iccas

UNIVERSITÄT LEIPZIG
Medizinische Fakultät

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

Denecke, Kerstin[1]

[1] Innovation Center Computer Assisted Surgery Leipzig

INTRODUCTION

Data and experiences on medical treatments and diagnosis are exchanged increasingly via instant messaging, blogs, social networking (e.g. Facebook) or video sharing (e.g. YouTube). In order to make use of the knowledge captured in this new information source, tools for automatic processing are necessary. Algorithms and tools are already available for mapping clinical and biomedical documents to concepts of medical terminologies and ontologies (e.g. MedLee, MetaMap [1], cTakes [2]). 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. It is still unclear whether the clinical NLP tools are suited to process medical social-media data given the different language characteristics. We will assess the extraction quality of MetaMap and cTakes on medical social media data through a qualitative study.

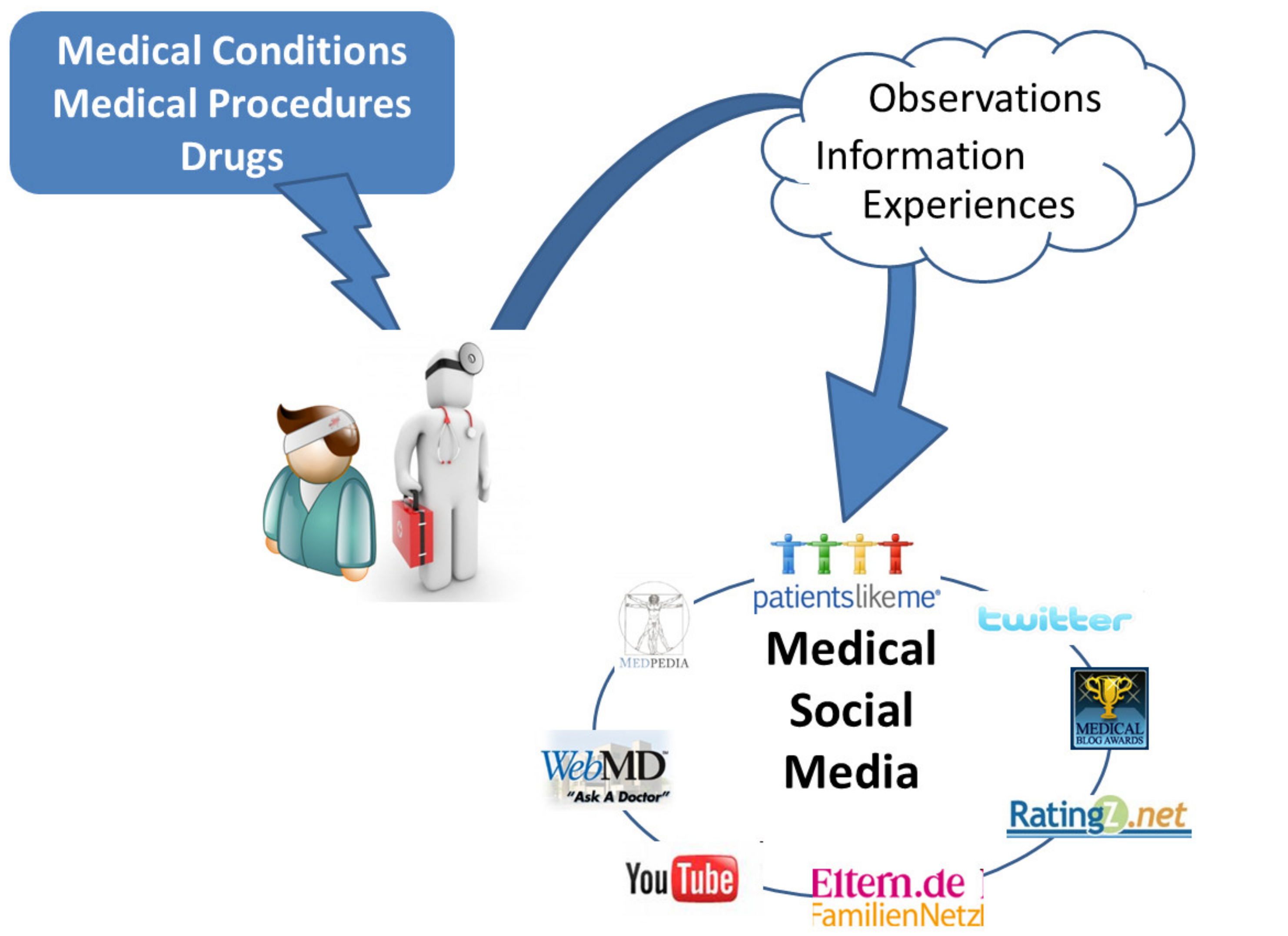

Figure 1 Medical social-media provides a rich source of information on diagnoses, treatment and experiences.

METHODS

cTAKES, and MetaMap are applied to a data set of medical social-media documents (ten texts from ”Health Day News” and ten blog postings from ”WebMD”). 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. Results were manually assessed with respect to 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).

Results

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. 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.

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%

Figure 2:Precision values per NE cateogry

Anatomical concepts occur sometimes in common language expressions (e.g. *don’t have to go hand in hand*) and lead to wrong extractions.

	Clinical texts	Medical Social Media
Sentence structure	Ungrammatical sentences; short, telegraphic phrases, often without verbs or other relational operators	Rather long sentences
Word usage	*Word compounds* formed ad hoc; modifiers are related to temporal information, evidential information severity information, body location	Adjectives, descriptive and narrative words
Spelling	Misspellings, abbreviations, acronyms	Abbreviations, misspellings
Language	Mix of Latin and Greek roots with corresponding host language (e.g. German, English), domain-specific language	Common language rather than domain-speific 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

Figure 3: Linguistic characteristics of clinical texts and medical social media

Conclusions

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 paraphrases or formulated in common language. 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.

[1] Aronson, A. R. (2001). Effective Mapping of Biomedical Text to the UMLs Metathesaurus: The MetaMap program. In *Proceedings of the AMIA 2001*
[2] 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.

Dr. Kerstin Denecke
Semmelweisstr. 14
04103 Leipzig
Tel: +49(0)341/97-12002
Fax: +49(0)341/97-12009
Email: kerstin.denecke@medizin.uni-leipzig.de

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

UNTERNEHMEN REGION
Die BMBF-Innovationsinitiative
Neue Länder