Social-Media Postings and the Need for Improved Natural Language Mapping Tools

Kerstin Denecke and Nazli Soltani

Abstract Medical social-media data provides a wealth of data generated by both healthcare professionals and patients alike. In fact, there are many medical social-media sites such as forums, where patients freely dialog with a healthcare professional or with other patients, often posing questions and responding to advice, or Weblogs, where groups of people describe their experiences with medical conditions and the various treatment plans to treat those conditions. All in all, one can no longer ignore the fact that social media has dramatically changed the structure of healthcare delivery in many ways. Simply from a medical data standpoint alone, social-media platforms have altered the way medical information is disseminated. That is, important medical information is no longer found exclusively in patients' clinical narratives, commonly shared by physicians and other healthcare workers at regular professional meetings and conferences. Instead, user-generated content on the Web has become a new source of useful information to be added to the conventional methods of collecting clinical data. The challenge we face, however, is to design information extraction tools that can make the rich resources of medical data found in social-media postings exploitable. In this chapter we analyze the linguistic features of medical social-media postings juxtaposed to the linguistic features of both clinical narratives (e.g., discharge summaries, chart reviews, and operative reports) and biomedical literature, for which there already exists tools for performing information extraction. We show the shortcomings of these mapping tools when applied to medical social-media postings, and propose ways to improve

K. Denecke (✉)
Innovation Center Computer Assisted Surgery, University of Leipzig, Leipzig, Germany
e-mail: kdenecke@web.de

N. Soltani
University Medical Center, Göttingen, Germany
e-mail: nazli.soltani@med.uni-goettingen.de

A. Neustein and J.A. Markowitz (eds.), *Where Humans Meet Machines: Innovative Solutions for Knotty Natural-Language Problems*, DOI 10.1007/978-1-4614-6934-6_2,

such tools so that the wealth of medical data located in medical social-media can be made available to healthcare providers, pharmaceutical companies, and government-supported epidemiological agencies.

Social Media and Its Use in Healthcare

As advances in the Internet and mobile technologies have improved the way how people access, use, and share information in the last few years, new ways of communicating about health have evolved, enabling a 24/7 and location-independent medical-information exchange. These new media comprise instant messaging (e.g., Twitter), blogs, online forums, social networking (e.g., Facebook), or video sharing (e.g., YouTube).[1] Mayo Clinic researchers have opined that social media has begun a process of "revolutionizing healthcare" by improving healthcare and quality of life (Aase et al. 2012). In fact, patients increasingly rely on the Internet when looking for medical information and advice. The Internet also facilitates patients' ability to share their personal experiences and opinions with others who have the same health concerns.

In analyzing the popularity of social media one cannot help realize that one distinct advantage is that the communication barriers are considerably lower than face-to-face communication, allowing patients to write in social-media platforms a bit more freely about their illnesses and their experiences with drugs and medical treatments than they would normally do in other settings. One of the most active groups of online health-information seekers are those who suffer from chronic conditions or from rare diseases. By surfing for medical information in cyberspace this group of users can learn what others have to say about quality of care, or about important issues regarding treatment and diagnosis. In addition to gaining knowledge, this group of online medical-information seekers is able to communicate with other persons who are suffering from the same disease, thus serving as a hedge against feelings of isolation and loneliness. All in all, social-media platforms offer a range of possibilities to facilitate the sharing of useful medical information and personal experiences related to healthcare, particularly when the information seeker is unable to obtain adequate support and feedback from fellow sufferers in their home community.

Another important observation worthy of mention here is that one cannot ignore the fact that social media has dramatically changed the structure of healthcare delivery in the modern world. From a medical data standpoint, important medical information is no longer exclusively found in patient clinical narratives, usually shared by physicians and other healthcare workers at regular professional meetings and conferences. Instead, user-generated content on the Web has become a new source of useful information to be added to the conventional methods of collecting clinical data.

[1]While younger populations were fast in adopting these new technologies, the number of older adults using social media is also growing fast.

Given the wealth of user-generated content, automatic methods are essential for extracting relevant information, for organizing and digesting the data for various user groups as well as for preparing it for statistical analysis. Medical social media has also created a burgeoning new class of empowered patients "armed with sophisticated technological tools" (Altarum Institute (2012)), yet we can't afford to lose this fountain of medical data generated daily by healthcare consumers and providers who avail themselves of social-media platforms for sharing their clinical experiences.

Against this ambitious task of extracting, organizing, and distilling medical social-media data automatically are a number of concerns that cannot be overlooked. For example, the existing approaches to information extraction primarily focus on clinical narratives or biomedical literature. It is still unclear whether these approaches and tools are suited for processing and analyzing medical social-media data. Furthermore, one must consider whether existing systems that understand the technical language of biomedical literature or the language of healthcare professionals appearing in clinical narratives for that matter can be adapted to the way ordinary people speak about their medical conditions. In this chapter, we explore the unique linguistic characteristics of medical social media and analyze two existing information-extraction tools for clinical narratives and biomedical literature to see if they can similarly be used to identify important medical concepts found in consumer blog postings. We also suggest important ways of modifying such tools so that they can perform better in analyzing medical social-media content.

Examples of Medical Social-Media Communication Platforms

Medical social media comprises Weblogs, forums, or social network platforms that deal with health-related issues. Forums are basically Internet message boards, where patients or friends and relatives of patients discuss their own experiences and personal thoughts or in the alternative ask questions and seek advice. A blog differs from a discussion forum in that all of the Web site entries are displayed in reverse chronological order; a blog often has only one author whereas many persons contribute to a forum. A blog is defined specifically as a medical blog, when its main topic is related to health or medicine (Boulos et al. 2006). The exact number of forums, blogs, and blog postings dealing with health issues is unknown. For blogs in general, Weblog hosting services have made some numbers available. WordPress,[2] one of the popular Weblog hosting services, reports on its Web site that WordPress.com users produce about 29.2 million new posts and 40.5 million new comments each month.

To bring together health bloggers, entire communities have been set up. *Blognation* and *HealthBlogger Network* are examples of such communities. *Blognation*[3] is

[2] http://en.wordpress.com/stats/.

[3] http://www.medical-blogs.org.

a network of blog directories that lists blogs for different categories including health, but also dealing with other topics, such as books, art, music, and lifestyle. The *HealthBlogger Network*[4] engages over 3,500 bloggers which can be broken down into two categories: healthcare professionals and patients. In general, about 50–60 % of the healthcare bloggers are healthcare practitioners and medical researchers, often from the leading American medical schools such as Harvard or Yale, while the remaining 40–50 % of bloggers comprise patients suffering from chronic or acute illnesses (Miller and Pole 2010).

Besides blogger networks, patients or healthcare professionals are forming communities where they share their knowledge, discuss, or learn from each other. A well-known example of a medical social-media communication platform is *PatientsLikeMe*.[5] This is a social network for patients that allows them to share health-related experiences and to compare various treatment plans. Such online conversations which can also be in the form of a medical diary often contain vast amounts of experiential knowledge. Other platforms, in contrast, try to make this kind of patients' first-hand experiences useable in some fashion. For example, patients' experiences and opinions extracted from social-media postings can be exploited for planning marketing strategies in the pharmaceutical industry or in the healthcare insurance industry. As a result, healthcare products may be improved based on patients' reported personal experiences found in medical social media.

Medical social media also represents a rich resource for learning about patient-compliance behavior, as well as their feelings, attitudes, and experiences with medical and surgical treatment. *Treato.com*, for example, is a social health site that analyzes online patient discussions, collects automatically information about what patients have to say about their medications and conditions found on the blogs and other social-media fora, and subsequently extracts and summarizes the relevant information from these postings. This is done in order to provide valuable insights into patients' opinions, attitudes, and experiences that would help pharmaceutical companies better improve their products.

Platforms such as *Webicina*[6] provide access to curated, medical social media which is defined as media that is filtered, selected, and reviewed. This helps to make more efficient use of the massive amount of social-media data that is out there. Many of those who rely on Webicina are physicians, pharmaceutical companies, and other healthcare professionals. Webicina is certainly not in a category all by itself as there are other enterprises that likewise try to make use of medical social-media data by effectively curating the information found in cyberspace (Hillan 2003; Himmel et al. 2008). Beyond getting information on health-related issues, sudden changes in the public health status can be identified in medical social-media data. This allows for a prompt reaction, early on, from health organizations to such critical events like a swine flu outbreak (Denecke et al. 2012).

[4] http://www.wellsphere.com/health-blogger.

[5] http://www.patientslikeme.com/.

[6] http://www.webicina.com.

Linguistic Characteristics of Medical Social Media

Medical social-media data is written for different purposes than clinical texts and biomedical literature, even though authors can be healthcare professionals as well as patients. Thus, the literary style of medical social media is markedly different than that of clinical texts and biomedical literature. Whereas the linguistic characteristics of clinical and biomedical texts have been analyzed in painstaking detail by other researchers (Kovic et al. 2008; Meystre et al. 2008; Friedman et al. 2002), the literary composition of medical social media has unfortunately not yet been analyzed with the same degree of precision.

Table 2.1 summarizes the linguistic characteristics of these three text types, showing how the first two differ from medical social media:

1. *Clinical texts* comprise documents produced by physicians (e.g., discharge summaries, chart reviews, and operative reports), which are mainly produced to fulfill the physician's reporting duties and to document their diagnosis and treatment of patients.
2. *Biomedical texts* refer to biomedical literature, where researchers are presenting the results of clinical studies.
3. *Medical social media* consists of textual content that is made available by either patients or healthcare professionals on the Web using social-media tools such as blogs or forums.

The content and language of medical social media depends on whether the author is a healthcare professional or a layperson and on their relationship to the disease, such as a treating physician, a patient undergoing care, or a relative or a friend of a patient. For example, one may find on the Web an assortment of clinical cases[7] written by healthcare professionals which are similar in content and language to clinical text (that is, short, abbreviated sentences that can be devoid of nouns and verbs). However, the majority of social-media postings are of a rather personal, as opposed to clinical, nature. As a result, they are often written in a narrative fashion. Since a Weblog consists of a kind of personal diary, it is not surprising that in this kind of media outlet one often finds that personal opinions are freely expressed. In such postings one can find a large number of personal pronouns (e.g., "*I saw ...*", "*I experienced ...*," *or* "*My daughter is sick*"). Furthermore, rather long sentences may also be prevalent, containing a wealth of adjectives to describe situations, experiences, impressions, etc. Postings are written in common everyday language—even using language that is modern and hip—so as to appeal to the average reader. An example from the multiple sclerosis blog *Stellarlife*[8] illustrates this point: "*Yesterday I finally got to see an orthopedic, okay hold the buggy,*

[7] e.g., http://clinicalcases.org.

[8] http://dj-astellarlife.blogspot.de/.

Table 2.1 Linguistic characteristics of clinical texts, biomedical texts, and medical social media

Text type	Clinical text	Biomedical text	Medical social media
Sentence structure	Ungrammatical sentences	Often literature style (long sentences)	Rather long sentences
	Short, telegraphic phrases (*Aspirin* or *Fever*)	Academic	
	Often without verbs or other relational operators	Grammatically correct formulated sentences	
Word usage	Word compounds (*high blood pressure*), formed ad hoc	Frequent use of passivization	Adjectives
	Modifiers are related to temporal information (e.g., *sudden*), evidential information (e.g., *rule out*, *no evidence*), severity information (*mild*, *extensive*), body location	Word compounds (e.g., *high blood pressure*), formed ad hoc	Descriptive and narrative words
		Nominalization and noun compounding	
Spelling	Misspellings	Correctly written words	Abbreviations
	Abbreviations, acronyms		Misspellings
Language	Mix of Latin and Greek roots with corresponding host language (German, English, ...)	Scientific language and writing	Common language, rather than domain-specific language or clinical terminology
	Domain-specific language	Mix of Latin and Greek roots with corresponding host language (German, English, ...)	Host language
		Domain-specific language	

I just googled the guy and he is a PA-C/MPAS!! *Shut the front door*! *That means*: *Physician's Assistant-Certified/Masters Physician's Assistant Studies—huh*."

While the intent is to use everyday language, the insertion of medical terminology in the posting actually depends entirely on the content of each posting. So, for example, when a person is writing about their experiences with a disease, the corresponding terminology may be used, sometimes even with explanations of the clinical terms. Here is a snippet from Diabetesmine[9] to illustrate this pattern: *No one is sure what causes this dead-in-bed syndrome*, *but the theory is that a nighttime low blood sugar—called a nocturnal low—episode triggers some kind of fatal cardiac arrhythmia*.

Another distinction between social-media postings and clinical texts is that in the former, abbreviations are often explained in the text of the posting itself. In addition, sometimes words or word phrases are typographically highlighted. This can be done via the use of full caps, quotation marks around a word or a word phrase, or other stylistic means of allowing part of the text to stand out more saliently. We draw again from Diabetesmine to make this point: *While many parents are likely relieved to* "*get a break*" *and have their kids back in class*, *this can be a very stressful experience for the parents of Children With Diabetes* (*CWDs*) *who have a LOT more to worry about than just textbooks and extracurricular activities*. In this sentence, the word "lot" was highlighted by using capital letters. For automatic language processing programs, the use of full caps can become a problem, since words that appear in upper case are often considered abbreviations by such programs.

Automatic language processing program can likewise become confused when bloggers insert in their postings a verbatim quote from another person, often a high-profile scientist or public official who has something to say about a major health issue. In *Diabetesmine*, we found that the blogger inserted a quote from a high-profile scientist to support her own contention about the danger of sugar levels dropping precipitously at night: "*We've now known for decades that* (*overnight*) *is the most common time for severe hypoglycemia*," *says Dr. Irl Hirsch*, *assistant professor and endocrinologist at the University of Washington*, *and a type 1 PWD himself*. In short, the problem with automatic processing of texts that contain verbatim quotes is that in order to have a correct reference to the quote a processing algorithm must correctly link the cited person, and not the blog author, to the quote itself. Co-reference resolution algorithms are required for this task.

Not only do blogs contain certain kinds of conversational features, as illustrated above, but they are also very much prone to short block paragraphs and bullet point itemization. Bloggers have been found likewise to give headings to their postings as a way of categorizing the content of their blogs. In health forums, categories are often formulated by the forum host in advance. In such cases, forum contributors just scroll down, adding their posting to one of these predefined categories. These categories may cover a wide spectrum of healthcare topics such as how patients (and their relatives, caretakers, and friends) cope with debilitating

[9] http://www.diabetesmine.com.

diseases, treatment plans, drugs and medical devices, health insurance coverage issues, or even suggestions of useful resource material on combatting illnesses and diseases, such as new books, informative radio and television programs, as well as television documentaries on the topic at hand.

In short, tools for the automatic processing of medical social media need to consider the range of stylistic preferences in the presentation of blog and forum content, the syntactic features characterizing blog communications—which can be notoriously long-winded, ungrammatical, and idiosyncratic as displayed by the use of full caps and italics for emphasis—along with other features emerging indigenously in this new form of communication. We briefly showed how medical social-media postings differ from clinical texts and biomedical text. Unfortunately, at present, the existing text-mining tools are best suited for processing clinical and biomedical texts, where the language usage and content differ significantly from medical social media.

Extracting Information from Text

Structured or coded data is required for extracting relevant information from medical social-media postings and to make it accessible to humans in a useful and understandable way. Consequently, well-designed natural language processing methods are sorely needed for the extraction of information. But first, let's understand what information extraction entails.

Information extraction identifies facts or information in texts (Grishman 1998). Named-entity recognition (NER) is a subfield of information extraction which aims at identifying within a collection of text all of the instances of a name for a specific type of thing (Cohen and Hersh 2005). Examples include names of diseases and illnesses, drugs, persons, or locations. A potential use case of NER is to acquire metadata from texts for user modeling (Barla and Bielikova 2010) or to support text classification, filtering, or information retrieval (Denecke 2012).

Entities can be recognized in natural language text in two ways:

1. A simple lexicon lookup
2. Extraction patterns that are either manually created or learned from training corpora using supervised machine learning techniques

Lexicon lookup approaches search for matches with words of a lexicon of named entities in a given text. Difficulties are found to arise, namely, because there is no complete dictionary for most types of medical or biomedical entities. Therefore, the simple text-matching algorithms that are commonly used in other domains are not sufficient here. In extraction pattern-based approaches, patterns such as "[Title] [Person]" for the extraction of a person name (e.g., "Mr. Warren") are generated either by hand or by supervised machine learning techniques. Manual rule-based approaches can be very efficient, but unfortunately such systems require manual efforts to produce the rules that govern them. Machine learning techniques on the

other hand that don't require costly human annotators do however require large training corpora to train their underlying models.

Despite these difficulties, there are tools available for extracting named entities from medical social-media postings which have been originally developed for clinical and biomedical texts. Such tools are premised on different methods and lexical resources. Below, we provide a brief overview of the lexical resources available for processing medical texts and present two NER tools which consider the linguistic peculiarities of such texts.

Knowledge Resources for Processing Medical Texts

Unified Medical Language System (UMLS[10]) is composed of three main knowledge components: Metathesaurus, Semantic Network, and SPECIALIST Lexicon.

- The UMLS Metathesaurus integrates vocabularies from the biomedical domain (e.g., Medical Subject Headings (MeSH), Systematized Nomenclature of Medicine (SNOMED CT)) and provides a mapping structure between them. Each concept has specific attributes that define the meaning of the concept. In the release version 2012AA, the Metathesaurus comprises 2,669,792 concepts in 21 languages. The vocabularies that are integrated into the UMLS contribute thesaural relationships between concepts (e.g., "child" or "parent" relationships). Each concept is assigned to at least one semantic type of the Medical Semantic Network (MSN). The UMLS MSN (McCray 2003) is a network of general semantic categories or types where semantic types are linked by relationships. It provides 134 semantic types that have been aggregated into a set of 15 semantic groups to reduce complexity (McCray et al. 2001) (e.g., the concept *atrial fibrillation* belongs to the semantic types *Finding and Pathologic Function* that in turn belong to the semantic group *Disorders*).
- Semantic network relationships connect UMLS semantic types to each other. For example, the semantic type *Body Part*, *Organ*, *or Organ Component* is associated with the semantic type *Body Location or Region* by the relations labeled *location of*, *has location*, and *adjacent to* with.
- The SPECIALIST Lexicon provides linguistic *knowledg*e. For example, syntactical information on (medical) terms, and natural language processing tools, such as a tokenizer which splits a sentence into tokens, are part of the SPECIALIST Lexicon.

As mentioned before, the UMLS integrates various vocabularies including MeSH and SNOMED CT.

[10] http://umlsinfo.nlm.nih.gov/.

MeSH[11] is the controlled vocabulary thesaurus of the National Library of Medicine which is used for indexing articles for their digital library known as PubMed. MeSH consists of sets of terms, naming descriptors in a hierarchical structure. At the most general levels of the hierarchical structure are the very broad headings, such as "*Anatomy*" or "*Mental Disorders*." More narrowly defined headings for these general terms are found at the more restricted levels of the hierarchy. For example, terms such as "*Ankle*" and "*Conduct Disorder*" reflect refinements of the categories of anatomy and mental disorders, respectively.

SNOMED CT[12] is a multilingual collection of medical terms. It contains more than 311,000 active concepts and around one million relationships. SNOMED CT provides general terminology for the electronic health record, consisting of concepts, descriptions, and relationships:

- Concept*s* represent clinical ideas, such as "*neoplasm*" or "*abscess*."
- Descriptions link appropriate human-readable terms to concepts.
- Relationships link each concept to other concepts that have a related meaning. As such, relationships provide formal definitions as well as other characteristics of concepts.

Tools for Extracting Medical Information

The vocabularies and terminologies introduced in Section "Knowledge Resources for Processing Medical Texts" can become relevant for identifying instances referring to medical entities by NER tools. Two tools are described below that are based on the UMLS.

1. The *MetaMap System* (Aronson 2001) is provided by the National Library of Medicine. The tool maps natural language text to concepts of the UMLS Metathesaurus. MetaMap follows a lexical approach and works in several steps. First, it parses a text into paragraphs, sentences, phrases, lexical elements, and tokens. From the resulting phrases, a set of lexical variants is generated. Candidate concepts for the phrases are retrieved by lexicon lookup from the UMLS Metathesaurus (version: UMLS 2012) and evaluated. The best candidates are organized into a final mapping in such a way as to best cover the text. Precision of MetaMap, which is the fraction of retrieved concepts that are relevant, was assessed for different text types already, namely, for respiratory findings (Chapman et al. 2004), mailing lists (Stewart et al. 2012), and figure captions in radiology reports (Kahn and Rubin 2009). The precision for these text types ranges between 56 and 89.7 %. Figure 2.1 shows an example mapping of MetaMap for a given sentence.

[11] http://www.nlm.nih.gov/mesh/meshhome.html.

[12] http://www.ihtsdo.org/snomed-ct/.

Example of MetaMap mapping result:
Input sentence: *Laboratory results in a patient with pneumonia, septic shock, and acute renal failure.*

Concept mapping:
laboratory (Laboratory domain) [Functional Concept]
Results (Result) [Functional Concept]
patient (Patients) [Patient or Disabled Group]
Pneumonia [Disease or Syndrome]
Septic Shock [Pathologic Function]
Acute Renal Failure (Kidney Failure, Acute) [Disease or Syndrome]

Fig. 2.1 Mapping example of MetaMap for sentence "Laboratory results in a patient with pneumonia, septic shock, and acute renal failure." Semantic types are shown in parentheses, while semantic groups are shown in *square brackets*

2. The *Dragon Toolkit* (Zhou et al. 2007) is a Java-based development package for information retrieval and text mining. It provides a linguistic parser, text clustering and classification algorithms, and an ontology-based biomedical text annotator. The biomedical text annotator MaxMatcher, which is part of the Dragon Toolkit, uses a generic extraction approach (referred to as "approximate dictionary lookup") to cope with term variations. The basic idea of this approach is to capture the significant words only, rather than all the words associated with a particular concept. MaxMatcher has already been evaluated on biomedical abstracts collected from MEDLINE. A precision rate of 71.6 % and a recall of 75 % were achieved (Zhou et al. 2007). The underlying terminology of MaxMatcher can be varied; and either the UMLS or the MeSH thesaurus (version UMLS 2004AA Version) can be used.

Tools in Practice

Ironically, while for online news a comparison of several NER tools (e.g., Alchemy API, DBpedia Spotlight, OpenCalais) has already been performed (Rizzo and Troncy (Rizzo and Troncy 2012)), there are yet no such similar comparisons made of NER tools for medical social media. As such, evaluation results of NER tools in the medical domain are only available for extraction from clinical or biomedical texts while not for medical social media. In this section, we present results of a qualitative comparison of the two described UMLS mapping tools, MetaMap and Dragon Toolkit, which are applied to medical social media.

Method

We applied the two tools to (1) twenty texts drawn from the forum "This MS[13]," where patients with multiple sclerosis are discussing their problems, and (2) blog postings from "WebMD[14]," where physicians are writing about topics related to health and medicine. The results were checked manually, sentence by sentence. The assessment of the output of the tools comprised:

- Judging presence of the detected named entity (*present* in the text or not)
- Judging relevance of the detected named entity (*relevant* or *irrelevant*)
- Judging the type of the detected named entity (*correct* or *incorrect*)

We identified words that are crucial for understanding the text or the sentence which could not be identified by either one of the tools used. The objective of the assessment was to give insights into the possibilities and limitations of these tools when they are applied to medical social-media data. MetaMap was run with UMLS 2012AA, while the available Dragon Toolkit was based upon UMLS 2004AA.

Observation on Quality of Information-Extraction Tools

The main observation for both tools is that the produced mappings do not contain concepts for all content-bearing terms that are used in medical social media. While medical terms are mostly reflected in the mappings, descriptive or concept-relating words are missing. Detailed observations are provided below in the discussion of both tools.

Mapping Observations of MetaMap

Terms from common language or consumer health vocabulary (CHV) referring to medical concepts are often mapped incorrectly by MetaMap or are even missing in the mapping altogether. Wrong mappings occur in particular for personal pronouns: "*I*" is mapped to "*Iodides* [*Inorganic Chemical*]"; "*my*" is mapped to "*Malaysia* [*Geographic Area*]"; *and* "*she*" is mapped to "*SHE gene* [*Gene or Genome*]." Verbs are often not mapped at all or are wrongly mapped. For example, the verb "*found*" in the sentence "*Keratin is found in your hair*" is mapped to (clinical) "*Finding*"; or the verb "*go*" is mapped to the concept "*GORAB gene* [*Gene or Genome*]." Keeping the meaning of verbs after mapping is extremely important for interpreting a text automatically (and also manually).

[13] http://www.thisisms.com/forum/daily-life-f35/topic20839.html (Section "Daily Life").

[14] 'http://rssfeeds.webmd.com/rss/rss.aspx?RSSSource = RSS_PUBLIC'.

Adjectives can also be incorrectly mapped or, like verbs, adjectives may not be mapped at all as in "*nasty*" or "*embarrassing*" which can drop off the mapping altogether.

Another class of wrongly mapped lexical items is that of words or word phrases used in free text (nonclinical texts) which are errantly mapped to clinical phrases. Some of the most common errors are words or word phrases such as "*of course*" which is mapped to "Course [Temporal Concept]," "*Hi*" which is mapped to "*ABCC8 gene* [*Gene or Genome*]," or "*Thanks*" which is mapped to "*TNFSF13B wt Allele* [*Gene or Genome*]." In addition, numeric expressions can stump mapping programs as they require separate processing. MetaMap, for example, destroys the expression "*about two to 2 1/2 months*" which is mapped to two concepts: "*Two* [*Quantitative Concept*]" and "*month* [*Temporal Concept*]."

And yet another problem is that MetaMap often provides multiple mappings which may differ significantly in regard to the underlying concepts of which they are comprised. The reason for this is that because words can have different meanings this often results in various possible mappings to concepts with different semantic types. No doubt, having multiple mappings available becomes a problem. MetaMap provides confidence values for these different mappings. Nevertheless, when there are several mappings with the same confidence value, it remains a question of how to select the "correct" mapping automatically.

Mapping Observations of Dragon Toolkit

For the mapping of the Dragon Toolkit or the underlying mapping algorithm MaxMatcher we have made the following observations: In general, Dragon Toolkit maps less terms to UMLS concepts than MetaMap. However, even though it maps fewer terms, the medical terms that it maps to concepts are for the most part correctly identified by Dragon Toolkit. For instance, "*cat scan*" is mapped to "*cat scan* [*Diagnostic Procedure*]," and "*MS*" (multiple sclerosis) is mapped to "*ms* [*Disease or Syndrome*]." Nevertheless, there are some abbreviations that are wrongly mapped as in "*edss*" (actually referring to *Expanded Disability Status Scale*) which is mapped to "*edss* [*Amino Acid*]."

Compared to MetaMap, it can be said that MaxMatcher fails in finding the correct medical concepts for compounded words. For example MetaMap maps "*chlamydia pneumoniae*" to "*Pneumonias*, *Chlamydial* (*Chlamydial pneumonia*) [*Disease or Syndrome*]" and "*Rickettsia*" to "*Rickettsia* [*Bacterium*]," whereas MaxMatcher identifies no exact matching for either term.

Similar to MetaMap, verbs are frequently not recognized or are incorrectly mapped by MaxMatcher. For example, "*write*" is mapped to "*write* [*Occupation or Discipline*]." In addition, nonmedical locations are also mapped wrongly, as in "*toilet*" which is mapped to "*toilet* [*Therapeutic or Preventive Procedure*]," or "*baths*" which is mapped to "*baths* [*Therapeutic or Preventive Procedure*]." Given that there are different meanings of terms and MaxMatcher only provides one mapping suggestion, it is not surprising that words can be incorrectly mapped.

This is in contrast to MetaMap which provides multiple possible mappings when several semantic types are possible.

Contractions may be wholly unrecognized by Dragon Toolkit. For example, "*don't*" is mapped to "*don* [*Organic Chemical*]." In addition, MaxMatcher also has difficulties in identifying terms referring to medical devices. For instance, the term "*pump*," as in "*I have a baclofen pump that usually controls the pain & spasms*," is not detected by Dragon Toolkit when in contrast MetaMap maps this term correctly as a medical device. Finally, qualitative concepts or adjectives often remain undetected by MaxMatcher. This is unfortunate because such modifiers can be of important medical significance as in "*increasing severe spasms & intense pain all day & night*" where the terms "severe" and "intense" carry much weight.

Discussion and Future Challenges

The assessment of the mappings showed that the NER tools still have problems in processing medical social-media data. In particular, both tools fail in mapping or produce wrong mappings for verbs, personal pronouns, adjectives, and connecting words. Clearly, these terms or at least their meaning and the relationships they infer are relevant for interpreting the content of a sentence and text. Since persons are describing their own personal experiences and observations in medical social-media data, the language they use inevitably includes to a large extent verbs that describe activities of persons and personal pronouns; consequently, it is crucial not to lose the meaning of these personal accounts from patients or healthcare professionals while engaging in automatic processing of blog or forum content. Whereas missing or wrong mappings are not necessarily an algorithmic problem, they might be a problem of the underlying knowledge resource. For example, there is no concept representing the verbs *warn*, *recommend*, and *cause* or the adjectives *horrible*, *miserable*, or *ineffective* in the UMLS, the language resource on which the tested tools are based. This is due to the fact that the terminology has been developed to formalize clinical knowledge, and thus the meanings of verbs or adjectives that are commonly used in medical social media are unfortunately not covered by this terminology.

One must take into consideration that authors of drug ratings, medical procedures, and other social-media content often have no medical training. As a result they often do not use the proper medical terms, but paraphrase these concepts instead. People frustrated with their medical conditions may use a metaphor to refer to their maladies. For example, a cancer patient wrote: "*The beast is going to kill me*." While the metaphor "beast" is not normally considered as synonym for *cancer*, this is what the patient used to refer to his illness. What we can see from patients' everyday usage of language to describe maladies is that the classical synonyms for medical terms that exist in biomedical ontologies may be wholly insufficient for the data considered here. Consideration of metaphors, paraphrases, and other ways that the lay population refers to illnesses and diseases could be a substantial extension

of these ontologies when applying such extraction tools that rely upon biomedical ontologies to mine medical social-media postings. Given that relevant meanings that conform to how patients articulate their symptoms are readily provided in more common vocabularies such as WordNet or CHV, some of the possible ways of improving the quality of mapping tools when processing medical social-media data are to consider additional knowledge resources or in the alternative to exploit a more general terminology. CHV which link everyday words and phrases about health (e.g., *heart attack*) to technical terms or jargon used by healthcare professionals (e.g., *myocardial infarction*) (Zeng and Tse 2006; Zeng et al. 2007) might in fact serve as a template for improving mapping tools for use in medical social media. In fact, the open-source, collaborative CHV initiative[15] tries to develop a CHV for consumer health applications which is intended to complement existing knowledge in the UMLS.

Interestingly enough, in addition to terminology extensions such as those found in the CHV that augment the nomenclature of the UMLS, other improvements can likewise be made to mapping tools that are used in the medical social-media setting: (1) By including general terminological resources such as WordNet, meanings of adjectives could be recognized and considered in the analysis. (2) Another possibility against wrong mappings of medical social-media postings is to enhance the underlying ontology, but this must be done cautiously as it is a very complicated process and could probably lead to problems in processing professional language. (3) A third possibility for an improved mapping or for improved NER is the extension of the mapping algorithm. Aronson et al. showed that it is possible to apply successfully an ensemble of classification systems originally developed to process medical literature on clinical reports (Aronson et al. 2007). Such approaches need to be assessed in the future to develop a better suited mapping tool for medical social media.

In fact, various mapping tools could be used together. For example, there are additional tools for mapping to medical vocabularies available, such as BioLabeler[16]or Open Biomedical Annotator[17] (Jonquet et al. 2009). There are also other NER tools that are using their own underlying terminology and do not refer to existing biomedical ontologies (e.g., OpenCalais,[18] LingPipe,[19] cTAKES[20]). Further, Open Information Extraction techniques (Etzioni et al. 2011) could help in identifying relevant relations as they are expressed by verbs in medical social media. This extraction paradigm learns a general model of how relations are expressed based on unlexicalized features such as part-of-speech tags (e.g., the identification of a verb in the surrounding context) and domain-independent regular expressions

[15]http://samwise1.partners.org/CHV.

[16]http:www.biolabeler.com.

[17]http://bioportal.bioontology.org/annotator.

[18]http://www.opencalais.com/.

[19]http://alias-i.com/lingpipe/.

[20]https://wiki.nci.nih.gov/display/VKC/cTAKES + 2.5.

(e.g., the presence of capitalization and punctuation) (Miller and Pole 2010). By making use of such extraction techniques, it would no longer be necessary to specify in advance the relevant terms or patterns found in social media. This approach may prove more practical given the fact that medical postings are fast-changing, thus making it simply impossible to continuously update the language of social media and their underlying lexical resources manually. Such an approach of open information extraction could help to identify relations expressed by verbs in medical social media which is so far impossible to do using existing mapping tools. To avoid wrong mappings of personal pronouns or connecting words, negative lists could be exploited, i.e., lists that instruct the algorithms not to map the listed words at all.

In sum, there are a number of obstacles that automatic processing tools must overcome in order to make better use of the richness of data found in medical social-media postings. Nevertheless, some of the methods we've analyzed in this chapter augur well for getting closer to meeting such challenges head on. In the end, better data extraction methods for medical blog content insure a healthier patient population and a more efficient healthcare delivery system.

References

Aase L, Goldman D, Gould M, Noseworthy J, Timimi F (2012) Bringing the Social-media Revolution to Health Care. Mayo Foundation for Medical Education & Research, United States, 2012

Altarum Institute (2012) Social-media and Health Care: Applications for Aging and Advanced Illness Populations. Highlights from Duke University's 07–08 May 2012, Durham, U.S., http://www.dukehsac.com/files/2012/09/CHAPI-Social-Media-and-Health-Care-Paper-1.pdf [downloaded October 25, 2012]

Aronson A (2001) Effective mapping of biomedical text to the UMLS metathesaurus: the metamap program. Proc AMIA Symp 2001:17–21

Aronson AR, Bodenreider O, Demner-Fushman D, Fug KW, Lee VK, Mork JG, Névéol A, Peters L, Roger WJ (2007) From indexing the biomedical literature to coding clinical text: experience with MIT and machine learning approaches. ACL, Workshop BioNLP, Prague, Czech Republic

Barla M, Bielikova M (2010) Ordinary web pages as a source for metadata acquisition for open corpus user modeling. In: White B, Isaías P, Andone D (eds.), Proceedings of the IADIS International Conference on WWW/Internet. (Timisoara, Romania). IADIS, 2010, pp 227–233

Boulos MNK, Maramba I, Wheeler S (2006) Wikis, blogs and podcasts: a new generation of web-based tools for virtual collaborative clinical practice and education. BMC Med Educ 6:41

Chapman WW, Fiszman M, Dowling JN, Chapman BE, Rindflesch TC (2004) Identifying respiratory findings in emergency department reports for biosurveillance using metamap. Stud Health Technol Inform 107:487–491

Cohen AM, Hersh WR (2005) A survey of current cork in biomedical text mining. Brief Bioinform 6(1):57–71

Denecke K (2012) An architecture for diversity-aware search for medical web content. Methods Inf Med 51(6):549–556

Denecke K, Dolog P, Smrz P (2012) Making use of social-media data in public health. In: Alain Mille et al (eds) Proceedings of the 21st World wide web conference, WWW 2012, Lyon, France, 16–20 April 2012, pp 243–246

Etzioni O, Fader A, Christensen J, Soderland S (2011) Open information extraction: the second generation, mausam. International joint conference on artificial intelligence, 2011, Barcelona, Catalonia, Spain

Friedman C, Kra P, Rzhetsky A (2002) Two biomedical sublanguages: a description based on the theories of Zellig Harris. J Biomed Inform 35:222–235

Grishman R (1998) Information extraction and speech recognition. In: Proceedings of the broadcast news transcription and understanding workshop, Lansdowne, VA, February 1998

Hillan J (2003) Physician use of patient-centered weblogs and online journals. Clin Med Res 1(4):333–335

Himmel W, Reincke U, Michelmann HW (2008) Using text mining to classify lay requests to a medical expert forum and to prepare semiautomatic answers, SAS global forum, San Antonio, TX

Jonquet C, Shah NH, Musen MA (2009) The open biomedical annotator. Summit on Translat Bioinform 2009:56–60

Kahn CEJ, Rubin DL (2009) Automated semantic indexing of figure captions to improve radiology image retrieval. J Am Med Inform Assoc 16:280–286

Kovic I, Lulic I, Brumini G (2008) Examining the medical blogosphere: an online survey of medical bloggers. J Med Internet Res 10(3):e28

McCray AT (2003) An upper level ontology for the biomedical domain. Comp Funct Genomics 4:80–84

McCray AT, Burgun A, Bodenreider O (2001) Aggregating UMLS semantic types for reducing conceptual complexity. Medinfo 10(1):216–220

Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF (2008) Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008:128–144

Miller EA, Pole A (2010) Diagnosis blog: checking up on health blogs in the blogosphere. Am J Public Health 100(8):1514–1519

Rizzo G, Troncy R (2012) NERD: a framework for unifying named entity recognition and disambiguation web extraction tools. System demonstration at the 13th conference of the European chapter of the association for computational linguistics (EACL'2012), Avignon, France, 23–27 April 2012

Stewart SA, von Maltzahn ME, Raza Abidi SS (2012) Comparing metamap to mgrep as a tool for mapping free text to formal medical lexions. In: Proceedings of the 1st international workshop on knowledge extraction & consolidation from social-media in conjunction with the 11th international semantic web conference (ISWC 2012), Boston, USA, 12 November 2012, pp 63–77

Zeng QT, Tse T (2006) Exploring and developing consumer health vocabularies. J Am Med Inform Assoc 13(1):24–29

Zeng QT, Tse T, Divita G et al (2007) Term Identification methods for consumer health vocabulary development. J Med Internet Res 9(1):e4

Zhou X, Zhang X, Hu X. Dragon toolkit: incorporating auto-learned semantic knowledge into large-scale text retrieval and mining. In: Proceedings of the 19th IEEE international conference on tools with artificial intelligence (ICTAI), Patras, Greece, 29–31 October 2007

http://www.springer.com/978-1-4614-6933-9

Where Humans Meet Machines
Innovative Solutions for Knotty Natural-Language Problems
Neustein, A.; Markowitz, J.A. (Eds.)
2013, XV, 315 p., Hardcover
ISBN: 978-1-4614-6933-9