Master of Science – Applied Analytics and Data Science

Clinical text processing

Dègninou Yehadji

Technological University Dublin

Ireland

B00108474[at]mytudublin.ie

Abstract

Clinical text is specific in the sense that it has syntax and grammatical errors, and abbreviations among many others. Consequently, clinical text processing can not be completed only with regular text processing methods. Some supplementary processing steps such as compound splitting, spell checking, syntactic analysis, abbreviation, negation and factuality detection, relative and temporal processing, relation extraction, and anaphora resolution are recommended for clinical text. Five of these supplementary steps have been presented in this review. Processed clinical text can be used for applications in detection, prediction, and prevention of diseases and adverse drug events, for applications to support clinicians, and for applications in clinical coding. However, the algorithms developed for clinical text processing are language specific. Future improvement in the area could explore the development of multi-language algorithms for clinical text processing.

Keywords: clinical text, clinical notes, electronic medical records, electronic health records, text mining, clinical text mining.

1. Introduction

Clinical data are collected during all healthcare operations to ensure evidence-based decision making, patient monitoring, and clinical research among other objectives. These data originate from many sources. Electronic medical records (EMR), also known as electronic health records (EHR), laboratory information systems (LIS), radiology information systems (RIS), prescription records, telemonitoring (telemedicine, telehealth) applications, internet of things (IoT) data, administrative data, insurance claims, patient and disease registries, health and demographic surveys, clinical trials data and social media are some of the most common data sources. EHRs are the most standardized forms of clinical data. They are collected at the point of care, specifically at a medical facility, hospital, clinic or practice and include sociodemographic information, laboratory, treatment, vital parameters, hospitalization, provider notes, and insurance among others (Maloy, 2012; Kubben, 2019).

As there are various data sources in healthcare, there are also various data type. These data are structured (tabular or time series), and unstructured (images, videos, and natural languages). Natural languages are formats data format in free text such as surveys, discharge summaries, daily logs by patients, clinician notes, and radiology reports. Also, data collected from social media are in form of free text. These clinical data are riddled with spelling and grammar errors, abbreviations, speculations, etc. For this reason, applying regular text processing methods to clinical text may limit the performance of models developed afterward. That is why this review is conducted to explore how text processing is performed on free text generated in healthcare setting. Specifically, five supplementary preprocessing tasks that are relevant to clinical text will be reviewed.

2. Overview of text mining

The purpose of data mining also known as knowledge discovery in databases (KDD) is to discover knowledge from structured, unstructured, or semi-structured data of various sources such as text, relational databases, web data, user log data, and others. Text mining is a field of KDD designated to perform data mining from text data collections. The goal is to discover knowledge, information, or patterns from text data. Text mining is also related to other advance computing, including machine learning, information retrieval, natural language processing, information extraction, statistics, pattern recognition, and artificial intelligence.

In text mining selection of relevant text from a collection is used as input to a meaningful pattern or knowledge discovery. The process is performed by starting with pre-processing, data transformation, actual text mining, and interpretation or evaluation

The objective of the pre-processing step is to remove noisy information and perform preliminary processes that facilitate the following ones. Generally, pre-processing includes text data tokenization, stop words removal, normalization, stemming, lemmatization (Cai and Sun, 2009). Other steps such as named entity recognition and part-of-speech (POS) tagging depending on the objective or the nature of the text.

The filtered tokens generated in the pre-processing steps are used for text transformation or attribute generation. The technics used are called text indexing which consists in representing the text by its words and their occurrences. Vector space is the main approach to text representation. Vector space is a statistical model for text documents representation as vectors of index terms to capture the relative weight of terms. Term frequency (TF) and term frequency-inverse document frequency (TF-IDF) are the most popular vector space schema used in text processing. The vector space modeling generates features (terms and their weight) that are passed to the feature selection step.

Text data is highly noisy and consequently, generates several redundant and irrelevant features that do not add any useful information for the data mining. Feature selection techniques are used to select subsets of relevant features from the original list. Information gain, correlation, and k-best discriminative terms are some of feature selection techniques used for text mining.

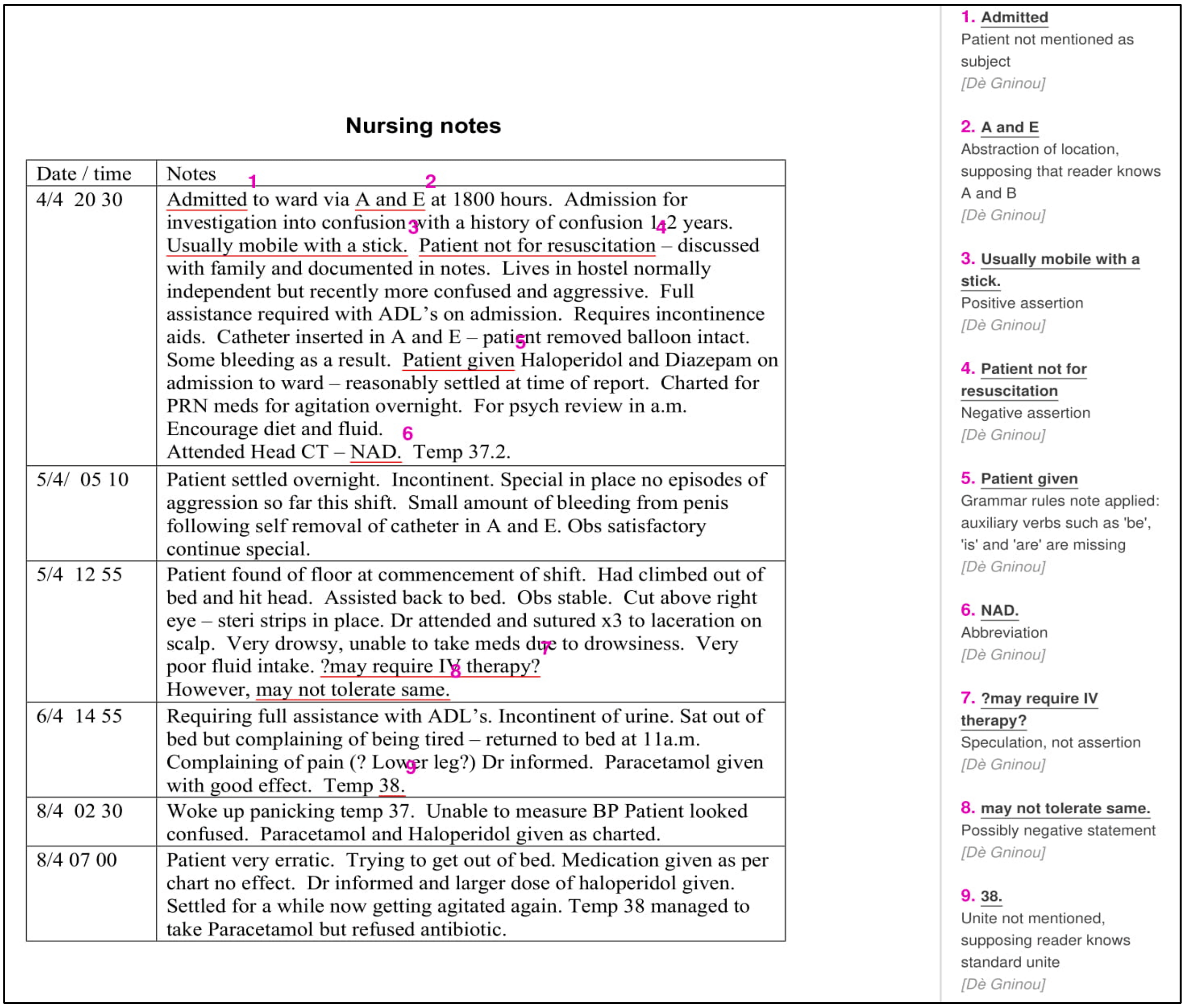
After presenting the general overview of text mining, the next section will be dedicated to the particular nature of clinical text that requires specific approaches to extract insight or knowledge.

3. Characteristics of clinical text

Standard text such as news, blog nots, novels, academic papers, or textbooks are written in plain language with well-structured sentences, which make them easier to be processed by standard natural language processing methods. Clinical text is written by health professionals throughout the continuum of care, for example to prepare progress or treatment notes, update patient charts, prepare letters for consultation and referral, and complete various administrative forms. The purpose of the notes is to provide factual information, including essential, accurate, and specific information about patient conditions, diagnosis, treatments, and prognoses (Hull, 2013).

Given the intent to provide factual and essential information, clinical text is particular in the sense that it is subject to spelling errors, abbreviations, acronyms, syntactic variations, word choices, compound words, negation and factuality expressions, and speculative cues (Figure 1).

Depending on whether it is physician note, discharge letter, nursing narrative, or pathology report, the writing style may be different. The style may also vary between different clinical units and specialties (Dalianis, 2018b). Little importance is given to linguistic and grammar in clinical notes. They are written in telegraphic style, sentences are often incomplete, and auxiliary verbs such as *"be"*, *"is"* and *"are"* are usually ignored (Allvin *et al.*, 2011). Pakhomov et al. (2005) studied clinical texts and reported 30% of non-word tokens, abbreviations, acronyms, and grammatical and spelling errors. Various spelling errors rate have been reported in clinical text depending on language – 10% in French, 2.3% in Australian English, and 7.6% in Swedish (Ruch, Baud and Geissbühler, 2003; Patrick and Nguyen, 2011; Nizamuddin and Dalianis, 2014).



Nursing notes

Date / time	Notes
4/4 20 30	Admitted to ward via A and E at 1800 hours. Admission for investigation into confusion with a history of confusion 1-2 years. Usually mobile with a stick. Patient not for resuscitation – discussed with family and documented in notes. Lives in hostel normally independent but recently more confused and aggressive. Full assistance required with ADL's on admission. Requires incontinence aids. Catheter inserted in A and E – patient removed balloon intact. Some bleeding as a result. Patient given Haloperidol and Diazepam on admission to ward – reasonably settled at time of report. Charted for PRN meds for agitation overnight. For psych review in a.m. Encourage diet and fluid. Attended Head CT – NAD. Temp 37.2.
5/4/ 05 10	Patient settled overnight. Incontinent. Special in place no episodes of aggression so far this shift. Small amount of bleeding from penis following self removal of catheter in A and E. Obs satisfactory continue special.
5/4 12 55	Patient found of floor at commencement of shift. Had climbed out of bed and hit head. Assisted back to bed. Obs stable. Cut above right eye – steri strips in place. Dr attended and sutured x3 to laceration on scalp. Very drowsy, unable to take meds due to drowsiness. Very poor fluid intake. ?may require IV therapy? However, may not tolerate same.
6/4 14 55	Requiring full assistance with ADL's. Incontinent of urine. Sat out of bed but complaining of being tired – returned to bed at 11a.m. Complaining of pain (? Lower leg?) Dr informed. Paracetamol given with good effect. Temp 38.
8/4 02 30	Woke up panicking temp 37. Unable to measure BP Patient looked confused. Paracetamol and Haloperidol given as charted.
8/4 07 00	Patient very erratic. Trying to get out of bed. Medication given as per chart no effect. Dr informed and larger dose of haloperidol given. Settled for a while now getting agitated again. Temp 38 managed to take Paracetamol but refused antibiotic.

1. Admitted
Patient not mentioned as subject
[Dè Gninou]

2. A and E
Abstraction of location, supposing that reader knows A and B
[Dè Gninou]

3. Usually mobile with a stick.
Positive assertion
[Dè Gninou]

4. Patient not for resuscitation
Negative assertion
[Dè Gninou]

5. Patient given
Grammar rules note applied: auxiliary verbs such as 'be', 'is' and 'are' are missing
[Dè Gninou]

6. NAD.
Abbreviation
[Dè Gninou]

7. ?may require IV therapy?
Speculation, not assertion
[Dè Gninou]

8. may not tolerate same.
Possibly negative statement
[Dè Gninou]

9. 38.
Unite not mentioned, supposing reader knows standard unite
[Dè Gninou]

Figure 1. Example of nursing notes annotated with some characteristics of clinical text. Clinical notes adapted from [www.wordtemplatesonline.net]

Clinical text contains a lot of compound words which are morphologically complex words composed of root words preceded by prefixes and followed by suffixes. These root words usually derive from Greek and Latin (Cotterill, 1996). The word *"epigastralgia"* for example is composed of *"epi"* (upon or upper), *"gastro"* (abdomen or stomach), *"algia"* (pain). Compound words are also present in non-English languages such as Swedish and German. Patient with diabetes is written "diabetespatient" in Swedish for example (Dalianis, 2018a).

Given the specificities of clinical text several clinical corpora – collection of clinical documents collected, organized and available (upon request) to health professionals and researchers. The Informatics for Integrating Biology & the Bedside (i2b2) clinical corpus a thousand notes in English, the Computational Medicine Center (CMC) corpus consisting of more than 2000 patient records in American English, the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC II) collection containing discharge summaries and nursing notes written in American English, are some of the clinical corpora available.

Clinical notation, terminology, domain-specific vocabulary, and classification systems have been developed to facilitate standardization reporting, administration, classification and explanation of diseases, treatments, and medications. The International Statistical Classification of Diseases and Related Health Problems (ICD) is a medical classification system maintained by the World Health Organization (WHO) that defines diagnosis codes. The latest version in deployment is the ICD-11 [https://icd.who.int]. SNOMED Clinical Terms (SNOMED CT) [https://browser.ihtsdotools.org] a collection of medical terms systematically organized and computer processable with codes, terms, synonyms, and definitions used in clinical documentation and reporting. The Unified Medical Language System (UMLS) hosted by the National Library of Medicine (NLM) [https://uts.nlm.nih.gov], is a set of files and software used to map many health and biomedical terminologies and standards to enable interoperability between systems. The Medical Subject Headings (MeSH) thesaurus is a controlled and hierarchically structured vocabulary also maintained by the NLM [https://meshb.nlm.nih.gov] and used for indexing, cataloging, and searching of biomedical literature.

4. Clinical text processing

Dalianis (2018a) summarized clinical text processing into three building blocks: morphological processing, syntactical analysis, semantic analysis and concept extraction. In morphological processing, beside lemmatization, stemming, and POS tagging which are already known in regular text processing, compound splitting (decompounding), abbreviation detection and expansion, and spell checking and correction can be added. Syntactical analysis or parsing includes shallow parsing (chunking) and use of grammar tools. In addition to NER which is already known in regular text processing, negation and factuality detection, relative and temporal processing, relation extraction, and anaphora resolution can be added to semantic analysis and concept extraction of clinical text. Five of these supplementary tasks performed in clinical text processing, namely, compound splitting, abbreviation and negation detection, spell checking, and syntactic analysis detection will be detailed in the following sub-sections.

4.1. Compound splitting

Compound splitting is the process of breaking compound (morphologically complex) words into their base components. Decompounding is performed using dictionaries with application of splitting rules. The approaches to decompounding can be organized into two main categories – monolingual and bilingual approaches. Brown (2002) introduced a corpus-driven approach for splitting compounds for bilingual – German-English translation of clinical documents. His method was based on comparison of terms in parallel corpora using the Levenshtein distance and German compounds are split when the corresponding (English) compound words are similar. Koehn & Knight (2003) described both a monolingual and bilingual approach. The monolingual approach is a frequency-based one for German, where a dictionary of compound-frequency tuples was produced from a monolingual corpus. They extended this frequency-based approach with a bilingual one where they use a dictionary extracted from parallel data to help finding better split options.

4.2. Abbreviation detection

The abbreviations found in clinical documents must be detected and resolved. In normalized writing, an abbreviation is usually preceded by its expanded form or definition at the first mention, which is not the case in clinical writing. The goal of this step is to correctly identify and interpret abbreviations in clinical documents. This step is challenging because simple abbreviation-definition patterns are not applicable. Although Hua et al. (2007) compared machine learning approaches to abbreviation detection and expansion and found that they achieved considerable results, the authors recommend the use of a manually created inventory, even in addition to the use of already existing external resources. Several studies have attempted to perform detection and expansion of abbreviations in clinical text. Siklósi & Novák (2013) developed two unsupervised methods to expand sequences of abbreviations in Hungarian medical records. The first method used lists of medical concepts and a manually created dictionary, and the second one used the dictionary with quality improved by using abbreviation interpretations automatically derived from the document collection.

4.3. Spell checking and correction

Spelling errors detection and correction help improving clinical text mining. In several clinical text mining projects, efforts are done to complete spell checking and error correction. For example, Nizamuddin & Dalianis (2014) developed a rule-based algorithm using external resources combined with other pre-processing tasks such as lemmatization and compound splitting for spelling error detection. The external resources were word lists such as a medical dictionary, a Swedish dictionary, and an abbreviations list. The algorithm was tested on a corpus of Swedish clinical documents and they concluded that the process required adequate word list and proper pre-processing such as lemmatization and compound splitting. Some open-source spell checkers are available for use in English as well as in other languages. Aspell or Hunspell are two example of tools that use lexicons for different languages. To use them for clinical text, it recommended customize them by adding medical domain specific lexicons. GNU Aspell is spell checker designed to be used as library or standalone spell checker. It is considered as more performant in suggesting possible replacements for misspelled words as compared to other spell checkers in English (GNU Aspell developers, 1998). Hunspell is the spell checker for LibreOffice, Mozilla Firefox, Google Chrome, macOS, InDesign, memoQ, Opera and SDL Trados among many others (Németh, 2016). For regular English, Peter Norvig developed a spell checker in python, which uses an algorithm based on the Levenshtein distance to detect permutations from the original word. The algorithm then compares permutations – transpositions, replacements, deletions, and insertions, to known terms in a term frequency list. To correct the misspelled words, those that are found more often in the frequency list have a higher probability to be the correct ones (Norvig, 2007; Barrus, 2018).

4.4. Syntactical analysis

This process is used to determine the grammatical structure of sentences. This involves determining and predicating the constitutive groups of the sentences (e.g., nouns, verbs, pronouns, etc) and their places. Lexicon are provided to determine POS for words and the analyser loop over the input sentence word by word to produce its structural description. The sentence *"The patient received Diazepam upon admission"* would be parsed as follows: *[NP The patient] [VP received] [NP Diazepam] [PP upon] [NP admission]* (Figure 4). Syntactical analysis can be performed by shallow parsing or by using grammar tools. Shallow parsing, chunking, or light parsing for sentence analysis is performed by identifying POS (nouns, verbs, adjectives, etc.) and linking them to grammatical syntactic structures (noun, verb, or adverb groups for example). The syntactic structure of a sentence, also known as parse tree, is important in NLP research in any field including medicine. A number of tools such as the Bikel parser, the Stanford parser, the Charniak parser, have been developed to facilitate parsing for general English (Charniak, 2001; Bikel, 2002; Chen and Manning, 2014). Jiang et al. (2015) evaluated these parsers with clinical notes. They concluded that it is critical to re-train parsers using clinical treebanks – corpus of parsed text used to annotate sentence structure, to improve performance on clinical text. They also recommended that combining clinical and open domain corpora might help achieving optimal performance for parsing clinical text. Indeed, Wang et al. (2015) adapted the Stanford unlexicalized probabilistic context-free grammar (CFG) parser for clinical text by expanding it with lexicon augmentation, statistics adjusting, and grammar rules modification based on operative reports. The found that the customized parser improved the F-score by 2.26% for on approach and by 3.81% for another approach. They also concluded that using statistics collected from clinical text tagged with POS taggers in addition to proper grammar modifications and lexicons of a general parser may improve parsing performance on specialized clinical text.

Several tools are available to create language structures or grammars for syntactic analysis. One of these language structures is the definite clause grammar (DCG). DCG is a set of rules in a notation based on Prolog logic programming. Lex (Lexical Analysis) and Yacc (Yet another compiler-compiler) are two other tools developed for building compilers for programming language which can be used to develop parsers.

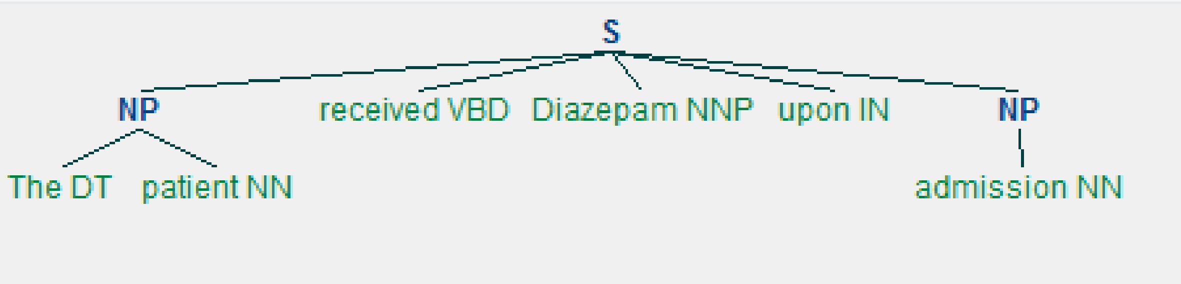


Figure 2. Example of sentence with its syntactical analysis.

DT: determiner; NN: singular noun; VBD: past tense verb; NNP: singular proper noun; IN: preposition or subordinating conjunction; NP: noun phrase.

4.5. Negation detection

The meaning of clinical statements is not always straightforward. The meanings are extremely altered by modifiers such as negation. That is why it important to procced to negation detection in clinical text, or otherwise, some statement may be taken as affirmative while they are made to rule out some hypothesis. In the sentence *"The patient had no fracture"* for example, focusing on *"fracture"* would lead to concluding in instead of ruling out fracture.

Efforts are done in the field of clinical text mining to ensure recognition of negations. In this sense, Chapman et al. (2001) developed simple algorithm to detect negation cue words without consideration to the sentences semantic structures. It is a simple regular expression algorithm named NegEx that implements several phrases and words such as *"absence of"*, *"declined"*, *"denied"*, *"no sign of"*, *"no"*, *"not"*, *"not demonstrate"*, *"doubt"*, *and "negative for"*, etc, that indicate negation, remove sentences with chunks that wrongly appear to be negation phrases, and determine the coverage of the negation phrases (Pizarro, 2019). Upon evaluation, the NegEx performed well with a specificity of 94.5% (versus 85.3% at baseline), a reasonable sensitivity of 77.8% (versus 88.3% at baseline) and a positive predictive value of 84.5% (versus 68.4% at baseline) (Chapman *et al.*, 2001).

Although the NegEx performed well, Mehrabi et al. (2015) found that this simplistic approach fails sometimes to correctly identify the negation status of clinical concepts in sentences with complex structure, due to its inability to take the contextual relationship between words within a sentence into account. They then developed a negation algorithm called DEEPEN to improve NegEx by decreasing false positives. They proceeded by considering the dependency relationship between negation words and concepts within sentences using Stanford dependency parser. They then assessed DEEPEN’s generalizability on clinical notes from two different institutions and found that it reduced the number of inaccurate negation assignment for patients with positive results, and therefore improve the identification of patients with the target clinical findings in patient records.

5. Conclusion

Throughout the continuum of care, not only structured data, but also unstructured data (mainly text data) are generated to maintain and keep track of patient records. On one hand, knowledge discovery is performed on the structured data using traditional data mining technics. On the other hand, knowledge discovery on clinical text required text mining technics. However, clinical text mining cannot be performed with common text processing tasks because these texts have specific characteristics – syntax errors, abbreviations, and neglect of grammatical rules among many others. Thus, clinical text mining requires supplementary pre-processing. In addition to lemmatization, stemming, POS tagging, and named entity recognition which are used for regular text processing, compound splitting, abbreviation, negation, and factuality detection, spell checking, syntactic analysis, relative and temporal processing, relation extraction, and anaphora resolution may be required as supplementary pre-processing steps for clinical text.

Compound splitting, abbreviation and negation detection, spell checking, and syntactic analysis have been reviewed in detail in this paper. However, it can be observed that most of the algorithms developed to perform the supplementary processing tasks on clinical text are language specific. This language specificity does not facilitate cross language application of clinical text processing, especially in the era of cross platform, team, and geographical sharing of clinical to improve patient care. Future research and improvement in the area could explore the development of multi language algorithms for clinical text processing.

After processing clinical text, the cleaned and selected features are used for several applications in healthcare. These applications rang from detection, prediction, and prevention applications, to applications for support to clinicians, and applications for clinical coding. The applications for detection of adverse drug events have been detailed in this review.

6. References

Allvin, H. *et al.* (2011) '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(3), p. S1. doi: 10.1186/2041-1480-2-S3-S1.

Barrus, T. (2018) *pyspellchecker: Pure python spell checker based on work by Peter Norvig*. Available at: https://github.com/barrust/pyspellchecker (Accessed: 24 December 2020).

Bikel, D. M. (2002) 'Design of a multi-lingual, parallel-processing statistical parsing engine', in *Proceedings of the second international conference on Human Language Technology Research*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. (HLT '02), pp. 178–182. Available at: https://dl.acm.org/doi/10.5555/1289189.1289191 (Accessed: 25 December 2020).

Brown, R. D. (2002) 'Corpus-driven splitting of compound words', in *Proceedings of the 9th international Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2002)*. Available at: https://core.ac.uk/display/20935025 (Accessed: 24 December 2020).

Cai, Y. and Sun, J.-T. (2009) 'Text mining', in LIU, L. and ÖZSU, M. T. (eds) *Encyclopedia of Database Systems*. Boston, MA: Springer US, pp. 3061–3065. doi: 10.1007/978-0-387-39940-9_418.

Chapman, W. W. *et al.* (2001) 'A simple algorithm for identifying negated findings and diseases in discharge summaries', *Journal of Biomedical Informatics*, 34(5), pp. 301–310. doi: 10.1006/jbin.2001.1029.

Charniak, E. (2001) 'Immediate-head parsing for language models', in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. USA: Association for Computational Linguistics (ACL '01), pp. 124–131. doi: 10.3115/1073012.1073029.

Chen, D. and Manning, C. D. (2014) 'A fast and accurate dependency parser using neural networks', in. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 740–750. doi: 10.3115/v1/D14-1082.

Cotterill, S. (1996) 'The components of medical terminology', in *Medical Terminology for Cancer*. CancerIndex. Available at: http://www.cancerindex.org/medterm/medtm4.htm (Accessed: 23 December 2020).

Dalianis, H. (2018a) 'Basic building blocks for clinical text processing', in Dalianis, H. (ed.) *Clinical text mining: secondary use of electronic patient records*. Cham: Springer International Publishing, pp. 55–82. doi: 10.1007/978-3-319-78503-5_7.

Dalianis, H. (2018b) 'Characteristics of patient records and clinical corpora', in Dalianis, H. (ed.) *Clinical text mining: secondary use of electronic patient records*. Cham: Springer International Publishing, pp. 21–34. doi: 10.1007/978-3-319-78503-5_4.

GNU Aspell developers (1998) *GNU Aspell*. Available at: http://aspell.net/ (Accessed: 24 December 2020).

Hull, M. (2013) ‘Learning and teaching clinical writing’, *Medical Writing*, 22, pp. 29–33.

Jiang, M. *et al.* (2015) ‘Parsing clinical text: how good are the state-of-the-art parsers?’, *BMC Medical Informatics and Decision Making*, 15(1), pp. 1–6. doi: 10.1186/1472-6947-15-S1-S2.

Koehn, P. and Knight, K. (2003) ‘Empirical methods for compound splitting’, in *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*. USA: Association for Computational Linguistics (EACL ’03), pp. 187–193. doi: 10.3115/1067807.1067833.

Kubben, P. (2019) ‘Data sources’, in Kubben, P., Dumontier, M., and Dekker, A. (eds) *Fundamentals of Clinical Data Science*. Cham: Springer International Publishing, pp. 3–9. doi: 10.1007/978-3-319-99713-1_1.

Maloy, C. (2012) *Data resources in the health sciences: clinical data*. Seattle, WA: University of Washington, Health Sciences Library. Available at: https://guides.lib.uw.edu/hsl/data/findclin (Accessed: 19 December 2020).

Mehrabi, S. *et al.* (2015) ‘DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx’, *Journal of Biomedical Informatics*, 54, pp. 213–219. doi: 10.1016/j.jbi.2015.02.010.

Németh, L. (2016) *Hunspell*. hunspell. Available at: https://github.com/hunspell/hunspell (Accessed: 24 December 2020).

Nizamuddin, U. and Dalianis, H. (2014) ‘Detection of spelling errors in Swedish clinical text’, in. *1st Nordic workshop on evaluation of spellchecking and proofing tools (NorWEST2014), SLTC 2014, Uppsala*. Available at: http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-110972 (Accessed: 22 December 2020).

Norvig, P. (2007) *How to write a spelling corrector*. Available at: https://norvig.com/spell-correct.html (Accessed: 24 December 2020).

Pakhomov, S., Pedersen, T. and Chute, C. G. (2005) ‘Abbreviation and acronym disambiguation in clinical discourse’, *AMIA ... Annual Symposium proceedings. AMIA Symposium*, pp. 589–593.

Patrick, J. and Nguyen, D. (2011) ‘Automated proof reading of clinical notes’, in *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*. *PACLIC 2011*, Singapore: Institute of Digital Enhancement of Cognitive Processing, Waseda University, pp. 303–312. Available at: https://www.aclweb.org/anthology/Y11-1032 (Accessed: 22 December 2020).

Pizarro, J. (2019) *negspacy: A spaCy pipeline object for negation.* Available at: https://github.com/jenojp/negspacy (Accessed: 26 December 2020).

Ruch, P., Baud, R. and Geissbühler, A. (2003) ‘Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record’, *Artificial Intelligence in Medicine*, 29(1–2), pp. 169–184. doi: 10.1016/s0933-3657(03)00052-6.

Siklósi, B. and Novák, A. (2013) ‘Detection and expansion of abbreviations in Hungarian clinical notes’, in *Advances in Artificial Intelligence and Its Applications*. *Mexican International Conference on Artificial Intelligence*, Springer, Berlin, Heidelberg, pp. 318–328. doi: 10.1007/978-3-642-45114-0_26.

Wang, Y. *et al*. (2015) ‘Domain adaption of parsing for operative notes’, *Journal of Biomedical Informatics*, 54, pp. 1–9. doi: 10.1016/j.jbi.2015.01.016.

Xu, H., Stetson, P. D. and Friedman, C. (2007) ‘A study of abbreviations in clinical notes’, *AMIA Annual Symposium Proceedings*, 2007, pp. 821–825.