Identifying pathological findings in German radiology reports using a syntacto-semantic parsing approach

Claudia Bretschneider[1,2], **Sonja Zillner**[1] **and Matthias Hammon**[3]
[1] Siemens AG, Corporate Technology, 81739 Munich, Germany
[2] University Munich, Center for Information and Language Processing, 80538 Munich, Germany
[3] University Hospital Erlangen, Department of Radiology, 91054 Erlangen, Germany
{claudia.bretschneider.ext,sonja.zillner}@siemens.com,
matthias.hammon@uk-erlangen.de

Abstract

In order to integrate heterogeneous clinical information sources, semantically correlating information entities have to be linked. Our discussions with radiologists revealed that anatomical entities with pathological findings are of particular interest when linking radiology text and images. Previous research to identify pathological findings focused on simplistic approaches that recognize diseases or negated findings, but failed to establish a holistic approach. In this paper, we introduce our syntacto-semantic parsing approach to classify sentences in radiology reports as either pathological or non-pathological based on the findings they describe. Although we operate with an incomplete, RadLex-based linguistic resource, the obtained results show the effectiveness of our approach by identifying a recall value of 74.3% for the classification task.

1 Introduction

In radiology, descriptions of the patient's health status are stored in heterogeneous formats. They range from radiology images - which are the primary source for radiologists - over dictated reports about the image findings up to written texts.

Although the various data items describe the same status, they are distributed in non-linked systems. This is hindering the radiologist's workflow. Especially when reading reports, radiologists want to link back from the described finding (in the text) to the related body location (in the images). Today, they establish the link manually. This is obviously time-consuming when state-of-the-art imaging modalities deliver a mass of stacked images.

In order to link radiology images and reports, each information source needs to be annotated with semantic meta-information about the anatomical entities they describe. The necessary semantic image annotations for the integration have been made available as a result of the Theseus MEDICO project (Seifert, 2010). Introduced algorithms automatically detect anatomical entities in radiology images and annotate those with the corresponding RadLex IDs (Seifert et al., 2009). The semantic annotations from the reports have to be in line with those image annotations. Therefore, the final result of the text analysis system need to be anatomical annotations based on RadLex. We introduce a mechanism that extracts those semantic annotations from the radiology reports to enable the integration.

We identified three challenges, which a text analysis system has to consider when extracting the relevant anatomical entities from text:

1. The linguistic characteristics of the reports differ significantly from standard free-text,
2. the underlying German linguistic resource (the RadLex taxonomy) is incomplete and
3. only a subset of the named anatomical entities in the reports are relevant for annotating.

First, the special *linguistic characteristics* of the handled German reports have to be taken into account. While the linguistic characteristics English radiology reports have been intensively studied (Friedman et al., 1994; Friedman et al., 2002; Sager et al., 1994), German ones are still a young research area. German reports are comparable to English ones when it comes to structural particularities. One can observe two characteristics in both languages: syntactic shortness and reduced semantic complexity. But the reports differ in richness of the language used. German language is rich in inflection form; the same is true for German medical language. Additionally, clinical texts extend the variety in inflection forms by introducing a huge amount of Greek- and Latin-rooted vocabulary. Further linguistic particularities will be introduced in a later section.

Second, the anatomical annotations will be established based on the controlled vocabulary of the *RadLex* taxonomy. Anatomical annotations of the images (based on RadLex) are already available and hence impose the mandatory condition to use RadLex annotations for the reports. We operate on German radiology reports that is why we use the German RadLex taxonomy. Compared to the English version, the German RadLex is lacking in terminology. This is an obstacle, we have to overcome.

Third, we have to find a way to filter *relevant anatomical annotations*. According to the radiologists we worked with, it is inappropriate to extract *all*

Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013), pages 27–35,
Sofia, Bulgaria, August 4-9 2013.

anatomical entities from the text to link them with the image annotations. A large portion of the anatomies is described with normal or absent findings, which do not describe pathologies. Those findings are included in the reports in order to exclude differential diagnoses. However, radiologists are interested in images of anatomical entities described with pathological findings. Thus, a crucial part of our work is to extract the anatomical entities with pathological findings in order to link only those with the image positions.

The core contribution of this paper is the description of a syntacto-semantic parsing approach to identify the sentences that describe pathological findings by using the German version of the RadLex taxonomy. The results of this approach are used to integrate relevant semantic information from heterogeneous data sources and support radiologists significantly in their work routine.

To introduce our solution, the remainder of this paper is organized as follows: Section 2 refers to related work in the field and shows where sub-problems are still unsolved. In Section 3, we analyze the linguistic characteristics of the reports. Section 4 introduces the text analysis system for integrating radiology text and images. The system handles both the linguistic particularities of the reports and the shortcomings of RadLex as linguistic resource and filters relevant anatomical entities from the reports. Section 5 evaluates and discusses the classification and extraction results. Finally, Section 6 concludes with possible future work.

2 Related work

Medical grammar-based text analysis systems Information extraction from medical texts is a well-researched task in medical natural language processing (Meystre et al, 2008). Especially radiology reports play an important role.

Theoretical work in the linguistic characteristics of the medical sublanguage has been conducted on the adaption of theories of Harris by (Friedman et al., 2002). Early systems of (Sager et al., 1994; Friedman et al., 1994) are adaptations of the theories and implement own (context-free) medical language grammar for radiology reports. They show that parsing of medical texts based on a combined semantic-syntactic grammar can be successfully conducted – but they conducted their research using English reports. Even today, advances in grammar-based parsing of medical texts are reached (Fan et al., 2011).

More recently, sophisticated semantic medical text analysis systems have integrated a component to parse texts. (Savova et al., 2010) They take the output of the parsing process to extract semantic relationships between the medical concepts described.

All those systems work with elaborated lexicons that fully cover the vocabulary used in English report.

Detecting diseases and Negated finding Most systems cover the problem of detecting pathological findings in the reports just partially: In order to detect pathologies, they automate the assignment of codes for diseases listed in ontologies such as UMLS (Aronson, 2001; Lindberg, 1990; Long, 2005) or ICD (Computational Medicine Center, 2007; Pestian et al., 2007).

Non-pathological findings are identified using negation detection algorithms. Available approaches range from simple algorithms based on dictionary lookup and regular expressions (Chapman et al., 2001; Mutalik et al., 2001) through machine learning (Goryachev et al., 2006) up to advanced approaches that apply a context-free "negation grammar" (Huang, 2007).

Gap analysis While the grammar-based analysis of radiology reports has proven to be successful with complete lexical resources, we have to face the shortcomings of an incomplete lexicon. Furthermore, in other systems the grammar is used to analyze the syntax of the reports. Our approach to use it for classification is novel and has not been applied so far.

Working with German clinical texts is another challenge in the field. English texts have been made available by a number of shared tasks and gained more and more interest in the last decade. Medical corpora in languages other than English are not available to that extent.

That is perhaps also the reason for the tremendous lack of German medical ontologies. While great effort is put into the advance of English ontologies, German language versions are rare.

Terminology acquisition and semantic classification Semantic classifications beyond the hierarchical information encoded in taxonomies and ontologies are still rare for ontology concepts. In particular, semantic classifications such as information about the pathological nature of the concepts are missing so far.

Several approaches address this lack of semantic information: Corpus-based approaches base their methods on statistical analyses about the coverage and usage frequency of UMLS ontology concepts (Liu et al., 2012; Wu et al., 2012). (Johnson, 1999) derives semantic classes from ontology mapping and disambiguates multiple senses in contexts of discharge summaries. (Campbell et al., 1999) applies pattern-based rules and combines them with UMLS concepts to acquire new and semantically classified terminology. However, this approach is limited to noun phrases.

Finally, (Zweigenbaum et al., 2003) introduce approaches to automatically extending the existing English UMLS ontology with non-English concepts based on statistical algorithms.

3 Corpus analysis

3.1 Reference corpus

Since a publicly available corpus of German radiology reports is missing, we build our own annotated corpus

based on 2713 de-identified reports from our clinical partner, the University Hospital Erlangen. The reports result from radiology examinations of lymphoma patients and range from April 2002 to July 2007. Each report contains two free-text sections: The first one describes findings observed in the images. In the second sections, the radiologist provides an overall evaluation about the findings, derives probable diagnoses and excludes differential diagnoses.

3.2 Development set of reports

From the corpus, we selected 174 reports for the development set. They are uniformly distributed across time and length.

The development set serves multiple purposes:

1. It is used for the linguistic analysis.
2. We use it for grammar derivation.
3. And pathology classifications and additional vocabulary are learned from the sentences.

A radiologist classified each of the contained sentences either as pathological or non-pathological. This is done based on the characteristics of the findings described in the sentence. Sentences describing normal or negated findings are classified as 'non-pathological' and those containing descriptions of abnormalities are classified as 'pathological'. In cases where sentences include both types of findings, they are classified as 'pathological'. Hence, each sentence in the development set was annotated with the classification information.

3.3 Statistics of the development set

The 174 reports in the development set contain 4295 sentences of which less than half are classified as 'pathological'. This ratio is in line with the radiologists' experience. As from their intuition, the majority of the findings described in radiology reports is noted as absent or has normal status. In the reports, they complement pathological findings in order to note the absence of finding and to exclude suspected diseases. However, those sentences classified as 'non-pathological' are irrelevant for our setting of linking the containing anatomies to the images.

Table 1 shows additional results of the statistical corpus analysis.

	Sentence classification	
Corpus characteristic	*pathological*	*non-pathological*
Sentences	1943	2352
Tokens used	16437	11572
Average sentence length	8.46	4.92
Distinct word types	2398	1581

Table 1: Results of statistical corpus analysis based on the development set

Another significant characteristic of the sentences is their average length. Pathological sentences are about as twice as long as non-pathological ones and thus are more complex in their syntax. The pathology classifier has to cover this complexity.

Furthermore, from comparing the distinct word types used, we conclude that the description of pathological findings requires a richer language than those of normal states and absent findings in non-pathological sentences. The linguistic resource has to cover this required rich language.

3.4 Semantic and syntactic characteristics

One of the most apparent syntactic characteristics of the reports is the elliptical style of the sentences. The texts are rich in omission of verbs; verbs are dispensable as they only underline the absence or presence of symptoms. An example that illustrates the facts is shown below.

General language
In der Lunge sind keine Ergüsse zu finden.
In the lung, there are no effusions found.
Radiologist's style
Lunge: Kein Erguss.
Lung: No effusions.

The observation of the syntactic structure of the sentences is in line with (Friedman et al., 2002) and will simplify the classification of the sentences.

The second observation we made is that the medical language uses a high amount of domain-specific vocabulary. This vocabulary is rarely used in every-day language and is highly connected with (implicit) medical domain knowledge. Thus, the linguistic handling of the reports requires a domain-specific lexicon. Furthermore, the vocabulary can be categorized into only a few semantic classes representing the content, such as measurements, dates, anatomies, modifier of the anatomies, diseases, etc.

Third, one feature of the medical language is very domain-specific: It uses a high amount of Greek- and Latin-rooted words. This is important, because those terms follow their own specific inflection forms. Furthermore, for many terms there exist both German and Latin-/Greek-rooted descriptions which are used interchangeably (e.g. descriptions of anatomical entities or diseases). However, most lexicons only contain a single term - not the complete list of synonyms.

Like the German language, the medical language is also rich in compound terms such as *Nasenseptumdeviation* (deviation of the nasal septum) or *Glukosestoffwechselsteigerung* (increase in glucose metabolism). Especially radiologists use a high number of compounds to describe pathological findings. They will be of particular importance for the identification of pathological findings. In many cases, only after determining the pathology classification of each

subtoken, the classification of the compound can be determined.

Systems that mine information from radiology reports have to consider the named syntactic and semantic characteristics and handle them as language-specifics. In particular, the short length of the sentences simplifies the development of a grammar with a limited number of rules.

4 Methods

4.1 Grammar-based classification approach

Based on the observations from the corpus analysis, we derive and apply a semantic context-free grammar (CFG) to classify sentences.

Using a grammar to classify the sentences may not seem intuitive for every-day language sentences. Nevertheless, the language used in radiology reports allows this approach. There are several facts that support the usage of a grammar.

1. The structure of the sentences created by radiologists differs significantly from the structure of general German language. To model this language an own (sublanguage) grammar is necessary.

2. Since the sentences are short in length, a relatively small number of grammar rules can represent their syntax. In particular, the omission of verbs allows us to create and use a simplified grammar.

3. As already researched by (Friedman et al., 2002), the sentences contain a limited number of semantic classes which are combined into few rules.

These observations support the approach to create a grammar with few rules to classify the sentences.

4.2 Overview of the building blocks of the text analysis system

After having analyzed the linguistic characteristics, we designed a text analysis system to extract the relevant information from the reports. The classification is based on a grammar whose components are setup first: The **grammar rules are created** and **lexicon is setup**. To overcome the incompleteness of the lexicon and to enhance the grammar with probabilities, we introduce an additional **learning step**. These first three steps can be regarded as preparation steps for the subsequent integration steps: Finally, the system is able to **classify** report sentences and **extracts** anatomical annotations from the sentences classified as 'pathological'. In the end, the semantic annotations from text and images are **linked** across the data sources. The described steps of the target system are shown in Figure 1.

This paper focuses on the details of the created grammar: how the parsing algorithm is adapted to learn new linguistic knowledge and how the probabilistic parsing algorithm is used to derive a classification for an input sentence.

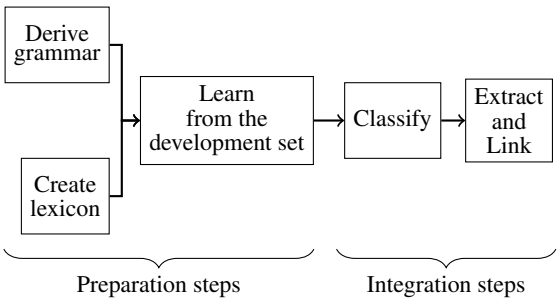

Figure 1: Processing steps in text analysis system

The following sections describe the details of the individual processing steps.

4.3 Derive grammar

The core component of the text processing system is the grammar. Our grammar has two functions:

1. It is used to describe the structure of a given input sentence, and

2. using the results of the parsing process, an input sentence can be classified as either 'pathological' or 'non-pathological'.

We use a *semantic* grammar for the description of the syntactic structure of the sentences. That means, instead of mapping syntactic categories from part-of-speech tags as non-terminal symbols, we use semantic representations of the content. E.g., the term *Niere* [spleen] gets assigned the non-terminal symbol ANATOMIE.

Following the proposal of (Friedman et al., 1994), we create semantic classes that represent the content of the radiology reports. However, we do not need their fine-grained semantic class definition. Our task of pathology classification requires only a reduced number of classes. We drop classes that do not change the pathology classification result (such as degree, quantity, technique, etc.) and introduce the generalized semantic classes MOD (modifier) and TERM. The list of semantic classes derived is shown in Table 2.

The grammar has to fulfill one condition to be able to classify sentences. Only non-terminal symbols used for classification are directly derived from the start symbol (S). We use the non-terminal symbols PATH for classifying sentences as 'pathological' and NOPATH for classifying as 'non-pathological'. Hence, the following unary rules designate the classification in our grammar:

$$S \rightarrow PATH$$
$$S \rightarrow NOPATH$$

Any subsequent rules have to be hierarchically embedded into those rules.

During the subsequent (manual) grammar derivation process, we use the listed semantic classes as non-terminal symbols and derive the grammar rules from

Structural non-terminals	
ROOT	
S	
KOMMA	
ENUM	
FIND_CONNECT	
Classification non-terminals	
PATH	*Constituents*
NOPATH	*(sentence-level,*
MOD_PATH	*modifier and term)*
MOD_NOPATH	*with pathology*
TERM_PATH	*classification*
TERM_NOPATH	*information*
FINDING_NOPATH	
FINDING_PATH	
Semantic non-terminals	
LOCATION	
DATE	*Non-terminals*
MEASUREMENT	*representing*
ANATOMIE	*constituents with*
NEGATION	*specific semantic*
DISEASE	*meaning*
Linguistic non-terminals	
ARTICLE	*Article non-terminal*
ARTICLE_GENITIV	
PREP_DATE	*Preposition*
PREP_LOCATION	*non-terminals*
PREP_MEASUREMENT	*indicating different*
	semantic units
Mapping semantic class - regular expression	
DATE_VALUE	
MEASUREMENT_VALUE	
IMAGE_VALUE	

Table 2: List of semantic non-terminals

the development corpus. Because of the limited number of semantic classes and the elliptical sentence style, a small set of 238 grammar rules suffices to describe the sentence syntax. The resulting grammar rules consider the syntactic complexity of the sentences describing pathological findings: 52% of the rules model the constituent structure of pathological sentences.

4.4 Create lexicon

The linguistic resource of our system is a lexicon, created based on the German version of the RadLex taxonomy.

RadLex (RSNA, 2012) is a taxonomy published by the Radiological Society of North America (RNSA) in order to deliver a uniform controlled vocabulary for indexing and retrieval of radiology information sources. The current English version 3.8 contains 39976 classes. A German version has been worked-out (Marwede et al., 2009) in 2007. The contained terms are organized in 13 major categories: anatomical entity as one among others such as treatment, image observation and imaging observation characteristics. But as the development of the German language version has been stopped, the latest version 2.0 contains only a subset of classes (n=10003). This lack in terminology is an obstacle to overcome.

Linguistic resource From the German RadLex we created a lexicon (n=9479), which we use as linguistic resource. Each entry is represented by a list of properties.

Besides the structural properties *label* and *RID*, we apply several steps of linguistic and semantic processing to enrich the lexical entries. The *normalized stem* of each entry results from an own tokenization, normalization and stemming algorithm.

The normalization aligns German and Latin style spellings (e.g. *Karzinom*/*Carzinom*, *Okzipitallappen*/*Occipitallappen*). The stemmer adapts the German Porter stemmer and incorporates additional rules for suffixes and inflection that are derived from Latin and Greek. E.g., this extension enabled the mapping of *Mediastinum* and *mediastinal* to the same stem *mediastin-*, which would not have been possible with the German Porter stemmer.

Furthermore, during lexicon setup each entry is enriched with *semantic classification* information. The semantic class is used during parsing. We use reasoning methods and the hierarchical *is-a* structure of the RadLex taxonomy in order to deduct a semantic class for each entry from the major categories. For example, this mechanism enables us to assign to deduct the semantic class ANATOMIE for sub-entities of the major category 'Anatomical entity' (such as *Prostata* [prostate]).

We apply a similar reasoning mechanism for the *pathology classification*. As the lexicon entries are initially unclassified according to their pathological information, we analyzed them and found the following mechanism: It is feasible to classify each of the major categories unambiguously either as 'pathological' or 'non-pathological'. For example, entries with semantic class ANATOMIE are classified as 'non-pathological'. This pathological classification information is added to 10 out of 13 major RadLex categories and inferred to all hyponyms. For three of the categories, the classification is ambiguous. The determination of the pathology classification results in the distribution shown in Table 3.

Classification	#	
non-pathological	6001	63.3%
pathological	1714	18.1%
not to be determined	1764	18.6%
	9479	100%

Table 3: Results of the initial pathology classification of RadLex-based lexicon entries

The algorithm is able to classify 81.4 % of the lexicon

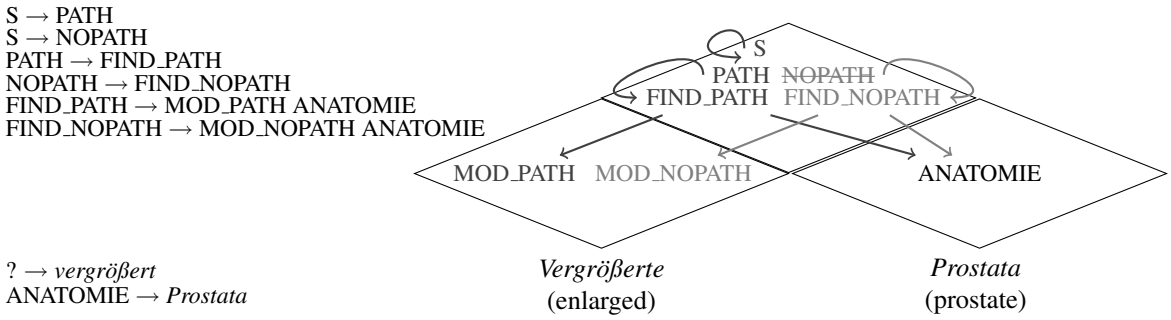

Figure 2: Learning lexical knowledge from sentence *Vergrößerte Prostata* (enlarged prostate)

entries. We have to find a way to classify the remaining unclassified entries. Only when all the lexical entries are classified, the sentence classification algorithm produces reliable results.

The finally derived lexical resource contains 9479 entries with 23588 tokens of which 6326 are distinct. Comparing this number with the distinct word types used in the development set (n=3172), one assumes that the lexicon could cover the vocabulary used in the reports. However, this is not the case. Important terms that occur quite frequently in the development set and have high relevance for the pathology classification are either not included in the lexicon (e.g. *Läsion/lesion*) or are included but are not classified (e.g. *sklerosiert* | RID 5906 [sclerosing]).

That is why we argue that an additional corpus-based learning step to extend the vocabulary and its classification is mandatory.

4.5 Learn from the development set

We introduce an additional learning step to extend the lexicon with missing items and to classify existing item missing a pathology classification. At the same time, the probabilities of the grammar rules are trained during this step. The learning is conducted using an adapted probabilistic CKY algorithm.

Extending the lexical resource How parsing is adapted to learn from the sentences is illustrated in Figure 2. The sentence *Vergrößerte Prostata* [Enlarged prostate] is input to the learning. From the sentence's annotation, we know that this sentence describes a pathological finding (PATH). The subset of the grammar necessary to parse this sentence is shown on the left-hand side of the figure. The non-terminal mapping of the words is shown below the grammar rules. Currently, only the mapping of the word *Prostata* to the non-terminal symbol ANATOMIE can be derived from the lexicon. Mapping *vergrößert* is not possible. The lexical entry has a semantic classification (*Modifier*) assigned, but no pathology classification. However, in this case both information items are necessary to determine the non-terminal mapping. In order to *learn* the missing pathology classification of this word, we apply an adapted CKY parsing algorithm.

The standard CKY algorithm (Kasami, 1965) operates bottom-up and uses two complete components to determine the parse tree of a given input sentence:

1. A complete lexicon to determine the non-terminal mapping of the words, and

2. a complete list of all grammar rules.

Our setting is missing the complete lexicon. That is why we adapt the standard algorithm and introduce a top-down analysis in order to extend the linguistic resource while parsing.

There are two possible non-terminal mappings for the word *vergrößert*: MOD_PATH (indicating a modifier for pathologies) or MOD_NOPATH (indicating a modifier not describing pathologies). Both of the options are used to determine the parse tree of the sentence. The ambiguity is resolved at the top-most parsing level: The sentence is annotated as 'pathological', hence, only rewritings that include the corresponding non-terminal symbol PATH are allowed. Finally, the parse tree of the sentence can be derived (as shown in Figure 3).

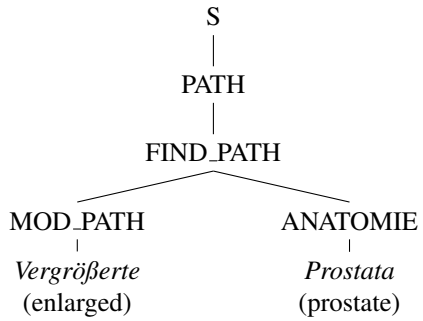

Figure 3: Parse tree derived from sentence *Vergrößerte Prostata* (Enlarged prostate)

In addition, the (formerly unknown) non-terminal mapping of the word *vergrößert* to MOD_PATH is deducted from the parse tree and the corresponding lexical entry is updated. Using this algorithm, we are also able to learn vocabulary that was not available in the lexicon before.

Training the grammar's probabilities The parse tree is also used as input for extending the grammar to a probabilistic context-free grammar. Each of the grammar rules used to form the parse tree is used to re-calculate the probabilities of the grammar rules.

After the learning step, the lexicon is extended to 10344 entries (before 9479). But even more important, the overall amount of lexicon entries classified as 'pathological' increased by 18.8 % to now 2036 entries (before 1714). We consider this a key success of the learning, as our classification depends on this encoded knowledge.

4.6 Classify

After conducting the previous steps,

1. the extended lexicon,
2. the trained P-CFG, and
3. the standard probabilistic CKY parsing algorithm

are applied to parse unclassified sentences.

The sentence classification is conducted based on the lexicon and the grammar rules. The lexicon helps to assign non-terminal symbols to the words in the sentence. Depending on non-terminal symbols assigned and the grammar rules applied during the subsequent parsing process, the parse tree will reveal the classification of the sentence.

As parsing algorithm we apply the standard probabilistic CKY (P-CKY) algorithm. It resolves both syntactic and classification ambiguities. In case, the sentence contains unknown words, the probabilistic parsing feature helps to disambiguate the non-terminal assignment. The derived parse tree describes both the syntactic structure of the sentence and the derived pathology classification.

4.7 Extract and Link

Finally, in case a sentence is classified as 'pathological', the contained anatomical entities are extracted. The sentences are annotated with the extracted anatomical information. An external system combines the anatomical annotations from images and reports. Thus, links are created successfully and the correlating image positions for pathological findings can be accessed from the text.

5 Evaluation

We evaluate the classification system using 40 randomly-chosen reports containing 1296 sentences.

5.1 Precision and recall measurements

We evaluate the classification results and the success of the alignment of radiology reports and images using precision and recall values. Only for sentences classified as 'pathological', the contained anatomical entities are extracted and anatomical annotations are created. That is why we prefer high recall values. If sentences are misclassified as 'pathological' – although they describe non-pathological findings (FP) – this is a minor issue. This misclassification results in alignment of anatomical entities in text and images without pathological findings. We accept lower precision values that yield those additional, but not intended alignments.

5.2 Baseline evaluation

We compare the evaluation results of the classification system with the results of a semantically-informed baseline algorithm. This algorithm detects negations and classifies the containing sentences as 'non-pathological'. Sentences containing diseases (determined based on Latin suffixes such as *-itis*, *-ose*, etc.) or a pathological RadLex concepts (as determined during the lexicon creation step) are classified as 'pathological'. Any remaining sentences are assumed to describe non-pathological findings.

The results of the baseline classification are shown in Table 4. The headings denote 'non-pathological' sentences (NOPATH) respectively 'pathological' sentences (PATH).

		expected classification	
		PATH	*NOPATH*
observed	*PATH*	17	0
classification	*NOPATH*	446	833

Table 4: Classification results using baseline algorithm

This baseline approach has the advantage of 100% precision value. However, it produces a low recall value of 3.67 %, which shows that this approach is not applicable for the alignment of text and images. The results show that the identification of pathologies is not feasible by only using (1) suffixes to determine diseases and (2) available pathology descriptions from the RadLex taxonomy.

5.3 Evaluation of the parsing-based classification results

Table 5 shows the system results of classifying the 1296 report sentences using the syntacto-semantic parsing approach.

		expected classification	
		PATH	*NOPATH*
observed	*PATH*	344	288
classification	*NOPATH*	119	545

Table 5: Sentence classification results using syntacto-semantic parsing approach

Taking into account the impact of the (still) incomplete lexicon, the recall value of 74.3 % indicates that the chosen approach to classify pathological sentences is successful. However, the precision value of 54.4 %

indicates that the classification of almost half of the 'pathological' sentences is incorrect.

Compared to the baseline, the acquisition of additional, pathology classified vocabulary and its incorporation into a parsing-based approach significantly improves the recall value. That is why we regard the enrichment of the lexicon at the crucial step for (further) improvement of the classification results. However, a large amount of sentences was classified incorrectly as 'pathological'. The error analysis will reveals some causes.

5.4 Error analysis

We identified four error types that produce incorrectly classified results.

1. Some of the pathology classification of the semantic knowledge acquired during learning is incorrect.

   Terms that do not describe pathological properties such as *Voraufnahme* [previous examination] or *Lymphknoten* [lymph node] were classified as 'pathological'; also, pathological findings such as *Läsion* [lesion] or *Infiltrat* [infiltrate] could not be classified correctly. Because of their high usage frequency (26, 116, 20, 7 times), these four terms are accountable for 169 of the misclassified sentences (both FP and FN) from the evaluation.

   The disambiguation of (word-level) pathology classification using sentence-level annotations is obviously very vague and imprecise. In order to improve the terminology acquisition results, we will include distribution information and probabilistic features into the learning process as future work.

2. The terminology acquisition leads to an extended lexicon, but still, terminology remains uncovered. In particular, the description of pathological findings requires a richer language, its lack inhibits their correct classification. Even though our corpus is limited to reports of lymphoma patients (i.e., contains limited medical vocabulary), still, the test set contains vocabulary that is not used in the training set. For a further elaborated lexicon, the training set has to be extended in size and also in content.

3. Furthermore, the majority of long sentences is not successfully parsed because of missing grammar rules. Those long sentences are more likely describing pathological findings, which leads to false negatives. We found that sentences longer than 8 tokens are rather incorrectly classified than correctly; nevertheless, this concerns only 8 % (99/1296) of all sentences. Thus, we regard this as a minor issues.

4. Finally, our assumption of covering the semantics with a limited number of non-terminals was disproven. The oversimplification of semantic classes is insufficient to parse the complex sentence structures in the reports. In particular, the structure of long sentences requires a wider range of non-terminals (and more grammar rules) in order to disambiguate the pathology classification. E.g., the defined semantic classes do nor distinguish modifiers of locations or size for anatomical entities or temporal modifier for pathologies. Their introduction will increase the resolution of dependencies in complex sentences and the overall classification.

The learning step is *the* crucial step for improvement of the classification results. It enriches the vocabulary. If the pathology classification of the learned vocabulary is optimized, the system will deliver even better results. The optimization of the vocabulary learning step will be future work.

6 Conclusion

We designed and implemented a system that aligns findings from radiology reports to findings in images based on semantic annotations. Providing the system, we assume to reduce the time necessary to find correlating descriptions of one finding in heterogeneous data sources.

We build our system on tailored NLP algorithms that extract relevant anatomical annotations with pathological findings. To identify sentences that describe pathological findings, we introduce a new, semantic grammar-based classification approach. To bridge the gap of the incomplete German terminology, a vocabulary acquisition step is introduced. Incorporating this newly learned vocabulary, the grammar-based classification delivers a recall value of 74.3%.

We identified a major issue relevant for further work on German clinical texts: The evaluation results reveal a large gap in coverage between the vocabulary used in non-English radiology texts and the controlled vocabulary delivered by RadLex. Furthermore, we believe that lexicons will be crucial resources for language processing in the medical domain. We will focus our future work on enriching existing lexicons and establishing new resources for linguistic analysis.

Acknowledgements

This research has been supported in part by the THESEUS Program in the MEDICO Project, which is funded by the German Federal Ministry of Economics and Technology under grant number 01MQ07016. The responsibility for this publication lies with the authors.

References

A. R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*, 17–21.

D. A. Campbell, S. B. Johnson. 1999. A technique for semantic classification of unknown words using UMLS resources.. *Proc AMIA Symp*, 716–20.

W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*, 34:301–310.

W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. 2001. Evaluation of negation phrases in narrative clinical reports. *Proc AMIA Symp.*, 105-109.

Computational Medicine Center. 2007. International Challenge: Classifying Clinical Free Text Using Natural Language Processing.. `http://www.computationalmedicine.org/challenge/index.php`.

J. W. Fan and C. Friedman. 2011. Deriving a probabilistic syntacto-semantic grammar for biomedicine based on domain-specific terminologies.. *J Biomed Inform*, 44(5):805–14.

C. Friedman, P. O. Alderson, J. H. M. Austin, J. J. Cimino, and S. B. Johnson. 1994. A General Natural-Language Text Processor for Clinical Radiology. *J Am Med Inform Assoc*, 1:161–174.

C. Friedman, P. Kra, and A. Rzhetksy. 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *J Biomed Inform*, 35:222–235.

S. Goryachev, M. Sordo, Q. T. Zeng, and L. Ngo. 2006. Implementation and evaluation of four different methods of negation detection.. Technical report, DSG.

Y. Huang and H. J. Lowe. 2007. A Novel Hybrid Approach to Automated Negation Detection in Clinical Radiology Reports. *J Am Med Inform Assoc*, 14:304–311.

S. B. Johnson. 1999. A semantic lexicon for medical language processing.. *J Am Med Inform Assoc*, 6(3):205-18.

T. Kasami. 1965. An efficient recognition and syntax-analysis algorithm for context-free languages. *Scientific Report AFCRL-65-758*, Air Force Cambridge Research Lab.

C. Lindberg. 1990. The Unified Medical Language System (UMLS) of the National Library of Medicine.. *J Am Med Rec Assoc*, 61(5):40–42.

H. Liu, S. T. Wu, D. Li, S. Jonnalagadda, S. Sohn, K. Wagholikar, P. J. Haug, S. M. Huff, and C. G. Chute. 2012. Towards a semantic lexicon for clinical natural language processing.. *AMIA Annu Symp Proc*, 568-576.

W. Long. 2005. Extracting diagnoses from discharge summaries. *AMIA Annu Symp Proc*, 470–4.

D. Marwede, P. Daumke, K. Marko, D. Lobsien, S. Schulz, and T. Kahn. 2009. RadLex - German version: a radiological lexicon for indexing image and report information. *Fortschr Röntgenstr*, 181(1): 38–44.

S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle. 2008. Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research. *Yearb Med Inform*, 24(11):128–144.

P. G. Mutalik, A. Deshpande, P. M. Nadkarni. 2001. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS.. *J Am Med Inform Assoc*, 8(6):598–609.

J. P. Pestian, C. Brew, P. Matykiewicz, D. J. Hovermale, N. Johnson, K. B. Cohen, and W. Duch. 2007. A Shared Task Involving Multi-label Classification of Clinical Free Text.. *BioNLP 2007: Biological, translational, and clinical language processing.*.

Radiological Society of North America. 2012. RadLex. `http://rsna.org/RadLex.aspx`.

N. Sager, M. Lyman, C. Bucknall, N. Nhan, and L. J. Tick. 1994. Natural Language Processing and the Representation of Clinical Data. *J Am Med Inform Assoc*, 1:142–160.

G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications.. *J Am Med Inform Assoc*, 17(5):507–13.

S. Seifert. 2010. THESEUS-Anwendungsszenario MEDICO. `http://www.joint-research.org/das-theseus-forschungsprogramm/medico/`.

S. Seifert, A. Barbu, K. Zhou, D. Liu, J. Feulner, M. Huber, M. Suehling, A. Cavallaro, and D. Comaniciu. 2009. Hierarchical Parsing and Semantic Navigation of Full Body CT Data. *SPIE Medical Imaging*.

S. T. Wu, H. Liu, D. Li, C. Tao, M. A. Musen, C. G. Chute, N. H. Shah. 2012. Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis.. *J Am Med Inform Assoc*, 19(1):149–56.

P. Zweigenbaum, R. Baud, A. Burgun, F. Namer, E. Jarrousse N. Grabar, P. Ruch, F. Le Duff, B. Thirion, and S. Darmoni. 2003. UMLF: a Unified Medical Lexicon for French. *AMIA Annu Symp Proc*, 1062.