Fundamentals of Clinical Text Extraction in NLP

Research Paper

.

presented at the Department IV
of Trier University

by

Chakravenu Gandham

s4chgand@uni-trier.de

Supervisor: Prof. Dr. Naumann, Sven

Trier, September 29, 2023

Contents

1 Introduction

Healthcare systems, particularly health record systems, encompass a wealth of both structured and unstructured data in the form of text. It is estimated that more than 40% of data within healthcare record systems consist of clinical text, which includes essential information related to symptoms, diagnoses, treatments, drug usage, and adverse events. This data benefits individual patients and holds great potential for improving healthcare outcomes on a broader scale. Additionally, healthcare professionals often provide their clinical reasoning within patient records [12].

However, within this trove of clinical text lies sensitive information, including personal details like names, phone numbers, and addresses of both patients and their relatives. To unlock the utility of clinical text for secondary use and research, this sensitive information must undergo pseudonymization, safeguarding patient confidentiality [35].

A significant portion of electronic patient records comprises unstructured free text. This clinical text, documented by various healthcare professionals, is often hurriedly composed, riddled with misspellings, non-standard abbreviations, jargon, incomplete sentences, and poses challenges for natural language processing (NLP) tools designed for standard text sources [2]; [28].

While substantial research has been conducted on clinical text processing in English, the same level of attention has not been dedicated to smaller and under-resourced languages, such as Swedish. Nevertheless, several projects have emerged, especially at DSV, Stockholm University, between 2007 and 2017, focusing on clinical text mining in the Swedish language [21].

This paper aims to explore the clinical text processing steps that involve Segmentation and Tokenisation, lemmatization, Stemming, Compound Splitting, Abbreviation Detection and Expansion, POS Tagging, SYntactical Analysis, Semantic Analysis and Concept Extraction, Relation Extraction and Anaphora resolution.

2 Fundamentals of Clinical Text Extraction

2.1 Definitions

In the realm of computer science and linguistics, Natural Language Processing (NLP) represents the conventional approach to intelligent text analysis. Within NLP, computer programs endeavor to decipher and comprehend natural language text or spoken language through the application of computational linguistic techniques. Synonymous terms for NLP encompass computational linguistics, language engineering, and language technology.

Information Retrieval (IR), while potentially leveraging NLP techniques, distinguishes itself by focusing on locating a specific document within a larger document collection. In contrast, Information Extraction (IE) concentrates on pinpointing particular pieces of information within a document or document collection. In contemporary discourse, the term "text mining" has gained prominence, denoting the process of uncovering previously undisclosed facts within a corpus of text or formulating hypotheses for subsequent validation. It's important to note that the usage of the term "text mining" can vary; at times, it implies the utilization of machine learning-based approaches. In the field of health informatics, "text mining" predominantly refers to the application of rule-based methods for processing clinical or biomedical text.

2.2 Basics of Clinical Text Processing

In clinical text processing, the fundamental building blocks employed for regular texts can generally be adapted. Nevertheless, clinical texts introduce additional challenges due to the presence of incomplete sentences, misspelled words, and non-standard abbreviations, which can complicate NLP tasks, more information in theses books: Mitkov [20]; Jurafsky and Martin, [18]; Clark [7]).

2.2.1 Segmentation and Tokenisation

Segmentation serves as an essential initial phase in Natural Language Processing (NLP). It encompasses the separation of sentences from one another (referred to as sentence segmentation) and also the isolation of individual words within those sentences (known as word segmentation). Typically, sentences are demarcated by punctuation marks such as periods, question marks, or commas. However, it's imperative to exercise caution because periods and commas can also appear within numerical values, necessitating a meticulous approach to segmentation.

In languages like Chinese and Japanese, where words lack spaces between them, the task of word segmentation becomes considerably more challenging. In such cases, the assistance of a word segmenter becomes indispensable.

A text comprises a continuous flow of words, alphanumeric characters, whitespace, inter-punctuations, and carriage returns, collectively known as tokens. Tokenization represents the subsequent step in natural language processing, involving the determination of what constitutes a token (or word) and what should be subject to analysis within the input character stream. Therefore, a tokenization tool is employed to extract these tokens from a sentence or text, preceding the actual natural language processing tasks.

Below 2.2.1 provides a visual representation of a tokenized sentence. Tokenizers face various decisions, such as whether to include sentence delimiters, how to handle constructions like "let's" (treated as one unit or split), and how to tokenize complex expressions like "400 mg/day" or "x-tra." Typically, a standard tokenizer, which is often integrated into many NLP tools, can be employed and customized for specific domains. Alternatively, a completely new tokenizer can be developed to suit particular requirements.

In general, whitespace and sentence delimiters like commas and question marks serve as reliable markers for words, facilitating the role of a basic tokenizer. However, clinical texts are notorious for their noise, replete with non-standard expressions and abbreviations. Consequently, tokenization in the clinical context can be a challenging endeavor.

The patient has signs of tuberculosis in his left lung, let's try a new treatment with x-tra Isoniazid, 400 mg/day.

This Sentence id tokenised as

"The" "patient" "has" "signs" "of" "tuberculosis" "in" "his" "left" "lung" "," "let" "'" "s" "try" "a" "new" "treatment" "with" "x" "-" "tra" "Isoniazid" "," "400" "mg" "/" "day" "."

Text 2.2.1 Example of Sentence and Tokenization

In the work of Patrick and Nguyen [23], the authors provide an insightful exploration of the challenges associated with tokenizing clinical text. They offer solutions, exemplifying

cases like "HR 72," which encompasses an acronym for heart rate measurement and the value "72," necessitating treatment as a single unit, namely "heart rate 72." Additionally, the article addresses the tokenization of dates, like "3/7/02," where the entire date needs to be treated as a single unit, reflecting the intricacies of clinical text processing [23].

2.2.2 Morphological processing

Following tokenization, the subsequent stage involves the morphological processing of each token. Morphological processing entails the analysis of a word's morphemes, encompassing various components like inflections such as prefixes, infixes, or suffixes. It's worth noting that certain tokens combine to form multiword expressions, consisting of two or more consecutive tokens that necessitate joint processing, such as "heart rate." In clinical text, there is a prevalent presence of combinations involving abbreviations and their corresponding full forms, further adding to the complexity of morphological processing.

Lemmatisation

It is the process of identifying the base form of a word, a particularly valuable task for languages with complex inflections like German, Polish, Russian, and Swedish. English, characterized by simpler morphology, doesn't require as sophisticated lemmatisation. A lemmatiser transforms words into their base or lemma forms, reducing the variation caused by different inflected forms of the same word to a single lemma. This simplification aids many natural language processing systems in word and meaning analysis. Numerous off-the-shelf lemmatisers are readily available, often integrated into taggers that determine a word's class or function 2.2.2.

Regarding inflected languages such as Swedish and German, where even nouns inflect, it becomes essential to lemmatise them to obtain their base or lemma forms.

Stemming

In some instances, stemmers can be used instead of lemmatisers. Stemmers provide a more basic form of lemmatisation by reducing words to their stems, which may not necessarily constitute real words (lemmas), but rather representations of the words. For instance, inflected words like "pathology," "pathologies" "pathological," and "pathologically" can be stemmed to "pathology," which indeed is a real word.

Useful stemmers, such as those available in the Snowball system, offer practical solutions for stemming, and the associated GitHub repository also provides stop word lists for various languages, aiding in corpus preprocessing. Stop words are non-functional words like "and", "or", "in", "on", "also", "with" etc., which often make up around 40%

of the words in a text. Stemming can impact recall and precision in retrieval settings, sometimes increasing precision as well [4].

Compound Splitting (Decompounding)

In languages like Swedish and German, which employ compounding, some tokens need to be decompounded to their base forms. Compound splitting, or decompounding, involves the use of dictionaries in conjunction with decompounding rules.

For example, "diabetespatient" in Swedish should be decompounded to "diabetes patient," and the plural form "diabetespatienter" to "diabetes patienter," which can then be lemmatised to "diabetes patient." Decompounding is essential in preserving the meaning of words in such languages.

Abbrevation Detection and Expansion

Clinical text typically contains abbreviations, making up 3% to 10% of the text, with a significant portion being ambiguous, necessitating disambiguation. The process involves abbreviation detection and subsequent expansion or normalization.

Several methods exist for detecting abbreviations, including dictionary matching, morphology-based analysis, and heuristic rule-based approaches. Dictionaries and rules are employed to identify abbreviations and expand them to their full forms.

Machine learning-based methods have also been utilized for abbreviation detection, demonstrating high precision and recall. Additionally, synonym extraction for abbreviations has been explored using distributional models and semantic spaces. These approaches have improved the understanding of abbreviations in clinical text.

Part-of-Speech Tagging (POS Tagging)

The subsequent stage in the NLP pipeline is part-of-speech tagging, a critical process that automatically assigns grammatical functions to words within a text. These functions encompass determiners, subjects, predicates, adjectives, adverbs, prepositions, and more. Additionally, other forms of tagging, such as semantic or thematic tagging, may be employed for specialized purposes.

While each word typically belongs to a specific word class, such as noun, adjective, adverb, determiner, or preposition, the way words function within a sentence depends on their relationships with other words. For instance, they may serve as the subject, predicate, direct object, or indirect object. It is the role of a part-of-speech tagger to discern and label these functions accurately.

Constructing part-of-speech taggers involves various methods, including rule-based or dictionary-based approaches. However, cutting-edge techniques employ machine learning methods. For many languages, specific taggers are available, often accompanied by language models tailored to each language. For English, one example is the TnT tagger, which can be trained on manually annotated corpora. Similarly, for Swedish, there are taggers like the Granska tagger and Stagger—the Stockholm Tagger—each designed for precise linguistic analysis (or tagging) of the respective language.

2.2.3 Syntactical Analysis

To accurately parse natural language, machine learning systems are trained on manually annotated text data, allowing them to automatically generate grammar rules. These grammar models are then applied to parse text. In this process, parsers also leverage information provided by taggers.

Parsers can employ various parsing strategies, including bottom-up, top-down, incremental parsing with noisy text, backtracking, left-to-right, and dependency-based parsing. The parser's output is a syntactic tree that delineates the different components of a sentence.

One viable parser option is MaltParser, known for its dependency-based parsing approach. MaltParser offers pre-trained models for several languages, including Swedish, English, French, and Spanish [22]

Hassel et al. (2011) [14] employed MaltParser with a pre-trained model for Swedish, applying it to Swedish clinical text. Remarkably, they achieved a commendable part-of-speech tagging accuracy of 92.4%. It's worth noting that this result is particularly noteworthy given the inherent noise in clinical text, which operates within a domain entirely distinct from the domain on which MaltParser's pre-trained model was originally trained. It's also essential to mention that MaltParser requires input text that has been morphosyntactically disambiguated using a tagger.

Chunking

Shallow parsing, also known as chunking or "light parsing," falls between part-of-speech tagging and full parsing. A shallow parser or chunker identifies constituent parts in sentences in the form of nominal phrases or verb phrases.

Grammer Tools

Numerous tools are available for creating grammars used in syntactic text analysis or parsing. One such grammar format is the definite clause grammar (DCG) 2.1 , imple-

```
sentence --> noun_phrase, verb_phrase.
noun_phrase --> det, noun.
verb_phrase --> verb, noun_phrase.
det --> [the].
det --> [a].
noun --> [cat].
noun --> [bat].
verb --> [eats].
```

Figure 2.1: Example of a toy DCG in Prolog, that can parse the sentence: "The cat eats the bat", and some variations on this, (from Wikipedia)

mented in the Prolog logic programming language. In DCG , a grammar can be expressed in an abstract form and then compiled into executable Prolog code. Prolog, originally developed in 1972, was initially intended for natural language processing tasks.

DCG-based grammars, and Prolog as a whole, can be viewed as sets of theorems, with lexical items serving as facts that need to be proven using a theorem prover, which is the built-in Prolog interpreter. If the facts align with the theorems, it signifies that the syntax of the analyzed sentence is correct. DCG offers the additional advantage of being easily extendable to generate syntax trees for subsequent operations. DCG bears a resemblance to the Backus-Naur Form (BNF), a notation used for describing grammars.

Other tools, originally designed for developing compilers for formal languages, such as Lex (Lexical Analysis) and Yacc (Yet another compiler-compiler), are also available and are built into the Linux operating system. These tools can be adapted for constructing parsing tools for natural languages.

2.2.4 Semantic Analysis and Concept Extraction

Semantic analysis, also known as text analytics, involves interpreting the meaning of identified entities within text. This analysis can encompass a range of tasks, from assigning meaning to parts of a syntactic parse tree to identifying named entities, negations, factuality, and uncertainty. More complex semantic tasks include relation extraction, temporal processing, and anaphora resolution.

Named Entity Recognition (NER)

Named Entity Recognition (NER) was initially defined as identifying personal names, locations, organizations, and time points or dates. In clinical text, NER extends to identifying and de-identifying text, including personal names, addresses, telephone numbers, symptoms, disorders (diseases), drugs, and body parts.

Traditionally, NER used name lists (Gazetteers) combined with regular expressions. However, modern approaches predominantly employ machine learning techniques, except for extracting numerical expressions like telephone numbers or drug dosages, where regular expressions are more efficient.

In a study by Skeppstedt et al. (2014) [27] , patient records from a Swedish internal medicine emergency unit were annotated with named entities such as disorders, findings, pharmaceutical drugs, and body structures. Inter-annotator agreement (IAA) scores were calculated for these categories, demonstrating the feasibility of NER in clinical texts.

Machine Learning for Named Entity Recognition

Machine learning approaches have become prominent in NER. Skeppstedt et al. (2014) [27] employed machine learning techniques, specifically the Conditional Random Fields (CRF) algorithm, for NER. They extracted features based on terminology from diagnosis codes, SNOMED CT, MeSH, pharmaceutical drugs, part-of-speech (POS) tagging, and lemmatization. These features were matched with tokens in the training text, and CRF++ was used for training. Their model achieved F-scores of 0.81 for disorders, 0.69 for findings, 0.88 for pharmaceutical drugs, 0.85 for body structures, and 0.78 for the combined category of disorders and findings.

The system developed by Skeppstedt et al. (2014) [27], named Clinical Entity Finder (CEF), has been used as a pre-annotation system in subsequent research of Henriksson et al. (2015) [16].

In 2.2 we can see previous study results, allowing for a comparison with several previous clinical Named Entity Recognition (NER) studies. In this diagram, the results from the same studies are linked by lines: solid lines represent the results from the present study, while dashed lines represent results from previous studies.

The diagram 2.2 is organized into columns, with the first column on the left displaying the outcomes of three rule-based studies. In contrast, the second column presents the findings of two machine learning studies, although the exact number of training instances used in these studies was not disclosed.

The remaining portion of the diagram illustrates the outcomes of several machine learning studies for which the number of training instances was reported. To facilitate comparisons between the present study and previous research, the entity names used in the present

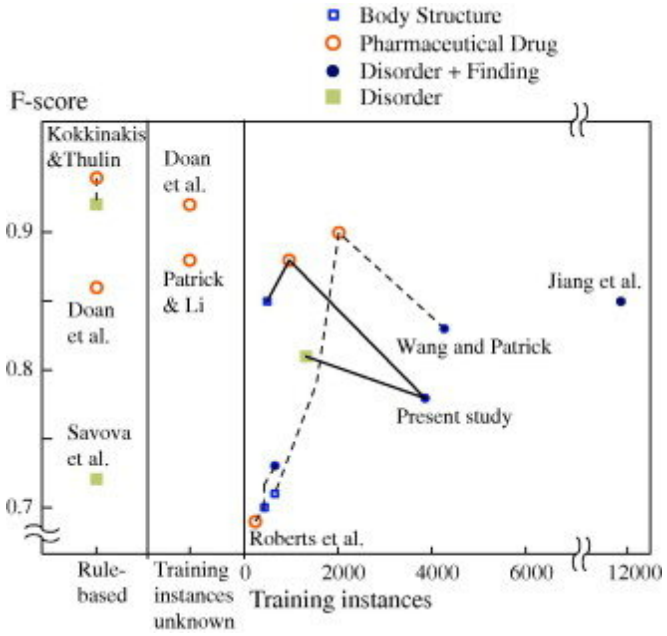

Figure 2.2: The clinical entity recognition systems are depicted in a diagram.Diagram and text cited from Skeppstedt et al. (2014) [27]. The present study is the study by Skeppstedt et al. (2014) [27]

study are utilized to denote equivalent entity types in the earlier studies.

In the 2010 i2b2/VA challenge, the best assertion extraction system for clinical named entity recognition utilized Conditional Random Fields (CRF) and achieved an F-score of approximately 0.90.

Regarding medication identification, a 2010 i2b challenge focused on drug names, brand names, dosages, modes, frequencies, and reasons. While rule-based approaches performed well, hybrid systems combining machine learning algorithms (e.g., CRF and SVM) with pattern-matching rules proved to be the most effective.

The Ph.D. thesis by Skeppstedt (2015) [26] provides a comprehensive overview of studies on clinical entity recognition, encompassing both rule-based and machine learning-based approaches.

Negation Detection and Trigger Lists

The NegEx algorithm, designed for detecting negations, operates through the utilization of regular expressions. NegEx employs three distinct negation trigger lists along with one list encompassing findings and disorders. These lists are cross-referenced with the input clinical text string to ascertain whether a concept (finding or disorder) within the string is negated (Chapman et al. 2001) [5].

The first trigger list, referred to as the pre-negation list, contains potential trigger phrases that should precede the negated word. For instance, it may include phrases like "no signs of." The second list, known as the post-negation list, comprises trigger phrases that should appear after the negated word, such as "unlikely." The third trigger list consists of possible pseudo-negations, which resemble negation triggers but are not indicative of negation, such as "not certain if."

These negation triggers are matched against the input text, and when a negation is identified, the distance from the negation is calculated as the number of words from the negation to the finding or disorder in the input string. This distance to the negation should not exceed six words from the finding or disorder. The list of findings and disorders is based on concepts from the Unified Medical Language System (UMLS).

The original English version of NegEx achieved a precision of 84.5% and a recall of 82.4% when tested on discharge summaries. Subsequently, NegEx was expanded into a version called ConText (Harkema et al. 2009), which also identified historical, hypothetical, and related conditions in patients.

NegEx was adapted for German, obtaining an F-score of 0.9 (Cotik et al. 2016) [9]. It was also adapted for Spanish by Costumero et al. (2014) [8], achieving an accuracy of 84.8%, and for use in French by Grouin et al.(2011) [13], who also extended it to handle conjunctions and potential negations by adding two more trigger lists. The French version of NegEx achieved an F-score of 0.863 when evaluated.

In a study by Chapman et al. (2013) [6], the English NegEx was further extended to accommodate Swedish, French, and German languages and was compared across these languages. The negation triggers "no" and "not" remained similar for all languages. French exhibited the highest diversity in triggers, while German had the least diversity. The presence of agglutination in Swedish and German posed challenges, as it affects the interpretation of negations. For example, in Swedish, "diabetesfri" means "diabetes-free," while "ej diabetesfri" means "has diabetes," a nuance that NegEx cannot easily process. Additionally, French negations vary depending on gender and number agreement, making the construction of trigger lists more complex.

Factuality Detection

In clinical text, many expressions fall into categories like negated, uncertain (speculative), or asserted. Uncertainty in clinical text implies that a statement isn't confirmed or factual and can vary in levels from slight to strong uncertainty. For instance, the English BioScope clinical corpus, consisting of 6383 sentences, was manually annotated for negation, speculation, and scope. Approximately 13.4% of the sentences contained speculative keywords (hedge sentences), and 13.6% contained negations, with some sentences containing both speculations and negations.

In a Swedish clinical corpus known as the Stockholm EPR Diagnosis Factuality Corpus,

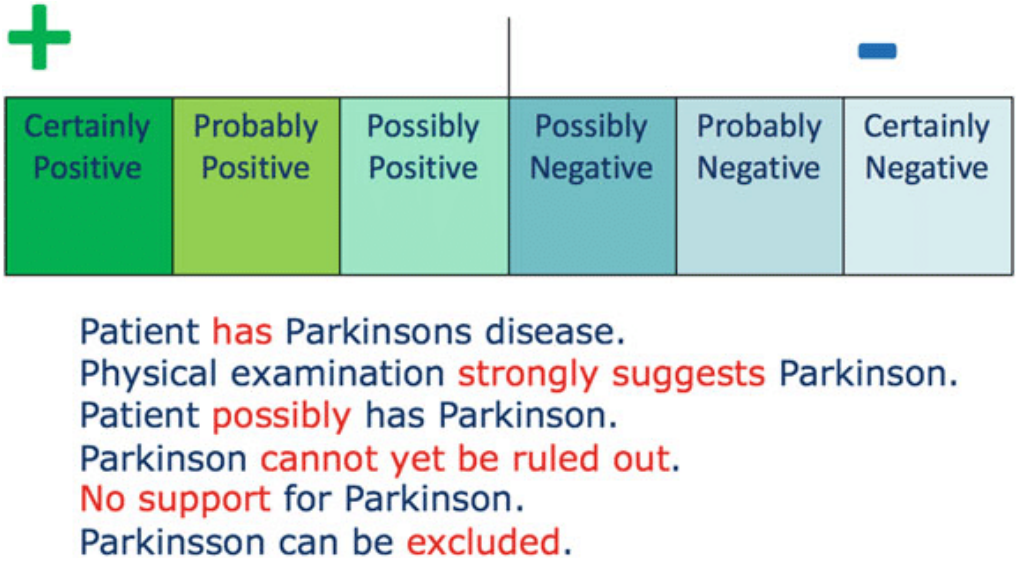

Figure 2.3: Examples of different levels of certainty, ranging from completely affirmed to negated(© 2012 The author—reprinted with permission from the Author. Published in Velupillai (2012) [36] p. 34)

developed to explore uncertainty levels related to diagnostic expressions, a total of 3846 documents containing 26,232 sentences were manually annotated. This corpus, originating from a medical emergency ward, focused on text under the heading "Assessment" or "Bedömning." During annotation, 6483 diagnoses were identified and categorized into six certainty levels: Certainly Positive, Probably Positive, Possibly Positive, Possibly Negative, Probably Negative, and Certainly Negative. See Figure: 2.3 for an example.

Family history Processing (Relative Processing)

In patient records, there are instances where findings and disorders may pertain to the patient's family members due to heredity concerns. When physicians diagnose a patient, they typically search for various symptoms. However, while reviewing the patient's record, if they come across a symptom that is linked to a family member, a family medical history, or an incident from a long time ago, it may not be directly relevant to the current case. Examples illustrating these scenarios are provided in Text 2.2.4.

His father had high hypertension, and now present with a fever of 39°C lasting two days.

or another possible option is:

Hypertension in the family (when it is related to the family of the patient).

or finally:

Had hypertension two years ago (which is a temporal relation).

Example Text 2.2.4 Clinical texts containing relatives or temporal entities.

This aspect of clinical text mining, which involves identifying relatives and distinguishing their symptoms from those of the patient under treatment, is often referred to as relative processing or experiencer analysis. In this context, a more detailed level of analysis can be performed to differentiate the degree of relatedness to the patient, such as identifying whether a family member is a 1st-degree relative or a 2nd-degree relative, as discussed by South et al. (2009) [29].

Relative named entity recognition is also employed in the de-identification of patient records, where the category "relative" is utilized. Entities falling under the "relative" class include terms like father, mother, son, daughter, brother, grandfather, and grandmother.

Temporal Processsing

In clinical text mining, temporal processing plays a crucial role in determining when symptoms, disorders, or events occurred in a timeline. It helps answer questions like whether a symptom is current or if it occurred weeks, months, or years ago. This temporal analysis is essential for physicians to understand disease progression, identify relevant symptoms for diagnosis, and remove non-relevant ones, similar to handling negated symptoms 2.2.4 in the text.

Understanding temporality involves assessing how symptoms relate to the timing of disorders. For instance, knowing that chest pain with radiation preceded angina pectoris helps establish that the former is a symptom of the latter.

Temporal processing in clinical text has its roots in artificial intelligence research. Allen (1984) [1]introduced **seven** time relations: before, meets, overlaps, is-finished-by, contains, starts, and equals. Various approaches have been developed since the 1990s to process temporal relations in clinical text, with notable studies by Zhou and Hripcsak (2007) [38] and Jung et al. (2011) [17].

76-year old woman with hypertension and angina pectoris. Possible heart attack 2 years ago. Admitted to hospital with central chest pain with radiation. Oct 23, underwent PCI,a and now present with a fever of 39°C lasting two days. (Note that admission date is October 18, 2012 obtained from an administrative data source).

Figure: 2.5.6

Figure 2.2.4 Illustrative clinical text depicting various temporal considerations that a computer program must address. In this context, aPCI refers to Percutaneous Coronary

Intervention, typically involving the placement of a stent in a narrowed coronary artery to alleviate angina pectoris or prevent a myocardial infarction (heart attack) from occurring.

Zhou and Hripcsak (2007) [38] categorized three main approaches:

1. Temporal reasoning based on AI theories and models.
2. Frameworks tailored to clinical application needs.
3. Addressing issues related to temporal granularity and uncertainty.

Handling absolute and relative time, as well as combining structured and unstructured time points, has also been explored. Clinical scenarios often involve questions like medication history before surgery, symptom frequency before treatment, or post-medication symptoms, as demonstrated in Sun et al.(2013b) [33].

Researchers, such as Zhou et al.(2005) [37], have developed approaches for temporal tagging of clinical narratives, inventing comprehensive temporal processing frameworks.

Temporal relation identification relies on two key components: recognizing clinical entities (e.g., findings, disorders) and temporal expressions (e.g., date, time, duration) and determining their chronological order. Derczynski (2017) [11] provides a comprehensive analysis of temporal ordering for events and time points, shedding light on this critical aspect of clinical text mining.

TimeML and TIMEX3 TIMEX, a format originating from the named entity research field, deals with annotating temporal expressions, particularly those related to dates and times. TIMEX evolved into TIMEX3 and was incorporated into TimeML, a specification language for temporal expressions in natural language. TimeML conforms to the ISO 8601 standard and addresses four key issues in the markup of event and temporal expressions (Pustejovsky et al., 2003) [24]:
1. Time stamping events,
2. Ordering events in relation to one another,
3. Handling contextually vague temporal expressions (like "last week" or "two weeks before"), and
4. Determining event persistence or duration.

HeidelTime Strötgen and Gertz (2010) [30] created HeidelTime, a rule-based temporal tagging system from Heidelberg University, Germany. It extracts temporal expressions and converts them into TIMEX3 format, following the TimeML standard. HeidelTime is versatile and can be adapted to multiple languages, currently supporting 13 languages, including English, German, Dutch, Vietnamese, Arabic, Spanish, Italian, French, Chinese, Russian, Croatian, Estonian, and Portuguese.

i2b2 Temporal Relations Challenge The i2b2 Temporal Relations Challenge focused on annotating temporal information in clinical records, particularly discharge summaries. It

involved 310 summaries and saw participation from 18 teams. HeidelTime, a rule-based system developed by Strötgen and Gertz in 2010 [30] , was adapted for English clinical text tagging in this challenge and outperformed other systems (Sun et al. 2013a [32]).

The challenge aimed to identify three main concepts: TIMEX3 temporal expressions (e.g., times, dates, duration), FREQUENCY (or SET), and EVENTs representing events or actions, as well as TLINKs denoting temporal relations between them (e.g., before, after, simultaneous). Styler et al. (2014) [31] categorized temporal relations into three types: relations between two events, two times, or a time and an event. Some TIMEX3s, like "twice daily" or "once a week at bedtime," denoted the frequency of an event.

2.2.5 Relation Extraction

Relation extraction is a crucial task in biomedical text mining, involving the identification of connections between entities. It typically follows two main steps: first, named entity recognition (NER) is employed to identify the entities of interest, such as clinical terms, and then relations are established between these entities. NER is a well-established field with high F-scores, often exceeding 0.90, while relation extraction is more challenging and usually achieves F-scores around 0.70 (Uzuner et al. 2011 [34]).

Relation extraction is prevalent in biomedical text mining for establishing connections between various entities, like molecules or chemical substances, to determine if they interact. This interaction can be crucial in understanding biological processes. Luo et al. (2016) [19] provide a comprehensive overview of different approaches to relation extraction in the biomedical domain.

In clinical text mining, one essential task involves detecting adverse drug events by establishing relations between drugs and findings or symptoms. This helps answer questions like whether there is an association between a drug and a particular medical condition.

2010 i2b2/VA Challenge Relation Classification Task

Uzuner et al. (2011) [34] reviewed multiple approaches for detecting clinical relations. However, their article didn't precisely define what constitutes an assertion or a concept, or the distinctions between them. Typically, relations exist between two assertions, two concepts, or between a concept and an assertion. Assertions and concepts often represent medical problems, tests, or treatments.

Among the systems evaluated, de Bruijn et al.(2011) [10] achieved the best results in both assertion detection (F-score of 0.936) and concept extraction (F-score of 0.852), which are prerequisites for relation assertion. They also obtained the second-best results for relation assertion (F-score of 0.731) De Bruijn et al.2011 [10]. employed semi-supervised learning through clustering, utilizing the Brown clustering algorithm on unlabelled clinical corpora for unsupervised learning. They applied the features produced by clustering to a

semi-Markov model.

For relation extraction in clinical text, Support Vector Machines (SVM) proved to be the most effective method in the 2010 i2b2/VA challenge (Uzuner et al. 2011) [34].

Other studies for Relation Extraction

Other studies have explored relation extraction in clinical text as well. Henriksson et al. (2015) [16] annotated a Swedish clinical text for adverse drug events (ADEs) and found F-scores around 0.84 for named entity annotations and lower F-scores around 0.65 for semantic relation annotations. Bejan and Denny (2014)[3] used discharge summaries, achieving high inter-annotator agreement (IAA) with an F-score of 0.979. They employed pre-annotation tools and LIBSVM for training, obtaining an F-score of 0.85. Roberts et al. (2008)[25] conducted one of the earliest attempts at relation extraction from clinical text, achieving moderate IAA and an F-score of 0.72 for a class of seven relation types using SVM.

2.2.6 Anaphora resolution

Anaphora resolution, also known as co-reference resolution, is the task of identifying which entity, such as a pronoun or a noun phrase, a given anaphor (pointing back reference) refers to. The entity being referred to is called the antecedent. This process is crucial in natural language understanding.

Anaphora resolution can occur within a sentence (intrasentential) or across sentence boundaries (intersentential). It's typically based on factors like gender and number agreement, but various methods can be employed, and sometimes resolution is not achieved. This area of computational linguistics is extensively studied but not entirely solved.

For instance, in a fictive clinical text, "She" might refer to either the daughter or the patient, and determining the correct reference can be challenging.

In the context of clinical text processing, He (2007) [15] conducted research on coreference resolution using a dataset of 47 hospital discharge summaries written in English, totaling 4978 lines of text. Two computer science students annotated the corpus with coreference chains and time stamps, achieving a high inter-annotator agreement.

The training set was relatively small, with 649 coreferent pairs, necessitating the use of numerous features. The author employed a machine learning algorithm, the supervised C4.5 Decision Tree, and achieved promising results in terms of precision and recall using specific evaluation metrics designed for coreference chains.

Additionally, the i2b2 Challenge in Coreference Resolution for Electronic Medical Records involved 20 teams from nine countries working on clinical records. This challenge demonstrated that both machine learning and rule-based approaches performed best when supplemented with external knowledge, leading to accurate predictions of coreference chains.

3 Conclusion and Outlook

Conclusion

This paper has provided an overview of the essential tools used in natural language processing (NLP), focusing on their adaptation to clinical natural language processing (Clinical NLP). These fundamental tools are crucial for effectively analyzing clinical text, which presents unique challenges due to its complexity, diversity, and sensitivity.

Our exploration began by discussing methods for segmenting and tokenizing strings of characters, processes that break down text into manageable units for analysis. We also addressed morphological processing, including lemmatization and stemming, which aids in understanding the base forms of words, enhancing semantic comprehension.

In addition to these basics, we considered challenges such as compound splitting and abbreviation handling, which are particularly relevant in clinical NLP due to the specialized terminology. Part-of-speech tagging was highlighted as a means of discerning word functions, further enriching our understanding.

To deal with misspelled words, we examined spell checking and correction techniques, which are essential for maintaining accuracy in clinical text analysis. We also emphasized the significance of syntactical analysis or parsing for comprehending structural nuances in clinical narratives.

Expanding into semantic analysis, we explored concept extraction and named entity recognition, vital for identifying medical entities and conditions within clinical text. We also discussed negation detection, recognizing its importance in handling negated clinical symptoms during diagnostic reasoning.

In conclusion, this paper has provided an in-depth overview of fundamental tools used in natural language processing, with a specific focus on their application in clinical text analysis. However, it is essential to acknowledge that NLP is an evolving field, and addressing the unique challenges of clinical text requires ongoing research and development.

Outlook

Looking forward, a central question remains: can the integration of artificial intelligence (AI) into patient record analysis truly enhance healthcare? The answer depends on how we perceive and employ AI in healthcare settings. AI should be viewed as a tool to enhance human capabilities rather than as an autonomous entity with its own will.

AI has the potential to transform healthcare by offering data-driven insights, aiding in diagnosis, predicting outcomes, and streamlining administrative tasks. However, responsible and ethical AI use in healthcare requires continued research, algorithm refinement, and collaboration among healthcare professionals, data scientists, and policymakers.

In the near future, AI-powered clinical NLP tools hold promise for improving the efficiency and accuracy of medical processes. However, these tools must be designed and implemented with care, aligning with clinical standards, safeguarding patient privacy, and contributing to better healthcare outcomes.

As we enter the era of AI-driven healthcare, the collaboration between humans and intelligent algorithms will be crucial to unlock the full potential of clinical NLP. This synergy promises benefits for healthcare providers and patients alike, provided it is approached with diligence, ethics, and a focus on improving the quality of care.

Bibliography

[1] J. F. Allen. Towards a general theory of action and time. *Artificial Intelligence*, 23(2):123–154, 1984.

[2] H. Allvin, E. Carlsson, H. Dalianis, R. Danielsson-Ojala, V. Daudaravicius, M. Hassel, et al. Characteristics of finnish and swedish intensive care nursing narratives: A comparative analysis to support the development of clinical language technologies. *Journal of Biomedical Semantics*, 2(Suppl 3):1–11, 2011.

[3] Cosmin A Bejan and Joshua C Denny. Learning to identify treatment relations in clinical text. In *AMIA Annual Symposium Proceedings*, volume 2014, page 282. American Medical Informatics Association, 2014.

[4] Joakim Carlberger, Hercules Dalianis, Martin Hassel, and Ola Knutsson. Improving precision in information retrieval for swedish using stemming. In *Proceedings of NODALIDA '01 - 13th Nordic Conference on Computational Linguistics*, page Page Numbers, Location, 2001.

[5] Wendy W Chapman, Will Bridewell, Philip Hanbury, Gregory F Cooper, and Bruce G Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310, 2001.

[6] Wendy W Chapman, Diane Hilert, Sumithra Velupillai, Maria Kvist, Maria Skeppstedt, Brian E Chapman, et al. Extending the negex lexicon for multiple languages. *Studies in Health Technology and Informatics*, 192:677, 2013.

[7] Alexander Clark, Chris Fox, and Shalom Lappin. *The Handbook of Computational Linguistics and Natural Language Processing*. Wiley, New York, 2013.

[8] Raquel Costumero, Fernando Lopez, Consuelo Gonzalo-Martín, Montserrat Millan, and Ernestina Menasalvas. An approach to detect negation on medical documents in spanish. In *International Conference on Brain Informatics and Health*, pages 366–375. Springer, 2014.

[9] Vitaly Cotik, Roy Roller, Feiyu Xu, Hans Uszkoreit, Klemens Budde, and Daniel Schmidt. Negation detection in clinical reports written in german. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2016), Held in Conjunction with Coling 2016*, pages 115–124, 2016.

[10] Berry de Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu.

Machine-learned solutions for three stages of clinical information extraction: The state of the art at i2b2 2010. *Journal of the American Medical Informatics Association*, 18(5):557–562, 2011.

[11] L. R. A. Derczynski. *Automatically Ordering Events and Times in Text*. Springer, Berlin, 2017.

[12] Kvist M. Velupillai S. Wirén M. Grigonyte, G. Improving readability of swedish electronic health records through lexical simplification: First results. in proceedings of the 3rd workshop on predicting and improving text readability for target reader populations – pitr, gothenburg, sweden. page 74–83, April 2014.

[13] Cyril Grouin, Louise Deléger, Arnaud Rosier, Lynda Temal, Olivier Dameron, Pascal Van Hille, et al. Automatic computation of cha2ds2-vasc score: Information extraction from clinical texts for thromboembolism risk assessment. In *AMIA Annual Symposium Proceedings*, pages 501–510. American Medical Informatics Association, 2011.

[14] Martin Hassel, Aron Henriksson, and Sumithra Velupillai. Something old, something new: Applying a pre-trained parsing model to clinical swedish. In *Northern European Association for Language Technology (NEALT)*, 2011.

[15] T. Y. He. Coreference resolution on entities and events for hospital discharge summaries. Master's thesis, Massachusetts Institute of Technology, 2007.

[16] Aron Henriksson, Maria Kvist, Hercules Dalianis, and Martin Duneld. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. *Journal of Biomedical Informatics*, 57:333–349, 2015.

[17] H. Jung, J. Allen, N. Blaylock, W. De Beaumont, L. Galescu, and M. Swift. Building timelines from narrative clinical records: Initial results based-on deep natural language understanding. In *Proceedings of BioNLP 2011 Workshop*, pages 146–154. Association for Computational Linguistics, 2011.

[18] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Pearson, London, 2014.

[19] Yuan Luo, Özlem Uzuner, and Peter Szolovits. Bridging semantics and syntax with graph algorithms–state-of-the-art of extracting biomedical relations. *Briefings in Bioinformatics*, 18(1):160–178, 2016.

[20] Ruslan Mitkov. *The Oxford Handbook of Computational Linguistics*. Oxford University Press, Oxford, 2005.

[21] A. Neveol, H. Dalianis, G. Savova, and P. Zweigenbaum. Clinical natural language processing in languages other than english: Opportunities and challenges. *Journal of Biomedical Semantics*, 9(12):1–13, 2018.

[22] Joakim Nivre, Johan Hall, and Jens Nilsson. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of the Fifth International Confer-*

ence on Language Resources and Evaluation, LREC 2006, pages 2216–2219, 2006. Accessed 11 Jan 2018.

[23] J. Patrick and D. Nguyen. Automated proofreading of clinical notes. In *Proceedings of the [Conference Abbreviation]*, 2011.

[24] J. Pustejovsky, J. M. Castano, R. Ingria, R. Sauri, R. J. Gaizauskas, A. Setzer, and et al. Timeml: Robust specification. 2003.

[25] Angus Roberts, Robert Gaizauskas, Mark Hepple, and Yikun Guo. Mining clinical relationships from patient narratives. *BMC Bioinformatics*, 9(11):1, 2008.

[26] Maria Skeppstedt. *Extracting Clinical Findings from Swedish Health Record Text*. PhD thesis, Department of Computer and Systems Sciences, Stockholm University, 2015.

[27] Maria Skeppstedt, Maria Kvist, Gunnar Nilsson, and Hercules Dalianis. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of Biomedical Informatics*, 49:148–158, 2014.

[28] K. Smith, B. Megyesi, S. Velupillai, and M. Kvist. Professional language in swedish clinical text: Linguistic characterization and comparative studies. *Nordic Journal of Linguistics*, 37(02):297–323, 2014.

[29] Brett R South, Shuying Shen, Marcus Jones, Jennifer Garvin, Matthew H Samore, Wendy W Chapman, and et al. Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. *BMC Bioinformatics*, 10(9):S12, 2009.

[30] Jannik Strötgen and Michael Gertz. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324. Association for Computational Linguistics, 2010.

[31] W. Styler IV, S. Bethard, S. Finan, M. Palmer, S. Pradhan, P. de Groen, et al. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154, 2014.

[32] W. Sun, A. Rumshisky, and Ö. Uzuner. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813, 2013.

[33] W. Sun, A. Rumshisky, and Ö. Uzuner. Temporal reasoning over clinical text: The state of the art. *Journal of the American Medical Informatics Association*, 20(5):814–819, 2013.

[34] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[35] Dalianis H. Hassel M. Nilsson G. H. Velupillai, S. Developing a standard for deidentifying electronic patient records written in swedish: Precision, recall and f-measure in a manual and computerized annotation trial. *International Journal of Medical Informatics*, 2009.

[36] Sumithra Velupillai. *Shades of Certainty: Annotation and Classification of Swedish Medical Records*. PhD thesis, Stockholm University, 2012.

[37] L. Zhou, C. Friedman, S. Parsons, and G. Hripcsak. System architecture for temporal information extraction, representation and reasoning in clinical narrative reports. In *AMIA Annual Symposium Proceedings*, pages 869–873, 2005.

[38] L. Zhou and G. Hripcsak. Temporal reasoning with medical data–a review with emphasis on medical natural language processing. *Journal of Biomedical Informatics*, 40(2):183–202, 2007.