Neural Facet Detection on Medical Resources

Thomas Steffek

April 2, 2019

Beuth Hochschule für Technik

BEUTH HOCHSCHULE
FÜR TECHNIK
BERLIN
University of Applied Sciences

Fachbereich VI - Informatik und Medien
Database Systems and Text-based Information Systems
(DATEXIS)

Bachelor's thesis

Neural Facet Detection on Medical Resources

Thomas Steffek

1. Reviewer Prof. Dr. habil. Alexander Löser
Fachbereich VI - Informatik und Medien
Beuth Hochschule für Technik
Database Systems and Text-based Information Systems

2. Reviewer Prof. Dr. Felix Gers
Fachbereich VI - Informatik und Medien
Beuth Hochschule für Technik

Supervisor M. Sc. Rudolf Schneider
Fachbereich VI - Informatik und Medien
Beuth Hochschule für Technik
Database Systems and Text-based Information Systems

April 2, 2019

Abstract

Physicians today rely on a variety of reference works and guidelines. However, traditional ways of information retrieval offer little more than retrieval of and access to whole documents. As a first step towards complex answer retrieval systems that could speed up this process by structuring results and extracting relevant sections we evaluate SECTOR as effective means of facet extraction on medical resources. Presented by Arnold et al. [Arn+19], SECTOR constitutes a novel approach for the joint task of segmenting documents into coherent sections and assigning topic labels to each section. We define two tasks for extracting normalized structural facets and ambiguous topical facets. To tackle the lack of German medical domain training data, we bootstrap our own using 7,553 doctors' letters from *Charité Berlin*. On this corpus we evaluated SECTOR in conjunction with varying language representation models. For segmentation and classification of 12 structural facets we report 98.97% F_1 and for the extended task of extracting 1,687 topical facets we note 87.07% F_1, both scored by SECTOR with bloom filter embeddings.

Abstract (German)

Heutzutage verlassen sich Ärzte auf eine Vielzahl von Nachschlagewerken und Richtlinien. Bisherige Formen des Information Retrieval bieten jedoch wenig mehr als die Möglichkeit, ganze Dockumente zu suchen und anzuzeigen. Complex Answer Retrieval Systems könnten diesen Prozess beschleunigen, indem sie Ergebnisse strukturieren und relevante Abschnitte hervorheben. Als ersten Schritt dorthin evaluieren wir die Effektivität von SECTOR zur Facet Extraction auf medizinischen Dokumenten. SECTOR von Arnold et al. [Arn+19] stellt einen neuen Ansatz für die zusammenhängenden Aufgaben dar, Dokumente in fortlaufende Abschnitte zu unterteilen und diese Abschnitte zu klassifizieren. Wir formulieren zwei Tasks zur Extrahierung von normalisierten strukturellen Facetten und mehrdeutigen thematischen Facetten. Aus Mangel an medizinischen, deutschen Trainingsdaten annotieren wir 7.553 Arztbriefe der *Charité Berlin*. Auf diesem Datensatz evaluierten wir SECTOR in Verbindung mit unterschiedlichen Language Representation Models und verzeichnen für SECTOR in Kombination mit Bloom Filter Embeddings bei der Segmentierung und Klassifizierung von 12 strukturellen Facetten einen Score von 98,97% F_1 und bei der erweiterten Aufgabe des Extrahierens von 1.687 thematischen Facetten 87,07% F_1.

Contents

1 Introduction

> *...there is nothing remarkable in being right in the great majority of cases in the same district, provided the physician knows the signs and can draw the correct conclusions from them.*
>
> — **Hippocrates of Kos**
> (Physician, father of medicine)

Hippocrates is said to be the first person to conjecture that diseases were not caused by mystical entities or superstition. Instead, he believed in diagnosing diseases and tried to prognose their course[1]. He claimed that physicians can be right most of the time, if only they knew the signs and drew the correct conclusions from them [Var84].
However, as medicine evolved over the ages and identified new diseases one after another, complexity of "knowing the signs" grew ever larger. Today, modern physicians rely on a variety of reference works and guidelines, both digital and nondigital. As medical knowledge and data expands current ways of retrieving information become unpractical [YM15]. New systems need to be developed and machine reading methods like facet extraction will be fundamental cornerstones of these approaches.

1.1 Facet Extraction on Medical Health Records

While traditional information retrieval systems like PubMed[2] have risen to be among the most important sources for up-to-date health care evidence [YM15], they offer little more than retrieval of and access to whole documents. The human reader then has to manually skim all results for the relevant sections and information she needs. *Complex answer retrieval* (CAR) systems like SMART-MD [Sch+18] could speed up this process and help clinicians to reach a better decision faster, by structuring results and highlighting relevant sections. As input for CAR systems a machine reading

[1] *Hippocrates*. en. Page Version ID: 890031202. Mar. 2019. URL: https://en.wikipedia.org/w/index.php?title=Hippocrates&oldid=890031202 (visited on Apr. 1, 2019).
[2] PubMed. *Home - PubMed - NCBI*. en. URL: https://www.ncbi.nlm.nih.gov/pubmed/ (visited on Mar. 31, 2019).

algorithm is necessary, which surpasses searching and identifying singular words, and instead analyzes the latent topic over the course of a document.
Arnold et al. [Arn+19] recently published such a method and achieved good results for Wikipedia articles. In a novel usage we apply this approach to clinical resources. In cooperation with *Charité Universitätsmedizin Berlin*'s *Medical Department, Division of Nephrology and Internal Intensive Care Medicine* we aim to devise a methodology as baseline for future adaption to other divisions or institutions.

1.2 Ethical Considerations

Technological advancements always come with a risk. Be it impacts on climate like with cars and planes or impacts on peoples' lives when they run into a pole while texting on their phone.[3]
Luckily, people are more considerate when it comes to the medical field. Not only based on the positive effects a new procedure entices, but also on the bad effects that error entails. Clinical decision support systems therefore need to be thoroughly analyzed and studied, not only regarding quality of the software but also regarding their ethical implications [Goo16].
However, facet extraction is but a small step towards clinical decision support. Yet this step includes ethical considerations as well. Working with medical resources means working with sensitive and private information. Aicardi et al. [Aic+16] claim that even anonymized data may not be anonymous forever: "data and material that are anonymized today may no longer be anonymous in the context of tomorrow's technologies and data resources."
This does not stop at raw data. Semantic neural representation of sensitive text is also sensitive. Based on the vector space distribution conclusions can be drawn about the data seen during training. Additionally, there has been a surge in research communities aiming for explainable neural networks [Sam+17]. Faessler et al. [Fae+14], who try to distribute neural models instead of medical datasets, are also aware of this issue. They stress the need of anonymized training data, even if the data is never to be published.
While technological advancements are without a doubt good, we should still employ "progressive caution" [Goo16].

[3] National Safety Council. *Pedestrian Safety*. URL: https://www.nsc.org/home-safety/safety-topics/distracted-walking (visited on Mar. 31, 2019).

1.3 Methodology

As mentioned in Chapter 1.1 our facet extraction method is SECTOR, a deep learning model. As with any deep learning approaches, suitable training data is necessary. Additionally, the text needs to be represented in a way that can be understood by neural networks. Finally, we define our task as training goals that best capture our intention.

Training data Charité Berlin supplies us with a set of doctors' letters, more specifically discharge summaries. As these are plain text, we further need annotated training data. We aim to find potential public datasets and examine their compatibility. If necessary, we annotate the letters using bootstrapping algorithms.

Text representation For text representation we assess promising language models. Since Sheikhshab et al. [She+18] and Lee et al. [Lee+19] observe improvements using specialized word embeddings, we additionally train specialized clinical models for comparison.

Modeling the task To formalize our task we consult existing facet classification approaches and identify what is most applicable. Afterwards we try to represent them using SECTOR's two variations, the *headings* and the *topics* model.

1.3.1 Hypotheses

Analyzing our approach, we formulate the two underlying hypotheses:

(i) **Specialized text embeddings perform better than general purpose text embeddings on medical domain**

(ii) **SECTOR as effective means of facet extraction on medical resources**

1.3.2 Limitations

We further note two limitations for our project:

(i) **Hardware requirements** Since we work with sensitive data, security measures need to be adhered. The doctors' letters are not allowed to be stored outside of Charité Berlin. Therefore, we have to use the available resources on-site.

(ii) **Time constraints** A bachelor's thesis prescribed timespan is three months. Taking research and writing overhead as well as hardware restrictions into account, model training time is critical. We will therefore constraint ourselves to efficient and fast language models.

1.4 Outline

The rest of this is thesis is structured as follows: Chapter 2 will give an introduction into basic terms and neural networks, before focusing on semantic text representation models using the example of language models. Finally it presents an overview of SECTOR's architecture. In Chapter 3 we will describe the contents of our doctors' letters and challenges we encountered in the field of clinical natural language processing. Chapter 3.3 and Chaper 3.4 focus on our bootstrapping method and definition of our tasks respectively. We discuss highlights of our implementation and the applied training parameters in Chapter 4. We further describe and evaluate our experiments both quantitatively and qualitatively and reexamine our hypotheses in Chapter 5. In closing, we summarize our methodology and insights and offer perspectives for future work in Chapter 6.

2 Basics and Related Work

In this chapter we introduce basic relevant terms and concepts. After clarifying some additional basics we give address neural networks, explaining their history, design and current impact on technology. In Chapter 2.3 we describe how neural networks understand and represent text in vector spaces and use language models as examples for semantic vectors spaces in particular. Chapter 2.4 contains an overview of SECTOR's architecture [Arn+19] and further defines facets [Mac+18].

2.1 Basics

Levenshtein distance The difference between two words can be described as edit distance, which describes the amount of editing operations that are necessary, to change one word into the other. Editing operations consist of *insertion*, *deletion* and *substitution* [JM09, p. 108]. To give an example, to transform *word* into *old*, one would delete *w* and substitute *r* for *l*.

Assigning each transformation a different cost can tweak the edit distance to different tasks. *Levenshtein distance* describes the simplest weighting factor: on all editing operations a cost of 1 is applied. Following our example, Levenshtein distance between "*word*" and "*old*" is 2.

Information retrieval The field of *information retrieval* (IR) contains a wide variety of topics that deal with "storage and retrieval of all manner of media" [JM09, p. 801]. The task described in this thesis, extracting facets on medical resources, is therefore an IR task.

Vector space model As necessary step for many IR tasks texts, documents or queries are often represented as vectors of features in a hyperdimensional space. These vectors can reflect the words in a text (see bag-of-words), but also more abstract extracted features, e.g. an underlying meaning or topic.

Bag-of-words *Bag-of-words* describes an unordered set of words. Its simplest form (also known as *one-hot* encoding) represents a text or sentence as "vector of features,

each binary feature indicating whether a vocabulary word w does or doesn't occur in the context" [JM09, p. 675].

For example, the bag-of-words representation of the sentence "*I eat fish.*", based on the vocabulary $[I, am, eat, eating, ham, fish]$, would be: $[1, 0, 1, 0, 0, 1]$.

TF-IDF TF-IDF (*term frequency-inverse document frequency*) measures the frequency of a term in the dataset in relation to its frequency in a document, penalizing more frequent words. It's a "numerical statistic to indicate how important a word is to a document with respect to a collection of documents." [PG17, p. 349]

N-grams Patterson and Gibson [PG17, p. 353] define a n-gram as "a contiguous sequence of *n* items from a given sequence of text or speech."

For example, the set of character n-grams of the word *where* with $n = 3$ would be: [*whe, her, ere*].

Stemming The practice of *stemming* describes "the process of collapsing together the morphological variants of a word" [JM09, p. 806]. In IR systems stemming is often used to allow query terms to match documents which contain the terms' inflected forms.

Stock and *Stocks* present a prominent example. Both terms would collapse to *Stock* after stemming. This also presents a limitation of simple stemming algorithms: the term *Stocking*, albeit having a different meaning, could also be reduced to the term *Stock*.

2.2 Introduction to Neural Networks

Neural networks are on the forefront of a wide variety of research fields. Among the most famous examples is Facebook's facial recognition software presented by Taigman et al. [Tai+14]. They made headlines by achieving close to human-level performance and setting new records for facial recognition. More recently, Alibert and Venturini [AV19] used neural networks in the field of astrophysics to compute the mass of forming planets, replacing differential equations. They also evaluated several machine learning approaches and proved deep neural networks to work best for the task.

In the field of *natural language processing* (NLP) neural networks revolutionized for example text translation, e.g. *Google's Neural Machine Translation system* by Wu et al.

[Wu+16], and language modeling, starting with *word2vec* [Mik+13] and continuing to this day [Boj+16; Pet+18; Dev+18].

Despite these recent innovations, neural networks aren't new. Inspired by our very brains, they have been around for at least 50 years. Fashioned after biological neurons in the mammalian brain they consist of interconnected *nodes*. Like in real brains, these connections can be strengthened or weakened to achieve a learning process [PG17, p. 1].

However, as both considerable computational power and large amounts of data are necessary for neural networks, they were not feasible back then. This changed in the early 2000s and they proved to be an indispensable tool when working with so-called *big data*. Ever since we witness a Renaissance of machine learning and neural networks.

2.3 Text Representation in Semantic Vector Spaces

As mentioned before, neural networks revolutionized language models. These translate terms to arrays of floating point values, *vectors*, which can be interpreted by other neural networks. They present a fundamental form of text representation and we use the most prominent ones as example for text representation in semantic vector spaces.

Bag-of-words captures words The simplest approach for language modeling poses bag-of-words (see Chapter 2). However, bag-of-words does not carry any semantic meaning. This means, its representations show no correlation between semantically similar but syntactically different words, e.g. *Queen* and *King*.

Word2Vec captures semantic meaning Word2Vec [Mik+13] tries to fix this problem by learning word representations based on each words surroundings. This follows the assumption, that words with similar meaning are also used similarly. For an example, both *Queen* and *King* could both be used in a sentence like "... rule over a country."
In their respective word2vec vector space representation, *Queen* and *King* are therefore close to each other[1]. Their semantic similarity is represented in their corresponding vectors. Mikolov et al. [Mik+13] even showed that, using simple algebraic

[1] Of course, only if both words were encountered in such a way during training phase.

operations, “for example $vector(”King”) - vector(”Man”) + vector(”Woman”)$ results in a vector that is closest to the vector representation of the word *Queen*.”

FastText understands morphological similarities A language model using the word2vec method knows only the words it has seen however. Given a text that does not contain the word *Queens*, the plural form of *Queen*, word2vec would not know its meaning. Bojanowski et al. [Boj+16] present *fastText*, a language model, “which takes into account subword information.” In addition to learning each word, fastText also learns vector representations for each character n-gram. To apply this to our example: fastText could infer the meaning of the unknown term *Queens*, simply because it shares most of its characters with the known term *Queen*.

ELMo recognizes contextual meaning Both word2vec and fastText struggle with polysemy. Given a text which contains both the royal *Queen* and the British rock band *Queen*, word2vec and fastText could not distinguish between the two and would infer both as the same vector. Peters et al. [Pet+18] recently addressed this problem with their *Embeddings from Language Models* (ELMo). ELMo additionally interprets the context of a term. Given the sentence "Queen went on stage with their instruments," ELMo could infer that this *Queen* is the band and present a different vector than for the royal *Queen*.

Another prominent and recent word representation model *Bidirectional Encoder Representations from Transformers* (BERT), was presented by Devlin et al. [Dev+18]. It features similar functions to ELMo, but is more complex and computationally more expensive.

However, text representation is not limited to word representation, e.g. Paragraph Vector [LM14] learns to represent variable-length pieces of text, such as sentences, paragraphs and documents. Other text representations, like SECTOR, don’t aim to represent the text itself, but its topics.

2.4 Facet Segmentation and Classification with SECTOR

Another form of text representation presents SECTOR [Arn+19]. Instead of representing terms like language models, SECTOR’s embeddings aim to capture the topics of sentences over the course of a document. It further segments these topics at topic shifts to create coherent sections. These sections could, for example, be queried in

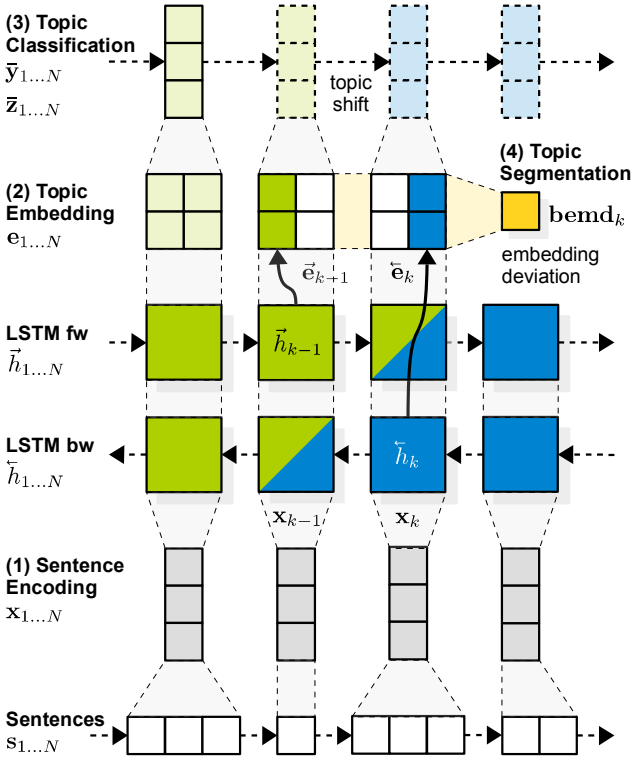

Fig. 2.1: Neural network architecture SECTOR. Image from Arnold et al. [Arn+19].

a information retrieval (IR) system like SMART-MD [Sch+18] to support medical personnel during decision-making.

The SECTOR architecture consists of four stages as shown by Figure 2.1. Arnold et al. [Arn+19] present two variations of SECTOR, which differ in their goal and architecture during the third stage.

(i) **Sentence encoding.** To represent the variable-length sentences as fixed-length vectors, Arnold et al. [Arn+19] offer two approaches: a weighted bag-of-words scheme as baseline, or a distributional sentence representation based on pre-trained word2vec models. For the weighted bag-of-words, the words of each sentence are represented by their one-hot encoded vector multiplied with their respective tf-idf (compare Chapter 2.1) score. The sentence representation follows the strategy of Arora et al. [Aro+17], which uses at its core a probability-weighted sum of word embeddings.

(ii) **Topic embedding.** The second stage produces a "dense distributional representation of latent topics for each sentence in the document" [Arn+19]. To achieve this, the architecture consists of two *long short-term memory* (LSTM) layers [HS97] with forget-gates [Ger+00].

LSTM networks add a memory cell to each node, that allows them to retain information over several time-steps [PG17, p. 150]. The addition of forget gates allowed them to learn to reset their internal state values at appropriate times. Before, the internal state values would grow indefinitely for continual input streams and cause the network to break down [Ger+00]. LSTMs are therefore excellent at capturing long-range dependencies. This means, that SECTOR's architecture accumulates topical meaning over the course of the document, a sentence is not seen as an atomic unit. Additionally, the LSTMs traverse the document opposite direction: one reads in forward direction, the other in backward direction. This is based on the assumption, that a human reader understands text not just based on the text before, but also the text afterwards. Graves and Schmidhuber [GS05] even show, that both directions are equally important and dub this configuration *bidirectional LSTM* (BiLSTM).

(iii) **Topic classification.** Arnold et al. [Arn+19] add a classification layer on top of the BiLSTM architecture. For the SECTOR *topics* task, which presents a multi-class single-label problem, this output layer uses softmax activation. For the *headings* task, a multi-class multi-label problem, it uses sigmoid activation and applies a ranking loss function. Multi-class single-label tasks aim to find a single label out of several classes as label for the input (*multinomial labeling systems*). Softmax activation represents this goal, since it returns a probability distribution over mutually exclusive output classes. Sigmoids on the other hand output an independent probability for each class. This matches the multi-class multi-label task, as its goal is to select several matching labels instead of just one [PG17, pp. 67-68].

(iv) **Topic segmentation** The last stage uses both the BiLSTM layers as well as the classification layer to segment the document. Arnold et al. [Arn+19] propose an edge detection approach that focuses on the difference of the left and right topic context over time. They call this "geometric mean of the forward and backward distance" the *bidirectional embedding deviation* (bemd).

Using these topics in an IR system allows for a more faceted answer when retrieving answers. This surpasses traditional factual answer retrieval and is known as complex answer retrieval (CAR). MacAvaney et al. [Mac+18] describe CAR as "process of retrieving answers to questions that have multifaceted or nuanced answers." Or simply put, questions that cannot be answered by a simple 'yes' or 'no'.
MacAvaney et al. [Mac+18] first characterized CAR facets. They discern two kinds of facets: structural and topical. Structural facets can apply to many similar topics, they represent the structure of a document. Topical facets on the other hand are specific to the question's topic. For example, given two headlines *diagnosis* and *gastroscopy* in

a Wikipedia article regarding disease: as presumably all diseases contain a *diagnosis* section, it presents a structural facet, while *gastroscopy* would be limited to articles about stomach disease or the like, it therefore presents a topical facet.

2.5 Summary

In Chapter 2 we discussed basic techniques and terms relevant to this thesis. Afterwards we gave a brief introduction into neural networks, showing their current impact on research, discussing their age and basic concept and explaining why they seem to be a recent innovation to many. We further explained the concept of text representation in semantic vector spaces, using three language models as example to show how different approaches can modify text representation. We finally describe SECTOR's architecture and two variations and differentiate between structural and topical facets.

3 Facet Detection on Medical Resources

On the following pages we describe the structure and appearance of the clinical resources in our data. We further explain why we can't use other available datasets and what challenges can be expected when trying to process clinical resources in Chapter 3.2.
Following that, Chapter 3.3 describes our approach to prepare and bootstrap training data using our raw clinical data. In Chapter 3.4 we build on the basics explained in the last chapter and expand on how we train our specialized word embeddings and how we use the two variations of SECTOR to solve the task of detecting structural and topical facets on medical resources.

3.1 Clinical Doctor's Letters from Charité Berlin

The dataset for this task consists of 7,553 discharge letters courtesy of *Charité Berlin*'s *Medical Department, Division of Nephrology and Internal Intensive Care Medicine*. They all feature letter structure: a head of the letter, a short salutary address, the body and a short, formal ending. See Table 3.1 for the most common structural headings in the body. A large proportion of these headings are auto-generated by software, but clinicians can alter and adjust them afterwards. This leads to a large variety of different headlines with similar meaning and a *vocabulary mismatch problem*.

3.1.1 Vocabulary Mismatch Problem

Furnas et al. [Fur+87] first observed "that people use a surprisingly great variety of words to refer to the same thing." They dubbed this the *vocabulary problem*[1]. Shekarpour et al. [She+17] name two key causes: *inflectional form* and *lexical form*. Inflectional forms include variations "of a word for different grammatical categories such as tense, aspect, person, number, etc.", while *lexical form* "relates words based on lexical categories" [She+17]. Both of these apply here, e.g. `Diagnose` (*diagnosis*) and `Diagnosen` (*diagnoses*) are inflectional forms, `Lungenfunktionsprüfung` (*lung function test*) and `Lungenfunktionsuntersuchung` (*lung function examination*)

[1] Later publications refer to the same problem as *vocabulary mismatch problem*.

structural section	description
"Diagnose/n"	a section regarding the patient's *diagnosis/diagnoses*
"Anamnese"	an *anamnesis* or medical history section
"Status bei Aufnahme"	a report of the patient's condition at arrival (*status at admission*)
"Labor"	a report of *laboratory* values, often a table or list of values
"Diagnostische Maßnahmen", "Bildgebende Verfahren"	any number of reports of *diagnostic measures* or *imaging methods*
"Konsil"	a report of a different medical department or specialist (e.g. an ophthalmology *consult)*
"Therapie und Verlauf"	a text describing *therapy and course* of the patients stay, treatment and sickness
"Medikation", "Medikation bei Entlassung"	and a report of the patient's *medication* or *medication administered at point of discharge*, often a table containing name, dose, and frequency of administration of each drug

Tab. 3.1: Most common structural section headings in order of most common appearance. The average letter contains 16.4 sections.

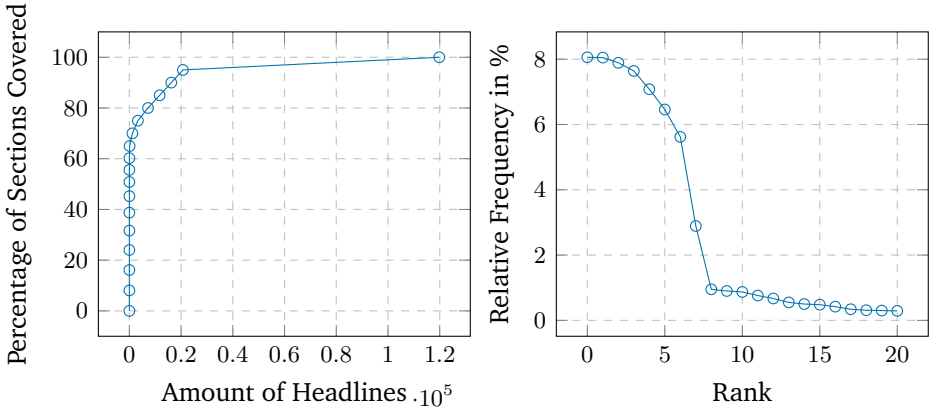

Fig. 3.1: Cumulative distribution of 119,839 headlines (Note that the function is only defined in steps of at least 5%.) and relative frequency of top 20 headings. We observe a *Zipfian* distribution: the top 18 out of 119,839 headlines cover 60% of all sections.

are lexical forms. Furthermore we observe abbreviations and erroneous writing: `Lungenfunktionsuntersuchung` is commonly shortened to `Lufu`, while `Lufo` is a typographical error. Kacprzyk and Fedrizzi [KF12] attributes the vocabulary problem to each environments own specialized terminology, while Furnas et al. [Fur+87] and Barrows Jr et al. [BJ+00] claim that these differences persist on an individual level.

We therefore observe a *Zipfian* distribution (compare Figure 3.1), that is defined by *Zipf's law*: if t_1 is the most common term in the collection, t_2 is the next most common, and so on, then the collection frequency cf_i of the ith most common term is proportional to $1/i$:

$$cf_i \propto \frac{1}{i} \tag{3.1}$$

This is a commonly used model of the distribution of terms in a collection [Man+08, p. 82].

3.2 Challenges in Clinical Facet Detection

During both training and evaluation phase a greater variation of data proves helpful. While more diversified training data helps the neural network in understanding underlying core concepts, cross dataset validation is an important step in validating any NLP method. In this chapter we elaborate several factors that pose a challenge when working with medical resources in general and with German medical resources in particular. We explain how general purpose natural language processing (NLP) datasets do not transfer well to clinical NLP tasks, why clinical training data is sparse

doctor's letter	Wikipedia article
Brief Kopf	Einteilung und Ursachen
Brief Anrede	Risikogruppen
Diagnose/n	Symptome
Anamnese	Untersuchungen
Status bei Aufnahme	Therapie
Labor	Pflege
Diagnostische Maßnahmen, Bildgebende Verfahren	Mögliche Komplikationen
Konsil	Prognose
Therapie und Verlauf	Vorbeugung
Medikation, Medikation bei Entlassung	Literatur
Brief Schluss	Weblinks

Tab. 3.2: Example outline for a doctor's letter and a Wikipedia article [Wik19b]. While some sections seem to have a counterpart (e.g. `Symptome` and `Status bei Aufnahme`, `Diagnostische Maßnahmen` and `Untersuchungen`) others are unique to their source (e.g. `Risikogruppen` and `Vorbeugung`, `Labor` and `Medikation`).

and how working with resources from a highly specialized field proves difficult to people from different fields of research.

3.2.1 Semantic Mismatch with `WikiSection`

A great complimentary dataset to enrich the small number of `doctors' letters` at hand would be the German diseases part of `WikiSection`. The dataset proposed by Arnold et al. [Arn+19] consists of articles from two domains of the English and German Wikipedia: *diseases* and *cities*. It is designed for the task of facet extraction and therefore annotated and segmented into structural and topical facets.
However, doctors' letters and Wikipedia articles serve a very different purpose: while Wikipedia articles are informative and describing, our discharge letters are summaries or reports. This leads both to *structural* and *vocabulary mismatches*.

Structural Mismatch The structural mismatches become apparent, when trying to match Wikipedia's articles' structural headings to those of doctors' letters as in Table 3.2. While Wikipedia articles lack any kind of letter structure the letters never contain an abstract or, e.g. sections about history, causes or epidemiology. Wikipedia articles also don't report data of individual patients like laboratory values or blood

pressure measurements. Some headings seem to be matching at first, however turn out to be vocabulary mismatches.

Vocabulary Mismatch The vocabulary mismatches are mainly of lexical forms (see Chapter 3.1.1). To serve an example, the word `Diagnose` (English: *diagnosis*) is used in both resources as frequent headline: While in clinical records `Diagnose` refers to the actual diagnoses of this patient (*what* was diagnosed), Wikipedia authors use it more along the lines of *diagnostics*, i.e. *how* to diagnose.

Together these mismatches constitute a great dissimilarity between the two resources and rule out the benefit of a combined training dataset for the task at hand.

3.2.2 Missing Training Data

As the combination of topic or facet segmentation and classification is a novel task presented by Arnold et al. [Arn+19] training data is sparse. They point out several datasets solving parts of the task, yet none of them are applicable for the joined task of topic segmentation. According to Chapman et al. [Cha+11] and Starlinger et al. [Sta+17], training data for medical NLP is even less accessible. The best matching dataset would be the discharge letters used by Tepper et al. [Tep+12], which was not released to the public and features English language.

No Publicly Available Datasets The lack of access to data and annotated datasets for clinical NLP is one of the major issues in this field. Chapman et al. [Cha+11] not only name patient privacy but also "worry about revealing unfavorable institutional practices" as reasons. Starlinger et al. [Sta+17] exclaim that this problem is even worse for German clinical NLP resources due to stricter European and German privacy regulations. While this issue is being addressed by Starlinger et al. [Sta+17], Schlünder [Sch15], and Chapman et al. [Cha+11], it is far from solved.

The doctors' letters used in this work fall under a non-disclosure agreement and cannot be made publicly accessible either.

3.2.3 Ambique Medical Language

Letters contain ambiguous medical terms Starlinger et al. [Sta+17] elaborate how the "clinical jargon used in a medical note not only depends on the respective medical specialty and document type, but also on the concrete individual institution." In the case of a decentralized organization such as the *Charité*, the particular clinic also factors in. Furnas et al. [Fur+87] and Barrows Jr et al. [BJ+00] note how

abbreviations and terms differentiate even at each respective clinician's level.
For an example in the dataset at hand, radiology department would call their findings and results `Kommentar` (English: *comment*), while the central laboratory would use the term `Befund` (*diagnostic findings*). However both these departments occasionally also use the term `Beurteilung` (*assessment*). Medical professionals will explain `Befund` and `Beurteilung` are fundamentally different, the first referring to the status of the patient and the latter to the medical assessment of said status. However `Befund/Beurteilung` is a common headline and proves how difficult distinction is even for medical experts.

Sections contain ambiguous content For non-medical personnel even classifying whole sections proves difficult. One example we encountered is the headline `Gastroskopie` (*gastroscopy*[2]). A typical `Gastroskopie` section in a doctor's letter consists of the clinician's description of his or hers visual observations. Given this knowledge and the choice between classification as image methods or as diagnostic measures, this might incline lay readers to classify it as imaging method. However, this procedure involves inserting an endoscope through the mouth of the patient to visualize the upper part of the gastrointestinal tract[3]. It is therefore, while producing images, undoubtedly an examination and classifies as diagnostic measure.
To present a second example, one of the most common headings in our data is `Labor` (*laboratory*). Several other headings, e.g. `Antibiogramm` (*antibiogram*), describe examinations that take place in a laboratory. However, since this is not the main laboratory, it does not belong to the structural class `Labor`, but rather to diagnostic measures. A distinction an individual unfamiliar with hospital operations could never make.[4]

Topical vs Structural Facets Even identifying `Labor` as structural facet proves difficult. MacAvaney et al. [Mac+18] define structural headings as "general question facets that could be asked about many similar topics." As similarity is a range, this definition is vague. Wikipedia articles about diseases seem similar to doctors' letters, yet differ in structure. On the other hand within the doctors' letters one might find a subset of patients with pulmonary diseases. In such a subset `lung function examination` could prove to be a structural facet. The aforementioned distinction between different laboratories might also be biased by the clinician, the medical division or hospital. Thus disambiguation between topical and structural facets is not just dependent on the dataset and the task, but also subjective to personal judgment.

[2] English Wikipedia features a section regarding alternative names for gastroscopy, adding to the point.
[3] Wikipedia. *Esophagogastroduodenoscopy*. en. Page Version ID: 877006453. Jan. 2019. URL: https://en.wikipedia.org/w/index.php?title=Esophagogastroduodenoscopy&oldid=877006453 (visited on Mar. 12, 2019).
[4] This example might be specific to the hospital at hand. If so, this supports the point even more.

3.2.4 Highly Specialized Domain Knowledge

To reach an educated decision deciding between structural and topical facets trained medical advisers are indispensable. German medical doctors regular course of study spans more than 6 years, often followed by another 5-6 years training to become a medical specialist.[5] During this time they acquire knowledge and vocabulary surpassing lay comprehension. As recent developments in medical and clinical sectors give patients access to their medical records, this knowledge gap is addressed by McCray [McC05] and Mossanen et al. [Mos+14]. Chen et al. [Che+18] tries to solve this problem by linking medical terms to lay definitions via NLP. However, until these efforts succeed highly specialized knowledge is necessary to understand and work with medical resources.

3.3 Preparing and Bootstrapping Training Data

As shown in Chapters 3.1 and 3.2 we don't have access to ready training data. Since manual generation of training data is costly and not feasible for large corpora, bootstrapping is common practice to generate labelled data from plain-text collections. Agichtein and Gravano [AG00] and Gupta and Manning [GM14] propose self-learning pattern detection algorithms to bootstrap entity extraction datasets. Our task of detecting headlines and segmenting doctors' letters proves considerably easier, as most of the headlines are presented in a standardized form.
Table 3.3) presents a shortened example. We first use simple rules to detect salutary address and formal ending of the letter. Afterwards we chunk the letter body into sections to receive a sectionized dataset annotated with their respective headlines. Additionally we strip all section headings, but keep newline characters.

Our further approach resembles Tepper et al. [Tep+12]'s procedure, but instead of random sampling we employ a frequency based algorithm.

3.3.1 Clinical Resources Require Specialized Structural Facets

Due to the structural mismatch with *Wikisection*, we can not rely on structural facets described by Arnold et al. [Arn+19]. Therefore, we need to formulate and normalize our own.

[5] Wissenschaftsrat. *6825-05.pdf*. Tech. rep. Drs. 6825/05. 2005; Charité-Universitätsmedizin Berlin. *Modellstudiengang Humanmedizin*. de. URL: https://www.charite.de/studium_lehre/studiengaenge/modellstudiengang_humanmedizin/ (visited on Mar. 11, 2019).

`letter head`	Epikrise von <name>, geb. <date> *Patient and hospital addresses omitted.* E P I K R I S E Patn. <name>, geb. <date>, wohnhaft <location>
`salutary address`	Sehr geehrte Kollegin, nachfolgend möchten wir über o. g. Patientin berichten, die sich am <date> in unserer stationären Behandlung befand.
`section`	Diagnosen: Dislokation des CAPD-Katheters bei Peritonealdialyse Terminale Niereninsuffizienz Renale Anämie
`section`	Anamnese vom <date>: Der Patient wurde am <date> auf die Station übernommen...
	More sections omitted.
`formal ending`	Mit freundlichen, kollegialen Grüßen Univ.-<name> <name> <name> Klinikdirektor Oberarzt Stationsarzt

Tab. 3.3: Shortened example doctor's letter. Underlined terms have been detected by regular expressions to segment and annotate the letter.

As first step of mapping original headlines to normalized structural facets we reduce the amount of classes using a custom algorithm inspired by stemming algorithms. Xu and Croft [XC98] propose a stemming algorithm using cooccurrence of word variants. We simplify this approach, but keep the main assumption: the correct stem a occurs in the same text windows. Our simplified approach sets the window size to contain all of the headlines and replaces the em metric with a frequency based scoring metric. Furthermore, we do not stem singular words but headlines. To reflect this decision and differentiate it from conventional stemming we call headline-stems *archetypes*.

More formally, for a headline $h \in H$ the set of possible archetypes is defined as follows:

$$A_h = \{\, a \in H \mid fsm(a, h) \,\} \tag{3.2}$$

where H is the set of all headlines and $fsm(a, h)$ equals a *fuzzy substring matching algorithm* such that a is a fuzzy substring of h[6]. We further define the following metric to score a headline $a \in H$ based on its occurrences as archetype:

$$score(a) = |\{\, h \in H \mid fsm(a, h) \,\}| \tag{3.3}$$

and define the archetype $arch_h$ as

$$arch_h = \arg\max_{a \in A_h} score(a) \tag{3.4}$$

See Chapter 4.1 for more information.
While this algorithm is an approximation and susceptible to error (as shown in Table 3.4) it provides an overview of the most common topics in the headlines. It further reduces them to an amount manageable for manual classification. We choose the 600 most frequent archetypes, which cover 94% of all headlines. We build our ontology with these 600 archetypes and the help of a medical professional.

3.3.2 Validation with a Medical Professional

As medical resources employ highly specialized domain knowledge and nomenclature, we enlist the help of a clinician. To build and validate our ontology, we meet her twice over a three step process:

(i) **Classify the headlines with a medical professional** Given the task to categorize the archetype headlines she would group headlines to categories. We do not instruct her regarding structural or topical facets. This guarantees that

[6] We allow a small Levenshtein distance to accommodate for typographical error.

archetype	CT
headlines	*CT- Low Dose-Technik*
	CT des Oberbauches und des Beckens
	CT des Bauchraumes und Beckens nach KM-Gabe
	CT des Bauchraumes und des Beckens
	CT des Bauchraumes (nativ)
	CT des Bauchraumes und Beckens
	CT des Bauchraums
	CT des Beckens
	Rectalabstrich

Tab. 3.4: Example showing a variety of original headlines reduced to the archetype headline CT (*CT scan*). Note how the typographical error Rectalabstrich (*rectal smear*) was also reduced to CT, albeit having nothing in common with a CT scan.

she creates categories unaffected by any notions of structural and topical. She therefore creates mixed categories. Examples can be seen in Table 3.5.

(ii) **Merge topical facets to structural ones** We decide whether a facet is topical or structural according to our knowledge of the doctors' letters and the insight gained by the meeting. We identify structural patterns in the remaining topical facets. When applicable, the German procedure classification of *Deutsches Institut für Medizinische Dokumentation und Information*[7] can provide structural categorization.

(iii) **Present our structural classification to a medical professional** Given our structural facets, we ask for any elements of a doctor's letter that she thinks lack representation and validate our merging step.

This process aims to allow the professional to create classes with as less bias as possible. At our first meeting she is not lead by possible interpretations of *structural* and *topical*, but instead produces valuable insight into more detailed classification.

We use both levels of detail to create our ontology. An overview of the process so far can be seen in Figure 3.2.

[7] *DIMDI - OPS Version 2019*. URL: https://www.dimdi.de/static/de/klassifikationen/ops/kode-suche/opshtml2019/ (visited on Mar. 16, 2019).

class	archetype headlines
`Labor`	*Labor, Schilddrüsenparameter, Schilddrüsenhormone, Virologie, Immunologie, Serologie, Autoantikörper*
`Diagnose`	*Diagnose, Diagnosen, Hauptdiagnose, Nebendiagnose, Weitere Diagnosen*
`Beurteilung`	*Kommentar, Beurteilung*
`CT`	*CT, CT des Kopfes*
`Szintigraphie`	*Szintigraphie, Skelettszintigraphie, Myokardszintigraphie, Schilddrüsenszintigraphie*

Tab. 3.5: Excerpt of 22 mixed classes defined by the medical professional. Later `CT` (*CT scan*) and `Szintigraphie` (*scintigraphy*) will be merged into the structural topic `Bildgebende Diagnostik` (*imaging methods*).

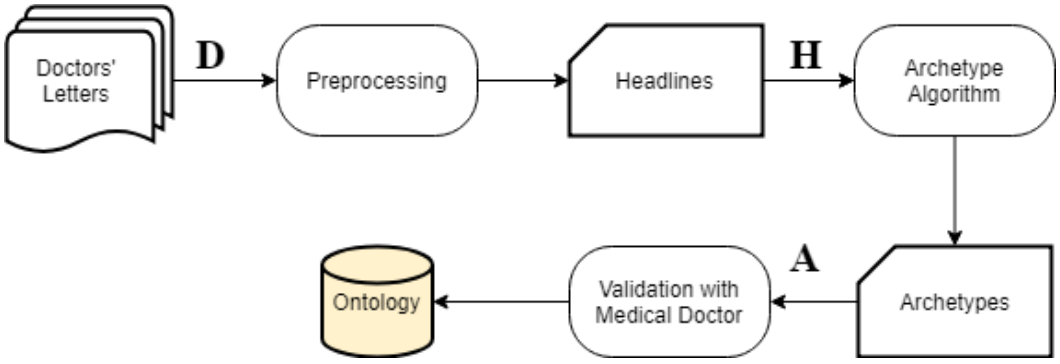

Fig. 3.2: Ontology creation process.

3.4 Extracting Medical Facets with SECTOR

Following our hypotheses we train specialized word embeddings for the clinical domain using the well established word2vec method [Mik+13] as well as the fastText method [Boj+16]. We further use our ontology to model our facet extraction tasks as problems suitable for SECTOR: multi-class single-label for structural facets and multi-class multi-label for topical facets.

3.4.1 Training German Clinical fastText Embeddings

As shown by Sheikhshab et al. [She+18] and Lee et al. [Lee+19] biomedical NLP tasks profit from domain-specific word or language representations. While both used more sophisticated language modeling approaches (ELMo and BERT) we opted for the computationally less expensive fastText [Boj+16] and word2vec [Mik+13] because of the available resources at Charité Berlin. We further assume that contextualized word embeddings provide little improvement when working with specialized documents like medical resources. Subword information as captured by fastText on the other hand has been proven by Bojanowski et al. [Boj+16] to perform better for small datasets and complex, technical and infrequent words. Both prominent characteristics in clinical resources, especially German ones.

Grave et al. [Gra+18] distribute a pre-trained German model trained on *Common Crawl*[8] and *Wikipedia*[9]. However, re-training or fine-tuning fastText embeddings is not possible as of this writing. We therefore train our own models. Since Charité's doctors' letters show little variance we use a joined dataset consisting of the letters and the German diseases Wikipedia articles presented by Arnold et al. [Arn+19]. This approach, albeit utilizing a considerably smaller dataset than the general purpose model (see Table 3.6), should give us highly specialized clinical embeddings combined with a general understanding of common biomedical terms.

3.4.2 Modeling Structural Facets as Multi-Class Single-Label Problem

Using the ontology created with the help of a medical professional we define 14 structural facets (compare Figure 3.3). These structural facets are distinct and in most cases contradict each other, e.g. a `letter head` can not be a `formal ending` or a `diagnosis` section at the same time. In other words, they are *mutually exclusive*

[8] *Common Crawl*. URL: http://commoncrawl.org/ (visited on Mar. 9, 2019).
[9] *Wikipedia*. URL: https://www.wikipedia.org/ (visited on Mar. 9, 2019).

dataset	# tokens	# words
doctor's letters	11,432,445	253,423
Wikipedia diseases [Arn+19]	2,220,688	31,134
Wikipedia + Common Crawl [Gra+18]	67,032,828,416	22,773,218

Tab. 3.6: Comparison of the size of the training corpora. *# words* displays the number of words that appear at least five times in the dataset.

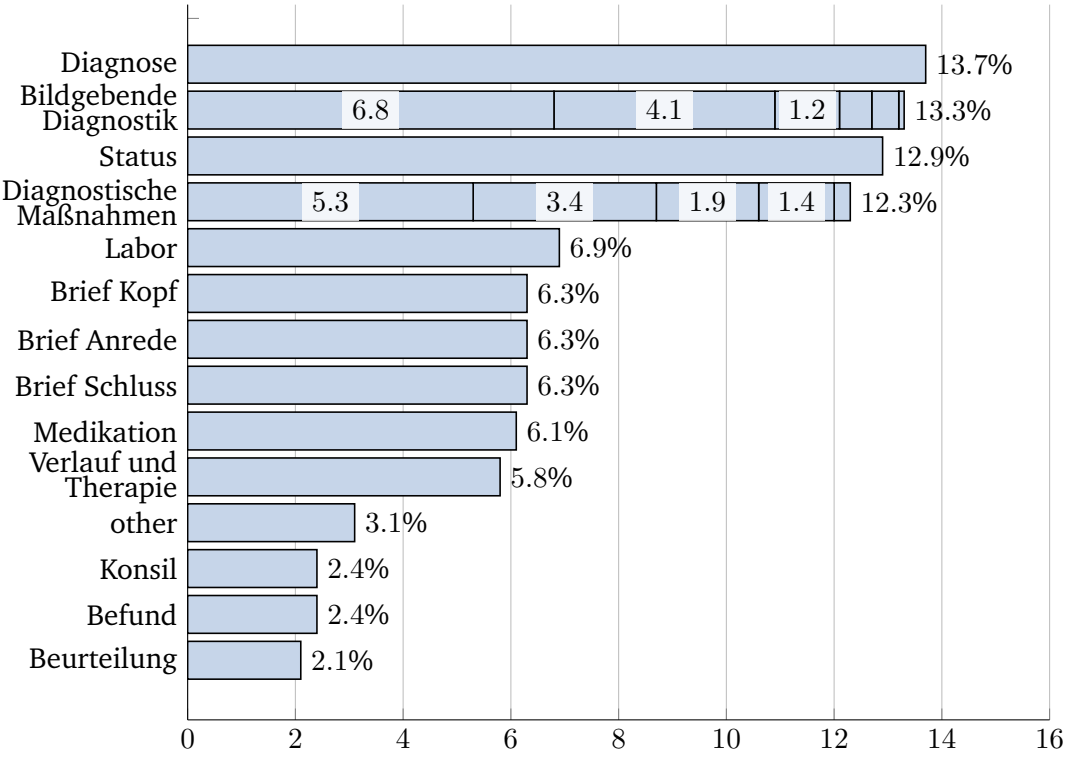

Fig. 3.3: The 14 structural facets and their relative text lengths. Note how during the first meeting the professional defined 22 facets (plus a catch-all category *other*). To gain pure structural facets we merged `Sonographie` (6.8%), `Röntgen` (4.1%), `CT` (1.2%), `MRT` (0.6%), `Szintigrpahie` (0.5%) and `Angiographie` (0.1%) to `Bildgebende Diagnostik` and `EKG` (5.3%), `Untersuchung` (3.4%), `Probe` (1.9%), `Anatomie` (1.4%) and `Histologie` (0.3%) to `Diagnostische Maßnahmen`.

and present a single-label problem [Man+08, p. 282].
The facets `Diagnostische Maßnahmen` (*diagnostic measures*) and `Bildgebende Diagnostik` (*imaging methods*) can be considered an exception, as they do not contradict each other. One might argue that `Bildgebende Diagnostik` is a subtopic of `Diagnostische Maßnahmen`. However, following the definition presented by MacAvaney et al. [Mac+18], structural facets are "headings that serve a structural purpose for an article—general question facets that could be asked about many similar topics." Since `Bildgebende Diagnostik` sections occur in nearly all letters and thus can be asked about all our doctors' letters we decide it to be a structural facet.

We therefore formally define the task of extracting structural facets following Arnold et al. [Arn+19]: we too define a document, a doctor's letter, $D = \langle S, T \rangle$ consisting of N consecutive sentences $S = [s_1, ..., s_N]$ and empty segmentation $T = \varnothing$ as input. We also assume a distribution of structural facets e_k for each sentence s_k that changes over the course of the document.

The task is to split D into a sequence of distinct structural sections $T = [T_1, ..., T_M]$, so that each predicted section $T_j = \langle S_j, y_j \rangle$ contains a sequence of coherent sentences $S_j \subseteq S$ and a structure label y_j that describes the prevalent structural facet in these sentences. For our example structure in Table 3.3, the sequence of structure labels is $y_{1...M} \approx$ *[Brief Kopf, Brief Anrede, Diagnose, Anamnese, ..., Brief Schluss]*.

In light of the downstream information retrieval task, when breaking all headlines down to 14 classes a lot of information is lost, e.g. any hierarchical structuring. To address this problem, we additionally model the original headlines as a multi-class multi-label problem.

3.4.3 Modeling Topical Facets as Multi-Class Multi-Label Problem

The original headlines contain essential detail to the content of their sections. We also gained insight into a more instinctive classification during our first meeting with the clinician. So using just the structural facets looses a lot of information and detail. `Bildgebende Diagnostik` for example combines a variety of different imaging methods like `Röntgen-Thorax` (*chest radiograph*), `CT` (*CT scan*) or `MRT` (*magnetic resonance imaging*, or *MRI*). To solve this problem, Arnold et al. [Arn+19] present SECTOR's second variation, the SECTOR *headings* task. They propose using all words in the original heading as multi-label bag.

However, original headings like `Röntgen-Thorax` do not reflect this hierarchy either. Different abbreviations or typographical errors further hinder this approach, e.g.

`Rö-Thorax` would not be matched to `Röntgen-Thorax`. Yet we can solve this problem using the more detailed information we gained during our first meeting with the medical professional and the finished ontology. By concatenating these three levels of detail `Rö-Thorax` becomes `Bildgebende Diagnostik | Röntgen | Rö-Thorax` and `Röntgen-Thorax` becomes `Bildgebende Diagnostik | Röntgen | Röntgen-Thorax`.

Using these modified headings we can now apply the headings approach of Arnold et al. [Arn+19]. We assign all words z in the heading h, such that $z_i \subset h_j$, as multi-label bag over the original heading vocabulary Z. We adjust our structural facet extraction task: Using the same doctor's letter $D = \langle S, T \rangle$ composed of N consecutive sentences $S = [s_1, ..., s_N]$ and empty segmentation $T = \varnothing$ as input, we assume a distribution of topical facets for each sentence s_k that, again, changes over the course of the document.

While the task to split D into a sequence of distinct structural sections $T = [T_1, ..., T_M]$ stays the same, each topic $T_j = \langle S_j, Z_j \rangle$ now consists of a sequence of coherent sentences $S_j \subseteq S$ and a ranked sequence Z_j, which contains all elements of Z. The ground truth labels for an example section `Bildgebende Diagnostik | Röntgen | Röntgen-Thorax` would then be {*bildgebende, diagnostik, röntgen, röntgen-thorax*}.

These modified headings do neither lose topical information nor suffer from a lack of normalization due to typographical errors or an individual's idiosyncratic terms when using individual words as multi-label tags. Therefore, we can capture the ambiguous topic facets present in each heading with this approach.

An overview of the process so far can be seen in Figure 3.4.

3.5 Summary

In this chapter we described the structure of our dataset of doctors' letters and showed that, although parts of the letters are computer generated, a vocabulary mismatch problem occurs. We further elaborated in Chapter 3.2 common issues when working with medical resources in general and in our case specifically: how the only other publicly available dataset for this task, albeit medical, does not match our dataset, why it is the only other accessible dataset, how medical terms are ambiguous and that you need medical expertise to understand clinical resources. As this induces a need for training data, we explained our approach to generate training data using Charité's doctors' letters and our validation process with the help of a

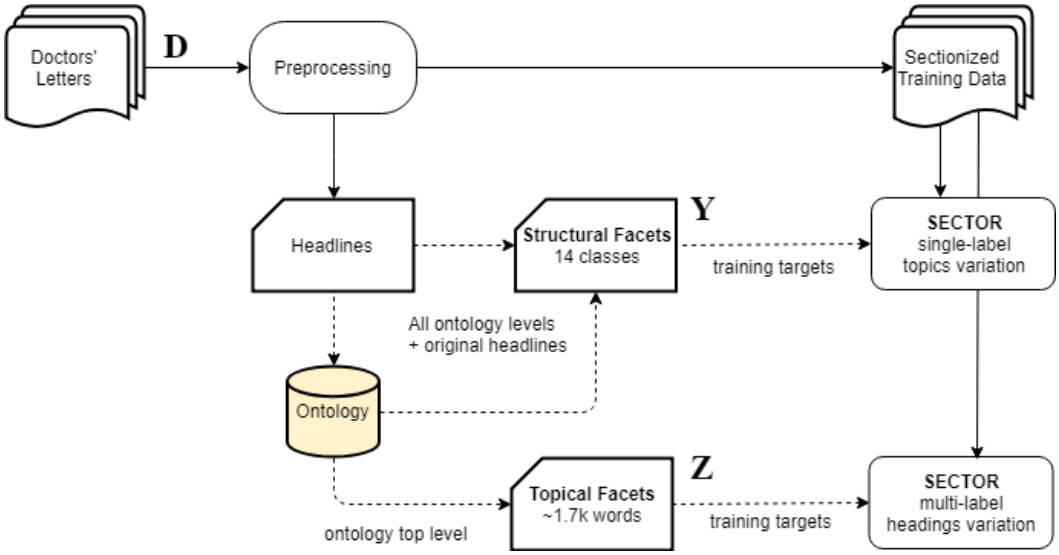

Fig. 3.4: SECTOR training process.

medical professional. In Chapter 3.4 we explained our methods for training word embeddings and used the annotated dataset and the insight gained by meeting the clinician to model our task into a multi-class single-label problem to segment and classify the structure of a doctor's letter as well as a multi-class multi-label approach to extract topical nuances.

4 Implementation

The preceding chapter formally defined our methodology. Basing on that, we now implement our approach and present relevant parts of our code. In Chapter 4.1 we elaborate our segmentation process and further explain our archetype algorithm using code examples. To utilize the bootstrapped training corpus we need to adjust the existing SECTOR training code to our needs. This is described in Chapter 4.2. Finally, we explain the training parameters for each model in Chapter 4.2.

4.1 Archetype Algorithm

Following the approach defined in Chapter 3.3, we implemented our custom headline level stemming algorithm, the archetype algorithm.

Headline reader As preprocessing step, we create a reader class as subclass of TeXoo's `RawTextDatasetReader`. Utilizing its basic structure, we override just the `readDocumentFromFile()` method. After reading the plain document, it detects start of header and footer and segments the region in between using regular expressions. It further annotates the segments using given labels or the detected headline. Additionally it allows to modify the detected headline.
Since sectionizing is based on regular expressions, we dub this reader the `RegexReader`. See Table 4.1 for the regular expressions used for annotating the doctors' letters.

Using the annotated letters we generate the list of all possible archetypes A_h per heading h using the `findArchetype()` method (Listing 4.1). A simple occurrence count in all lists produces our $score(h)$. We assume the highest scoring possible archetype to be the correct archetype $arch_h$. For a small example, refer to Table 4.2.

Label detection
`/(^(.*):\s *$\s *)((?:[\s \S ]*?(?=^.*:\s *$))|[\s \S ]*)/m`
Header detection
`/(Sehr geehrte[\s \S ]+?)(?=^.*:\s *$)/mi`
`/.*(?:Zeichen[\s \S ]+?)^(?:Patn?\..+geb.+$)\s + ([\s \S ]+?)(?=^.*:\s *$)/mi`
Footer detection
`/(Mit (?:freundlichen|kollegialen)[\s \S ]*)/i`
`/(^(?:Univ\.-)[\s \S ]*)/mi`
Header Modification
`/(?:\h *(?:vom|v\.|zum|am|bis|von)*\h * (?:(?:[x\d ]{1,2}\.){1,2}(?:\d {2}){0,2}|heute)-?){1,2}/i`
`/,?(?:\h *\d {1,2}:\d {2}(?:\h *uhr(?:\h *(?:bis|-))?)?){1,2}/i`

Tab. 4.1: Regular expressions used for segmenting the raw letters. Header detection aimed for a common salutary address or an auto-generated line right before the header. As footer the algorithm detected common endings of a letter. Header modification removed time and date specification.

```
public static String findArchetype(String str1, String str2) {
  if (str1.equals(str2)) return str1;
  if (str1.isEmpty() || str2.isEmpty()) return null;

  final String shorter, longer;
  if (str1.length() < str2.length()) {
    shorter = str1;
    longer = str2;
  } else {
    shorter = str2;
    longer = str1;
  }

  if (containsWithDifference(longer, shorter,
          (int) Math.min(shorter.length() * 20.0 / 100.0, 1)))
    return shorter;
  else
    return null;
}
```

Listing 4.1: The findArchetype() method. For longer words we allowed for a small Levenshtein distance to accommodate typographical errors. Note that `containsWithDifference()` equates to $fsm(a, h)$ in Chapter 3.3.

headline h	**possible archetypes** A_h	$score(h)$
`Röntgen`	{Röntgen}	3
`Röntgen Thorax`	{Röntgen, Röntgen Thorax}	2
`Röntgen-Thorax bed side`	{Röntgen, Röntgen Thorax, Röntgen-Thorax bed side}	1

Tab. 4.2: Example for the archetype algorithm. Since `Röntgen` features the highest score and is a possible archetype for all three, each one simplifies to `Röntgen` despite typographical errors and additional words.

4.2 SectorTrain

Since we utilize the TeXoo[1] framework a working `TrainSectorAnnotator` class was given. However, at the beginning of this project TeXoo's TrainSectorAnnotator was lacking command line options to adjust language models, bloom filters or switch between each variation. We therefore modified command line options and adjusted the class accordingly. These adjustments have been integrated into TeXoo in release 1.1.2.

4.2.1 K-Fold Evaluation

We further integrate k-fold evaluation as new feature. To ensure that no lingering side effects between each training iteration persist, we terminate the *Java Virtual Machine* after each training.
We therefore split the training data into k fragments with our target test set size and one fragment containing the remainder. We then use a shell script to train SECTOR using all fragments but one which we use as test set.

4.3 Training Parameters

Training parameters, or hyper-parameters, adjust several aspects of the training as well as of the architecture of a neural model. They often make the difference between mediocre and state-of-the-art results [Hut+14]. Due to our limitations (see Chapter 1.3.2), we focus on optimizing just one of our models, fastText.

[1] Sebastian Arnold. *TeXoo – A Zoo of Text Extractors. Contribute to sebastianarnold/TeXoo development by creating an account on GitHub*. original-date: 2018-07-19T08:42:32Z. Feb. 2019. URL: https://github.com/sebastianarnold/TeXoo (visited on Mar. 30, 2019).

hyperparameter	topics	headings	word2vec
LSTM layer size	256	256	n/a
embedding layer size	128	128	256
learning rate	0.01	0.001	0.025
min learning rate	n/a	n/a	0.001
dropout	0.5	0.0	n/a
batchsize	16	16	16
epochs	6	6	1
iterations	n/a	n/a	5
minimum word frequency	n/a	3	3
window size	n/a	n/a	10
negative sample	n/a	n/a	10

Tab. 4.3: SECTOR and word2vec hyperparameter settings. A value of 'n/a' indicates that the hyperparameter is not applicable.

SECTOR parameters For our SECTOR trainings, we choose TeXoo's default settings as of release 1.0.2. See Table 4.3 for details. We additionally use the dos Santos ranking loss function presented by Arnold et al. [Arn+19] and identity activation.

Bag-of-words parameters We follow Arnold et al. [Arn+19] and apply a layer size of 4096 and 5 independent hash functions.

Word2vec parameters We again follow the TeXoo defaults. See Table 4.3 for details.

FastText parameters As newest and most promising language model, we focus on optimizing fastText. We start with the default parameters presented by the official fastText implementation[2]. Grave et al. [Gra+18] describe their settings when learning word vectors for 157 languages. Additionally, we apply TeXoo's default word2vec settings. Finally, we combine both settings following Chiu et al. [Chi+16] recommendations. Table 4.4 shows each approach's settings.

Grave et al. [Gra+18] present a new CBOW model that uses position dependent weights in order to better capture positional information. However, this model has not been made public as of this writing. We therefore train both skipgram and CBOW models.

[2] *fastText*. URL: https://fasttext.cc/index.html (visited on Apr. 1, 2019).

hyperparameter	default	Grave et al. [Gra+18]	word2vec	mix
model	skip/CBOW	pCBOW	n/a	skip/CBOW
embedding layer size	100	300	256	300
loss function	ns	ns*	ns*	ns/hs
min length of n-gram	3	5	3*	3
max length of n-gram	6	5	6*	6
learning rate	0.05	0.05*	0.025	0.05
learning update rate	100	100*	100*	100
subsampling	0.0001	0.0001*	0.0001*	0.0001
number of epochs	5	10	1	10
minimum word frequency	5	5*	3	3
window size	5	5*	10	10
negative sample	5	10	10	10

Tab. 4.4: FastText hyperparameter settings. A * indicates that the value was not described. We assume default values. They also feature a new CBOW algorithm using position dependent weights. To differentiate from traditional CBOW, we dub this 'pCBOW'.

4.4 Summary

This chapter gave an overview of relevant parts of the code. Chapter 4.1 further elaborated the archetype algorithm first described in Chapter 3.3 as well as the segmentation process. As the TeXoo framework featured most of the necessary code, only minor adjustments had to be made to train SECTOR. We elaborated these adjustments as well as our solution for k-fold evaluation in Chapter 4.2. Finally, we summarized and justified our hyperparameter settings for each neural model we trained in Chapter 4.2.

5 Evaluation

On the following pages we first revisit our hypothesis first noted in Chapter 1.3.1. We present further evidence to support our line of argument. In Chapter 5.2 and Chapter 5.3 we describe and evaluate our experiments both quantitatively and qualitatively before discussing our results in Chapter 5.4 in light of our hypotheses.

5.1 Hypotheses

In Chapter 1.3.1 we conjecture useful methods to extract facets from medical resources. Following we will further elaborate on these methods and justify our reasoning:

5.1.1 Specialized Text Embeddings Perform Better than General Purposed Text Embeddings on Medical Domain

Howard and Ruder [HR18] show that fine-tuning your language model improves their quality. This applies also in the biomedical domain, as shown by Sheikhshab et al. [She+18] and Lee et al. [Lee+19]. Thus we also expect specialized word embeddings to perform better than their general purpose counterpart. Especially clinical resources, which are not just beyond lay comprehension [McC05; Mos+14], but also highly variable depending on the hospital, medical division and even clinician [Fur+87; BJ+00; Sta+17], should benefit from custom word embeddings that capture their unique semantic nuances.

5.1.2 SECTOR as Effective Means of Structural Facet Extraction

Extracting structural facets on sentence level for whole documents requires a neural network architecture capable of comprehending long-term dependencies. Also, when grasping the structure of a document, humans instinctively incorporate both the text

before and after the specific location. Graves and Schmidhuber [GS05] even show that both directions are equally important.

Bidirectional LSTMs as used in SECTOR excel at both. LSTMs [HS97] with forget gates [Ger+00] solved the issue of RNNs' internal signals either "blowing up or vanishing", thus making them prime candidates for long-term dependency extraction. Additionally, since SECTOR employs bidirectional LSTMs, both the text before and after the current position is accounted for. Arnold et al. [Arn+19] further improve segmentation by utilizing the bidirectional embedding deviation, which helps in detecting topic shifts.

5.1.3 SECTOR as Effective Means of Topical Facet Extraction

Since extracting topical and structural facets are fundamentally similar tasks, we assume SECTOR is an effective means for topical facet extraction for the same reasons. Moreover, since topical facets are ambiguous and do not present a clear answer, modeling them as multi-label task seems reasonable and solves the problems Tepper et al. [Tep+12] faced with their combined categories.

5.2 Quantitative Evaluation

To evaluate our hypotheses we conduct four experiments on both tasks described in Chapter 3.4.

The structural facet extraction task aims to extract a small number of mutually exclusive classes (11-14 structural facets). The topical facet extraction task performs multi-label classification with a larger, ambiguous target vocabulary (1.6k topical facets). We experiment with four word representation approaches. A baseline bag-of-words scheme and two word embedding architectures, word2vec and fastText. Both are trained on a specialized domain specific dataset as described in Chapter 3.4.1. Additionally, we evaluate a general purpose fastText model.
We k-fold evaluate the best performing models with $k = 5$.

Evaluation datasets As described in Chapter 3.2, public clinical domain data is sparse. Our only available dataset are therefore the Charité doctors' letters (Chapter 3.1). Following our methodology of Chapter 3.3, we normalize the headings using our ontology, reducing all headlines to 14 structural facets including an *other*-class. We further enrich the original headlines to add hierarchical structure and use

these modified headlines as topical facets. As our ontology features two levels of granularity, we dub the datasets containing both levels *level two letters* (`L2L`). We further present the `L2.1L` dataset (level 2.1 letters), which incorporates changes made due to qualitative evaluation, as described in Chapter 5.3.3.

Quality measures. Following Arnold et al. [Arn+19] we measure using P_k on sentence level and micro-averaged MAP and F_1 at segment-level. We further display micro-averaged precision and recall at first and third position for a more detailed insight. The *probabilistic* P_k *error score* measures segmentation by calculating the probability of a false boundary in a window of size k. A lower P_k score equals better segmentation. Just like Arnold et al. [Arn+19] we set k to half of the average segment length. For classification, we match predicted sections with ground truth sections using maximum boundary overlap. We further note the *Mean Average Precision* (MAP), "which evaluates the average fraction of true labels ranked above a particular label" [Arn+19]. Following Manning et al. [Man+08, pp. 142-144] the F_1 score is the harmonic mean of precision and recall:

$$F_1 = \frac{2PR}{P+R} \tag{5.1}$$

with precision being defined as the fraction of retrieved sections that are relevant:

$$Precision = \frac{\#(relevant\ items\ retrieved)}{\#(retrieved\ items)} = P(relevant|retrieved) \tag{5.2}$$

and recall being defined as the fraction of relevant sections that are retrieved:

$$Recall = \frac{\#(relevant\ items\ retrieved)}{\#(relevant\ items)} = P(retrieved|relevant) \tag{5.3}$$

5.2.1 Experiments

We present the four best performing models on the `L2L` dataset (Table 5.1) and further evaluate the three best performing approaches on `L2.1L` (Table 5.2).

Bag-of-words with bloom filters outperforms word embeddings. Except for the L2L-topical task, the bag-of-words representation performed best for both tasks with all measures. It improves F_1 measure for the L2.1L-topical by 0.64% and the structural task by an average of 0.42% compared to the second best performing model. It further improves MAP by averaged 0.29% on both structural tasks and 0.21% on the L2.1L-topical.
This matches the observation made by Arnold et al. [Arn+19]. They too noted

best performing model	P@1	P@3	R@1	R@3	F_1	P_k	MAP
L2L dataset: 14 structural facets as single-label task							
SEC>T+bow*	**95.21**	**32.68**	**95.21**	**98.04**	**95.08**	**2.40**	**96.74**
SEC>T+W2V@WD+DL	94.72	32.60	94.72	97.79	94.83	2.56	96.55
SEC>T+fT@CC	94.08	32.51	94.08	97.53	94.35	3.10	96.26
SEC>T+fT@WD+DL	94.58	32.59	94.58	97.77	94.65	2.82	96.50
L2L dataset: 1,670 topical facets as multi-label task							
SEC>H+bow	85.49	45.20	61.90	84.58	77.90	10.15	88.74
SEC>H+W2V@WD+DL*	**95.16**	**52.20**	**65.22**	**91.19**	**82.25**	8.91	**94.45**
SEC>T+fT@CC	93.42	50.52	64.66	89.71	81.48	9.16	93.10
SEC>H+fT@WD+DL	94.89	51.63	65.12	90.53	82.20	**6.36**	93.89

Tab. 5.1: Best performing models per task and text representation. Following Arnold et al. [Arn+19], numbers are given as P_k on sentence level, micro-averaged F_1 and MAP at segment-level. We further note micro-averaged precision and recall. A model marked with * has been k-fold evaluated.

best performing model	P@1	P@3	R@1	R@3	F_1	P_k	MAP
L2.1L dataset: 12 structural facets as single-label task							
SEC>T+bow	**98.72**	**33.25**	**98.72**	**99.74**	**98.97**	**0.96**	**99.41**
SEC>T+W2V@WD+DL	98.68	33.25	98.68	99.75	95.6	3.21	97.59
SEC>T+fT@WD+DL	97.79	33.15	97.79	99.44	98.39	1.69	99.02
L2.1L dataset: 1,687 topical facets as multi-label task							
SEC>H+bow	**99.13**	**52.90**	**69.33**	**93.92**	**87.07**	**5.8**	**97.36**
SEC>H+W2V@WD+DL	97.68	52.23	68.68	93.32	86.43	7.64	97.15
SEC>H+fT@WD+DL	97.5	51.51	68.67	92.58	86.45	7.15	96.70

Tab. 5.2: Models trained after updating our ontology as result of qualitative evaluation. Again, following Arnold et al. [Arn+19], numbers are given as P_k on sentence level, micro-averaged F_1 and MAP at segment-level. We further note micro-averaged precision and recall.

that bloom filters performed on par with word embeddings. We also observed significantly longer training duration with a factor of 2.9.

General purpose model performed worst. The fastText model trained on *Common Crawl* and *Wikipedia* was distributed by Grave et al. [Gra+18]. It features a newer weighted fastText approach, which is not yet publicized. Grave et al. [Gra+18] observed its most significant impact on the German dataset and improved performance by 10%. Other languages were not impacted that much.
While it consistently performed worst, it's only an average of 1.1% F_1 and 0.92% MAP behind the best performing model.

Specialized domain word embeddings are equal. Both word2vec (*W2V*) and fastText (*fT*) trained on the joined dataset of German **W**ikipedia **d**iseases articles [Arn+19] and our **d**octors' **l**etters (*WD+DL*) performed equally well for classification. While word2vec scored slightly better F_1 difference averages at 0.07% for L2L-structural and -topical, as well as L2.1L-topical. For the L2.1L-structural however, W2V performed 2.79% F_1 worse for segment-level and doubled the probability of a wrong segment boundary P_k but improved sentence-level F_1 by 0.9%. This is in contrast to the topical best, where W2V improved segment-level MAP by average 0.51%.

5.2.2 Conclusion

Following these observations, we draw three conclusions:

1. As the general purpose embedding scored consistently worst, a specialized word embedding seems to perform better on clinical resources.

2. Since both word2vec's and fastText's performance is nearly identical, with fastText performing slightly worse, fastText's subword information does not seem to benefit word representation in doctors' letters.

3. Since bag-of-words performed best and fastText did not benefit from its subword information, we assume that the words themselves are more important than their semantic meaning.

Class	#Examples	TP	FP	Acc	Prec	Rec	F1
Diagnose	2082	2032	84	97.6	96.03	97.6	96.81
Bildgebende Diagnostik	753	717	230	95.22	**75.71**	95.22	84.35
Status	981	575	61	58.61	90.41	58.61	71.12
Diagnostische Maßnahmen	1732	1424	194	82.22	88.01	82.22	85.01
Labor	23131	23041	1439	99.61	94.12	99.61	96.79
Brief Kopf	3393	3393	0	100	100	100	100
Brief Anrede	491	476	3	96.95	99.37	96.95	98.14
Brief Schluss	1588	1588	4	100	99.75	100	99.87
Medikation	6431	6425	3	99.91	99.95	99.91	99.93
Verlauf und Therapie	888	699	17	78.72	97.63	78.72	87.16
other	799	328	23	41.05	93.45	41.05	57.04
Konsil	82	70	31	85.37	**69.31**	85.37	76.5
Beurteilung	458	62	8	**13.54**	88.57	**13.54**	**23.48**
Befund	276	137	21	**49.64**	86.71	**49.64**	**63.13**
[macro-avg]	43085	40967	2118	95.08	91.36	78.46	81.38

Tab. 5.3: Evaluation for the best performing model (bag-of-words) on L2L-structural per class. Bold numbers indicate the two worst results per column. As *other* represents a catch-all class, we ignore its results.

5.3 Qualitative Evaluation

Quantitative evaluation is a good way of comparing models. However, to diagnose errors of our methodology and bootstrapping method, as well as to reveal weaknesses of our approach a more in-depth analysis is necessary. To this end we single out and categorize potential flaws before analyzing them.

Considering the results in Table 5.3 and Table 5.4, four issues are apparent: two classes, *Beurteilung* and *Befund*, show poor recall, while *Konsil* and *Bildgebende Diagnostik* show worse, albeit still good, precision compared to the rest. We therefore focus on *Beurteilung* and *Befund*. As recall is the fraction of relevant sections that are retrieved out of all possible relevant sections [Man+08, pp. 142-144], we concentrate our analysis on false negatives. We observe 63% false negatives for *Befund* and 86% false negatives for *Beurteilung*. We choose 50 random samples from both focus sets and 20 samples from the third worst performing class, *Status*. We further pick 50 random samples from *Bildgebende Diagnostik*'s false positives to address its precision deficit.

ground truth \ predicted	bi-di	di-ma	bef	beu	anr	kop	sch	dia	konsil	lab	med	oth	sta	v-u-t
Bildgebende Diagnostik	0.96	0.01	0.00	0.01	-	-	-	-	0.00	0.00	0.00	0.00	0.00	0.02
Diagnostische Maßnahmen	0.03	0.73	0.00	0.00	-	-	-	0.02	0.00	0.15	0.00	0.01	0.04	0.01
Befund	0.36	0.09	0.37	0.00	-	-	-	-	-	0.14	-	0.00	0.02	0.00
Beurteilung	0.27	0.13	-	0.14	-	-	-	-	-	0.43	-	0.02	-	0.01
Brief Anrede	-	-	-	-	0.51	0.19	-	0.28	-	-	-	-	0.01	-
Brief Kopf	-	-	-	-	0.00	0.99	-	0.00	-	-	-	-	0.00	-
Brief Schluss	-	-	-	-	-	-	1.00	-	-	-	0.00	-	-	-
Diagnose	0.00	0.02	-	-	0.01	-	-	0.95	-	0.00	-	-	0.01	0.00
Konsil	0.08	0.05	-	-	-	-	-	-	0.84	0.00	-	-	-	0.03
Labor	0.00	0.00	0.00	-	-	-	-	0.00	0.00	0.99	0.00	0.00	0.00	0.00
Medikation	0.00	-	-	-	-	-	0.00	-	-	-	0.99	-	-	0.00
other	0.08	0.05	0.01	0.01	-	-	-	0.01	0.01	0.36	0.00	0.45	0.02	0.01
Status	0.00	0.00	0.00	0.00	0.00	-	-	0.01	-	0.07	-	0.00	0.92	0.00
Verlauf und Therapie	0.00	-	-	-	-	-	0.00	0.00	0.01	0.06	0.00	0.00	0.00	0.91

Tab. 5.4: Confusion matrix for the best performing model (bag-of-words) on L2L-structural in %. Columns display retrieved values, while rows show ground truth.

5.3.1 Common Error Types

We formulate the following error classes based on our observations:

Hierarchical error Sections that are identified as atomic units, but actually constitute a subcategory of the preceding section.
Example: A section `Beurteilung` (*assessment*) that does not represent an assessment on document level, but on section level.

Bootstrapping error Sections that are wrongfully labeled due to errors during bootstrapping process.
Examples: Missing line breaks lead the bootstrapping algorithm to fail to recognize the starting or ending point of a new section.
`Befundbericht Virologie` (*virology report*) being identified as `Befund` (*report*) instead of `Virologie` (*virology*), which belongs to the class `Labor` (*laboratory*).

Ambiguity error Sections whose contents seem to belong to a specific class, but belong to another.
Example: `Gastroskopie` (*gastroscopy*) sections contain a description of the clinician's visual observation, which make it seem like an imaging method, but are an examination and therefore `Diagnostische Maßnahmen` (*diagnostic measures*).

5.3.2 Error Analysis

Hierarchical error While our ontology contains supercategories to hierarchically order and cluster different headlines, it fails to express hierarchical structures within sections of the doctors' letters.
For example, we observe the original headline `Kommentar` (*commentary*) to be a subcategory for `Bildgebende Diagnostik`. We initially identified it as `Beurteilung` (*assessment*), since its content is equivalent to an assessment. It may prove to be a structural facet among imaging method sections, but should be considered a topical subsection on letter level.
Hierarchical errors make up 70.5% of our samples. Nearly all *Bildgebende Diagnostik* false positives are of this kind (97.2%), but none of the *Status* samples.

Bootstrapping error As Bootstrapping uses rules and assumptions to generate approximated labeled data, it inevitably produces errors. It occasionally fails to recognize starting or ending point of a section or fails to realize the focus of a headline when several trigger words are present.
As example for the first case we observe missing line breaks before or after a

headline. An example for the latter presents `Befundbericht Virologie` (*virology report*). Virology actually belongs to the `Labor` class (*laboratory*), but was labeled as `Befund` (*report*).
We observe 21.9% bootstrapping errors in our samples. 70% of all *Status* errors belong to this group as well as 28.8% of *Beurteilung*. We observe only 14.2% for *Befund* and none for *Bildgebende Diagnostik*.

Ambiguity error Chapter 3.2 describes ambiguity as one of the main challenges when processing clinical or medical text. Due to this certain sections are nearly indistinguishable from one another, but describe fundamentally different things.
An example offers `Gastroskopie`. While it actually presents an examination and is therefore categorized as diagnostic measure, its section content consists of the clinicians visual observations during the procedure and is therefore hardly distinguishable from an imaging method.
Only 7.6% of all examined errors are due to ambiguity. Most of which are false negatives of *Status* (70%).

5.3.3 Conclusion

Following our error analysis, the main source of error is not the SECTOR model. As the ambiguity error class constitute for only a small fraction of all sampled errors, we recognize two leading flaws:

1. **Structural hierarchy.** Together with the medical professional we created a purely topical ontology based on the vocabulary in the headlines, but failed to account for structural hierarchy that may be present within sections.

2. **Flawed bootstrapping.** Bootstrapping algorithms present an approximation, and thus will always be flawed. They are no replacement for manual annotated training data.

As solution for the first flaw, we present dataset `L2.1L`. We removed the structural facets `Beurteilung` and `Befund`. We thus reduced the 14 structural facets to 12 and modified the topical facets of relevant sections to reflect the subordination to preceding sections.

5.4 Findings and Discussion

Quantitative evaluation presents us with insight regarding the effect different language models exert on the model. Qualitative evaluation grants us more detailed understanding of our methodology and potential flaws. On basis of both evaluations we reexamine our hypotheses formulated in Chapter 1.3.1:

Specialized Text Embeddings Perform Better than General Purposed Text Embeddings on Medical Domain Following the results on the L2L dataset, we can confirm this hypothesis. The general purpose word embeddings trained on the whole German Wikipedia and Common Crawl, albeit having a 5,000 times larger training corpus (Chapter 3.6) and an improved fastText model, performed consistently worse than both specialized word2vec and fastText embeddings. We can therefore confirm the conclusion presented by Sheikhshab et al. [She+18] and Lee et al. [Lee+19] for fastText as well.

SECTOR as Effective Means of Structural Facet Extraction With an F_1 score of 98.97% and P_k of 0.96% SECTOR performed exceptionally well for both segmentation and classification. Additionally, since our qualitative evaluation revealed several issues with the bootstrapped data, the remaining error might be due to incorrectly labeled data, not SECTOR. We therefore assume a near perfect result for this dataset and can confirm this hypothesis. However, as Starlinger et al. [Sta+17] noted, "highly accurate results" within singular institutions have been reported repeatedly and cross-institutional validation is necessary to allow further judgment.

SECTOR as Effective Means of Topical Facet Extraction For topical facet extraction, we observe lower values with 87.07% F_1 and 5.8% P_k. However, with 1,687 topical instead of 12 structural facets, the task is considerably harder. Thus, while we confirm this thesis, we again note the need for validation on a broader corpus.

We further noticed during evaluation:

Bag-of-words encoding with bloom filter performs better than word embeddings Arnold et al. [Arn+19] observed bloom filter encoding to perform worse by 0.7% F_1 for German datasets. We noted an increase of 0.6% F_1 for both structural and topical tasks. We assume that the semantic meaning of each word is less important than the word itself. This might be due to the doctors' letters being semi-auto-generated or to clinicians being used to write the same words repetitively. Alternatively, the German Wikipedia diseases corpus was not fit for this task and either a bigger, biomedical Wikipedia corpus is necessary, or no additional dataset at all.

5.5 Summary

In this chapter we evaluated our hypotheses, bootstrapping and methodology. In Chapter 5.1 we formalized and justified our three hypotheses: that specialized word embeddings perform better than general purpose ones and that SECTOR constitutes effective means of facet retrieval, both structural and topical. During Chapter 5.2 we measured SECTOR's performance when given different word embeddings quantitatively. We noticed that bag-of-words performed consistently better, word2vec and fastText performed equally well and the general purpose fastText performed worst. Based on these observations we concluded, that the words themselves are more important than their semantic meaning for facet extraction. In Chapter 5.3, we pinpointed potential weaknesses and examined examples for each. We identified two main sources of error: our bootstrapping algorithm and that we failed to take structural hierarchy among headlines into account. We present a solution for the second error in form of a new dataset: `L2.1L`.
Finally we reevaluated our hypotheses in light of our observations and were able to confirm them all in Chapter 5.4 with some reservations due to the small and invariant training corpus.

6 Summary and Future Work

During this final chapter we recap the goal, findings and results of this thesis as well as the methodology we employed to reach it. We further discuss future potential steps to improve on this work.

6.1 Summary

Our goal for this thesis was evaluating SECTOR as effective means of facet extraction on medical resources. This is to be used as upstream task for a complex question answer system in clinical context. Additionally we hypothesized that a specialized clinical language model performs better than general purpose ones.
Following we summarize the challenges we encountered, our most important steps and findings as well as our conclusion.

Challenges working with medical resources We identified several challenges when working with medical resources. Looking for training data, we first evaluated the `WikiSection` corpus presented by Arnold et al. [Arn+19]. We observed structural and vocabulary mismatches due to the differing nature of Wikipedia articles and our clinical resources. Furthermore, we observed a lack of publicly available datasets for clinical resources in general and German ones in particular due privacy regulations. Additionally the ambiguity of medical language and highly specialized domain knowledge necessary hindered our work and makes professional medical assistance indispensable.

Bootstrapping training data To tackle these challenges we employed the help of a medical doctor and bootstrapped our own training corpus based on 7,553 discharge letters courtesy of *Charité Berlin's Medical Department, Division of Nephrology and Internal Intensive Care Medicine*. In the process we identified relevant structural classes and developed an ontology to capture hierarchical structures in the original headlines.

Modeling the tasks We split the facet extraction task into subtasks of extracting either structural or topical facets. As structural facets are mutually exclusive, we model the classification problem as multi-class single-label. To reflect hierarchical

structure and ambiguity in topical headings we extended the original headings using the different levels of our ontology and defined the task as multi-class multi-label. These problems aligned with SECTOR's two variations, the multi-class single-label *topics* and the multi-class multi-label *headings* task.

Training the models We leveraged the TeXoo project [Arn19] as implementation of SECTOR. Using varying language models we conducted several experiments. After a first evaluation, we identified flaws in our ontology and retrained most promising approaches on the updated training corpus. We then k-fold validated our best results.

Quantitative and qualitative evaluation To evaluate our results, we applied several quantitative evaluation metrics and presented the results for *precision@1*, *precision@3*, *recall@1*, *recall@3*, F_1, P_k and MAP. For our best performing models we observed near perfect results of 98.97% F_1 and 99.41% MAP for the structural facet extraction task and 87.07% F_1 and 97.35% MAP for the topical one.
We further identified the worst performing structural classes for our best performing model. Inspecting random samples of falsely classified sentences we identified two main flaws in our approach: the lack of structural hierarchy in our ontology and errors in our bootstrapping algorithm due to not uniform input data.

Conclusion In conclusion we were able to confirm all of our hypotheses.
Although the bag-of-words baseline performed best, our specialized language models achieved better results than the general purpose one. Even though the general purpose model featured more training data and a potentially better algorithm.
Although lacking comparable results or cross-institutional validation, our SECTOR model performed exceptionally well and is without a doubt an effective means of extracting facets on medical resources.

6.2 Future Work

During this project we noticed several approaches that could further improve this task:

Neural approach for creating the ontology Future applications would profit from removing the need for a medical professional. However, identifying structural and topical hierarchies within the source data proves difficult for untrained individuals. Automation of this step could improve both speed and costs for the performing team. A well performing method might potentially remove the need of seeing sensitive

data altogether. Possible approaches could include identifying and clustering the contents of each section or headline based on paragraph or term based language models such as *paragraph vectors* [LM14] or *word2vec* [Mik+13].

Combining topics and headings task As we have seen during qualitative evaluation, topical and structural facets are interwoven and disambiguation dependent on the context. Topical facets sometimes represent structural facets of structural facets. Thus defining the task as multi-task learning problem might leverage synergy effects.

Using contextual language models Due to hardware and time constraints for this project we had to restrict ourselves to fast and efficient word representations. However, Peters et al. [Pet+18] showed that contextual language models like ELMo can improve a variety of NLP tasks.

Specialized sentence and token segmentation Further experimentation could include adjusting the sentence and token segmentation step. Faessler et al. [Fae+14] distribute a variety of NLP models trained on "a well-balanced mixture of medical document types such as discharge summaries, pathology reports but also medical textbook excerpts, all written in German language." Given the highly ambique medical language, a specialized tokenization or sentence segmentation might prove useful.

Cross-institutional datasets SECTOR achieved near perfect results for structural facet extraction. However, Starlinger et al. [Sta+17] noted that several approaches achieved good scores on datasets within singular institutions. Tepper et al. [Tep+12] reported results within 80-90% F_1, too, for a nearly identical task on English discharge letters. Therefore testing on datasets from other medical divisions or institutions is necessary to truly evaluate SECTOR's effectiveness.

Bibliography

[AG00] Eugene Agichtein and Luis Gravano. „Snowball: Extracting Relations from Large Plain-text Collections“. In: *Proceedings of the Fifth ACM Conference on Digital Libraries*. DL '00. event-place: San Antonio, Texas, USA. New York, NY, USA: ACM, 2000, pp. 85–94 (cit. on p. 19).

[Aic+16] Christine Aicardi, Lorenzo Del Savio, Edward S. Dove, et al. „Emerging ethical issues regarding digital health data. On the World Medical Association Draft Declaration on Ethical Considerations Regarding Health Databases and Biobanks“. In: *Croatian Medical Journal* 57.2 (Apr. 2016), pp. 207–213 (cit. on p. 2).

[Arn+19] Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. „SECTOR: A Neural Model for Coherent Topic Segmentation and Classification“. In: *Transactions of the Association for Computational Linguistics (TACL)*. arXiv: 1902.04793. Feb. 2019 (cit. on pp. v, 2, 5, 8–10, 16, 17, 19, 24–27, 32, 36–39, 44, 47).

[Aro+17] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. „A Simple but Tough-to-Beat Baseline for Sentence Embeddings“. In: *5th International Conference on Learning Representations*. 2017 (cit. on p. 9).

[AV19] Yann Alibert and Julia Venturini. „Using Deep Neural Networks to compute the mass of forming planets“. In: *Astronomy and Astrophysics* (Mar. 2019). arXiv: 1903.00320 (cit. on p. 6).

[BJ+00] R. C. Barrows Jr, M. Busuioc, and C. Friedman. „Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches.“ In: *Proceedings of the AMIA Symposium* (2000), pp. 51–55 (cit. on pp. 15, 17, 35).

[Boj+16] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. „Enriching Word Vectors with Subword Information“. In: *Transactions of the Association for Computational Linguistics (TACL 2016)*. arXiv: 1607.04606. July 2016, pp. 135–146 (cit. on pp. 7, 8, 24).

[Cha+11] Wendy W. Chapman, Prakash M. Nadkarni, Lynette Hirschman, et al. „Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions“. eng. In: *Journal of the American Medical Informatics Association: JAMIA* 18.5 (Oct. 2011), pp. 540–543 (cit. on p. 17).

[Che+18] Jinying Chen, Emily Druhl, Balaji Polepalli Ramesh, et al. „A Natural Language Processing System That Links Medical Terms in Electronic Health Record Notes to Lay Definitions: System Development Using Physician Reviews“. en. In: *Journal of Medical Internet Research* 20.1 (2018), e26 (cit. on p. 19).

[Chi+16] Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. „How to Train good Word Embeddings for Biomedical NLP“. In: *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 166–174 (cit. on p. 32).

[Dev+18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding“. In: *CoRR* abs/1810.04805 (Oct. 2018). arXiv: 1810.04805 (cit. on pp. 7, 8).

[Fae+14] Erik Faessler, Johannes Hellrich, and Udo Hahn. „Disclose Models, Hide the Data— How to Make Use of Confidential Corpora without Seeing Sensitive Raw Data“. en. In: *European Language Resources Association (ELRA)*. May 2014, p. 8 (cit. on pp. 2, 49).

[Fur+87] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. „The Vocabulary Problem in Human-system Communication“. In: *Commun. ACM* 30.11 (Nov. 1987), pp. 964–971 (cit. on pp. 13, 15, 17, 35).

[Ger+00] Felix A. Gers, Jürgen A. Schmidhuber, and Fred A. Cummins. „Learning to Forget: Continual Prediction with LSTM“. In: *Neural Comput.* 12.10 (Oct. 2000), pp. 2451–2471 (cit. on pp. 9, 10, 36).

[GM14] Sonal Gupta and Christopher Manning. „Improved Pattern Learning for Bootstrapped Entity Extraction“. en. In: *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. Ann Arbor, Michigan: Association for Computational Linguistics, 2014, pp. 98–108 (cit. on p. 19).

[Goo16] Kenneth W Goodman. „Ethical and Legal Issues in Decision Support“. en. In: *Clinical Decision Support Systems - Theory and Practice*. Springer International Publishing, 2016, pp. 131–146 (cit. on p. 2).

[Gra+18] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. „Learning Word Vectors for 157 Languages“. In: vol. abs/1802.06893. arXiv: 1802.06893. Feb. 2018 (cit. on pp. 24, 25, 32, 33, 39).

[GS05] Alex Graves and Jürgen Schmidhuber. „Framewise phoneme classification with bidirectional LSTM and other neural network architectures“. eng. In: *Neural Networks: The Official Journal of the International Neural Network Society* 18.5-6 (July 2005), pp. 602–610 (cit. on pp. 10, 36).

[HR18] Jeremy Howard and Sebastian Ruder. „Universal Language Model Fine-tuning for Text Classification“. In: *ACL*. arXiv: 1801.06146. Jan. 2018 (cit. on p. 35).

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. „Long Short-Term Memory“. In: *Neural Computation* 9.8 (Nov. 1997), pp. 1735–1780 (cit. on pp. 9, 36).

[Hut+14] Frank Hutter, Holger Hoos, and Kevin Leyton-Brown. „An Efficient Approach for Assessing Hyperparameter Importance“. en. In: *International Conference on Machine Learning*. 2014, pp. 754–762 (cit. on p. 31).

[JM09] Dan Jurafsky and James H. Martin. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition*. en. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International, 2009 (cit. on pp. 5, 6).

[KF12] J. Kacprzyk and Mario Fedrizzi. „Multiperson Decision Making Models Using Fuzzy Sets and Possibility Theory“. en. In: *Multiperson Decision Making Models Using Fuzzy Sets and Possibility Theory*. Google-Books-ID: HWfoCAAAQBAJ. Springer Science & Business Media, Dec. 2012, p. 131 (cit. on p. 15).

[Lee+19] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, et al. „BioBERT: a pre-trained biomedical language representation model for biomedical text mining“. en. In: *CoRR* abs/1901.08746 (2019) (cit. on pp. 3, 24, 35, 44).

[LM14] Quoc Le and Tomas Mikolov. „Distributed Representations of Sentences and Documents“. In: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. ICML'14. event-place: Beijing, China. JMLR.org, 2014, pp. II–1188–II–1196 (cit. on pp. 8, 49).

[Mac+18] Sean MacAvaney, Andrew Yates, Arman Cohan, et al. „Characterizing Question Facets for Complex Answer Retrieval“. In: *SIGIR*. arXiv: 1805.00791. May 2018 (cit. on pp. 5, 10, 18, 26).

[Man+08] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008 (cit. on pp. 15, 26, 37, 40).

[McC05] Alexa T. McCray. „Promoting Health Literacy“. en. In: *Journal of the American Medical Informatics Association* 12.2 (Mar. 2005), pp. 152–163 (cit. on pp. 19, 35).

[Mik+13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. „Efficient Estimation of Word Representations in Vector Space“. In: *CoRR* abs/1301.3781 (Jan. 2013). arXiv: 1301.3781 (cit. on pp. 7, 24, 49).

[Mos+14] Matthew Mossanen, Lawrence D. True, Jonathan L. Wright, et al. „Surgical pathology and the patient: a systematic review evaluating the primary audience of pathology reports“. eng. In: *Human Pathology* 45.11 (Nov. 2014), pp. 2192–2201 (cit. on pp. 19, 35).

[Pet+18] Matthew E. Peters, Mark Neumann, Mohit Iyyer, et al. „Deep contextualized word representations“. In: *Proc. of NAACL* (Feb. 2018). arXiv: 1802.05365 (cit. on pp. 7, 8, 49).

[PG17] Josh Patterson and Adam Gibson. *Deep Learning: A Practitioner's Approach*. O'Reilly, 2017 (cit. on pp. 6, 7, 10).

[Sam+17] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. „Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models“. en. In: *CoRR* abs/1708.08296 (Aug. 2017) (cit. on p. 2).

[Sch+18] Rudolf Schneider, Sebastian Arnold, Tom Oberhauser, et al. „Smart-MD: Neural Paragraph Retrieval of Medical Topics“. In: *Companion Proceedings of the The Web Conference 2018*. WWW '18. event-place: Lyon, France. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2018, pp. 203–206 (cit. on pp. 1, 9).

[She+17] Saeedeh Shekarpour, Edgard Marx, Sören Auer, and Amit Sheth. „RQUERY: Rewriting Natural Language Queries on Knowledge Graphs to Alleviate the Vocabulary Mismatch Problem“. In: *31st AAAI Conference on Artificial Intelligence (AAAI-17)*. 2017 (cit. on p. 13).

[She+18] Golnar Sheikhshab, Inanç Birol, and Anoop Sarkar. „In-domain Context-aware Token Embeddings Improve Biomedical Named Entity Recognition“. In: *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*. Brussels, Belgium, 2018, pp. 160–164 (cit. on pp. 3, 24, 35, 44).

[Sta+17] Johannes Starlinger, Madeleine Kittner, Oliver Blankenstein, and Ulf Leser. „How to improve information extraction from German medical records“. en. In: *it - Information Technology* 59.4 (Jan. 2017) (cit. on pp. 17, 35, 44, 49).

[Tai+14] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. „DeepFace: Closing the Gap to Human-Level Performance in Face Verification“. en. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, OH, USA, June 2014, pp. 1701–1708 (cit. on p. 6).

[Tep+12] Michael Tepper, Daniel Capurro, Fei Xia, Lucy Vanderwende, and Meliha Yetisgen-Yildiz. „Statistical Section Segmentation in Free-Text Clinical Records“. In: *LREC*. 2012 (cit. on pp. 17, 19, 36, 49).

[Var84] Various. „Prognosis“. English. In: *Hippocratic Writings*. Ed. by G. E. R. Lloyd. Trans. by J. Chadwick. Reprint edition. Harmondsworth: Penguin Classics, Mar. 1984, pp. 170–185 (cit. on p. 1).

[Wu+16] Yonghui Wu, Mike Schuster, Zhifeng Chen, et al. „Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation“. In: *CoRR* abs/1609.08144 (Sept. 2016). arXiv: 1609.08144 (cit. on p. 6).

[XC98] Jinxi Xu and W. Bruce Croft. „Corpus-based Stemming Using Cooccurrence of Word Variants“. In: *ACM Trans. Inf. Syst.* 16.1 (Jan. 1998), pp. 61–81 (cit. on p. 21).

[YM15] Illhoi Yoo and Abu Saleh Mohammad Mosa. „Analysis of PubMed User Sessions Using a Full-Day PubMed Query Log: A Comparison of Experienced and Non-experienced PubMed Users“. en. In: *JMIR Medical Informatics* 3.3 (2015), e25 (cit. on p. 1).

[Wis05] Wissenschaftsrat. *6825-05.pdf*. Tech. rep. Drs. 6825/05. 2005 (cit. on p. 19).

Websites

[Arn19] Sebastian Arnold. *TeXoo – A Zoo of Text Extractors. Contribute to sebastianarnold/-TeXoo development by creating an account on GitHub*. original-date: 2018-07-19T08:42:32Z. Feb. 2019. URL: https://github.com/sebastianarnold/TeXoo (visited on Mar. 30, 2019) (cit. on pp. 31, 48).

[Noaa] *Common Crawl*. URL: http://commoncrawl.org/ (visited on Mar. 9, 2019) (cit. on p. 24).

[Noab] *DIMDI - OPS Version 2019*. URL: https://www.dimdi.de/static/de/klassifikationen/ops/kode-suche/opshtml2019/ (visited on Mar. 16, 2019) (cit. on p. 22).

[Noac] *fastText*. URL: https://fasttext.cc/index.html (visited on Apr. 1, 2019) (cit. on p. 32).

[Noad] *Hippocrates*. en. Page Version ID: 890031202. Mar. 2019. URL: https://en.wikipedia.org/w/index.php?title=Hippocrates&oldid=890031202 (visited on Apr. 1, 2019) (cit. on p. 1).

[Noae] *Wikipedia*. URL: https://www.wikipedia.org/ (visited on Mar. 9, 2019) (cit. on p. 24).

[Sch15] I. Schlünder. *GMS | 14. Deutscher Kongress für Versorgungsforschung | Datenschutzkonforme Lösungen für die Versorgungsforschung*. de. Sept. 2015. URL: http://www.e-gms.de/static/de/meetings/dkvf2015/15dkvf064.shtml (visited on Mar. 10, 2019) (cit. on p. 17).

[Cha] Charité-Universitätsmedizin Berlin. *Modellstudiengang Humanmedizin*. de. URL: https://www.charite.de/studium_lehre/studiengaenge/modellstudiengang_humanmedizin/ (visited on Mar. 11, 2019) (cit. on p. 19).

[Nat] National Safety Council. *Pedestrian Safety*. URL: https://www.nsc.org/home-safety/safety-topics/distracted-walking (visited on Mar. 31, 2019) (cit. on p. 2).

[Pub] PubMed. *Home - PubMed - NCBI*. en. URL: https://www.ncbi.nlm.nih.gov/pubmed/ (visited on Mar. 31, 2019) (cit. on p. 1).

[Wik19a] Wikipedia. *Esophagogastroduodenoscopy*. en. Page Version ID: 877006453. Jan. 2019. URL: https://en.wikipedia.org/w/index.php?title=Esophagogastroduodenoscopy&oldid=877006453 (visited on Mar. 12, 2019) (cit. on p. 18).

[Wik19b] Wikipedia. *Lungenentzündung*. de. Page Version ID: 185868927. Feb. 2019. URL: https://de.wikipedia.org/w/index.php?title=Lungenentz%C3%BCndung&oldid=185868927 (visited on Mar. 17, 2019) (cit. on p. 16).

List of Figures

List of Tables