Extracting process graphs from medical text data

An approach towards a systematic framework to extract and mine medical sequential processes descriptions from large text sources.

Andreas Niekler [1,*]

[1]University of Leipzig, Department of Computer Science, Natural Language Group, Augustusplatz 10, 04109 Leipzig

ABSTRACT

In this work a natural language processing workflow to extract sequential activities from large collections of medical text documents is developed. The approach utilizes graph structures to process, link and assess activities found in the documents.

*** Contact:** aniekler@informatik.uni-leipzig.de

relation extraction, natural language processing, graph processing, process models

1 INTRODUCTION

Medical publications, surgical procedure reports or medical records typically contain procedural descriptions. For example, all activities included in a medical study must be documented for reproducibility purposes, in surgical reports a stepwise description of included procedures is documented and in medical records a history of medical treatment is listed. Additionally, related studies or reports describe alike activities with some alterations or rely on preceding activities that may be described in other documents. Consider for example the preparation steps before DNA could be sequenced which are described in scientific papers. They are often the same but need to be documented for each study. Such redundant activity descriptions can be found amongst many documents describing research within the same domain or field of research. Nevertheless, differences amongst the activities in related documents also exist. A complete overview of activities from a defined document collection provides an easy insight to workflows and paradigms within a domain or field of study.Thus, finding this link between the documents and aligning the activities w.r.t. redundant activities helps to structure and analyze procedural knowledge from topical- or domain-related medical texts. For example, early stages or parts of a larger process might be documented separately to other parts or later stages.

In this work a general natural language processing workflow to extract sequential activities from large collections of medical text documents is developed. A graph-based data structure is introduced to merge extracted sequences which contain similar activities in order to build a global graph on procedures which are described in documents on similar topics or tasks.

2 TEXT MINING METHODOLOGY FOR PROCESS EXTRACTION

In this section we describe a methodology which extracts and links activities from medical text documents. The described system follows a sequence of procedures in order to create an activity graph as a result. First, the text sources have to be processed in order to access the entity items in the text. Different entities in a sentence are related and form an expressed activity. Therefore, the extraction of valid relations that form activities is introduced to the text processing step. The second step in our proposed methodology is the creation of a directed graph structure which can be further used for the representation of the activities contained within a text collection.

2.1 Text processing and classification for activity extraction

The text sources must be separated into sentences and tokens first by using state of the art tools. Additionally, POS-Tagging must be applied to the text sources. To extract the procedural knowledge from the texts, named entity recognition (NER) is required as a pre-processing step. In the separated and preprocessed sentences multiple entities may form an activity. Consider the sentence "Real-time_JJ PCR_NNP was_VBD done_VBN using_VBG the_DT fluorescent-labelled_JJ oligonucleotide_NN probes_NNS". "Real-time PCR" and "fluorescent-labelled oligonucleotide probes" are the identified entities which form the activity "done". This activity can be part of a chain of activities throughout multiple documents.

The characteristics of activities or relations between entities change within different domains or described procedures. Thus, the process for identifying and connecting entities to activities within the sentences should not be fixed or static. To answer this fact the identification of relations or activities is defined as classification task using a Support Vector Machine (SVM) along with word- and POS-Tag-level features GuoDong et al. [2005]. If a sentence contains an entity E_1 and E_2 the two words before E_1, the two words after E_2 and all words between E_1 and E_2 are extracted as features. Furthermore, the POS-tags of the extracted words are used as features for the SVM.

Before the training process is applied the user must define the type and the form of the desired relation. On the basis of this definition training examples are collected from the data. For this purpose an active learning procedure is introduced where the user iteratively collects training data with the support of an automatic classification. An initial search for sentences that include a minimum of entities and verbs that indicate an activity is conducted. The set of matching sentences which contain this custom pattern is presented to the user. Correct entities are selected from the proposed sentences along with the definition whether there is a relation between them or not. The features are extracted automatically and the set of positive and negative examples is used to train an initial SVM model. The trained model is used to identify additional examples in the data. The user judges on those examples and with every batch of new examples the classifier can be refined. If the training quality of the SVM does not change with new examples a final model is trained and applied to all documents. The result is a set of sentences from a document collection where each sentence contains an activity or valid relation between entities.

2.2 Process graphs for activity representation and processing

In the next processing step a data structure is constructed on the basis of the set of activities that were identified by the classification process. For each activity the two entities E_1 and E_2, the Verb V (past participle) between them, a document identifier and a sentence identifier are stored. A graph structure A, a directed graph, is introduced where all identified activities are represented as vertices. All vertices that build a sequence of activities within a document are connected with directed edges, e.g. consecutive activities will be connected as a chain of activities within the graph structure. This procedure creates a chain of connected vertices for every document represented in A. The main target for the further processing of A is the linking of different activity chains from multiple documents. This will produce a graph structure which represents networks of activities that supplement each other. In A the connected components can be understood as a summary of activities which come from, or lead to, similar activities. For example, multiple surgical reports contain many redundant descriptions for a certain type of surgical procedure. In some cases there might have been complications and the surgeon had to react on those. A graph which merges different sequential activities from different documents should introduce a cycle of activities describing such additional complications. In a later review of the graph such cycles represents differences from standard procedure described in the document collection.

To detect similar relations throughout different documents a similarity operation $sim(R_{D_1}, R_{D_2})$ is defined. This similarity operation can be constructed on the basis of word level similarity or semantic similarity. With a preprocessing of the corpus like word2vec or a co-occurrence analysis each of the relation components can be augmented by semantic vectors representing the associated vocabulary, e.g. the semantic embedding Mikolov et al. [2013], Bordag [2008]. This allows to compare entities semantically and conceptual similarities between entities can be used to find alike relations. Note, that the similarity function is an exchangable component of the described information extraction approach. The similarity between all activities is calculated for E_1, E_2 and V separately which results in three different similarity matrices which are transformed to adjacency matrices by applying a threshold to the similarity values. It is also imaginable to set three different thresholds for each single similarity matrix or to weight the matrices for further processing. All resulting adjacency matrices are multiplicated element-wise in order to create a single adjacency matrix S of similar activities, e.g. two activities where E_1, E_2 and V are similar between the two activities are represented by the value of 1 in the final matrix.

In the following step the activities considered to be similar are collapsed using the adjacency matrix S resulting in a graph A'. Starting from graph A all edges from similar vertices are taken over to a single vertex and all vertices where the edges where taken from are deleted. That means similar vertices are collapsed to a single vertice and the associated ingoing and outgoing edges of those relations are merged. The resulting graph connects the sequences of different documents where similar activities build single vertices with more than one incoming or outgoing edge ($d_G^+(v) > 1$ or $d_G^-(v) > 1$). Activities with this property are identified more frequently than other activities in the data and thus are of some importance for the overall activity summarization. In summary, it can be said, A' is an unconnected graph where a set N of connected components can be identified. This set represents different graphs where the interaction and coherence of related processes, described in different documents, is encoded.

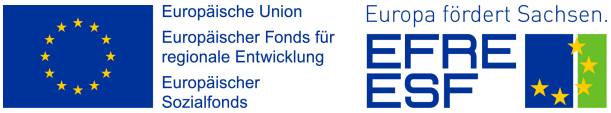
2.3 Future work

In order to quantitatively judge on the quality of the extraction process an evaluation dataset and evaluation strategy needs to be developed as prequisite for future work. More research on suitable similarity functions for relations which can also handle semantic similarities will optimize the quality of the graph merging process. Future work will also include the adoption of domain knowledge from knowledge graphs. Those has been described as very helpful resources in order to adopt to a domain in Roberts and Harabagiu [2011]. The links and dependencies between entities and their possible representations in the data can be encoded in those data structures by domain experts. This will add supervision and control to the graph creation process and thus allows for a higher precision of the graph. Additionally, anaphora resolution can be modeled with knowledge bases to connect graph structures where the relations represent processes which produce other entities as results. Such edges can't be established with character or semantic comparison of the relations. In the moment a connection can only be established if the producing process is encoded within a single document. Furthermore, the introduction of manual corrections steps to refine the graph and the optimization of the relation extraction classification may also optimize the quality.

ACKNOWLEDGEMENT

ExB Labs GmbH kindly helped to compile and preprocess the corpus for this work.

Funding: The project is funded by European Regional Development Fund (ERDF/EFRE) and the European Social Fund (ESF).



REFERENCES

Stefan Bordag. A comparison of co-occurrence and similarity measures as simulations of context. In *Computational Linguistics and Intelligent Text Processing*, pages 52–63. Springer, 2008.

Zhou GuoDong, Su Jian, Zhang Jie, and Zhang Min. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 427–434, 2005.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

Kirk Roberts and Sanda M Harabagiu. A flexible framework for deriving assertions from electronic medical records. *Journal of the American Medical Informatics Association*, 18(5):568–573, 2011.