# Extracting Medical Concepts from Medical Social Media with Clinical NLP Tools: A Qualitative Study

**Kerstin Denecke**
University of Leipzig
Semmelweisstr. 14, Leipzig, Germany
kerstin.denecke@iccas.de

**Abstract**

Medical social-media provides a rich source of information on diagnoses, treatment and experiences. For its automatic analysis, tools need to be available that are able to process this particular data. Since content and language of medical social-media differs from those of general social media and of clinical document, additional methods are necessary in particular to identify medical concepts and relations among them. In this paper, we analyse the quality of two existing tools for extracting clinical terms from natural language that were originally developed for processing clinical documents (cTakes, MetaMap) by applying them on a real-world set of medical blog postings. The results show that medical concepts that are explicitly mentioned in texts can reliably be extracted by those tools also from medical social-media data, but the extraction misses relevant information captured in paraphrase or formulated in common language.

**Keywords:** Information Extraction, Natural Language Processing, Medical Social Media, Sublanguage Analysis

## 1. Introduction

The advances in internet and mobile technologies changed the way how people access, use and share information. Data and experiences 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. Mayo Clinic researchers have opined that social media has begun a process of "revolutionising healthcare" by improving healthcare and quality of life (Aase et al., 2012).

In order to make use of the knowledge captured in this new information source, tools for automatic processing are necessary. Natural language as used in clinical documents or medical social-media is unstructured. In contrast, structured or normalised data is required for automatically analysing the content and to enable further interpretation and processing of the data. Extracting concepts (such as drugs, symptoms, and diagnoses) from clinical narratives constitutes a basic enabling technology to unlock the knowledge within texts and support more advanced reasoning applications such as diagnosis explanation, disease progression modelling, and intelligent analysis of the effectiveness of treatment (Jonnalagadda et al., 2012).

Algorithms and tools are already available for mapping clinical and biomedical documents to concepts of medical terminologies and ontologies (e.g. MedLee, MetaMap (Aronson, 2001), cTakes (Savova et al., 2009)). Once applied to a document they provide for extracted terms concepts of clinical terminologies that can be used to describe the content of a document in a standardised way. Existing tools for concept extraction were designed specifically to process clinical documents, i.e. they are specialised to the linguistic characteristics of these documents. The linguistic characteristics of clinical and biomedical texts have been analysed in painstaking detail by other researchers (Friedman et al., 2002), (Kovic et al., 2008), (Meystre et al., 2008). In contrast, the literary composition of medical social-media data has unfortunately not yet been analysed with the same degree of precision. Clinical texts such as radiology reports contain short, telegraphic phrases resulting in a compact description of facts and observations written in medical terminology. In contrast, medical social media texts can consist of complete, complex sentences (e.g. in blogs and forums) or can be very short without using complete sentences (e.g. Twitter). Consumer health vocabulary, clinical terms and common language is exploited to deliver information. Language in medical social-media differs from language in clinical documents. Table 1. summarises the linguistic characteristics of the two text types.

It is still unclear whether the clinical NLP tools are suited to process medical social-media data given the different language characteristics. This question will be addressed in this paper. We will assess the extraction quality of such tools through a qualitative study. The quality of two named entity recognition tools originally designed for processing clinical texts is compared when they are applied to medical social-media text.

The paper is structured as follows. Section 2. provides an overview on natural language processing in general and concept extraction in particular from clinical documents. Further, corresponding methods and tools will be reviewed. Then, we describe the data set and analysis method that was applied to study the extraction quality for the two tools MetaMap and cTakes that are applied to medical social media (section 3.). Section 4. describes the results of the assessment. In section 5., the results are discussed and observations on the content and linguistic characteristics of medical social-media are summarised. The paper finishes with conclusions and remarks on future work.

## 2. Extracting Information from Texts

In this section, methods and tools from extracting information from unstructured documents in general and from clinical documents in particular will be summarised.

| Text type | Clinical text | Medical Social-Media |
|---|---|---|
| **Sentence structure** | ungrammatical sentences; short, telegraphic phrases (e.g.*Aspirin or Fever*); often without verbs or other relational operators | rather long sentences |
| **Word usage** | word compounds (*high blood pressure*), formed ad hoc; modifiers are related to temporal information (e.g. sudden), evidential information (e.g. rule out, no evidence), severity information (mild, extensive), body location | adjectives; descriptive and narrative words |
| **Spelling** | misspellings; abbreviations, acronyms | abbreviations, misspellings |
| **Language** | mix of Latin and Greek roots with corresponding host language (German, English); domain-specific language | common language, rather than domain-specific language or clinical terminology; host language |
| **Semantic categories of words** | Procedures, Disorders, Anatomy, Concepts and Ideas | Living Beings, Disorders, Chemicals and Drugs, Concept and Ideas |

Table 1: Linguistic characteristics of clinical texts, and medical social-media

## 2.1. Methods for Information Extraction and NER

Information extraction identifies facts or information in texts (Grishman, 1998). An information extraction system is often specialised on a specific domain (e.g. medicine) (Grisham, 2002) and composed of several modules, normally working in a pipeline fashion (Cunningham, 2002). It requires lexical resources, that provide background knowledge and associated terms as well as domain knowledge. In the medical domain, multiple standardised vocabularies and ontologies are available (e.g. UMLS[1], SNOMED CT[2]). This knowledge is exploited by extraction tools to identify meanings in sentences and to identify relevant text snippets given an extraction task. Further, knowledge for interpreting the data is necessary.

Information extraction comprises several tasks, including named entity recognition, coreference resolution, relation extraction and template filling. Named-entity recognition (NER) 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) and is focus of this work. Examples of named entity categories in the medical domain include diseases and illnesses, drugs. More general named entity categories are person names, organizations, or locations. Recognised medical entities can be mapped to concepts of a medical terminology such as UMLS or SNOMED CT to enable a normalised representation of extracted information.

A concept represents a single meaning. Due to the flexibility in language usage, the same meaning can be expressed in different ways, e.g. through a noun, its synonym, an abbreviation etc. Through mapping of terms to concepts of a terminology, texts can be represented semantically and become interpretable for computer algorithms. For example, the UMLS Metathesaurus is organised by concepts: each concept has specific attributes defining its meaning. It is linked to the corresponding concept names in the various source vocabularies.

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

1. a simple lexicon lookup; and
2. extraction patterns that are either manually created or learnt 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 do not require costly human annotators do however require large training corpora to train their underlying models.

For named entity recognition from unstructured texts, several tools exist. Stanford NLP tools[3], Alchemy API[4], LingPipe[5] or OpenCalais[6] are some examples for NLP tools that can be exploited for extracting named entities from unstructured text. However, these systems were mainly designed for processing news articles and often specifically trained on news data sets. They support detection of entities referring to persons, organizations or locations and are not designed for extracting diagnoses, medical conditions or medical procedures.

For these purposes, specialised tools were developed. Existing information extraction systems designed for processing clinical documents or biomedical literature are based on (1) pattern matching techniques such as regular expressions

[1] http://www.nlm.nih.gov/research/umls/

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

[3] http://nlp.stanford.edu/software/CRF-NER.shtml

[4] http://www.alchemyapi.com

[5] http://www.alias-i.com/lingpipe

[6] http://opencalais.com

(e.g. the REgenstrief eXtraction tool (Friedlin and McDonald, 2006)), (2) full or partial parsing (e.g. LifeCode system (Mamlin et al., 2003), (3) a combination of syntactic and semantic analysis (e.g. MedLEE (Friedman et al., 1994) and MetaMap (Aronson, 2001)). They were mainly developed to extract information from textual documents in the electronic health record, among others from chest radiography reports, radiology reports, echocardiogram reports and discharge summaries. Evaluations showed that the current natural language processing tools for clinical narratives are effectively enough for practical use (Friedman et al., 2013).

### 2.2. Clinical NER: cTakes and MetaMap

In the following, two tools that were developed to process clinical text or biomedical literature are described in more depth. These tools are freely available and are used in our qualitative study.

The Apache Clinical Text Analysis and Knowledge Extraction System (**cTAKES**, (Savova et al., 2009)) is an open-source natural language processing system for extracting information from documents. The algorithms were specifically trained to process clinical documents. Among others, the system provides a recognizer that identifies clinical named entities in text using a dictionary-lookup algorithm. Through lexicon-lookup, each named entity is mapped to a concept of a terminology, e.g. the Unified Medical Language System (UMLS). The recognition concentrates on concepts of semantic types: diseases, sign/symptoms, procedures, anatomy and drugs. cTAKES was built using the Apache UIMA Unstructured Information Management Architecture engineering framework and OpenNLP natural language processing toolkit. Its components are specifically trained for the clinical domain out of diverse manually annotated datasets, and create rich linguistic and semantic annotations that can be utilised by clinical decision support systems and clinical research. cTAKES has been used in a variety of use cases in the domain of biomedicine such as phenotype discovery, translational science, pharmacogenomics and pharmacogenetics. In evaluations with clinical notes, the algorithm achieved F-scores from 0.715 to 0.76 (Kipper-Schuler et al., 2008).

The **MetaMap System** (Aronson, 2001) is provided by the National Library of Medicine (NLM). The tool maps natural language text to concepts of the UMLS Metathesaurus. MetaMap uses a knowledge-intensive approach based on symbolic, natural-language processing (NLP) and computational-linguistic techniques. Besides being applied for both IR and data-mining applications, MetaMap is one of the foundations of NLM's Medical Text Indexer (MTI) which is being used for both semi-automatic and fully automatic indexing of biomedical literature at NLM. MetaMap was originally developed to extract information from MEDLINE abstracts, but has been applied to clinical documents as well (e.g. for pathology reports (Schadow and McDonald, 2003), for respiratory findings (Chapman et al., 2004)). 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 and evaluated. The best candidates are organised 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%.

Ironically, while for online news a comparison of several NER tools (e.g., Alchemy API, DBpedia Spotlight, OpenCalais etc.) has already been performed (Rizzo and Troncy (Rizzo and Troncy, 2011)), 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. On the other hand, reliable technologies for analysing the textual content of medical social-media are necessary. Thus, we will analyse the mapping quality for two example clinical NLP tools on medical social-media data.

## 3. Methods

In this section, we describe our study design of a qualitative comparison of the two clinical named entity recognition tools (cTAKES, MetaMap), when they are applied to medical social-media documents. The objective is to clarify whether the tools extract relevant information from social media correctly and to determine which information remains unconsidered. The results of the study are important for the development of social media processing tools, in particular to decide whether existing technology is sufficient or whether and which adaptations are necessary to achieve good analysis results.

### 3.1. Data Set

We applied the two tools to 1) ten texts drawn from "Health Day News"[7] and 2) ten blog postings from "WebMD"[8]. The Health Day news service provides daily health news for both consumers and medical professionals. Content is provided by professional writers. Our data collection concentrated on information provided for consumers. WebMD provides valuable health information, tools for managing health, and support to those who seek information. The blog postings are collected from physician blogs made available through WebMD website, where physicians are writing about topics related to health and medicine. The postings were collected and HTML code was removed from the postings.

### 3.2. Study Design

The mapping results produced by MetaMap and cTakes when applied to the data set were checked manually, sentence by sentence by two persons, a computer scientist specialised in medical informatics and a medical doctor. They assessed different parts of the data set, thus no annotator agreement could be determined. The assessment of the output of the tools concerned the correctness of the extraction,

[7] http://consumer.healthday.com/

[8] http://www.webmd.com

the relevance of the extracted concepts and limitations and possibilities of the extraction tools with respect to processing medical social-media data. More specifically, the annotators had to judge the

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

We identified words that are crucial for understanding the text or sentence which could not be identified by either one of the tools used. Correct and incorrect annotations were counted. Extracted concepts were labeled as *wrong* when they did not represent the actual meaning of the underlying term or even when the extraction was incomplete (e.g. when for the phrase *breast cancer* only the concept referring to *breast* is provided). Further, observations on reasons for errors were collected.

The objective of the assessment is to give insights into the possibilities and limitations of these tools when they are applied to medical social-media data. Unfortunately, the systems use different versions of the UMLS. MetaMap was run with UMLS 2013AB, while cTAKES is distributed with UMLS 2011AB. However, this is not supposed to critically influence the mapping quality in our study.

MetaMap processing was restricted to identify concepts of semantic types that are referring to medical conditions, procedures, medications or anatomy. This restriction was made to achieve comparability with the cTakes results. CTakes only determines concepts referring to these semantic types. More specifically, MetaMap processing was restricted to the semantic types: Therapeutic or Preventive Procedure, Sign or Symptom, Physiologic Function, Pharmacologic Substance, Laboratory or Test Result, Laboratory Procedure, Injury or Poisoning, Disease or Syndrome, Diagnostic Procedure, Daily or Recreational Activity, Clinical Drug, Body System , Body Substance, Body Space or Junction, Body Part, Organ, or Organ Component, Body Location or Region, Behavior, Anatomical Structure, Activity, Acquired Abnormality. Table 2 shows to the cTakes categories the corresponding MetaMap categories as considered in this work.

| **cTakes Category** | **MetaMap Categories** |
|---|---|
| Diseases | Disease or Syndrome; Acquired Abnormality |
| Sign / Symptom | Sign or Symptom; Physiologic Function; Laboratory or Test Result; Injury or Poisoning |
| Procedures | Therapeutic or Preventive Procedure; Laboratory Procedure; Diagnostic Procedure |
| Anatomy | Body System; Body Substance; Body Space or Junction; Body Part, Organ, or Organ Component; Body Location or Region; Anatomical Structure |
| Drugs | Pharmacologic Substance; Clinical Drug |

Table 2: cTakes categories and MetaMap correspondence

## 4. Results

This section describes the evaluation results and observations of the annotators. Precision values determined in the evaluation are listed in Table 3.

cTakes achieved an average precision of 94% for the data set. Annotations of type *DiseaseDisorder, SignSymptom, Drug and Procedure* are correct with around 93%; and *Anatomy* annotations correct with an precision of 98%.

Compared to the cTakes results, MetaMap's results are more often incomplete and wrong. Symptoms are recognized best with an precision of 75.1%, followed by concepts referring to *procedures* with 69% precision. The precision values are significantly lower than those of cTakes.

One problem of MetaMap is that certain phrases such as *breast cancer* are not mapped to a disease or finding. Only the location (e.g. *breast*) is annotated. The annotators were asked to consider such mappings as incorrect.

Table 4 shows the proportion of extracted concepts per category. The tools show clear differences. cTakes extracted in total 1399 concepts. Half of them are referring to diagnoses and signs and symptoms. In contrast, MetaMap extracted 1020 concepts from the same data set and the majority of concepts refer to procedures and medication.

Relevant terms that were not annotated by MetaMap with corresponding concepts are for example: *fever, pregnant, autism, inflammation, cancer, tumor, blood pressure*. The pronoun *his* is mapped to the concept *Histidine*. Further, the verbs *said* and *led* are mapped incorrectly.

Other errors occur in both systems. It could be recognised that named entities referring to job positions, journals, or organisations used in the texts led to wrong or rather misleading annotations in both tools. For example from the phrase *director of the virus hepatitis program* the phrase *virus hepatitis* is annotated as disease occurrence. The term *division* in phrase *a member of the faculty at the division of global health at the University of California* is annotated as *Procedure Mention*.

Anatomical concepts occur sometimes in common language expressions (e.g. *don't have to go hand in hand*). The term *hand* in this phrase is annotated with the category *anatomical site mention* which is correct in general, but in that particular phrases no anatomical concept is meant. Another observation is that the annotation normally concentrates on medical concepts and looses the context. For example, cTakes annotates the phrase *drop in estrogen levels* with a concept referring to *estrogen level*, but the information captured in the complete phrase that this level is dropping is lost by such annotation. Another example is the annotation of the phrase *lack of sleep* where the annotation misses *lack*. Given the fact that cTakes concentrates

on extraction specific medical concepts, it is not surprising that also qualitative judgements such as *It's a very effective treatment* remain unconsidered in the mapping.
Additionally, to the categories that were actually target of the evaluation, cTakes provides annotations of type *Roman Numeral Annotation*. It could be recognized that abbreviations, personal pronouns or measurement units are mapped to concepts of that annotation type (e.g. the abbreviation *CDC* or the personal pronoun *I*). Almost all mappings to that type are wrong. The pronoun *I* is always annotated as Roman Numeral Mention which is a false positive annotation. Interestingly, number expressions such as *300ml* were correctly extracted and annotated.

| Category | MetaMap | cTakes |
|---|---|---|
| **Disease** | 59,6% | 92,9% |
| **Sign, Symptom** | 75,2% | 92,9% |
| **Procedure** | 69,05% | 93,7% |
| **Anatomy** | 54,08% | 98,1% |
| **Drug** | 66,54% | 93,8% |

Table 3: Precision per category for both systems

| Category | MetaMap | cTakes |
|---|---|---|
| **Disease** | 17.5% | 35.7% |
| **Sign, Symptom** | 12.6% | 27.1% |
| **Procedure** | 28.8% | 19.1% |
| **Anatomy** | 19.2% | 15.5% |
| **Drug** | 21.9% | 2.4% |

Table 4: Precision per category for both systems

## 5. Discussion and Conclusions

In this section, we discuss the results of the evaluation and limitations of the study design.

### 5.1. Discussion of the Evaluation Results

In summary, both tools are able to extract concepts from medical social media when medical conditions or procedures are explicitly mentioned and described by nouns. In particular the cTakes mappings are already sufficient to get an overview on the medical content of the posting. For a deeper understanding some additional information need to be provided.
In particular, both tools often 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, cause* or for the adjectives *horrible, miserable, or ineffective* in the UMLS, the language resource on which the tested tools 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 medical 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.
Extraction quality and completeness depends clearly on the content. When the text is dealing with diseases and clearly mentioned symptoms, both tools provide appropriate annotations. However, medical social-media texts discuss sometimes also issues related to nutrition or wellness. In that case, the tools are not performing well. The data set comprised two text types: one personal blog and health news. The blog postings were rather dealing with personal experiences. However, we could not recognise any quality differences in the mapping. When medical concepts are explicitly mentioned using the proper medical terms the tools identify them properly. In the medical social media data, verbs are bearing information, but the tools are not identifying them due to missing background knowledge on their meaning. E.g. the verbs *infected, elevated, transmitted* would be relevant to be determined.
Ambiguity of terms led to errors in both system. For example, the phrase *hand in hand* led to annotations of an anatomical concepts. Avoiding such errors is difficult and requires consideration of the larger context. For both systems, phrases referring to person names, organisations, journals or job positions led to errors. For example, the tools labeled concepts referring to *disease* or *prevention* when processing the phrase *center of disease prevention and control*. However, a correct annotation would determine the complete phrase.

### 5.2. Discussion of the Study Design

The problem of the large number of concepts not extracted with MetaMap originates in the restriction to some selected semantic type. The concept *breast cancer* belongs to the category *Neoplastic Process* which was not selected in the MetaMap settings. The same holds true for other terms where UMLS concepts exist, but our restriction in the semantic types led to missing mappings. We assume, that the mapping quality increases when some additional categories are included. This problem is obviously due to the study design. However, the restriction to certain semantic types is important to reduce mapping errors. MetaMap would otherwise identify too many irrelevant and wrong concepts similar to the extraction of type *Roman Numeral Annotation* in cTakes where almost each mapping was wrong. No information was found which UMLS semantic types are un-

derlying the cTakes categories. Otherwise, exactly the same types would have been included into the MetaMap settings. MetaMap provides mapping candidates and sometimes provides multiple, ranked mappings. We considered in the evaluation the best mapping or the first one in the list when multiple mappings had the same rank. It might be, that the correct mapping was not the first one.

We did not compare the mappings of the two tools directly, i.e. it was not identified to what extent the tools provided the annotation to the same words. Such comparison would be interesting for future analysis. Additional annotations of cTakes were not analysed in depth (e.g. annotation type "Predicate" that annotates verbs).

Given the huge amount of different medical social media sources, the selection of a representative data set for a study as presented in this paper is difficult. A broader spectrum of authors and multiple sources should be considered in future, in particular when improving the tools to ensure generality.

### 5.3. Potential Extensions of Clinical NLP Tools

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 consumer health vocabularies, some of the possible ways of improving the quality of mapping tools when processing medical social-media data is to consider additional knowledge resources or in the alternative to exploit a more general terminology. Consumer health vocabularies (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 consumer health vocabulary initiative 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 named entity recognition is the extension of the mapping algorithm. Aronson et al. (Aronson et al., 2007) showed that it is possible to apply successfully an ensemble of classification systems originally developed to process medical literature on clinical reports. 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 to achieve a more complete extraction and annotation. Our evaluation showed that tools that extract organization or person names could help in avoiding mapping errors. Further, Open Information Extraction techniques 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 (e.g., the presence of capitalization and punctuation). 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 have 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 insures a healthier patient population and a more efficient healthcare delivery system.

We learned from the assessment that medical social media data contains named entities to a large extent referring to person names or organisations. Recognising person and organisation names in advance could help in reducing errors in concept mapping to clinical concepts. The named entities referring to persons and organisations could be filtered out before processing.

## 6. Conclusions

In this paper, we applied two existing clinical NLP tools MetaMap and cTakes to medical social media texts and assessed the mapping quality. Surprisingly, the number of wrong mappings were very low for the cTakes system. However, not all information relevant for an automated analysis and interpretation is made available by the cTakes mappings. Regarding linguistic characteristics of medical social media we learned, that in those texts named entities referring to persons and organisations occur frequently and require additional processing which is so far not realized by clinical NLP tools. In future, we will combine existing clinical mapping tools with general named entity recognition tools and concentrate also on relation extraction among

concept mentions.

## 7. References


Aase, L., Goldman, D., Gould, M., Noseworthy, J., and Timimi, F. (2012). *Bringing the Social-media Revolution to Health Care*. Mayo Clinic Center for Social-media.

Aronson, A. R., Bodenreider, O., Demner-Fushman, D., Fung, K. W., Lee, V. K., Mork, J. G., Neveol, A., Peters, L., and Rogers, W. J. (2007). From indexing the biomedical literature to coding clinical text: experience with MTI and machine learning approaches. In *Biological, translational, and clinical language processing*, pages 105–112, Prague, Czech Republic, June. Association for Computational Linguistics.

Aronson, A. R. (2001). Effective Mapping of Biomedical Text to the UMLs Metathesaurus: The MetaMap program. In *Proceedings of the AMIA 2001*.

Chapman, W., Fiszman, M., Dowling, J., Chapman, B., and Rindflesch, T. (2004). Identifying respiratory findings in emergency department reports for biosurveillance using MetaMap. In *Stud Health Technol Inform.*, pages 487–91.

Cohen, A. and Hersh, W. (2005). A survey of current work in biomedical text mining. *Brief Bioinform*, 6(1):57–71, January.

Cunningham, H. (2002). GATE, a general architecture for text engineering. *Computers and the Humanities*, 36:223–254.

Friedlin, J. and McDonald, C. (2006). A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports. *AMIA Annu Symp Proc.*, pages 269–73.

Friedman, C., Alderson, P., Austin, J., Cimino, J., and Johnson, S. (1994). A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc.*, 1(2):161174.

Friedman, C., Kra, P., and Rzhetsky, A. (2002). Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35:222–235.

Friedman, C., Rindflesch, T. C., and Corn, M. (2013). Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the national library of medicine. *Journal of Biomedical Informatics*, 46(5):765 – 773.

Grisham, R. (2002). Chapter 30: Information extraction. In *Mitkov R: The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Grishman, R. (1998). Information extraction and speech recognition. In *Proceedings of the Broadcast News Transcription and Understanding Workshop, Lansdowne, VA*.

Jonnalagadda, S., Cohen, T., Wu, S., and Gonzalez, G. (2012). Enhancing clinical concept extraction with distributional semantics. *Journal of Biomedical Informatics*, 45(1):129 – 140.

Kahn, C. and Rubin, D. (2009). Automated semantic indexing of figure captions to improve radiology image retrieval. *Journal of the American Medical Informatics Association*, 16:280–286.

Kipper-Schuler, K., Kaggal, V., Masanz, J., Ogren, P., and Savova, G. (2008). System evaluation on a named entity corpus from clinical notes. In *Language Resources and Evaluation Conference, LREC 2008*, pages 3001–7.

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

Mamlin, B., Heinze, D., and McDonald, C. (2003). Automated extraction and normalization of findings from cancer-related free-text radiology reports. *AMIA Annu Symp Proc.*, pages 420–4.

Meystre, S., Savova, G., Kipper-Schuler, K., and Hurdle, J. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook Med Informat*, page 12844.

Rizzo, G. and Troncy, R. (2011). NERD: A framework for evaluating named entity recognition tools in the Web of data. In *ISWC 2011, 10th International Semantic Web Conference, October 23-27, 2011, Bonn, Germany*, Bonn, ALLEMAGNE, 10.

Savova, G., Bethard, S., Styler, W., Martin, J., Palmer, M., Masanz, J., and Ward, W. (2009). Towards temporal relation discovery from the clinical narrative. In *AMIA Annual Symposium Proceedings*, volume 2009, page 568. American Medical Informatics Association, American Medical Informatics Association.

Schadow, G. and McDonald, C. J. (2003). Extracting structured information from free text pathology reports. *AMIA Annu Symp Proc.*, page 584588.

Stewart, S., von Maltzahn, M., and Raza Abidi, S. (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 and Consolidation from Social-media in conjunction with the 11th International Semantic Web Conference (ISWC 2012), Boston, USA, November 12, 2012*, pages 63–77.

Zeng, Q. and Tse, T. (2006). Exploring and developing consumer health vocabularies. *J Am Med Inform Assoc*, 13(1):2429.

Zeng, Q., Tse, T., Divita, G., and et al. (2007). Term identification methods for consumer health vocabulary development. *J Med Internet Res*, 9(1).