Term-Based Extraction of Medical Information: Pre-Operative Patient Education Use Case

Martin Wolf[1], **Volha Petukhova**[2] **and Dietrich Klakow**[2]
Computational Linguistics[1], Spoken Language Systems[2], Saarland University, Germany
martinw@coli.uni-saarland.de;
{v.petukhova;dietrich.klakow}@lsv.uni-saarland.de

Abstract

The processing of medical information is not a trivial task for medical non-experts. The paper presents an artificial assistant designed to facilitate a reliable access to medical online contents. Interactions are modelled as doctor-patient Question Answering sessions within a pre-operative patient education scenario where the system addresses patient's information needs explaining medical events and procedures. This implies an accurate medical information extraction from and reasoning with available medical knowledge and large amounts of unstructured multilingual online data. Bridging the gap between medical knowledge and data, we explore a language-agnostic approach to medical concepts mining from the standard terminologies, and the data-driven collection of the corresponding seed terms in a distant supervision setting for German. Experimenting with different terminologies, features and term matching strategies, we achieved a promising F-score of 0.91 on the medical term extraction task. The concepts and terms are used to search and retrieve definitions from the verified online free resources. The proof-of-concept definition retrieval system is designed and evaluated showing promising results, acceptable by humans in 92% of cases.

1 Introduction

Nowadays, digital online services possess the dominant role delivering widely accessible applications at limited costs. For instance, recent technological advances make the provision of various eHealth services feasible. Using these applications, patients can stay informed searching content outside hospital business hours in a more convenient manner. A doctor, who conducts 120,000 - 160,000 interviews in the course of a 40-year career (Lipkin et al., 1995), meets then 'competent' patients who understand their medical needs and potential consequences of medical decisions. In the healthcare sector, language barriers and domain complexity may result in poor understanding of diagnosis, low compliance with recommendations, a significantly greater likelihood of a serious medical event and lower patient satisfaction (Bonacruz Kazzi and Cooper, 2003; Cohen et al., 2005; Pitkin Derose et al., 2009). The mainstream online services are therefore required to be reliable, accessible and to account for the diversity in individual needs, educational backgrounds, personal preferences, cognitive and physical limitations.

This paper addresses the needs in the reliable access to verified multilingual complex medical information. As a use case, we simulate pre-operative Question Answering (QA) sessions between doctors and patients. As a core part of these medical encounters, Patient Education Forms (PEFs) need to be filled in and the patient's informed consent signed. It is of chief importance that the forms are properly understood, medical procedures and risks are explained. PEFs contain many medical terms including those in Latin and as abbreviations. These terms have to be detected and corresponding definitions retrieved from available electronic medical documents. Although a number of biomedical entities recognition systems (Zhang and Elhadad, 2013; Björne et al., 2013; Sahu and Anand, 2017) and medical resources exist (Gurulingappa et al., 2010; Ohta et al., 2012), they are mostly built for English. We explore a language-agnostic approach to medical concepts mining based on the existing de-facto standard terminologies and dictionaries, and the collection of the corresponding German seed terms in a distant supervision setting. The extracted concepts and terms are used to search and retrieve definitions from the verified online free text sources. The proof-of-concept definition retrieval system is designed and evaluated.

The paper is structured as follows. In Section 2, we provide an overview of the state-of-the-art in Biomedical Named Entity Recognition (BM-NER). In Section 3, we discuss the conceptual design of a medical domain. Section 4 defines the medical term extraction task, assesses various resources for medical information extraction and presents the overall QA system architecture. Section 5 proposes the experimental design by specifying the collected and simulated data, and discusses the obtained results. Finally, Section 6 summarizes our findings and outlines directions for the future research and development.

2 Biomedical Named Entity Recognition

In 1995, the 6th Message Understanding Conference (MUC-6) focused on the Information Extraction (IE) from unstructured textual data (Grishman and Sundheim, 1996) and defined the Named Entity Recognition and Classification (NERC) task, see (Nadeau and Sekine, 2007) for a comprehensive overview. The relevant entities comprised names of persons, organizations and locations defined as ENAMEX (Entity Name Expression), extended later with TIMEX (Time Expressions) and NUMEX (Numerical Expressions).

In the early 2000s, the interest in bioinformatics lead to enriching the categories with concepts from biomedical domains focusing on the recognition of biological and genetic terms, disease and drug names and other medical or clinical entities (Settles, 2004; Shen et al., 2003). Biomedical named entity recognition is a key step in biomedical language processing.

Early (BM-)NER approaches were largely *rule-based* detecting entities based on the observed contextual and orthographic patterns. Such systems are especially useful if no or little training examples are available (Sekine and Nobata, 2004), are often straightforward to implement, suited for the entity classes or domains where the regularities in orthography or morphology can be exploited, and have other important advantages (Chiticariu et al., 2013). Although they achieve a rather high precision, recall is often low as the rule sets are rarely exhaustive. AbGene system of Tanabe and Wilbur (2002) uses a POS tagger extended to include gene and protein names as tag types. The system was trained on the manually labelled biomedical text. In its second iteration, it applies manually defined post-processing rules.

Another successful approach underlies the so-called *dictionary-based* systems. Here, the decision whether an entity is of an interest is made by matching against the entries in a dictionary, i.e. gazetteer or word list. To expand the coverage, linguistic methods (e.g. stemming or lemmatization), as well as fuzzy or exact matching strategies are used. cTakes of Savova et al. (2010) is an open-source information extraction tool from Electronic Health Records (EHR) which NER component is based on a dictionary look-up approach.

Dictionaries are also used supplementary to *machine learning* approaches (Tsuruoka and Tsujii, 2003), which are particularly useful if there is a high variability in entities observed. *Supervised* models like Hidden Markov Models (Zhou and Su, 2002), Support Vector Machines (Björne et al., 2013), Conditional Random Fields (Settles, 2004) and Neural Networks (Sahu and Anand, 2017) are reported to show the state-of-the-art performance. These approaches rely on large amounts of the annotated training data. To perform BM-NER some resources are created: the NCBI Disease Corpus (Doğan et al., 2014), the GENIA corpus (Kim et al., 2003) for molecular biology, the i2b2[1] corpus of clinical notes. The data for languages other than English is still an issue. Techniques which allow to automatically generate labelled training data like bootstrapping and distant supervision (Mintz et al., 2009) methods are proposed to build models in *semi-supervised* or *weakly supervised* way. For example, Dembowski et al. (2017) extract word lists from Wikipedia to label the data for an NER model training. The trained classifier outperforms the simple dictionary baseline.

Unsupervised approaches do not require any labelled training data, but rely on external resources like knowledge-bases or semantic nets (Alfonseca and Manandhar, 2002), lexical patterns (Evans and Street, 2003), and distributional semantics. Zhang and Elhadad (2013) applied a distributional semantics method to clinical notes and biological texts. The final system yields competitive results

[1] https://www.i2b2.org/NLP/DataSets/

Category	Frequency (in %)	Seed Terms Examples	
		German	English
Body-organ	19.6	Bronchien, Kopf	bronchia, head
Body-related	6.5	Atem, Hören	breath, hearing
Condition	4.0	gesund, schläfrig	healthy, sleepy
Disease	3.1	Hepatitis, Schlaganfall	hepatitis, stroke
Drug	8.0	Aspirin, Schlafmittel	aspirin, sleeping pills
Effect	2.9	Wärmegefühl	warm sensation
Institution	0.5	Intensivstation, Aufwachraum	intensive care unit, recovery room
Instrument	7.8	Nadel, Larynxmaske	needle, laryngeal mask
Person	2.5	Arzt, Pflegepersonal	doctor, nursing staff
Procedure	17.6	Eingriff, Narkose	intervention, narcosis
Procedure-related	2.2	intravenös, operativ	intravenous, operative
Purpose	0.7	muskelentspannend, Schmerzausschaltung	muscle relaxing, pain relief
Symptom	21.9	Atemnot, Juckreiz	shortness of breath, itching
Misc	2.7	medizinisch, peripher	medical, peripheral

Table 1: The taxonomy and distribution (relative frequencies, in %) of semantic concepts categories illustrated with examples of German and English seed terms.

on the i2b2 biomedical dataset of clinical notes and the GENIA corpus of biological literature, and outperforms the dictionary-matching baseline. This approach incorporates the collection of *seed terms*. The seed term sets are gathered from external terminologies and grouped into entity classes that represent the domain the best. For a QA application, it means that the classes of domain-specific semantic concepts can be used to generate signature vectors and the semantic similarity with the signature vectors of the answer candidates can be computed for retrieval and ranking. The concept classes can be also translated into the Expected Answer Types (EATs) to query the structured or unstructured data to retrieve an answer in a supervised or rule-based way.

3 Conceptual Domain Modelling

To facilitate an accurate information extraction from and reasoning with large amounts of (un-)structured data, it is important to specify and model real world entities and relations between them. This knowledge is often represented as ontologies, terminologies with semantic concepts groupings and taxonomic relations between them, and semantic networks. In many knowledge-based QA systems, high level semantic representations are used to query databases or other types of structured data. For example, Wilensky et al. (1988) developed the *Berkeley Unix Consultant*, for the domain related to the UNIX operating system where questions are analysed and transformed into an internal representation which are used to generate hypothesis about the user's information needs. A knowledge-based QA system as used by Apple Siri[2] and Wolfram Alpha[3] also first builds a query representation and then maps it to structured data like ontologies, gazeteers, etc. The Watson a DeepQA system of IBM Research (Ferrucci et al., 2010) incorporates content acquisition, question analysis, hypothesis generation, etc. Inside the hypotheses generation, it relies on NE detection, triple store and reverse dictionary look-up to generate candidate answers.

Alternative approaches advocate that intelligent behaviour is a result of the processing of stimuli rather than symbols. Sub-symbolic modelling is based on uninterpreted input and distributed representations by dynamic connection weights, e.g. Artificial Neural Networks comprise interconnected networks of simple processing units. In QA, so-called Neural Question Answering currently dominates the field, see e.g. (Weston et al., 2015). Based on neural network models, the systems involve relatively small pipeline, but require a significant amount of annotated data.

Recently, a number of approaches have been devised proposing a combination of symbolic and sub-symbolic processing. It has been shown that fundamental to human cognitive abilities is the capacity to process *concepts* which emerge from a distributed connectionist representation at a lower level where stimuli are processed, and are combined to form symbolic structures at the highest level to support understanding and reasoning, see e.g. (Gärdenfors, 2004). For a QA system, it implies that questions understanding and answers retrieval can be modelled at a higher level of semantic abstraction mining key concepts from available

[2] http://www.apple.com/ios/siri/

[3] www.wolframalpha.com

(e.g. medical) taxonomies, mapping them to (parts of) EATs, which on their turn can be used to perform data-driven entity recognition and semantic relation classification tasks.

Over the past few years, the community proposed different approaches to generate taxonomies which range from flat lists of mutually exclusive categories to hierarchical taxonomies with coarse categories subdivided into fine-grained classes. For example, Srihari and Li (1999) used the defined MUC NER categories to derive their taxonomy. Kim et al. (2001) created a taxonomy for semantic categorization of questions and candidate answers based on the WordNet categories. Chuang and Chien (2003) clustered queries with similar information needs into groups. Hereby, higher ranked results from web search engines were used as features to create multi-way trees via hierarchical clustering.

Chilton et al. (2013) applied crowdsourcing techniques to generate taxonomies based on three Human Intelligence Tasks (HITs) where different groups of participants: (1) generate a category for each shown item; (2) decide which items and the generated categories fit the best; and (3) decide for each category whether an item fits in it or not.

The conceptual complexity of medical domains, can make it difficult for users of information systems to comprehend and interact with the knowledge embedded in those systems (Wickens et al., 1998). To give an example, the Unified Medical Language System (UMLS)[4] integrates over 2 million names for 900 000 concepts from more than 60 families of biomedical vocabularies, as well as 12 million relations among these concepts. The UMLS semantic network reduces the complexity of this construct by grouping concepts according to the semantic types that have been assigned to them McCray et al. (2001). Medical knowledge bases, ontologies, standard terminologies and lexicons can facilitate many NLP and AI tasks, and are exploited in this work.

4 Methodology

4.1 Medical Term Extraction: The EAT Taxonomy and Seed Terms

Ideally, doctors want to meet competent patients who understand their medical needs and potential consequences of medical decisions, so that doctors can be sure that the patient's consent is well-informed. It is a common practice nowadays that before meeting a doctor who will plan an operative medical procedures, patients often have to fill in Patient Education Forms (PEF) to understand procedures and risks involved, and ask their doctors more precise and in-depth questions. To model system's QA behaviour for our use case, the reference PEF[5] was analysed to extract the domain-specific semantic concepts and grouped them into 14 categories using the UMLS semantic network. The form consists of 1,886 tokens, from which 448 tokens (261 unique tokens) were identified as medical entities. Thus, in theory a patient can ask 261 question to the system requesting additional information or explanation. The resulted taxonomy (Table 1) was used to annotate 64 PEFs in German, cluster dictionary terms and to define the EAT to classify questions and retrieve definitions for the system's answers.

Source	Trustworthy	Available	Accessible
Pschyrembel	+	(+)	-
Wikipedia	(+)	+	+
Wiktionary	(+)	+	+
Roche	+	(+)	-
MedlinePlus	+	+	(+)

Table 2: Overview of the assessed medical online resources. (+ stands for ‘yes’, – for ‘no’, and (+) for ‘partially’. see Section 4.2

The semantic categories were populated with relevant seed terms. For this, the dictionary was created using a medical word list available on Wiktionary[6]. We first matched the PEFs seed terms to the lexicon entries using dictions (i.e. case, number), lemma, and/or stem, and enriched it further with synonyms, hyponyms, and hypernyms using the Wiktionary relations.

To improve the coverage of the proposed term set, we augmented the list with entities from available online unstructured medical data in a distant supervision setting. The classifiers, Naive Bayes (NB) and Multinominal Naive Bayes (MNB), were trained operating on different types of lexical and linguistic (e.g. words, lemmas, stems and as-

[4] https://www.nlm.nih.gov/research/umls/

[5] The form in English, German, French, Italian, Serbian and Turkish can be found here: https://www.oegari.at/arbeitsgruppen/arge-praeoperatives-und-tagesklinisches-patientenmanagement/937.html

[6] https://de.wiktionary.org/wiki/Verzeichnis:Deutsch/Medizin An XML dump of the German Wiktionary was used: https://dumps.wikimedia.org/dewiktionary/20181001/

signed POS tags), orthographic (e.g. capitalization information and word length), morphological (e.g. inflections) and contextual features where preceding and following words as well there POS tags are encoded as bi- and tri-grams.

4.2 Resources for Definition Retrieval

Users of medical QA systems want to get medical information which is accurate, not misleading or fake. The verified sources, in our view, need to fulfil the following criteria:

1. **Trustworthiness**: the resource is accepted as medical information source;
2. **Availability**: the resource is distributed under non-exclusive license agreements with no costs associated with its use;
3. **Accessibility**: the resource can be crawled from the website or there are APIs available.

For example, the **Pschyrembel**[7] is the most referred clinical German database. Although there exists a free online test version, a license is required for a complete access. The website can not be crawled. Pschyrembel is an excellent source for medical terminology even for laymen, since the definitions are very well explained and concise, include synonyms and an English translation.

Wikipedia[8] and **Wiktionary**[9], published by the *Wikimedia Foundation*[10], are freely available databases. Generally, the Wikimedia databases are good information sources, however can not be considered as trusted medical resources. Both resources are available in many different languages enabling terms alignment and translation. Wiktionary definitions are mostly one-sentence short explanations capturing the term meaning in general, whereas Wikipedia often provides long and detailed descriptions of multiple related aspects. There are interfaces available to access the data.

Other surveyed medical resources are **Roche Lexikon Medizin**[11] for German and **MedlinePlus**[12] of the *US National Institute of Health* for English. However, their trustworthiness comes with a price, see Table 2 for a comparison.

Dataset	#texts	#tokens	#NE
Training set			
PEFs	64	55,280	7,333
Wikipedia articles	6,865	4,017,388	262,337
Full training dataset	6,929	4,072,668	263,568
Test set			
PEF	1	1,886	448

Table 3: Training and test datasets.

4.3 QA System Architecture

The designed QA system consists of three core modules performing pre/post-processing, term extraction and definition retrieval. A general overview of the system is depicted in Figure 1.

Patient's input and available resources are *preprocessed*, e.g. tokenized, segmented; language models and (multilingual) word embeddings are computed. As output, vectors representing questions and documents are generated.

The next step is concerned with medical entities recognition. The *medical term extractor* exists in two versions (Section 5.2.1). The dictionary-based (DB) extractor annotates tokens depending on their presence in the dictionary. The module takes different parameters specific for the matching process such as word, dictions, lemma, stem, and case-(in-)sensitive matching. The machine learning (ML) classifier operates on the computed features discussed above, applies the trained prediction models and extracts the relevant entities.

To query online resources either unstructured online contents or available medical knowledge bases, queries are formulated containing the EAT concepts extended with the collected (multilingual) seed terms. The expanded queries are also transformed into the signature vectors to measure the semantic similarity with the previously computed document vectors. Multiple *definitions* can be *retrieved* and ranked. For the generation of system's answers, definitions can be summarized (Hardy et al., 2002), simplified or lexically/syntactically adapted, see e.g. (Wang et al., 2016).

5 Experiments

5.1 Data

The data used in our IE experiments is of two types: (1) dictionaries, and (2) medical free texts as training and test data for classifiers.

[7] https://www.pschyrembel.de/

[8] https://de.wikipedia.org/wiki/Wikipedia:Hauptseite

[9] https://de.wiktionary.org/wiki/Wiktionary:Hauptseite

[10] https://wikimediafoundation.org/

[11] https://www.roche.de/lexikon/index.htm?loc=www.roche.de&content=/lexikon/suche.html

[12] https://medlineplus.gov/healthtopics.html

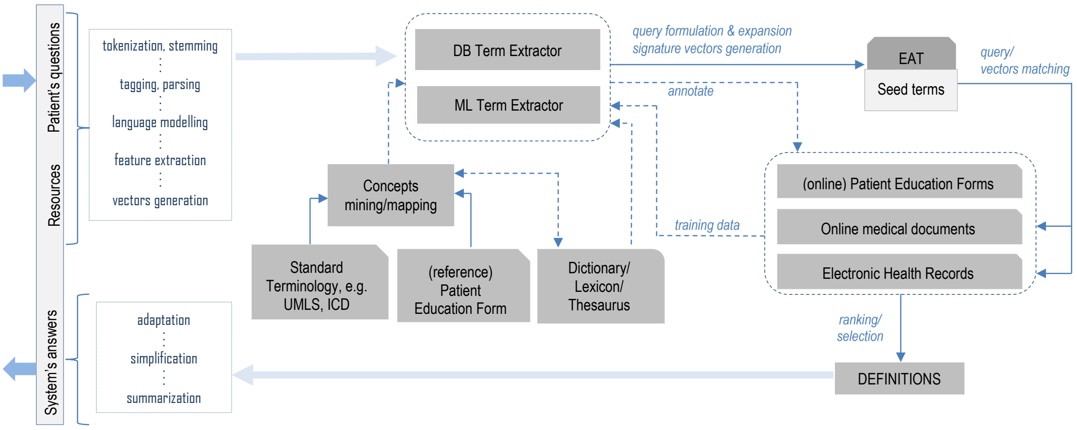

Figure 1: Proposed QA system architecture. From left to right: patient's questions and available medical documents are processed. Medical terms are extracted using available concepts taxonomies, terminologies, medical dictionaries and are learned from the annotated data. Concepts and terms are mapped to the EAT to formulate and expand the query. Signature vectors for questions and answer candidates are computed. The verified (un-)structured data/knowledge sources are queried to retrieve definitions which are ranked and post-processed before returning to the patient.

Dictionary comprises the initial word list of 2,035 Wiktionary medical term entries.[13] The word list covers different fields of medical work from general medicine to dentistry, and contains a mix of Latin and German names of medical procedures, tools and events. We augmented this list with the Wiktionary technical terms[14] and Wikipedia medical terms[15]. The resulting cleaned term list comprises 12,711 terms, see Table 4.

Word list	#terms	P	R	F1
Wiktionary medical	2035	0.947	0.120	0.213
Wiktionary medical + technical	2485	**0.949**	0.125	0.220*
Wikipedia medical	11041	0.928	0.285	0.436*
Complete list	12711	0.915	**0.312**	**0.465***

Table 4: Results of word list experiments. Here and in the further Tables, P stands for precision, R - for recall, F1 - for F-scores. *differs significantly from the baseline obtained on the smallest word list according to the McNemar's test, $\alpha < 0.05$

The **training data** consists of the 64 online PEFs and 6,865 Wikipedia articles[16] extracted with the Wikipedia term list. The test data as described above was constructed from a single PEF, see Table 3 for details.

5.2 Results

5.2.1 Medical NE Recognition

We conducted: (1) the dictionary-based, and (2) machine learning NER experiments. For evaluation, the standard metrics of precision, recall, and F-scores were used. The McNemar's tests were performed to measure statistical significance (McNemar, 1947).[17]

Relation depth: synonym,hypernym,hyponym	P	R	F1
0,0,0	0.92	0.31	0.47
0,0,1	0.93	0.41	0.57*
1,1,1	0.85	0.48	**0.61***
1,1,2	0.85	0.48	**0.61***
3,2,2	0.63	0.48	0.55*
2,2,3	0.76	0.48	0.59*
3,1,1	0.63	0.48	0.55*

Table 5: Results of the dictionary-based experiments: the assessment of the relation depth levels. *differs significantly from the baseline 0,0,0 setting according to the McNemar's test, $\alpha < 0.05$

Dictionary-based (DB) term recognition experiments were performed in two steps. First, the dictionary was gradually expanded to improve its coverage. We evaluated the performance using the word lists compiled from *Wiktionary* medical terms and technical medical terms, *Wikipedia* medical terms and combinations of those. Further, we experimented with the depth of Wik-

[13] https://de.wiktionary.org/wiki/Verzeichnis:Deutsch/Medizin

[14] https://de.wiktionary.org/wiki/Verzeichnis:Deutsch/Medizin/Fachwortliste

[15] https://de.wikipedia.org/wiki/Portal:Medizin/Index

[16] The Wikipedia articles dump of September 10, 2018 https://dumps.wikimedia.org/dewiki/20181001/ was used.

[17] The null hypothesis for our tests states that two algorithms, applied to the same data, retrieve the same results. The test statistic has a distribution of χ^2 with one degree of freedom. A significance level of $\alpha = 0.05$ was set.

Setting	P	R	F1
baseline	**0.855**	0.477	0.611
+diction	0.810	0.635	0.712*
+lemma	0.787	0.666	0.721*
+stem	0.807	0.641	0.715*
-case	0.847	0.481	0.614
+diction -case	0.802	0.639	0.711*
+lemma -case	0.789	0.673	**0.726***
+stem -case	0.713	0.657	0.684*
+diction+lemma +stem-case	0.703	**0.679**	0.691*
Naive Bayes	0.639	0.670	0.653
Multinominal Naive Bayes	**0.853**	**0.859**	**0.851***

Table 6: Results of the **dictionary-based** experiments assessing of various matching strategies with relation depth 1,1,1 (best results) and the **classification** performance. *differs significantly from the baseline according to the McNemar's test, $\alpha < 0.05$

tionary *synonyms*, *hypernyms* and *hyponyms* relations. For example, a depth of 2 in the hypernym relation means that the hypernym of the word, and the hypernym of the hypernym is added to the dictionary.

Subsequently, different matching strategies were tested tuning parameters like *lemma*, *stem* and different *inflections* types and combinations of those. The matching was also conducted in *case-sensitive* and *case-insensitive* setting.

From the results presented in Table 4 can be observed that larger dictionaries result in a better system performance in terms of higher F-scores. The expanded dictionary coverage leads to a higher recall, as more relevant terms can be found. The precision, by contrast, drops slightly when larger dictionaries are used, due to the larger amount of false positives. For our use case, we assume that the system's acceptance will depend on its ability to explain as many terms as possible than missing many relevant of them.

In the second set of experiments, we assessed the impact of the relation depth on the term extraction performance. As Table 5 shows that encoding the relation information of 1,1,1 and 1,1,2 types resulted in the best performance (F-scores of 0.611). We concluded that recall increases with the increased relation depth. Deeper relations, however, generate more out-of-domain terms causing the precision drop. For example, the further up the hypernym relation gets, the more general the terms become. Considering synonyms of all word senses introduce further noise in the training data, e.g. the German word ‘Nase / nose’ is also a fish and the synonym list does not only contain ‘Riechorgan / olfactory organ’ or ‘Zinken / beak’, but also ‘Näsling / common nase’, which is a kind of carp and is unlikely to occur in PEFs.

Features	P	R	F1
word	0.875	0.879	0.876
+POS	0.876	0.880	0.877
+Suffix	0.879	0.882	0.879*
+Prefix	**0.882**	**0.885**	**0.883***
+nextBigramPOS	0.877	0.880	0.877
+prevBigramPOS	0.879	0.882	0.880*
Best features	**0.909**	**0.909**	**0.909***

Table 7: Classification performance on different feature sets. Note: only features that improved the previously obtained results are reported here. *differs significantly from the word baseline according to the McNemar's test, $\alpha < 0.05$

In the final dictionary experiments, we assessed the impact of the word-based matching strategies. The experiments showed that using lemmas and case-insensitive strategies yielded the best results. The best overall F-score of 0.726 was achieved using the complete Wikipedia and Wiktionary word list, a relation depth set to 1,1,1 for synonyms, hypernyms, and hyponyms respectively, lemmatising and ignoring capitalizations in the input. The performance of the best dictionary-based extractor outperforms the Wiktionary medical baseline by broad margins, compare Tables 4 and 6.

For our **machine learning** (ML) experiments, we generated the training data using the distant supervision approach and the best version of the dictionary-based extractor. The MNB classifier outperformed the NB classifier by broad margins, achieving F-scores of 0.85 comparing to 0.65, consider two last rows of Table 6.

Finally, the impact of different feature combinations on the classifier performance was evaluated. For this, each feature was tested individually in combination with the *word* feature. Results showed that only few features contributed to the improvement of the overall classification performance, see Table 7 for an overview. The best feature combination was found to be a combination of word features and POS information of previous, current and next word, as well as the morphological information concerning prefixes and suffixes.

Our experiments showed that the built classifiers outperformed the dictionary-based extractors. The overall F-scores improvement of 0.183 was achieved. More importantly, the recall was drastically improved from 0.236 (dictionary baseline)

Dictionary-based NE recognition		Machine-learning based NE recognition	
Configuration	**F1**	**Configuration**	**F1**
Baseline: Wiktionary medical data	0.213	Baseline: PEF training data	0.851
Best Lexicon: Wikipedia & Wiktionary data	0.465	Best training data: PEF & Wikipedia articles	0.876
Best relation depth: 1,1,1	0.611	Best feature pair: word+prefix	0.883
Best matching strategy: +lemma, -case	0.726	Best feature combination: word, +POS trigram, +inflexion	0.909

Table 8: Summary of the best obtained results for medical entities extraction.

Source	**# retrieved definitions** (in % of all PEF terms)	**# accepted definitions** (in % of all retrieved)
Wiktionary	133 (51.0)	123 (92.4)
Wikipedia	124 (47.5)	93 (75.0)
Both resources	134 (51.3)	123 (92.4)

Table 9: Coverage and quality of the retrieved Wiktionary and Wikipedia definitions.

and from 0.673 (best dictionary-based system) to 0.909 of the best classification model. Table 8 summarizes the key experimental results.

5.2.2 Definition Retrieval

The proof-of-concept definition retrieval was implemented using Wiktionary and Wikipedia resources that contain clear understandable definitions and are available in many different languages. The methods developed for German and English can be used for many other languages.

On a technical note, the Wikipedia and Wiktionary APIs are available to retrieve the summary part of the corresponding Wikipedia article, and the sense of Wiktionary. **Coverage** of the reference PEF medical terms and the **quality** of the retrieved definitions were evaluated.

Both resources covered 51.3% of the annotated PEF terms: 47.5% for Wikipedia and 51.0% for Wiktionary. The retrieved definitions were evaluated on their acceptability: whether the definition is *correct*, *clear* and *sufficient*. The evaluation was performed by three human raters. Out of the 133 Wiktionary definitions, 123 (92.4%) definitions were evaluated as acceptable: wording and sentence structure were simple, i.e. not containing other complex terminology and more than one subordinate clause. The retrieved Wikipedia definitions were, by contrast, evaluated as less acceptable: multi-sentence definitions are frequent with complex sentence structures using other medical expressions. The assessment results for the definition coverage and quality from the respective sources can be found in Table 9.

6 Conclusions and Future Work

In this paper, we addressed medical terms and definitions extraction simulating Patient Education QA sessions. We assessed two core methods to medical terms extraction: based on the standard medical terminologies and available dictionaries, and applying a machine-learning approach to extract German seed terms in a distant supervision setting expanding the system's coverage. We also proposed criteria to test and select medical resources for a QA application. A proof-of-concept definition retrieval systems was implemented and evaluated. The work contributes to a closed-domain QA system design to facilitate access to verified multilingual medical information.

The baseline DB and ML-based extraction techniques are assessed considering various dictionaries/datasets sizes, word matching strategies and different feature combinations. The distant supervision is a viable method to overcome the shortage of manually annotated monolingual data and can be successfully applied to automatically and productively generate large sets of the annotated multilingual seed terms. The proposed term-based information extraction opens perspectives for multi- and cross-lingual QA application design. The concepts categories populated with terms in multiple languages enable cross-lingual mappings. If the language is available on Wiktionary, the relational connections can be used as well.

Our future work will pursue multiple goals. To improve the quality, a larger annotated corpus for German will be collected. Larger data sets will also allow to train machine learning classifiers on noisy labelled data. Different search and retrieval methods will be explored, i.e. based on machine translation, cross-lingual language models and multilingual embeddings. In particular, we are interested in training new neural networks in multi- and cross-lingual term extraction and definition retrieval settings. We also plan to invest into the adaptation and simplification of the retrieved definitions where the complex medical terms will be translated into common terms. This can be achieved in a dictionary-based setting augmenting seed terms collections, but also defining the task as a machine translation one.

References

Enrique Alfonseca and Suresh Manandhar. 2002. An unsupervised method for general named entity recognition and automated concept discovery. In *Proceedings of the 1st international conference on general WordNet, Mysore, India*. pages 34–43.

Jari Björne, Suwisa Kaewphan, and Tapio Salakoski. 2013. Uturku: drug named entity recognition and drug-drug interaction extraction using svm classification and domain knowledge. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. volume 2, pages 651–659.

G Bonacruz Kazzi and Carolyn Cooper. 2003. Barriers to the use of interpreters in emergency room paediatric consultations. *Journal of paediatrics and child health* 39(4):259–263.

Lydia B Chilton, Greg Little, Darren Edge, Daniel S Weld, and James A Landay. 2013. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pages 1999–2008.

Laura Chiticariu, Yunyao Li, and Frederick R Reiss. 2013. Rule-based information extraction is dead! long live rule-based information extraction systems! In *Proceedings of the 2013 conference on empirical methods in natural language processing*. pages 827–832.

Shui-Lung Chuang and Lee-Feng Chien. 2003. Automatic query taxonomy generation for information retrieval applications. *Online Information Review* 27(4):243–255.

Adam L Cohen, Frederick Rivara, Edgar K Marcuse, Heather McPhillips, Robert Davis, et al. 2005. Are language barriers associated with serious medical events in hospitalized pediatric patients? In *peds*. volume 2005:0521, page 116.

Julia Dembowski, Michael Wiegand, and Dietrich Klakow. 2017. Language independent named entity recognition using distant supervision. In *Proceedings of Language and Technology Conference (LTC)*.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics* 47:1–10.

Richard Evans and Stafford Street. 2003. A framework for named entity recognition in the open domain. *Recent Advances in Natural Language Processing III: Selected Papers from RANLP* 260(267-274):110.

D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, J. Murdock, E. Nyberg, J. Prager, N. Schlaefer, and C. Welty. 2010. Building watson: An overview of the deepqa project. *AI Magazine* 31(3):59–79.

Peter Gärdenfors. 2004. *Conceptual spaces: The geometry of thought*. MIT press.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. volume 1.

Harsha Gurulingappa, Roman Klinger, Martin Hofmann-Apitius, and Juliane Fluck. 2010. An empirical evaluation of resources for the identification of diseases and adverse effects in biomedical literature. In *2nd Workshop on Building and evaluating resources for biomedical text mining (7th edition of the Language Resources and Evaluation Conference)*.

Hilda Hardy, Nobuyuki Shimizu, Tomek Strzalkowski, Liu Ting, Xinyang Zhang, and G Bowden Wise. 2002. Cross-document summarization by concept classification. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pages 121–128.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. 2003. Genia corpusa semantically annotated corpus for bio-textmining. *Bioinformatics* 19(suppl_1):i180–i182.

Soo-Min Kim, Dae-Ho Baek, Sang-Beom Kim, and Hae-Chang Rim. 2001. Question answering considering semantic categories and co-occurrence density. In *AUTHOR Voorhees, Ellen M., Ed.; Harman, Donna K., Ed. TITLE The Text REtrieval Conference (TREC-9)(9th, Gaithersburg, Maryland, November 13-16, 2000). NIST Special Publication. INSTITUTION National Inst. of Standards and Technology, Gaithersburg, MD.; Advanced Research Projects Agency (DOD), Washington, DC.*. Citeseer, page 261.

Mack Lipkin, Richard M Frankel, Howard B Beckman, Rita Charon, and Oliver Fein. 1995. Performing the interview. In *The medical interview*, Springer, pages 65–82.

Alexa T McCray, Anita Burgun, and Olivier Bodenreider. 2001. Aggregating umls semantic types for reducing conceptual complexity. *Studies in health technology and informatics* 84(0 1):216.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12(2):153–157.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*. Association for Computational Linguistics, pages 1003–1011.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30(1):3–26.

Tomoko Ohta, Sampo Pyysalo, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Open-domain anatomical entity mention detection. In *Proceedings of the workshop on detecting structure in scholarly discourse*. Association for Computational Linguistics, pages 27–36.

Kathryn Pitkin Derose, Benjamin W Bahney, Nicole Lurie, and José J Escarce. 2009. Immigrants and health care access, quality, and cost. *Medical Care Research and Review* 66(4):355–408.

Sunil Kumar Sahu and Ashish Anand. 2017. Unified neural architecture for drug, disease and clinical entity recognition. *arXiv preprint arXiv:1708.03447* .

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association* 17(5):507–513.

Satoshi Sekine and Chikashi Nobata. 2004. Definition, dictionaries and tagger for extended named entity hierarchy. In *LREC*. Lisbon, Portugal, pages 1977–1980.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*. Association for Computational Linguistics, pages 104–107.

Dan Shen, Jie Zhang, Guodong Zhou, Jian Su, and Chew-Lim Tan. 2003. Effective adaptation of a hidden markov model-based named entity recognizer for biomedical domain. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13*. Association for Computational Linguistics, pages 49–56.

Rohini Srihari and Wei Li. 1999. Information extraction supported question answering. Technical report, CYMFONY NET INC WILLIAMSVILLE NY.

Lorraine Tanabe and W John Wilbur. 2002. Tagging gene and protein names in biomedical text. *Bioinformatics* 18(8):1124–1132.

Yoshimasa Tsuruoka and Jun'ichi Tsujii. 2003. Boosting precision and recall of dictionary-based protein name recognition. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13*. Association for Computational Linguistics, pages 41–48.

Tong Wang, Ping Chen, John Rochford, and Jipeng Qiang. 2016. Text simplification using neural machine translation. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698* .

Christopher D Wickens, Sallie E Gordon, Yili Liu, et al. 1998. *An introduction to human factors engineering*. Longman New York.

Robert Wilensky, David N Chin, Marc Luria, James Martin, James Mayfield, and Dekai Wu. 1988. The berkeley unix consultant project. *Computational Linguistics* 14(4):35–84.

Shaodian Zhang and Noémie Elhadad. 2013. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of biomedical informatics* 46(6):1088–1098.

GuoDong Zhou and Jian Su. 2002. Named entity recognition using an hmm-based chunk tagger. In *proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pages 473–480.