Automated Classification of Selected Data Elements from Free-text Diagnostic Reports for Clinical Research*

Martin Löpprich[1]; Felix Krauss[2]**; Matthias Ganzinger[1]; Karsten Senghas[1]; Stefan Riezler[2,3]; Petra Knaup[1]**

[1]Institute of Medical Biometry and Informatics, Heidelberg University, Heidelberg, Germany;
[2]Department of Computational Linguistics, Heidelberg University, Heidelberg, Germany;
[3]Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Heidelberg, Germany

Keywords
Medical informatics, medical writing, natural language processing, classification, multiple myeloma

Summary
Objectives: In the Multiple Myeloma clinical registry at Heidelberg University Hospital, most data are extracted from discharge letters. Our aim was to analyze if it is possible to make the manual documentation process more efficient by using methods of natural language processing for multiclass classification of free-text diagnostic reports to automatically document the diagnosis and state of disease of myeloma patients. The first objective was to create a corpus consisting of free-text diagnosis paragraphs of patients with multiple myeloma from German diagnostic reports, and its manual annotation of relevant data elements by documentation specialists. The second objective was to construct and evaluate a framework using different NLP methods to enable automatic multiclass classification of relevant data elements from free-text diagnostic reports.
Methods: The main diagnoses paragraph was extracted from the clinical report of one third randomly selected patients of the multiple myeloma research database from Heidelberg University Hospital (in total 737 selected patients). An EDC system was setup and two data entry specialists performed independently a manual documentation of at least nine specific data elements for multiple myeloma characterization. Both data entries were compared and assessed by a third specialist and an annotated text corpus was created. A framework was constructed, consisting of a self-developed package to split multiple diagnosis sequences into several subsequences, four different preprocessing steps to normalize the input data and two classifiers: a maximum entropy classifier (MEC) and a support vector machine (SVM). In total 15 different pipelines were examined and assessed by a ten-fold cross-validation, reiterated 100 times. For quality indication the average error rate and the average F1-score were conducted. For significance testing the approximate randomization test was used.
Results: The created annotated corpus consists of 737 different diagnoses paragraphs with a total number of 865 coded diagnosis. The dataset is publicly available in the supplementary online files for training and testing of further NLP methods. Both classifiers showed low average error rates (MEC: 1.05; SVM: 0.84) and high F1-scores (MEC: 0.89; SVM: 0.92). However the results varied widely depending on the classified data element. Preprocessing methods increased this effect and had significant impact on the classification, both positive and negative. The automatic diagnosis splitter increased the average error rate significantly, even if the F1-score decreased only slightly.
Conclusions: The low average error rates and high average F1-scores of each pipeline demonstrate the suitability of the investigated NPL methods. However, it was also shown that there is no best practice for an automatic classification of data elements from free-text diagnostic reports.

Correspondence to:
Martin Löpprich
Heidelberg University
Institute of Medical Biometry and Informatics
Im Neuenheimer Feld 305
69120 Heidelberg
Germany
E-mail: martin.loepprich@med.uni-heidelberg.de

* Supplementary material published on our website http://dx.doi.org/10.3414/ME15-02-0019
** These authors contributed equally to this work

Methods Inf Med 2016;0: ■–■
http://dx.doi.org/10.3414/ME15-02-0019
received: December 15, 2015
accepted in revised form: April 25, 2016
epub ahead of print: ■■■

Fundings
The Multiple Myeloma disease registry has been funded by the Dietmar-Hopp-Stiftung, Walldorf, Germany. CLIOMMICS is funded by the German Ministry of Education and Research within the e:Med initiative. Grant ID: 01ZX1309A.

1. Introduction

In 1964 Yoder et al. claimed clinical data "must satisfy the dual requirements of providing the physician with information which he needs for the care and management of his patient, and at the same time, of supplying information needed by scientists for research purposes" [1]. Since then, it is a challenging task for medical informatics to support dual use of data for patient care and research. Well known approaches use routine data for decision-support systems [2, 3], quality management [4], the detection of epidemics [5] or result improvement of randomized controlled trials [6]. Nevertheless, clinical data are still most often collected, managed and stored multiple times in separate documentation systems for routine care and research, even if data elements overlap [7]. As a result, the same clinical information is entered into an electronic patient record (EPR) within the hospital information systems (HIS) for routine patient care, and again for clinical research purposes into case report forms (CRFs) of dedicated electronic data capture (EDC) systems.

To overcome multiple data handling, the single source approach, has been propagated [8–10]. The idea is to collect patient data for both, routine and research, within the HIS and to export the data into the research EDC database [9]. The benefits of single source are obvious, starting from reducing the documentation burden for clinicians, nurses, and researchers [11], over improving patient recruitment for clinical studies [12, 13], to supporting health assessments of controlled trials [14].

Since HIS primarily focus on supporting patient care, some data are stored in free-text to fulfill the requirements of flexible documentation in clinical processes. Such unstructured information in the HIS cannot be directly used for research without extracting the data required for research. Typically, the process of extracting data elements from free-text is carried out manually by documentation specialists who open the document in the EPR, select relevant information and re-enter them in a structured format into the research EDC. A task that is tedious and prone to transcription errors. An automated multiclass classification of free-text clinical reports by advanced techniques of natural language processing (NLP) seems attractive and may help reduce the time and cost spend by the manual documentation process [15, 16]. Different studies have been published on methods applied to clinical reports in cancer [17], lung cancer [18], breast cancer [19], neuroradiology [20] or imaging reports [21] with varying, but consistent positive results. Making data from free-text documents available for research in a structured form is a challenging task since routine data often consist of abbreviations, acronyms, spell-errors and inconsistently used formatting, punctuations and enumerations [22].

2. Objectives

In the context of the disease registry for patients with multiple myeloma (a cancer of plasma cells) at Heidelberg University Hospital, most data are extracted from discharge letters. Due to the complexity and long duration of the treatment, it would be efficient to support the extraction process by automatic methods. Another research project, for which the data extracted will be used as a data source is the systems medicine project "Clinically-applicable, omics-based assessment of survival, side effects, and targets in multiple myeloma" [23]. Our aim was to use NLP methods for multiclass classification of free-text diagnostic reports to automatically document the diagnosis and state of disease of myeloma patients. Because multiple myeloma is a rare disease with approximately 1.3% of all new cancer cases in Germany [24], and annotated corpora and NLP tools are mostly available in English language, our first objective was to create a corpus consisting of free-text diagnosis paragraphs of patients with multiple myeloma from German diagnostic reports, and its manual annotation of relevant data elements by documentation specialists. The text corpus can be used to train and test NLP methods in the context of automated classification.

The second objective was to construct and evaluate a framework using different NLP methods to enable automatic multiclass classification of relevant data elements from free-text diagnostic reports.

3. Methods

3.1 Creation of an Annotated Text Corpus

In total, one third, or rather 737 patients treated in the Section of Multiple Myeloma at Heidelberg University Hospital in the Department of Hematology, Oncology, and Rheumatology and at the National Center for Tumor Diseases (NCT) Heidelberg, were randomly selected from the multiple myeloma research database. From each patient one clinical report, signed by a senior physician, was randomly picked and the paragraph with main diagnoses was extracted. The main diagnoses paragraph summarizes current and previous diagnoses and characterizes the state of disease. It is mostly written in German, but may also include English or Latin expressions. Each paragraph may contain multiple diagnoses or information relevant in the research topic of multiple myeloma, or none at all. No changes of the main diagnoses paragraph were made through the acquisition process, besides a random alternation of all dates in order to ensure anonymity of patient data.

Further, REDCap (Research Electronic Data Capture) was used as a web-based database tool to conduct the study [25]. A CRF was created that consists of a textbox, used for the extracted diagnoses paragraph, and several specific data elements for multiple myeloma characterization. The manual documentation task was performed by two Medical Informaticians, who had experience in data entry of myeloma related information for several years. Both conducted the documentation independent from each other and used only data available from the diagnoses paragraph. In rare cases with multiple diagnoses and partly missing data, the data entry specialists might have gained additional information through background knowledge on multiple myeloma.

Data quality and accuracy of the manual documentation was ensured by pre-coded data elements (e.g., drop-down menus) and conditional logical statements to detect inconsistency.

The results of the manual documentation task from the first data entry person (E1) and the second data entry person (E2)

were combined by a third Medical Informatician to the final annotated text corpus (ATC). Cases, with no consensus between E1 and E2 and those with consensus but disagreement were reviewed and clearly decided together.

3.2 Construction of a Framework for NLP

The framework was constructed using open source software. It consists of the following parts, applying different NLP methods:

- A basic preprocessing step, comprising tokenizing, the removing of uppercases and of stop words.
- A self-developed package that splits an input sequence into several subsequences if multiple diagnosis are contained.
- Several specific preprocessing steps to enable an automatic classification, to normalize the input data or to supply additional features for improved classification performance.
- Two different classifiers from the Machine Learning for Language Toolkit (MALLET) [26]: a maximum entropy classifier (MEC) and a support vector machine (SVM).

The multi-diagnosis splitting is done by checking for diagnosis dates inside the input sequence and splitting at the next tab stop. Tokenization was executed using OpenNLP [27] in combination with a special model [28] that was trained on FRAMED, a German language clinical text corpus [29]. In addition, all tokens were lowercased and stop words were removed. Further orthographic normalization, e.g., the replacement of the German vowel 'ä' to 'ae', was not applied, since literature search showed no clear advantage of such normalization. The resulting bag-of-words feature set served as baseline for each classification. Since the available free-text diagnosis paragraphs contain a lot of abbreviations, a list of the most frequent ones was manually created and fed to a self-developed module for abbreviation resolution. Because distinction of sequences with different meanings depending on the context is problematic when using only the bag-of-words feature set, regular expressions were applied that detect some relevant cases with a flexibility to some extent.

The spelling correction tool Hunspell was applied to detect and correct spelling errors [30]. Hunspell is able to make spelling suggestions based on one or more provided lexicons. An original German lexicon was extended with medical terms and an English lexicon was added. To broaden the coverage further, a self-developed module was deployed to check for entries in the online database of the Unified Medical Language System (UMLS). An algorithm was used to select the best suggestion based on either corpus frequency information or the lowest editing distance (Levenshtein distance). Finally, an OpenNLP NP-Chunker was integrated into the system to provide detected noun phrases as additional features for the classification. Since Part-of-Speech (POS) tagging is required for a successful application of NP-Chunking, the OpenNLP POS tagger was used in combination with another freely available FRAMED model. Furthermore, the tagger was extended with a manually created POS lexicon containing the POS tags for words that appear at least 5 times in the created dataset. As no FRAMED trained Chunker model was available, an alternative model, trained on the German TIGER corpus, was used [31].

For classification the MALLET library was used for training and testing. Based on previous experience, the MEC and SVM were selected as classifiers.

The framework allows to apply the described specific preprocessing steps and classifiers in any desired combination. An overview of the developed framework is given in ▶Figure 1.

3.3 Validation of the Framework for NLP

In total, 15 different pipelines were examined: The MEC and SVM as stand-alone classifiers were analyzed with and without one of the four specific preprocessing

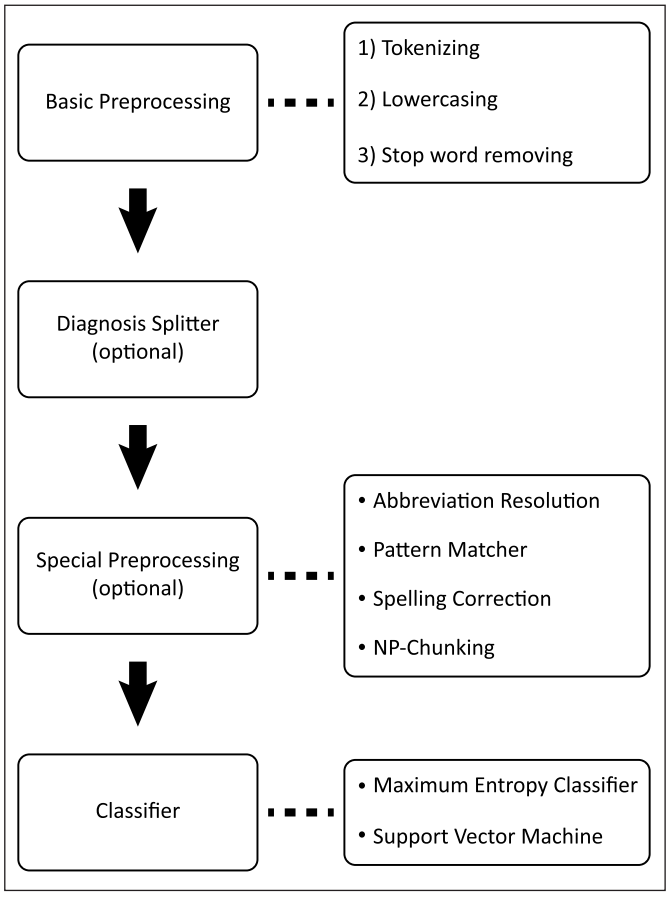

Figure 1
The developed framework for NLP. The framework is constructed as pipeline with a basic preprocessing (comprising a tokenizing, the removing of uppercases and of stop words), an optional diagnosis splitter, one of four optional specific preprocessing methods and one of two classifiers.

methods. The diagnoses splitter, an optional, self-developed package to automatically separate the diagnoses paragraph into multiple subsequences with only one relevant diagnosis, was tested with the MEC. To avoid interference between different preprocessing methods and to explore the performance of each method, each preprocessing method was examined individually and no combinations.

The pipelines were validated by a tenfold cross-validation. Hereby 9/10th of the dataset was used to train the pipeline and the remaining 1/10th to test its performance. This is repeated ten times so that each part will be used nine times for training and one time for testing. The final result is the average value over all validation results. The quality of the automatic classification of the test set was assessed by the rate of incorrectly classified data elements and by the F1-score, as the harmonic mean of precision and recall. All these steps were reiterated 100 times, each time with a randomly ordered dataset to rule out the influence of advantageous or disadvantageous arrangements in the data.

For significance testing between the results of the stand-alone classifiers (baseline) and the enabled specific preprocessing methods, the approximate randomization test was used [32, 33].

4. Results

4.1 Annotated Text Corpus of Multiple Myeloma Diagnosis

From the free-text diagnosis paragraphs of the 737 discharge letters the following data elements were derived within REDCap and exported as CSV file for training and testing of the NLP methods:

- no_diagnoses: The total number of relevant diagnoses. In the Multiple Myeloma research registry only "multiple myeloma" (MM), "monoclonal gammopathy of undetermined significance" (MGUS), "smoldering MM" and "solitary plasmacytoma of bone" are of relevance. Additional diagnoses were not considered. According to the total number of relevant diagnoses the following data elements appear between zero and three times.
- date_initial_diagnosis_(1 to 3): The date of the first histological incidence of the diagnosis in the format mm/yyyy. In German clinical reports the term "Erstdiagnosedatum" (date of first diagnosis) or its abbreviation "ED" is used.
- diagnosis_(1 to 3): One of the four relevant diagnosis (MM, MGUS, smoldering MM, plasmacytoma) was chosen. The diagnoses "symptomatic myeloma" was coded as MM, "asymptomatic myeloma" as smoldering MM.
- heavy_chain_(1 to 3): The class of immunoglobulin (Ig) produced by the myeloma disease with the following options: IgA, IgG, IgD, IgE, IgM, the biclonal type IgA-IgG or, if no immunoglobulin was present, light chain myeloma. Bence Jones protein was used synonymously to light chain myeloma. If the information was missing the option "other or not available" was chosen. The type of heavy chain does not change over the course of progression from plasmacytoma over MGUS and smoldering MM to MM. In cases where multiple diagnoses were present, and the heavy chain was not specified for each diagnosis, the same option was applied to all diagnoses.
- light_chain_(1 to 3): The type of free light chains produced by the myeloma disease. The options were: "kappa", "lambda", "kappa and lambda", and "other or not available". The deductive reasoning as for heavy chain applies here too.
- salmon_durie_staging_(1 to 3): The staging system according to Salmon and Durie to classify the clinical stage, designated with roman numbers I to III. If the information was missing the option "other or not available" was chosen. The staging system was only applied to the diagnoses smoldering MM and MM.
- creatinine_level_(1 to 3): Classification of the serum creatinine in the classes A, B, or "other or not available" depending on the laboratory value. The creatinine level is an additional specification of the Salmon-Durie-staging-system and stated in diagnosis reports mostly together, e.g. as "IA" or "IIIB". The creatinine level was also only applied to the diagnoses smoldering MM and MM.
- crab_(1 to 3): The diagnostic criteria applied to the symptomatic condition. The options were: C (hypercalcemia), R (renal failure), A (anemia), B (bone lesions), pain and other, like focal soft tissue swellings. Multiple options could be applied. In cases with only the diagnosis MM and no further information, the CRAB criteria were set to "not specified". The various synonyms, spellings and notations in German clinical letters and the corresponding coding are listed in ▶Table 1.

Table 1 CRAB criteria and the corresponding synonyms, spellings and notations.

CRAB criteria coding option	Synonyms, variations in spelling or notation
C (hypercalcemia)	Hyperkalzämie, Hypercalcämie, Hyperkalziämie, Hypercalciämie, hyper calcemia
R (renal failure)	Nierenversagen, Niereninsuffizienz, Nierenfunktionsverschlechterung, Nierenwertverschlechterung, Nierenfunktionseinschränkung, Cast-Nephropathie
A (anemia)	Anämie, Blutarmut, Blutmangel, anemia
B (bone lesions)	Osteolysen, Osteoporose, Knochendestruktion, Knochenerkrankung, Osteopenie, Osteodestruktion, Knochenschädigung, Knochenkomplikationen, knöcherne Komplikationen, Frakturen
Pain	Schmerz, Myelom-assoziierte Schmerzen
Other	Weichteilherde, Weichteiltumore

▶Table 2 provides an example of an extracted free-text diagnosis paragraph from a clinical report and of the annotation of the data elements of interest.

4.2 Evaluation of the Annotated Text Corpus of Multiple Myeloma Diagnosis

The data entry persons performed the manual documentation task and created

independently the two datasets E1 and E2. Both datasets were compared with each other by a third person. In total, E1 consists of 7709 data items and E2 of 7697, from which 7642 were consensus. Of the consensus items, 7573 (99.1%) were equal, resulting in an absolute error of 69 items (0.9%). Distributed on all data elements, the CRAB criteria C (hypercalcemia) was equal in all items, and the error rate of the CRAB criteria "not specified" was with 2.2% the highest. The Cohen's kappa coefficient as a measure of the inter-rater agreement, was on average 0.96, indicating an almost perfect agreement between E1 and E2.

The ATC was created by a third person, reviewing the 69 unequal items and the non-consensus items of E1 and E2. The final ATC consists of 7722 data items. The completeness of E1 was 99.5% (36 missing items) and of E2 99.4% (44 missing items). The 69 unequal items between E1 and E2 were distributed as 31 errors from E1, 37 errors from E2 and one error from both.

The results of the evaluation of E1 and E2 and of the ATC are available as ►supplementary online file.

The final ATC consists of 737 different diagnoses paragraphs. The total number of relevant, coded diagnoses is 867. The number of relevant diagnoses is distributed as follows (total number, percentage): 0 (18, 2.4%), 1 (591, 80.2%), 2 (108, 14.7%), 3 (20, 2.7%). The diagnosis MM is with 468 and 54.0% the most frequent choice. MGUS (295, 34.0%), smoldering MM (86, 9.9%) and plasmacytoma (18, 2.1%) follow. For the heavy chain, only the options IgG (562, 64.8%) and IgA (165, 19.0%) occur frequently, the other options only occasionally: IgM (27), IgD (4), IgA-IgG (3), IgE (0). For the light chain, "kappa" (545, 62.9%) is in front of "lambda" (296, 34.1%). The biclonal type "kappa and lambda" only occurred in six cases. The Salmon-Durie-staging, only relevant for diagnoses MM and smoldering MM, is rated as stage I in 122 (22.0%), stage II 50 (9.0%) and stage III 355 cases (64.1%). The level of serum creatinine is distributed as follow: A (442, 79.8%), B (77, 13.9%) and other or NA (35, 6.3%). The diagnostic criteria CRAB, only applied to the diagnosis MM and a multiple choice of options, is indicated in 265 (56.6%) cases with at least a single answer and is missing in 203 (43.4%) cases. In cases where the CRAB criteria is specified in the diagnosis paragraph (total of 376 selected options), the option were distributed as follows: B (bone lesions, 220, 58.5%), A (anemia, 74, 19.7%), R (renal failure, 43, 11.4%), C (hypercalcemia, 22, 5.9%), other (14, 3.7%) and pain (3, 0.8%).

The ►supplementary online files contain the complete annotated corpus as CSV file for training and testing of NLP methods.

Table 2 Example of a free-text diagnosis paragraph and of the annotated data elements. The number in brackets represents the corresponding coding.

Free-text diagnosis paragraph	Multiples Myelom Typ IgG kappa Stadium III A nach Salom und Durie, ED 01/10, symptomatisch; Monoklonale Gammopathie vom Typ IgG kappa ED 12/09
no_diagnoses	2 (2)
date_initial_diagnosis_1	01/2010
diagnosis_1	multiple myeloma (1)
heavy_chain_1	IgG (2)
light_chain_1	Kappa (1)
salmon_durie_staging_1	III (3)
creatinine_level	A (1)
crab_1	not specified (-99)
date_initial_diagnosis_2	12/2009
diagnosis_2	MGUS (2)
heavy_chain_2	IgG (2)
light_chain_2	Kappa (1)

4.3 Evaluation of the Framework for NLP

In total, 15 different pipelines were examined. The MEC and SVM were tested without any specific preprocessing step as well as with one of the four specific preprocessing methods. The diagnosis splitter was evaluated with the MEC on the corpus, containing 737 main diagnoses paragraphs. The performance of the MEC and SVM was evaluated without the diagnosis splitter on a subset of the corpus containing 591 instances with one coded diagnosis.

Evaluation of the classifiers showed a good to very good overall performance. The quality of automatic classification was however slightly better for the SVM with an average rate of incorrectly classified data elements of 0.84 on 1/10th of the dataset and an average F1-score of 0.92 compared to the MEC with an average error rate of 1.05 and average F1-score of 0.89.

The results for both quality indicators varied widely depending on the data element the classifiers were applied to, e.g. for the F1-score of the SVM between 0.67 (for the CRAB option "other") and 0.98

Table 3 Total number of data elements with decreased (positive change) or increased (negative change) error rate caused by preprocessing method on the maximum entropy classifier (MEC) and support vector machine (SVM). Total number of significant changes, if present, is added in brackets.

	Abbreviation Resolution		Pattern Matcher		Spelling Correction		NP-Chunking	
	MEC	SVM	MEC	SVM	MEC	SVM	MEC	SVM
Positive change	9 (6)	5 (5)	7 (4)	5 (3)	9 (9)	6 (4)	4 (4)	3 (3)
Negative change	1	5 (2)	3 (1)	6 (4)	2 (2)	6 (3)	7 (7)	8 (7)

(for the CRAB option "anemia"). This was also true for the effect of preprocessing. While a particular method had a positive effect on a single data element, the effect had been negative on another. For example, the pattern matcher prior to the SVM lowered the error rate on five data elements and increased it on six – four times, even significantly. ▶Table 3 shows the effect of the specific preprocessing method on the data elements. In total, 23 positive and 10 negative significant changes were caused to the MEC by the specific preprocessing compared to 15 positive and 16 negative changes to the SVM.

Detailed examples of the negative effect of different preprocessing methods are available as ▶supplementary online file.

For most data elements the SVM was the better of both classifiers. For the specific preprocessing method, no clear trend was observable. ▶Table 4 lists all classified data elements together with the best classifier and specific preprocessing method according to the highest average F1-score and lowest average error rate.

The error rate of the automatic diagnosis splitter, evaluated with the MEC and without any specific preprocessing method, was for all data elements on average 3.93 on 1/10th of the dataset (minimum 0.3 for data element CRAB "pain"; maximum 7.97 for data element light chain). Through specific preprocessing the error rate was reduced significantly in 25 cases and increased significantly in 8 cases. Compared to the MEC tested on the single diagnosis dataset, the average error rate of the MEC tested on the multiple diagnosis dataset was significantly higher, on baseline as well as with preprocessing. For some data elements, such as the heavy chain, the error rate increased by almost 7 additional errors. The F1-score decreased only slightly, as a maximum for the light chain from 0.84 to 0.76.

The results of all evaluated classifier with and without specific preprocessing method are available as ▶supplementary online file.

5. Discussion

The aim of this study was to automatically classify the diagnosis and state of disease of free-text diagnostic reports by using advanced techniques of NLP. As a first step, an annotated text corpus was created that contains free-text paragraphs along with the specific annotation. To our knowledge NLP is broadly used in medical disciplines [17–21], but not yet in hematological malignancies such as leukemia, lymphoma or multiple myeloma. The reason for this could be that rare diseases are of minor research interest in the context of NLP, and that annotated text corpora are not publicly available. Aggravating this situation is the rare existence of clinical corpora in languages other than English. And yet expectations regarding exactness and reliability of automatic classified German diagnostic reports in the area of multiple myeloma were high.

In order to limit complexity, the developed NLP framework processes only the main diagnoses paragraph, instead of the diagnostic report in total. This restriction was decided, as the main diagnoses paragraph in the discharge letters for multiple myeloma patients of Heidelberg University Hospital is more structured with a higher information density compared to continuous free-text paragraphs like anamnesis or epicrisis. A higher frequency of data elements, concerning the state of disease and its condition, cannot be found in other parts of a diagnostic report. Automatic extraction and classification of entire clinical reports, where medical conditions are described in natural language, requires combination of NLP, information retrieval and heuristic approaches and additional research and training corpora.

Sebastiani stated that the availability of an annotated corpus, necessary to train a classifier, is a major challenge [15]. Since multiple source is the norm, clinical data exist as unstructured, narrative text in free-text documents in the HIS, and as structured data elements in a research EDC database. Often without any possibility to merge both data sources, the clinical document with the associated data elements. Additional effort and cost for the creation of an annotated text corpus prior to the actual task of automatic classification is required. Therefore, the created German corpus for multiple myeloma classification is provided for further research and method optimization.

The free-text diagnoses paragraphs were annotated by two independent persons, and differences were assessed by a third. The evaluation and the low number of errors and high Cohen's kappa coefficient underlines, that the manual annotation was executed reliably and that the annotated text corpus is of high quality.

The developed framework offers a flexible environment and several useful tools for training and testing classifiers. The functionality ranges from reading training data and converting it to a structure that can easily be further processed. It can also be conveniently extended with additional

Table 4 Classifier and specific preprocessing method with the best performance on the automatic classification of the different data elements.

Data element	Best practice classifier	Best practice specific preprocessing
Diagnosis	SVM	Abbreviation Resolution
Heavy chain	SVM	Spelling Correction
Light chain	SVM	Spelling Correction
Salmon-Durie-staging	MEC	Pattern Matcher
Creatinine level	SVM	Baseline
C (hypercalcemia)	SVM	Spelling Correction
R (renal failure)	SVM	Spelling Correction
A (anemia)	MEC	NP-Chunking
B (bone lesions)	SVM	Abbreviation Resolution
Pain	MEC/SVM	Baseline
Other	SVM	NP-Chunking
Not specified	SVM	Spelling Correction

classification algorithms and specific preprocessing steps.

The evaluation of the different pipelines suggest a good performance of the investigated classifiers and NLP methods, even at the baseline. Normalization of the data and the extraction of additional features showed significant improvements when compared to the baseline. As an example the pattern matching preprocessing step reduces the average error rate by around 43% and increases the average F1-score by 0.02 when using the SVM for classification of the Salmon-Durie-staging. However, it is hard to tell in advance which combination of classifier and specific preprocessing produces the best results for a certain data element since preprocessing may also have a negative effect on classification performance. Abbreviation resolution, for example, increases the average error rate by almost 40% when classifying the creatinine level with the SVM.

The better evaluation results of the SVM compared to the MEC were apparently rooted to its better ability to deal with few occurrences in the training data. The data element "C (hypercalcemia)" for example is present in only 22 of the 591 instances in the dataset used to evaluate single diagnosis performance. The MEC showed a reduced F1-score of 0.61 caused by low recall of only 0.55 as positive instances are falsely classified negative. The SVM accomplishes a distinctly higher recall of 0.79 resulting in a F1-score of 0.84.

The applied approach for diagnosis splitting uses a simple heuristic on the diagnosis dates, which may explain the drop in performance when compared to the single diagnosis results. Since the absence of regular punctuation and the ungrammatical sentences of the free-text paragraphs made the multi-diagnosis splitting difficult, no known NLP technique could be applied. An automatic separation of the diagnoses paragraph into multiple subsequences, each containing one relevant diagnosis, should be subject to further investigations.

To conclude, the manual annotation and the error rate of 0.9% may indicate that an automatic classification is not needed. It should, however, be noted, that the data entry persons had only a (small) free-text diagnoses paragraph to annotate and in routine patient care they have to deal with a multi-page long clinical report to extract the relevant information. Only if NPL methods are applied to a whole report, or record, with similar promising results as ours, they would have a real benefit on quality, time and effort.

6. Conclusions

The low average error rates and high average F1-scores of each pipeline demonstrate the suitability of the investigated NPL methods. However, it was also shown that there is no best practice for an automatic classification of data elements from free-text diagnostic reports. Rather, the performance of an automatic classification depends on the properties of a data element, such as its character length and uniqueness, its frequency distribution in the training and test set and the degree of improved quality through the preprocessing method of the framework.

Acknowledgment

The authors would like to thank the Section of Multiple Myeloma at the Department of Hematology, Oncology, and Rheumatology from Heidelberg University Hospital and of the National Center for Tumor Diseases (NCT) Heidelberg. Especially we thank the Head of Section Professor Hartmut Goldschmidt and the workgroup for scientific documentation and data management around Dr. Maria Pritsch and Kerstin Scherbaum-Lawrenz.

References

1. Yoder RD, Swearingen DR, Schenthal JE, Sweeney JW, Nettleton WJ. An Automated Clinical Information System. Methods Inf Med. 1964; 3(2): 45–50.
2. Brigl B, Ringleb P, Steiner T, Mann G, Leiner F, Grau A et al. Multiple Verwendbarkeit Klinischer Dokumentationen am Beispiel eines wissensbasierten klinischen Arbeitsplatzsystems in der Neurologie. Informatik, Biometrie und Epidemiologie in Medizin und Biologie 1995; 26(3): 240–9.
3. Thomson R. DILEMMA: Decision Support in Primary Care, Oncology and Shared Care. In: Laires, MF et al., editors. Health in the New Communications Age. Amsterdam: IOS Press; 1995. p. 208–217.
4. Georgiou A, Pearson M. The role of health informatics in clinical audit: part of the problem or key to the solution? J Eval Clin Pract. 2002; 8(2): 183–8.
5. Lazarus R, Kleinman K, Dashevsky I, Adams C, Kludt P, DeMaria A et al. Use of Automated Ambulatory-Care Encounter Records for Detection of Acute Illness Clusters, Including Potential Bioterrorism Events. Emerg Infect Dis. 2002; 8(8): 753–60.
6. Lewsey JD, Leyland AH, Murray GD, Boddy FA. Using routine data to complement and enhance the results of randomised controlled trials. Health Technol Assess. 2000; 4(22): 1–55.
7. Herzberg S, Rahbar K, Stegger L, Schäfers M, Dugas M. Concept and Implementation of a Single Source Information System in Nuclear Medicine for Myocardial Scintigraphy (SPECT-CT data). Appl Clin Inform. 2010; 1(1): 50–67.
8. Holm MB, Rogers JC, Burgio LD, McDowell BJ. Observational data collection using computer and manual methods: which informs best? Top Health Inf Manage. 1999; 19(3): 15–25.
9. Dugas M, Breil B, Thiemann V, Lechtenbörger J, Vossen G. Single source information systems to connect patient care and clinical research. Stud Health Technol Inform. 2009; 150: 61–5.
10. Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods Inf Med. 2009; 48(1): 38–44.
11. Ammenwerth E, Spötl H. The time needed for clinical documentation versus direct patient care. A work-sampling analysis of physicians' activities. Methods Inf Med. 2009; 48(1): 84–91.
12. Embi PJ, Jain A, Clark J, Bizjack S, Hornung R, Harris CM. Effect of a clinical trial alert system on physician participation in trial recruitment. Arch Intern Med. 2005; 165(19): 2272–7.
13. Dugas M, Lange M, Berdel WE, Müller-Tidow C. Workflow to improve patient recruitment for clinical trials within hospital information systems – a case-study. Trials 2008; 9: 2.
14. Williams JG, Cheung WY, Cohen DR, Hutchings HA, Longo MF, Russell IT. Can randomised trials rely on existing electronic data? A feasibility study to explore the value of routine data in health technology assessment. Health Technol Assess. 2003; 7(26): iii, v-x, 1–117.
15. Sebastiani F. Machine learning in automated text categorization. ACM Computing Surveys 2002.
16. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc. 2004; 11(5): 392–402.
17. Hanauer DA, Miela G, Chinnaiyan AM, Chang AE, Blayney DW. The registry case finding engine: an automated tool to identify cancer cases from unstructured, free-text pathology reports and clinical notes. J Am Coll Surg. 2007; 205(5): 690–7.
18. Jouhet V, Defossez G, Burgun A, Le Beux P, Levillain P, Ingrand P et al. Automated classification of free-text pathology reports for registration of incident cases of cancer. Methods Inf Med. 2012; 51(3): 242–51.
19. Buckley JM, Coopey SB, Sharko J, Polubriaginof F, Drohan B, Belli AK et al. The feasibility of using

natural language processing to extract clinical information from breast pathology reports. J Pathol Inform. 2012; 3: 23.
20. Elkins JS, Friedman C, Boden-Albala B, Sacco RL, Hripcsak G. Coding neuroradiology reports for the Northern Manhattan Stroke Study: a comparison of natural language processing and manual review. Comput Biomed Res. 2000; 33(1): 1–10.
21. Yadav K, Sarioglu E, Smith M, Choi H. Automated outcome classification of emergency department computed tomography imaging reports. Acad Emerg Med. 2013; 20(8): 848–54.
22. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008: 128–44.
23. Ganzinger M, Gietzelt M, Karmen C, Firnkorn D, Knaup P. An IT Architecture for Systems Medicine. Stud Health Technol Inform. 2015; 210: 185–9.
24. Kaatsch P, Spix C, Hentschel S, Katalinic A, Luttmann S, Stegmaier C et al. Krebs in Deutschland 2009/2010. 9th ed. Robert Koch-Institut, Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V., editors. Berlin; 2013.
25. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap) – a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009; 42(2): 377–81.
26. MALLET: A Machine Learning for Language Toolkit. University of Massachusetts Amherst; 2002. Available from: http://mallet.cs.umass.edu.
27. OpenNLP; 2013. Available from: https://opennlp.apache.org/.
28. Faessler E, Hellrich J, Hahn U. Disclose Models, Hide the Data – How to Make Use of Confidential Corpora without Seeing Sensitive Raw Data. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14); 2014. p. 4230–4237.
29. Wermter J, Hahn U. An Annotated German-Language Medical Text Corpus as Language Resource. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04); 2004. p. 473–6.
30. Hunspell; 2014. Available from: http://hunspell.sourceforge.net/.
31. OpenNLP_1.5.1-German-Chunker-Tiger-Corps07.zip: OpenNLP German Chunker Tiger Corpus; 2011. Available from: http://gromgull.net/blog/2010/01/ noun-phrase-chunking-for-the-awful-german- language/.
32. Eric W. Noreen. Computer Intensive Methods for Testing Hypotheses. An Introduction. New York: Wiley; 1989.
33. Riezler S, Maxwell JT. On Some Pitfalls in Automatic Evaluation and Significance Testing for MT. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, Michigan: Association for Computational Linguistics; 2005. p. 57–64.