Semi-automatic Terminology Generation for Information Extraction from German Chest X-ray Reports

Jonathan KREBSa,[[1]](#footnote-1), Hamo COROVICb, Georg DIETRICHa, Max ERTLc, Georg FETTEc, Mathias KASPARc, Markus KRUG a, Stefan STOERKc, Frank PUPPEa

a Würzburg University, Chair of Computer Science 6; Germany

b Bamberg Hospital, Department of Radiology, Germany

c Würzburg University and University Hospital, CHFC, Germany

**Abstract.** Extraction of structured data from textual reports is an important subtask for building medical data warehouses for research and care. Many medical and most radiology reports are written in a telegraphic style with a concatenation of noun phrases describing the presence or absence of findings. Therefore a lexico-syntactical approach is promising, where key terms and their relations are recognized and mapped on a predefined standard terminology (ontology). We propose a two-phase algorithm for terminology matching: In the first pass, a local terminology for recognition is derived as close as possible to the terms used in the radiology reports. In the second pass, the local terminology is mapped to a standard terminology. In this paper, we report on an algorithm for the first step of semi-automatic generation of the local terminology and evaluate the algorithm with radiology reports of chest X-ray examinations from Würzburg university hospital. With an effort of about 20 hours work of a radiologist as domain expert and 10 hours for meetings, a local terminology with about 250 attributes and various value patterns was built. In an evaluation with 100 randomly chosen reports it achieved an F1-Score of about 95% for information extraction.

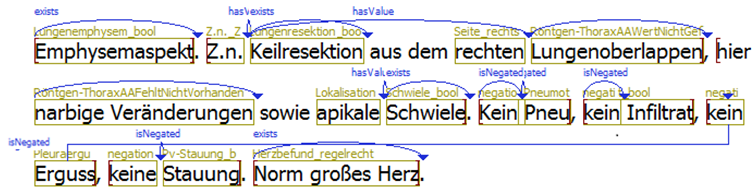
**Keywords.** Information Extraction, Terminology Generation, Data Warehouse, Radiology Reports.

# Introduction

Developing information extraction (IE) applications from medical text documents, like reports and discharge letters, is a laborious task and requires adapting IE-tools to the structure and terminology (ontology) of the respective domain. Although standard terminologies like ICD, SNOMED, LOINC, RadLex etc. exist, those terms are often not easy to match to the terms used in text documents, written or dictated by physicians. In addition, IE of languages other than English faces many challenges [1], e.g. the standard terminologies may be partially or not at all translated. As for RadLex, a translation into German is announced for 2017 by the German X-ray Society (Deutsche Röntgengesellschaft)[[2]](#footnote-2). Although the German X-ray Society offers templates for structured reporting[[3]](#footnote-3), they are often not used by physicians, since dictating is a faster way for documentation. For the same reason, physicians often dictate noun phrases in a telegraphic style instead of using grammatically correct sentences. Similar observations can be made in other domains like sonography, pathology etc.

Our goal is to support the process of building an IE terminology from a sample of reports, since this is the most time consuming part in the development of IE applications. Examples of IE systems from German reports include radiology reports [2], German Patient Records [3], transthoracic echocardiography reports [4], and lung function tests [5]. Building large terminologies usable in an IE pipeline is very time-consuming. Such pipelines are often implemented using natural language processing frameworks that offer many useful components, e.g. GATE[[4]](#footnote-4) and UIMA[[5]](#footnote-5). Some well-known systems for clinical information extraction are HITEx ([6]; based on GATE) and Apache cTAKES ([7]; based on UIMA). We used a special IE pipeline similar to [4] based on UIMA consisting of the following main steps for reports:

1. Anonymization (if necessary).
2. Segmentation of documents in sections and section classification (if necessary).
3. For each section, information extraction with section specific terminologies:
   1. Segmentation.
   2. Shallow parsing.
   3. Extraction of clinical concepts (attribute-value-pairs) with IE terminology.
   4. Special processors for detecting negations, time dependencies etc.
   5. Disambiguation of clinical concepts using context information.
   6. Postprocessing operations e.g. for information aggregation.



**Figure 1.** Example for a chest X-ray report (in German) with inferred annotations.

# Methods

Our goal is not to develop a new IE-algorithm, but a semi-automatic terminology generation algorithm with a set of reports as input and a list of attribute-value pairs with synonyms and regular expressions as output. This knowledge is used by the IE pipeline described above to generate annotations as attribute-value pairs (see Fig. 1).

The first step is anonymization using a two-pass algorithm eliminating known named entities and using heuristics to detect names, addresses, phone-numbers etc. Since many phrases are stereotypical, the second step is to segment a report in phrases and aggregate phrases from different reports into one document without phrase duplication. This step uses mainly punctuation marks as separators with special treatment of abbreviations, enumerations (like 4. in Fig. 2), and numbers. The main part of the terminology generation algorithm operates on the aggregated document (see Fig. 2).

1. Herz normal groß ((152)).
2. Kein Infiltrat ((141)).
3. Aortensklerose ((103)).
4. Normaler Herz-, Mediastinal- und Lungenbefund ((49)).
5. Aortenelongation ((43)).

6. Aorta elongiert und sklerosiert ((2)).

7. Postoperative Veränderungen und Minderbelüftungen im linken Mittelfeld ((1)).

**Figure 2.** Excerpt of an aggregated document with some frequent (1-5) and some rare phrases (6-7) from 3000 chest X-ray reports (in German). The numbers in double parentheses represent the frequency of the phrases in the original reports.

After anonymization and segmentation the main steps of the semi-automatic terminology generation algorithm are:

1. Candidate terminology generation.
   1. Extract nouns from phrases and aggregate them by stemming and edit distance.
   2. Suggest an attribute type (e.g. Boolean, choice or number) for each noun.
   3. Filter nouns which are unlikely to be an attribute.
   4. Expand nouns occurring within enumerations (like 4 in Fig. 2).
   5. Display all relevant nouns with additional information in a terminology table.
2. Candidate terminology validation and elaboration (manual step by domain specialist, usably by making simple entries in the generated terminology table).
   1. Mark irrelevant nouns
   2. Group equivalent nouns.
   3. Add synonyms and regular expressions for relevant nouns.
   4. Check and modify the types of the attributes if necessary.
   5. Mark values for laterality, location, degree of severity, etc. for each attribute
   6. Define value templates for categories, e.g. for laterality: right, left, both sides.
3. Terminology update: The program processes the notes of step 2 and generates an updated terminology table. Steps 2 and 3 may be iterated several times.
4. Terminology application: The table is translated into a data structure suitable for the IE, loaded into the tool ATHEN[[6]](#footnote-6) (Annotation and Text Highlighting Environment) and applied. ATHEN highlights all extracted entities, see Figure 1.
5. Reference standard definition: The domain specialist defines a reference standard of correct relations in ATHEN de novo or based on the suggestions of the IE when applying the terminology.
6. Evaluation with previously unseen documents. If necessary, steps 2-6 are repeated.

# Results

We measured both, the precision and recall of the IE result and the time spent by the physician for terminology development. The physician (HC) spent about 20 hours of work mainly by editing and correcting documents in MS-Excel generated by the terminology extraction algorithm described above. He spent additional 10 hours for meetings with the computer science group (including the time spent for the evaluation). The computing time on a standard workstation took just a few minutes for steps 1 and 4.

The terminology was built based on a training corpus of 3000 chest X-ray reports from Würzburg university hospital. Its final version consists of 258 attributes with the following value categories: negation, laterality (right, left, both sides; relevant for 137 of the 258 attributes), location (38 possible values; relevant for 47 attributes), degree of severity (5 values; relevant for 131 attributes), condition-after ("Zustand nach"; relevant for 67 attributes), and progression note (3 values; relevant for 73 attributes).

With this terminology fed into the IE pipeline we performed an evaluation with 100 randomly chosen chest X-ray documents that were not already included in the training set. The manually defined reference standard contained 735 attribute-value pairs for the 352 segments (roughly half of them were negated) within the 100 documents. Each value category for an attribute (existence/negation, laterality, location, severity, condition-after, progression) was counted separately, if the report mentioned a value for the attribute in the respective category. In addition, since secondary findings are often mentioned in chest X-ray reports, we annotated the core attributes (i.e., the clinically most important attributes) and measured their performance separately. Further on, we measured the recognition rate restricted on the positive findings (not counting negations). The results are shown in Table 1 with an F1 value of 95.1% for all attributes. The core attributes had a precision of 100% in the evaluation sample and an F1 value of 98.9%. If only non-negated, i.e. pathological, attributes are considered, F1 values drop from 95.1% to 90.8% for all attributes and remain constant for the core attributes, since negations are recognized quite well and non-pathological statements have a lower variance than the non-core pathological statements. The accuracy on document level (number of documents with all attribute-value extractions correct) is 70%.

**Table 1.** Evaluation results with 100 previously unseen chest X-ray documents (TP = True Positive, FP = False Positive, FN = False Negative) on attribute level (# = number of attributes)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Types of attributes included** | **#** | **TP** | **FP** | **FN** | **Precision** | **Recall** | **F1** |
| All attributes | 258 | 735 | 4 | 65 | 99.4% | 91.1% | 95.1% |
| All attributes without negation | 258 | 278 | 3 | 53 | 98.8% | 83.9% | 90.8% |
| Core attributes | 119 | 351 | 0 | 8 | 100% | 97.8% | 98.9% |
| Core attributes without negation | 119 | 130 | 0 | 3 | 100% | 97.7% | 98.9% |

An error analysis for the 4 FP and 69 FN attributes in Table 1 revealed the following error categories (with frequencies):

* Missing regular expressions for an existing attribute or value: 25
* Segmentation errors: 23
* Missing attributes or values: 12
* Misspellings in report: 4
* Other: 5

Missing regular expressions are the most common error type. In general, it requires little effort for correction. Although it is difficult to cover all variations for a specific attribute, a few iterations would reduce the quantity of this error type considerably. A similar argument holds for missing attributes or values. Segmentation errors are more difficult to correct. Abbreviations ending with "." were a frequent source of error if they finish their segment needing a special disambiguation. Other segmentation errors occurred if an attribute-value was listed in a different segment than the attribute itself.

# Discussion

From a practical point of view, the results of this rather efficient approach to building hospital specific information extraction models for particular reports are good enough for use in a data warehouse, since in particular the core attributes were found with high precision and recall. The evaluation results are comparable to those of other publication (e.g. [2-6]), but were achieved with a much lower manual effort by the domain expert. To the best of our knowledge, the effort for knowledge engineering is not reported in these and other respective publications. We did not investigate machine learning approaches offering also a potential reduction of manual work. Our approach supports such IE-approaches as well, since they need a gold standard to generalize from and the approach presented is useful to reduce the effort of defining or adapting the gold standard for a particular domain.

# Summary and future Work

An efficient method for building an IE terminology for a specific domain was presented. The next step is to map this local terminology to a standard terminology like Radlex, if it is translated in German. We plan to apply this method for other domains (e.g. MRT, CT, etc.) in order to populate a data warehouse with structured information from unstructured reports. Further on we will investigate how well such local termi­no­lo­gies can be transferred from one hospital to another with manual adaptions and/or machine learning techniques and assess the efforts required therefore.

# Conflict of Interest

The authors state that they have no conflict of interests.

References

1. J. Starlinger, M. Kittner, O. Blankenstein, U. Leser. How to Improve Information Extraction from German Medical Notes, it - Information Technology 58 (10/2016).
2. C. Bretschneider, S. Zillner, M. Hammon. Identifying pathological findings in german radiology reports using a syntacto-semantic parsing approach.In: Proc. Workshop on Biomedical Natural Language Processing. 2013. p. 27–35. <http://www.aclweb.org/anthology/W13-1904>.
3. H.-U. Krieger, C. Spurk, H. Uszkoreit, F. Xu, Y. Zhang, F. Mueller, and T. Tolxdorff. Information extraction from german patient records via hybrid parsing and relation extraction strategies. In LREC, pages 2043-2048, 2014.
4. M. Toepfer, H. Corovic, G. Fette, P. Kluegl, S. Stoerk, and F. Puppe. Fine-grained information extraction from german transthoracic echocardiography reports. BMC medical informatics and decision making, 15(1):1, 2015.
5. M. Toepfer, D. Schmidt, G. Dietrich, M. Ertl, G. Fette, M. Kaspar, S. Störk, F. Puppe, Extraktion von Lungenfunktionsparametern aus Arztbriefen, GMDS Jahrestagung 2015,
6. Q. Zeng , S. Goryachev, S. Weiss, M. Sordo, S. Murphy, R. Lazarus. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inf Decis Making. 2006;6:30.
7. G. Savova, J. Masanz, P. Ogren, J. Zheng, S. Sohn, K. Kipper-Schuler et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17(5):507–13. 938.

1. Corresponding Author, Jonathan Krebs, Würzburg University, Institute for Informatics, Am Hubland, 97074 Würzburg, Germany; E-mail: jonathan.krebs@uni-wuerzburg.de. [↑](#footnote-ref-1)
2. http://www.befundung.drg.de/de-DE/2910/deutsches-radlex (visited 18.3.2017) [↑](#footnote-ref-2)
3. http://www.befundung.drg.de/de-DE/2909/befundvorlagen (visited 18.3.2017) [↑](#footnote-ref-3)
4. https://gate.ac.uk/ [↑](#footnote-ref-4)
5. https://uima.apache.org/ [↑](#footnote-ref-5)
6. https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/Athen [↑](#footnote-ref-6)