Medical Information Extraction in the Age of Deep Learning

Udo Hahn¹, Michel Oleynik²

¹ Jena University Language & Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Jena, Germany
² Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria

Summary

Objectives: We survey recent developments in medical Information Extraction (IE) as reported in the literature from the past three years. Our focus is on the fundamental methodological paradigm shift from standard Machine Learning (ML) techniques to Deep Neural Networks (DNNs). We describe applications of this new paradigm concentrating on two basic IE tasks, named entity recognition and relation extraction, for two selected semantic communities: medicine and medical informatics, natural language processing, as well as neural networks and artificial intelligence.

Methods: For the time period from 2017 to early 2020, we searched for relevant publications from three major scientific communities: medicine and medical informatics, natural language processing, as well as neural networks and artificial intelligence.

Results: In the past decade, the field of Natural Language Processing (NLP) has undergone a profound methodological shift from symbolic to distributed representations based on the paradigm of Deep Learning (DL). Meanwhile, this trend is, although with some delay, also reflected in the medical NLP community. In this section, we report how the paradigm has been shifting from feature-engineered ML to DNNs. We briefly sketch the literature in Section 2 and deal with the challenges and applications of DNNs in medical NLP in Section 3.

Conclusions: The paradigm shift from (feature-engineered) ML to DNNs changes the fundamental methodological rules of the game for medical NLP. This change is by no means restricted to medical IE but should also deeply influence other areas of medical informatics, either NLP- or non-NLP-based.

Keywords
Neural networks, deep learning, natural language processing, information extraction, named entity recognition, relation extraction

1 Introduction

The past decade has seen a truly revolutionary paradigm shift for Natural Language Processing (NLP) as a result of which Deep Learning (DL) (for a technical introduction, cf. [1]; for comprehensive surveys, cf. [2] and [3]) became the dominating mind-set of researchers and developers in this field (for surveys, cf. [4, 5]). Yet, DL is by no means a new computational paradigm. Rather it can be seen as the most recent offspring of neural computation in the evolution of computer science (cf. the historical background provided by Schmidhuber [6]). But unlike in previous attempts, it now turns out to be extremely robust and effective for adequately dealing with the contents of unstructured visual [7], audio/speech [8], and textual data [9].

The success of Deep Neural Networks (DNNs) has many roots. Perhaps the most important methodological reason is that, with DNNs, manual feature selection or (semi-)automated feature engineering is abandoned. This time-consuming tuning step was at the same time mandatory and highly influential on the performance of earlier generations of ML systems in NLP based on Markov Models (MMs), Conditional Random Fields (CRFs), Support Vector Machines (SVMs), etc. In a DL system, however, the relevant features (and their relative contribution to a classification decision) are automatically computed as a result of thousands of iterative training cycles.

The ultimate reason for the success behind DNNs is a pragmatic criterion, that is, system performance. Compared with results in biomedical Information Extraction (IE), obtained in previous years with standard ML methods, DL approaches changed profoundly the rules of the game. In a landslide manner, for the same task and domain, performance figures jumped up to levels unprecedented so far and DL systems consistently outperformed by large margins non-DL state-of-the-art (SOTA) systems for different tasks. Section 3 provides ample evidence for this claim and features the new SOTA results with a deeper look at IE, a major application class of medical NLP (for alternative surveys, cf. [10–12]).

Despite specialized hardware at disposal now, training DNNs still requires tremendous computational resources and processing time. Luckily, for general NLP, huge collections of language models (so-called embeddings) have already been trained on huge corpora (comprised of hundreds of millions of Web scrapped documents, including newspaper and Wikipedia articles) so that these pre-compiled model resources can be readily reused when dealing with general-purpose language. But medical (and biological) language mirrors special-purpose language characteristics and comprises a large variety of sublanguages of its own. This becomes obvious in Section 3 where we deal with scholarly scientific writing (with documents typically taken from PubMed). Here, differences to general language are mostly due to the use of a highly specialized technical vocabulary (covered by numerous terminologies, such as MeSH, SNOMED-CT, or ICD). Even more challenging are clinical notes and reports (with documents often taken from the MIMIC³ (Medical Information Mart

³ https://mimic.physionet.org/
for Intensive Care) clinical database) which typically exhibit syntactically ill-formed, telegraphic language with lots of acronyms and abbreviations as an additional layer of complexity (cf. the seminal descriptive work distinguishing both these sublanguage types by Friedman et al. [13]). Newman-Griffis and Fosler-Lussier [14] investigated different sublanguage patterns for the many varieties of clinical reports (pathology reports, discharge summaries, nurse and Intensive Care Unit notes, etc.), while Nunez and Carenini [15] discussed the portability of embeddings across various fields of medicine reflecting characteristic sublanguage use patterns. These constraints have motivated the medical NLP community to adapt embeddings originally trained on general language to the medical language. Table 1 lists those medically informed embeddings, many of which are the basis for the IE applications discussed in Section 3.

Our survey emphasizes the fundamental methodological paradigm shift of current NLP research from symbolic to distributed representations as the basis of DL. It thus complements earlier contributions to the International Medical Informatics Association (IMIA) Yearbook of Medical Informatics which focused exclusively on the role of social media documents [23], had a balanced view on the relevance of both Electronic Health Records (EHRs) and social media posts [24], or dealt with the importance of shared tasks for the progress in medical NLP [25]. The last two Yearbook surveys of the NLP section most closely related to medical IE were published in 2015 [26] and 2008 [27]. The survey by Velupillai et al. [28] dealt with opportunities and challenges of medical NLP for health outcomes research, with particular emphasis on evaluation criteria and protocols.

We also refer readers to alternative surveys of DL as applied to medical and clinical tasks. Wu et al. [29] reviewed literature for works using DL for a broader view of clinical NLP, whereas Xiao et al. [30] and Shickel et al. [31] performed systematic reviews on the applications of DL to several kinds of EHR data, not only text. Miotto et al. [32] and Esteva et al. [33] further extended that perspective to include clinical imaging and genomic data beyond the scope of classical EHRs. From an even broader perspective of the huge amounts of biomedical data, Ching et al. [34] examined various applications of DL to a variety of biomedical problems—patient classification, fundamental biological processes, and treatment of patients—and discussed the unique challenges that biomedical data pose for DL methods. In the same vein, Rajkomar et al. [35] used the entire EHR, including clinical free-text notes, for clinical predictive modeling based on DL (targeted, e.g., at the prediction of in-hospital mortality or patient’s final discharge diagnoses). They also demonstrated that DL methods outperformed traditional statistical prediction models.

### Table 1: An Overview of Common Embeddings—Biomedical Language Models

| Model          | Base model | Dimensions | Source corpus                         | Size (in words) |
|----------------|------------|------------|--------------------------------------|-----------------|
| BioWord2Vec    | Word2vec   | 200        | PubMed + PMC                         | 5.58            |
| BioSent2vec    | fastText + node2vec | 200      | PubMed + MeSH                        | 3.78            |
| flair(pubmed-X) | Flair      | 200        | PubMed + MIMIC-III                   | 4.4B + 539M     |
| SciBERT        | BERT base  | 768        | Semantic Scholar                     | 3.1B            |
| BioBERT        | BERT base  | 768        | PubMed + MIMIC-III                   | 4.5B + 13.5B    |
| ClinicalBERT   | BERT large | 768        | MIMIC-III                            | 539M            |
| BlueBERT       | BERT base / BERT large | 768/1024 | PubMed + MIMIC-III                   | >4B + >500M     |

2 Design and Goals of this Survey

In this survey, we concentrated on publications within the time window from 2017 to early 2020 and screened the contributions from three major scientific communities involved in medical IE:

- Medicine and medical informatics are covered by PubMed;
- Natural language processing is covered by the ACL Anthology, the digital library of the Association for Computational Linguistics;
- Neural networks are covered by the major conference series of the neural network community (Neural Information Processing Systems (NIPS/NeurIPS)) whereas the artificial intelligence community gets in via the Association for the Advance ment of Artificial Intelligence (AAAI) Digital Library which keeps the records from the AAAI and IJCAI conferences.

We also included health-related publications from the digital libraries of the Association for Computing Machinery (ACM) and the Institute of Electrical and Electronics Engineers (IEEE). When necessary, we also referred to e-preprint archives such as arXiv.org, since they have become a new, increasingly important distribution channel for the most recent research results in computer science (yet, in that state typically without peer review) and thus foreshadow future directions of research.

We searched these literature repositories with a free-text query that can be approximated as follows: (information extraction OR text mining OR named entity recognition OR relation extraction) AND (deep learning OR neural network) AND (medicine* OR clinical* OR health)

For this setting, we found approximately 1,000 unique publications, screened them for relevance, and, finally, included roughly 100 into this survey.
3 Deep Neural Networks for Medical Information Extraction

In this section, we introduce applications of DNNs to medical NLP for two different tasks, Named Entity Recognition (NER) and Relation Extraction (REX). The focus of our discussion lies on studies dealing with English as reference language since the vast majority of benchmark and reference data sets are in English. After a brief description of each task, we summarize the current SOTA in tables which generalize often subtle distinctions in experimental design and workflows. Our main goal is to show the diversity of major benchmark datasets, DL approaches, and embeddings being used. For these tables, we extracted all symbolic (e.g., corpus or DL approach) and numerical information (e.g., about annotation metadata, performance scores) directly from the cited papers.

The assessment of different systems for the same task is centered around their performance on gold data in evaluation experiments. We refrain from highlighting minor differences in the reported scores because of different datasets being used for evaluation, changing volumes of metadata, and sometimes even the genres they contain. Hence, from a strict methodological perspective, the reported results have to be interpreted with utmost caution for two main reasons [37]. First, the choice of pre-processing steps, such as tokenization, inclusion/exclusion of punctuation marks, stop word removal, morphological normalization/lemmatization/stemming, n-gram variability, entity blinding strategies, and, second, the calibration of training methods (split bias, pooling techniques, hyperparameter selection (dropout rate, window size, etc.)) have a strong impact on the way a chosen embedding type and DL model finally performs, even within the same experimental setting. However, the data we report give valuable comparative information of the SOTA, though with fuzzy edges. This situation might be remedied by a recently proposed common evaluation framework for biomedical NLP, the BLUE (Biomedical Language Understanding Evaluation) benchmark [22], which consists of five different biomedical NLP tasks (including NER and REX) with ten corpora (including BC5CDR, DDI, and i2b2 that also occur in the tables below), or the one proposed by Chauhan et al. [37] ”enabling a more lucid comparison of various training methodologies, pre-processing, modeling techniques, and evaluation metrics.

For the tables provided in the next subsections, we used the F1 score as the main ordering criterion for the cited studies (from highest to lowest). We usually had to select among a large variety of experimental conditions (with different scores). The final choices we made were led by the criterion to favor comparability among all studies. This means that higher (and lower) outcomes may have been reported in the cited studies for varying experimental conditions. Still, the top-ranked system(s) in each of the following tables defines the current SOTA for a particular application.

3.1 Named Entity Recognition

The task of Named Entity Recognition (NER) is to identify crucial medical named entities (i.e., spans of concrete mentions of semantic types such as diseases or drugs and their attributes) in running text. For a recent survey of DL-based approaches and architectures underlying NER as a generic NLP application, see [38].

3.1.1 Diseases

A primary target of NER in the medical field is the automatic identification of diseases in scientific articles and clinical reports. For instance, textual occurrences of disease mentions (e.g., “Diabetes II” or “cerebral inflammation”) are mapped to a common semantic type, Disease6. The crucial role of recognizing diseases in medical discourse is also emphasized by a number of surveys dealing with the recognition of special diseases. For instance, Sheikhalishahi et al. [40] discussed NLP methods targeted at chronic diseases and found that shallow ML and rule-based approaches (as opposed to more sophisticated DL-based ones) prevail. Koleck et al. [41] summarized the use of NLP to analyze symptom information documented in EHR free-text narratives as an indication of diseases and similar to the previous survey found little coverage of ML methods in this application area as well. Savova et al. [42] reviewed the current state of clinical NLP with respect to oncology and cancer phenotyping from EHR. Datta et al. [43] focused on an even more specialized use case—the lexical representation required for the extraction of cancer information from EHR notes in a frame-semantic format.

The research summarized in Table 2 is strictly focused on Disease recognition and, for reasons of comparability, based on the use of shared data sets and metadata (gold annotations). Two benchmarks are prominently featured, BC5CDR [44] and NCBI [45]. BC5CDR is a corpus made of 1,500 PubMed articles, with 4,409 annotated chemicals, 5,818 diseases, and 3,116 chemical-disease interactions, created for the BioCreative V Chemical and Disease Mention Recognition Task [44]. As an alternative, the NCBI Disease Dataset [45] consists of a collection of 793 PubMed abstracts annotated with 6,892 disease mentions which are mapped to 790 unique disease concepts (thus, this corpus can also be used for grounding experiments).

---

2 Wu et al. [29, Table 3(b)] found that 71% of the corpora they screened were English, 20% Chinese, 2% Spanish, Japanese or Finnish and all other languages ranked below 1%. For a survey on medical NLP dealing explicitly with languages other than English, see [36].

4 https://github.com/geetickachauhan/rela-

tion-extraction

5 We disregard here the common distinction between strict and partial matching; numbers given in the tables typically reflect the strongest condition, i.e., strict (complete) match between system prediction and gold standard data.

6 Even more ambitious is the task of linking (or grounding) textual mentions and semantic types to unique identifiers of a given terminology or ontology (such as SNOMED-CT, ICD, or the Human Disease Ontology, https://www.ebi.ac.uk/ ols/ontologies/oid): an issue we will not elaborate on in this survey, cf. e.g., [39].

7 Concrete numbers in the column “Number of Mentions,” indicating the number of named entity mentions (possibly split into training, development, and test set, if provided), may slightly differ for the same corpus because of data cleansing (e.g., removal of duplicates), different pre-processing (e.g., tokenization), and other version issues.
The current top performance for Disease recognition comes close to 90% F1. Lee et al. [20] use a Transformer model with in-domain training (BioBERT), but also (attention-based) BiLSTMs perform strongly in the range of 88–89% F1 score. For the choice of embeddings being used, self-trained ones might be a better choice than pre-trained ones, e.g., those provided by bio.nlplab.org [16]. The incorporation of (large) dictionaries does not provide a competitive advantage in the experiments reported here. Though multi-task learning and transfer learning seem reasonable choices ([39, 46] and [47], respectively) to combat the sparsity of datasets, they generally do not boost systems to the top ranks.

Interesting though are differences for the same approach on different evaluation data sets. For the second-best system by Sachan et al. [47], F1 scores differ for BC5CDR and NCBI by 2.0 (for the third-best [46] by 2.7) percentage points, whereas for the best non-DL approach by Lou et al. [48], this difference amounts to remarkable 4.1 percentage points. This hints at a strong dependence of the results of the same system set-up on the specific corpus; these results have been worked out and, thus, limits generalizability. On the other hand, corpora obviously cannot be blamed for intrinsic analytical hardness since cross-rankings occur: the system by Lee et al. [20] gets the over-all highest F1 score for NCBI but underperforms for BC5CDR, whereas for the tagger used by Sachan et al. [47] the ranking is reversed— their system performs better on BC5CDR than on NCBI (differences are in the range of 2 percentage points). The most stable system in this respect is the one by Zhao et al. [39]. Finally, the distance between the best- and second-best-performing DL systems ([20] and [47], respectively) and their best non-DL
Table 3  Medical Named Entity Recognition: Drugs. Benchmark Datasets: n2c2 [56]; i2b2 2009 [57]; MADE 1.0 [59]; DDI [60].

| Citations          | NE type | Corpus             | # of Mentions train (+ dev) / test | DL Approach                  | Embeddings (source)                        | F1-score % |
|--------------------|---------|--------------------|-----------------------------------|------------------------------|--------------------------------------------|------------|
| Wei et al. [62]    | Drug    | n2c2               | 16,225 / 10,575                   | BiLSTM-CRF                  | Word2Vec embeddings self-trained on MIMIC-III [58] | 95.6       |
| Gligic et al. [63] | Drug    | i2b2 2009          | —                                 | RNN                         | Word2Vec embeddings self-trained on i2b2 2009 [57] | 94.6       |
| Zang et al. [64]   | Drug    | DDI                | 11,260 / 3,689                    | BiLSTM-CRF                  | word & character embeddings pre-trained on Wikipedia | 92.0       |
| Unanue et al. [65] | Drug    | DDI: DrugBank      | 9,715 / 180                       | BiLSTM-CRF                  | GloVe [53] pre-trained on CommonCrawl + GloVe self-trained on MIMIC-III [58] | 91.8       |
| Li et al. [66]     | Drug    | MADE 1.0           | —                                 | BiLSTM-CRF                  | pre-trained [16]                            | 91.3       |
| Wunnava et al. [67]| Drug    | MADE 1.0           | 17,008 /                         | BiLSTM-CRF                  | pre-trained on Wikipedia, EHR notes, and PubMed [68, 69] | 90.4       |
| Dandala et al. [70]| Drug    | MADE 1.0           | 13,507 / 2,395                   | BiLSTM-CRF                  | pretrained character & word embeddings     | 90.0       |
| Tao et al. [71]    | Drug    | i2b2 2009          | 7,988 / 8,440                    | SVM                         | (as feature for SVM: GloVe [53] embeddings self-trained on MIMIC-III [58]) | 88.9       |
| Chapman et al. [72]| Drug    | MADE 1.0           | 13,508 / 2,395                   | CRF                         | n/a                                        | 88.6       |
| Unanue et al. [65] | Drug    | DDI: MEDLINE       | 1,574 / 171                      | BiLSTM-CRF                  | GloVe [53] pre-trained on CommonCrawl + GloVe self-trained on MIMIC-III [58] | 75.6       |

For medication information, four gold standards had a great impact on the field in the past years. The most recent one came out of the 2018 n2c2 Shared Task on Adverse Drug Events and Medication Extraction in Electronic Health Records [56], a successor of the 2009 i2b2 Medication Challenge [57], now with a focus on Adverse Drug Events (ADEs). It includes 505 discharge summaries (303 in the training set and 202 in the test set), which originate from the MIMIC-III clinical care database [58]. The corpus contains nine types of clinical concepts (including drug name), eight attributes (reason, ADE, frequency, strength, duration, route, form, and dosage – from which we chose five for comparison), and 83,869 concept annotations. Relations between drugs and the eight attributes were also annotated and summed up to 59,810 relation annotations (see Section 3.2.1). The third corpus, MADE 1.0 [59], formed the basis for the 2018 Challenge for Extracting Medication, Indication, and Adverse Drug Events (ADEs) from Electronic Health Record (EHR) notes and consists of 1,092 de-identified EHR notes from 21 cancer patients. Each note was annotated with medication information (drug name, dosage, route, frequency, duration), ADEs, indication (symptom as reason for drug administration), other signs and symptoms, severity (of disease/symptom), and relations among those entities, resulting in 79,000 mention annotations. Finally, the DDI corpus [60], originally developed for the Drug-Drug Interaction (DDI) Extraction 2013 Challenge [61], is composed of 792 texts selected from the (semi-structured) DrugBank database and other 233 (unstructured) MEDLINE abstracts, summing up 1,025 documents. This fine-grained corpus has been annotated with a total of 18,502 pharmacological substances and 5,028 drug-drug interactions. Hence, 9. Tables 3 and 4 provide an overview of the SOTA on this topic.

9 See, e.g., the HL7 FHIR Medication Statement at https://www.hl7.org/fhir/medicationstatement.html#MedicationStatement

Notes and consists of 1,092 de-identified EHR notes from 21 cancer patients. Each note was annotated with medication information (drug name, dosage, route, frequency, duration), ADEs, indication (symptom as reason for drug administration), other signs and symptoms, severity (of disease/symptom), and relations among those entities, resulting in 79,000 mention annotations. Finally, the DDI corpus [60], originally developed for the Drug-Drug Interaction (DDI) Extraction 2013 Challenge [61], is composed of 792 texts selected from the (semi-structured) DrugBank database and other 233 (unstructured) MEDLINE abstracts, summing up 1,025 documents. This fine-grained corpus has been annotated with a total of 18,502 pharmacological substances and 5,028 drug-drug interactions. Hence, 10 https://www.drugbank.ca/

11 The DDI corpus is actively maintained and enhanced leading to a large number of versions. Hence, comparisons based on DDI have to be carried out very carefully.
| Citations         | NE type | Corpus    | # of Mentions train (+dev) / test | DL Approach | Embeddings (source)                                                                 | F1-score % |
|--------------------|---------|-----------|-----------------------------------|-------------|------------------------------------------------------------------------------------|------------|
| Wei et al. [62]    | Dr-Freq | n2c2      | 6,281 / 4,012                     | BiLSTM-CRF  | Word2vec embeddings self-trained on MIMIC-III [58]                                | 97.5       |
| Tao et al. [71]    | Dr-Freq | i2b2      | 3,881 / 3,925                     | CRF         | (as feature for CRF: GloVe [53] embeddings self-trained on MIMIC-III [58])       | 92.4       |
| Gligic et al. [63] | Dr-Freq | i2b2 2009 | —                                 | RNN         | Word2vec embeddings self-trained on i2b2 2009 [57]                               | 90.9       |
| Li et al. [66]     | Dr-Freq | MADE 1.0  | —                                 | BiLSTM-CRF  | pre-trained [16]                                                                   | 86.3       |
| Dandala et al. [70]| Dr-Freq | MADE 1.0  | 4,147 / 659                       | BiLSTM-CRF  | pretrained character & word embeddings                                            | 86.3       |
| Chapman et al. [72]| Dr-Freq | MADE 1.0  | 4,147 / 659                       | CRF         | n/a                                                                                | 85.8       |
| Wunnava et al. [67]| Dr-Freq | MADE 1.0  | 5,050 /                            | BiLSTM-CRF  | pretrained on Wikipedia, EHR notes, and PubMed [68, 69]                           | 84.3       |
| Gligic et al. [63] | Dr-Route| i2b2 2009 | —                                 | RNN         | Word2vec embeddings self-trained on i2b2 2009 [57]                               | 96.9       |
| Wei et al. [62]    | Dr-Route| n2c2      | 5,476 / 3,513                     | BiLSTM-CRF  | Word2vec embeddings self-trained on MIMIC-III [58]                                | 95.6       |
| Tao et al. [71]    | Dr-Route| i2b2      | 3,052 / 3,299                     | SVM         | (as feature for SVM: GloVe [53] embeddings self-trained on MIMIC-III [58])       | 94.1       |
| Wunnava et al. [67]| Dr-Route| MADE 1.0  | 2,862 /                            | BiLSTM-CRF  | pre-trained on Wikipedia, EHR notes, and PubMed [68, 69]                           | 92.4       |
| Chapman et al. [72]| Dr-Route| MADE 1.0  | 2,278 / 389                       | CRF         | n/a                                                                                | 91.9       |
| Li et al. [66]     | Dr-Route| MADE 1.0  | —                                 | BiLSTM-CRF  | pretrained character & word embeddings                                            | 91.7       |
| Dandala et al. [70]| Dr-Route| MADE 1.0  | 2,278 / 389                       | BiLSTM-CRF  | pretrained character & word embeddings                                            | 91.7       |
| Wei et al. [62]    | Dr-Dose | n2c2      | 4,221 / 2,681                     | BiLSTM-CRF  | Word2vec embeddings self-trained on MIMIC-III [58]                                | 94.8       |
| Gligic et al. [63] | Dr-Dose | i2b2 2009 | —                                 | RNN         | Word2vec embeddings self-trained on i2b2 2009 [57]                               | 93.0       |
| Tao et al. [71]    | Dr-Dose | i2b2      | 4,132 / 4,371                     | CRF         | (as feature for CRF: GloVe [53] embeddings self-trained on MIMIC-III [58])       | 91.5       |
| Wunnava et al. [67]| Dr-Dose | MADE 1.0  | 5,978 /                            | BiLSTM-CRF  | pre-trained on Wikipedia, EHR notes, and PubMed [68, 69]                           | 88.0       |
| Li et al. [66]     | Dr-Dose | MADE 1.0  | —                                 | BiLSTM-CRF  | pre-trained [16]                                                                   | 88.0       |
| Chapman et al. [72]| Dr-Dose | MADE 1.0  | 4,893 / 801                       | CRF         | n/a                                                                                | 87.5       |
| Dandala et al. [70]| Dr-Dose | MADE 1.0  | 4,893 / 801                       | BiLSTM-CRF  | pretrained character & word embeddings                                            | 87.1       |
| Wei et al. [62]    | Dr-Dur  | n2c2      | 592 / 378                         | BiLSTM-CRF  | Word2vec embeddings self-trained on MIMIC-III [58]                                | 86.2       |
| Li et al. [66]     | Dr-Dur  | MADE 1.0  | —                                 | BiLSTM-CRF  | pre-trained [16]                                                                   | 77.6       |
| Wunnava et al. [67]| Dr-Dur  | MADE 1.0  | 926 /                              | BiLSTM-CRF  | pre-trained on Wikipedia, EHR notes, and PubMed [68, 69]                           | 76.7       |
| Dandala et al. [70]| Dr-Dur  | MADE 1.0  | 765 / 133                         | BiLSTM-CRF  | pretrained character & word embeddings                                            | 75.0       |
| Chapman et al. [72]| Dr-Dur  | MADE 1.0  | 765 / 133                         | CRF         | n/a                                                                                | 72.2       |
| Gligic et al. [63] | Dr-Dur  | i2b2 2009 | —                                 | RNN         | Word2vec embeddings self-trained on i2b2 2009 [57]                               | 63.0       |
| Tao et al. [71]    | Dr-Dur  | i2b2      | 508 / 499                         | CRF         | (as feature for CRF: GloVe [53] embeddings self-trained on MIMIC-III [58])       | 61.8       |
| Wunnava et al. [67]| Dr-ADE  | MADE 1.0  | 1,807 /                            | BiLSTM-CRF  | pre-trained on Wikipedia, EHR notes, and PubMed [68, 69]                           | 63.5       |
| Yang et al. [73]   | Dr-ADE  | MADE 1.0  | 1,807 /                            | BiLSTM-CRF  | pre-trained on Wikipedia, EHR notes, and PubMed [68, 69]                           | 60.5       |
| Dandala et al. [70]| Dr-ADE  | MADE 1.0  | 1,509 / 431                       | BiLSTM-CRF  | pretrained character & word embeddings                                            | 57.7       |
| Li et al. [66]     | Dr-ADE  | MADE 1.0  | —                                 | BiLSTM-CRF  | pre-trained [16]                                                                   | 55.4       |
| Wei et al. [62]    | Dr-ADE  | n2c2      | 959 / 625                         | BiLSTM-CRF  | Word2vec embeddings self-trained on MIMIC-III [58]                                | 53.0       |
| Chapman et al. [72]| Dr-ADE  | MADE 1.0  | 1,509 / 431                       | CRF         | n/a                                                                                | 51.1       |
the medication NER task not only comes with a higher entity type complexity but also with text genres different from the disease recognition task—while the former puts emphasis on clinical reports, the latter focuses on scholarly writing.

Except for route and ADE, all top scores for NER were achieved on the n2c2 corpus. For drug names, the current SOTA exceeds 95% F1 score established by Wei et al. [62]. As to the subtypes, their system also compares favorably to alternative architectures by a large F1 margin ranging from 8.6 percentage points (for duration) down to 1.0 (for drug name). For route, the distance to the best system is marginal (around 1 percentage point)12, whereas for ADE it is huge (more than 10 percentage points, a strong outlier). Overall, frequency, route, and dosage recognition reach outstanding F1 scores in the range of 95 up to 97%, while for duration information top F1 scores drop remarkably by at least 10 to 20 percentage points. Still, the recognition of ADEs seems to be the hardest task, with the best system by Wunnava et al. [67] peaking at around 64% F1 on MADE 1.0 data (here the top performing system by Wei et al. [62] plummed down to 53% F1). Interestingly, ADEs are verbally the least constrained type of natural language utterance compared with all the other entity types considered here.

In terms of DL methodology, BiLSTM-CRFs are the dominating approach. Yet, the type of embeddings used by different DL systems varies a lot ranging from pre-trained Word2vec embeddings and those self-trained on MIMIC-III (for the top performers) to GloVe embeddings pre-trained on CommonCrawl, Wikipedia, EHR notes, and PubMed. There seems to be no generalizable winner for either choice of embeddings given the current state of evaluations, but self-training on medical raw data, such as MIMIC-III, challenge data sets, or, more advisable, using the now available BioSentVec [18] and BlueBERT [22] embeddings pre-trained on MIMIC-III, might be advantageous.

Studies in which the same system configuration was tested on different corpora are still lacking so that corpus effects are unknown (unlike for diseases; see Table 2). Yet, there is one interesting though not so surprising observation: Unanue et al. [65] explored the two slices of the DDI corpus, with a span of F1 scores of more than 16 percentage points. This obviously witnesses the influence of a priori (lack of) structure—DrugBank data is considerably more structured than MEDLINE free texts and, thus, the former gets much higher scores than the latter.

Comparing DL approaches vs. non-DL ones (a CRF architecture) on the same corpus (MADE 1.0), we found that for the core entity type (Drug), the recognition performance differs by almost 3 percentage points, for frequency, route and dose marginally by less than 1, yet for duration and ADE it amounts to roughly 5 and 12 percentage points, respectively—consistently in favor of Deep Neural Networks (DNNs).

3.2 Relation Extraction

Once named entities have been identified, a follow-up question emerges: does some sort of semantic relation hold among these entities? We surveyed this Relation Extraction (REX) task with reference to results that have been achieved for information related to medication attributes and drug-drug interaction.

3.2.1 Medication-Attribute Relations

In Section 3.1.2, we already dealt with single named entity types typically associated with medication information, namely drug names and administration frequency, duration, dosage, route, and ADE, yet in isolated form only. In this subsection, we are concerned with making the close associative ties between Drugs and typical conceptual attributes, such as Frequency, Duration, Dosage, Route, ADE, and Reason (for prescription), explicit. Hence, the recognition of the respective named entity types (Drugs, Dr-Freq, Dr-Dur, Dr-Dose, Dr-Route, Dr-ADE, and Dr-Reason) turns out to be a good starting point for solving this REX task. Not surprisingly, the benchmarks for this task are a subset of the ones in Tables 3 and 4 depicting the results for medication-related NER. Table 5 provides an overview of the experimental results for finding medication-attribute relations in medical, in effect, clinical, documents.

The overall results from medication-focused NER are mostly confirmed for the REX task. The n2c2 corpus is the reference dataset for top performance. The group who achieved top F1 scores for the medication NER problem also performed best for the medication-attribute REX task [62], with extraordinary figures for Frequency, Route, and Dosage relations (in the upper 98% F1 range), a superior one for the Duration relation (93% F1), and good ones on the (hard to deal with) Adverse and Reason relations (85% F1). Still, the distances to the second-best system for the same corpus (n2c2) are not so pronounced in most cases, ranging by 1 percentage point (for Frequency, Route, Dosage, and Duration), yet increased up to 3 (for Adverse) and 7 (for Reason) percentage points.

For the MADE 1.0 corpus, a similar picture emerges. From a lower offset (typically around 3 F1 percentage points compared with n2c2), differences between the best and second-best systems were on the order of (negligible) 1 percentage point for Frequency, Route, and Dosage, yet increased by roughly 3, 5, and 7 percentage points for Reason, Duration, and Adverse events, respectively. Yet, in 4 out of 6 cases (Frequency, Dosage, Duration, and Adverse events) non-DL systems (CRFs, SVMs) outperformed their DL counterparts with small margins (in the range of (again, negligible) 1 percentage point) for Frequency and Dosage, yet with higher ones for Duration and Adverse events (5 and 7 percentage points, respectively). In cases where the DL approach ranked higher than a non-DL one, differences ranged between 1 and 3 percentage points (for Route and Reason, respectively). Thus, the MADE 1.0 corpus constitutes a benchmark where well-engineered standard ML classifiers can still play a competitive role. However, we did not find this pattern of partial supremacy of non-DL approaches for the n2c2 benchmark.

The top performers for the medication attribute REX task [62] employed a joint learning approach based on CNN-RNN

12 Interestingly, the transfer learning approach advocated by Gligic et al. [63] performs well for some medication NER tasks, but fails to deliver competitive results for the medication relation task (cf. Table 5).

IMIA Yearbook of Medical Informatics 2020
Hahn et al.
### Table 5: Medical Relation Extraction: Medication-Attribute Relations (including ADEs). Benchmark Datasets: n2c2 [56]; MADE 1.0 [59].

| Citations     | Relation type | NE type 1 | NE type 2 | Corpus     | # of Mentions train (+ dev) / test | DL Approach                      | Embeddings (source)                                                                 | F1-score % |
|---------------|---------------|-----------|-----------|------------|----------------------------------|----------------------------------|-----------------------------------------------------------------------------------|------------|
| Wei et al. [62] | Frequency     | Drug      | Dr-Freq   | n2c2       | (6,310 / 4,034)                  | CNN-RNN                          | Word2vec embeddings self-trained on MIMIC-III [58]                                | 98.7       |
| Christopoulou et al. [74] | Frequency     | Drug      | Dr-Freq   | n2c2       | 4,986 ( + 1,312)                  | Att-BiLSTM-CRF  + Transformer Network | word and relative position embeddings                                               | 97.6       |
| Chapman et al. [72] | Frequency     | Drug      | Dr-Freq   | MADE 1.0   | 4,419 / 730                       | non-DL: CRF                      | n/a                                                                               | 94.7       |
| Dandala et al. [70] | Frequency     | Drug      | Dr-Freq   | MADE 1.0   | 4,419 / 730                       | Att-BiLSTM                       | pretrained character & word embeddings                                            | 93.7       |
| Wei et al. [62] | Route         | Drug      | Dr-Route  | n2c2       | (5,538 / 3,546)                  | CNN-RNN                          | Word2vec embeddings self-trained on MIMIC-III [58]                                | 98.6       |
| Christopoulou et al. [74] | Route         | Drug      | Dr-Route  | n2c2       | 4,327 ( + 1,208)                  | Att-BiLSTM-CRF  + Transformer Network | word and relative position embeddings                                               | 97.8       |
| Dandala et al. [70] | Route         | Drug      | Dr-Route  | MADE 1.0   | 2,551 / 455                       | non-DL: CRF                      | n/a                                                                               | 95.3       |
| Chapman et al. [72] | Route         | Drug      | Dr-Route  | MADE 1.0   | 2,551 / 455                       | non-DL: CRF                      | n/a                                                                               | 94.1       |
| Wei et al. [62] | Dosage        | Drug      | Dr-Dose   | n2c2       | (4,225 / 2,695)                  | CNN-RNN                          | Word2vec embeddings self-trained on MIMIC-III [58]                                | 98.6       |
| Christopoulou et al. [74] | Dosage        | Drug      | Dr-Dose   | n2c2       | 3,299 ( + 921)                    | Att-BiLSTM-CRF  + Transformer Network | word and relative position embeddings                                               | 98.0       |
| Chapman et al. [72] | Dosage        | Drug      | Dr-Dose   | MADE 1.0   | 5,177 / 866                       | non-DL: CRF                      | n/a                                                                               | 96.0       |
| Dandala et al. [70] | Dosage        | Drug      | Dr-Dose   | MADE 1.0   | 5,177 / 866                       | non-DL: CRF                      | n/a                                                                               | 95.2       |
| Wei et al. [62] | Duration      | Drug      | Dr-Dur    | n2c2       | (643 / 426)                       | CNN-RNN                          | Word2vec embeddings self-trained on MIMIC-III [58]                                | 92.9       |
| Chapman et al. [72] | Duration      | Drug      | Dr-Dur    | MADE 1.0   | 906 / 147                         | non-DL: CRF                      | n/a                                                                               | 92.4       |
| Christopoulou et al. [74] | Duration      | Drug      | Dr-Dur    | n2c2       | 518 ( + 124)                      | Att-BiLSTM-CRF  + Transformer Network | word and relative position embeddings                                               | 91.7       |
| Dandala et al. [70] | Duration      | Drug      | Dr-Dur    | MADE 1.0   | 906 / 147                         | Att-BiLSTM                       | pretrained character & word embeddings                                            | 87.8       |
| Wei et al. [62] | Adverse       | Drug      | Dr-ADE    | n2c2       | (1,107 / 733)                     | CNN-RNN                          | Word2vec embeddings self-trained on MIMIC-III [58]                                | 85.0       |
| Christopoulou et al. [74] | Adverse       | Drug      | Dr-ADE    | n2c2       | 891 ( + 216)                      | Att-BiLSTM-CRF  + Transformer Network | word and relative position embeddings                                               | 78.3       |
| Chapman et al. [72] | Adverse       | Drug      | Dr-ADE    | MADE 1.0   | 2,082 / 530                       | non-DL: CRF                      | n/a                                                                               | 73.1       |
| Dandala et al. [70] | Adverse       | Drug      | Dr-ADE    | MADE 1.0   | 2,082 / 530                       | non-DL: CRF                      | n/a                                                                               | 66.0       |
| Wei et al. [62] | Reason        | Drug      | Dr-Reason | n2c2       | (5,169 / 3,410)                  | CNN-RNN                          | Word2vec embeddings self-trained on MIMIC-III [58]                                | 84.9       |
| Christopoulou et al. [74] | Reason        | Drug      | Dr-Reason | n2c2       | 4,069 ( + 1,090)                  | Att-BiLSTM-CRF  + Transformer Network | word and relative position embeddings                                               | 82.2       |
| Dandala et al. [70] | Reason        | Drug      | Dr-Reason | MADE 1.0   | 4,554 / 876                       | Att-BiLSTM                       | pretrained character & word embeddings                                            | 80.9       |
| Yang et al. [73] | Reason        | Drug      | Dr-Reason | MADE 1.0   | 4,530 / 871                       | non-DL: SVM                      | n/a                                                                               | 77.9       |
(thus diverging from the most successful architectures for medication NER; see Tables 3 and 4) and rule-based post-processing that outperformed a simple CNN-RNN. Summarizing, the CNN-RNN approach seems more favorable than an (attention-based) BiLSTM, with preferences for self-trained in-domain embeddings.

### 3.2.2 Drug-Drug Interaction

The second type of medication-focused relation we consider here are drug-drug interactions as featured in the DDI challenge (for surveys on the impact of DL on recent research on drug-drug interactions, cf. [82, 83], for a survey on drug-drug interaction combining progress in data and text mining from EHRs, scientific papers, and clinical reports but lacking in-depth coverage of DL methods, cf. [84], for the NLP-focused recognition of ADEs also lacking awareness of DL contributions to this topic, cf. [85]). Four main types of relations between drugs are considered: pharmacokinetic Mechanism, drug Effect, recommendation or Advice regarding a drug interaction, and Interaction between drugs without providing any additional information. Overall, the DDI corpus on which these evaluations were run is divided into 730 documents taken from DrugBank and 175 abstracts from MEDLINE and contains 4,999 relation annotations (4,020 train, 979 test).

Recognition rates for these relations (cf. Table 6) are considerably lower than for the medication-related attributes when linked to drugs (cf. Table 5). The best systems peak at 85% F score for Advice (a distance of more than 13 percentage points to the top recognition results for medication-attributes), they slip to 78%\(^{13}\) and 77% for Mechanism and Effect, respectively, and plummet to 59% for Interaction\(^{14}\). Differences between the first and second-ranked systems are typically small, yet become larger on subsequent ranks (roughly between 3 to 4 percentage points relative to the top-ranked system).

As with medication attributes, drug-drug interactions can also be recognized in a competitive way by CNN-RNN architectures, but attention-based LSTMs perform also considerably well. Again, self-trained embeddings using in-domain corpora seem to be advantageous for this relation class. Reflecting the drop in performance, one may conclude that drug-drug interactions constitute a markedly harder task than the conceptually much closer medication-attribute relations.

Finally, Table 6 most drastically supports our claim that DL approaches outperform non-DL ones. The difference between both approaches amounts to 5 percentage points for Mechanism, 7 for Effect and Interaction, and 8 for Advice.

### 4 Conclusions

We have presented various forms of empirical evidence that (with one exception only) Deep Learning-based neural networks outperform non-DL, feature engineered, approaches for several information extraction tasks. However, despite their success, Deep Neural Networks and their embedding models have their shortcomings as well.

One of the most problematic issues is their dependence on huge amounts of training data: SOTA embedding models are currently trained on hundreds of billions of tokens [89]. This magnitude of data volume is out of reach for any training effort in the medical/clinical domain [90]. Also, embeddings are very vulnerable to malicious attacks or adversarial examples—small changes at the input level may result in severe misclassification [5]. Another well-known problem relates to the instability of word embeddings. Word embeddings depend on their random initialization and the processing order of the underlying examples and therefore they do not necessarily converge on exactly the same embeddings even after several thousands of training iterations [91, 92]. Finally, although DL is celebrated for not requiring manual feature engineering, the effects of proper hyperparameter tuning on DNNs [93] remain an issue for DL [94]. Apart from these intrinsic problems, Kalyan and Sangeetha [95] and Khattak et al. [96] refer to extrinsic drawbacks of neural networks, such as opaque encodings (resulting in lacking interpretability) or limited transferability of large models (hindering knowledge distillation for smaller models).

Still, the sparsity of corpora and special linguistic phenomena of the medical (clinical) sublanguage(s) create intrinsic problems for data-greedy DL approaches that have to be overcome by special learning strategies for neural systems, such as transfer learning or domain adaptation. Research on adapting general language models to medical language constraints is just in its beginning. Yet, there is no simple solution to this problem. Wang et al. [97] evaluated Word2vec embeddings trained on private clinical notes, PMC, Wikipedia, and the Google News corpus both qualitatively and quantitatively and showed that the ones trained on Electronic Health Record data performed better on most of the tested scenarios. However, they also found that word embeddings trained on biomedicalexternal corpora do not necessarily have better performance than those trained on general domain corpora for any downstream biomedical NLP task (other experimental evidence of the effects of in- and out-of-domain corpora and further parameters, such as corpus size, on word embedding performance is reported by Lai et al., [98]).

While this survey focused on the application domain of medical IE to demonstrate the outstanding role of DL for medical Natural Language Processing, one might be tempted to generalize this trend to other applications as well. There is, indeed, plenty of evidence in the literature that other application fields, such as question answering (and the closely related area of machine reading), summarization, machine translation, and speech pro-

\(^{13}\) Xu et al. [86] even reach slightly more than 79% F score for Mechanism (using UMLS-based concept embeddings with a Bi-LSTM approach), but substantially fall below the results for the other three relation types in comparison with all the systems mentioned in Table 6.

\(^{14}\) Dewi et al. [87] and Sun et al. [88] report on 86.3% and 84.5% F\(_1\) scores, respectively, for the overall relation classification task both using a multi-layered CNN architecture, yet unfortunately fail to provide details on each of the four single relations under scrutiny here. Both results exceed the overall result of the best-performing system depicted in Table 6 [76] (77.3%) by a large margin of 9 and 7 percentage points, respectively.
### Table 6  Medical Relation Extraction: Drug-Drug Interaction. Benchmark Dataset: DDI [60].

| Citation          | Relation type | NE type 1 | NE type 2 | Corpus  | # of Mentions train (+ dev) / test | DL Approach                           | Embeddings (source)                                                                 | F1-score |
|-------------------|---------------|-----------|-----------|---------|------------------------------------|---------------------------------------|-------------------------------------------------------------------------------------|----------|
| Sun et al. [75]   | Mechanism     | Drug      | Drug      | DDI     | 1,319 / 299                        | Recurrent hybrid CNN with focal loss function + semantic & position embeddings self-trained on DDI | pre-trained on PubMed & Wikipedia: [16] + semantic & position embeddings self-trained on DDI | 78.3     |
| Zhang et al. [76] | Mechanism     | Drug      | Drug      | DDI     | 1,319 / 302                        | Att-BiLSTM                             | word embeddings self-trained on Drug subset of PubMed + DDI corpus                  | 77.5     |
| Wang et al. [77]  | Mechanism     | Drug      | Drug      | DDI     | 1,319 / 302                        | BiLSTM                                | word embeddings: [78] self-trained on PubMed + dependency layers                  | 75.4     |
| Lim et al. [79]   | Mechanism     | Drug      | Drug      | DDI     | 1,260 / 301                        | binary tree-LSTM                       | pre-trained on PubMed & Wikipedia: [16]                                          | 75.1     |
| Zhang et al. [80] | Mechanism     | Drug      | Drug      | DDI     | 1,319 / 302                        | hierarchical                          | word embeddings self-trained on Drug subset of PubMed                            | 74.0     |
| Raihani and Laachfoubi [81] | Mechanism | Drug      | Drug      | DDI     | 1,319 / 302                        | non-DL: SVM                           | n/a                                                                               | 73.5     |
| Zhang et al. [76] | Effect        | Drug      | Drug      | DDI     | 1,687 / 360                        | Att-BiLSTM                             | word embeddings self-trained on Drug subset of PubMed + DDI corpus                | 76.6     |
| Sun et al. [75]   | Effect        | Drug      | Drug      | DDI     | 1,669 / 360                        | Recurrent hybrid CNN with focal loss function + semantic & position embeddings self-trained on DDI | pre-trained on PubMed & Wikipedia: [16] + semantic & position embeddings self-trained on DDI | 73.5     |
| Lim et al. [79]   | Effect        | Drug      | Drug      | DDI     | 1,592 / 357                        | binary tree-LSTM                       | pre-trained on PubMed & Wikipedia: [16]                                          | 72.9     |
| Zhang et al. [80] | Effect        | Drug      | Drug      | DDI     | 1,687 / 360                        | hierarchical                          | word embeddings self-trained on Drug subset of PubMed                            | 71.8     |
| Wang et al. [77]  | Effect        | Drug      | Drug      | DDI     | 1,687 / 360                        | BiLSTM                                | word embeddings: [78] self-trained on PubMed + dependency layers                  | 69.5     |
| Raihani and Laachfoubi [81] | Effect | Drug      | Drug      | DDI     | 1,687 / 360                        | non-DL: SVM                           | n/a                                                                               | 69.5     |
| Zhang et al. [76] | Advice        | Drug      | Drug      | DDI     | 826 / 221                          | Att-BiLSTM                             | word embeddings self-trained on Drug subset of PubMed + DDI corpus                | 85.1     |
| Lim et al. [79]   | Advice        | Drug      | Drug      | DDI     | 814 / 221                          | binary tree-LSTM                       | pre-trained on PubMed & Wikipedia: [16]                                          | 82.7     |
| Wang et al. [77]  | Advice        | Drug      | Drug      | DDI     | 826 / 221                          | BiLSTM                                | word embeddings: [78] self-trained on PubMed + dependency layers                  | 80.9     |
| Sun et al. [75]   | Advice        | Drug      | Drug      | DDI     | 822 / 221                          | Recurrent hybrid CNN with focal loss function + semantic & position embeddings self-trained on DDI | pre-trained on PubMed & Wikipedia: [16] + semantic & position embeddings self-trained on DDI | 80.5     |
| Zhang et al. [80] | Advice        | Drug      | Drug      | DDI     | 826 / 221                          | hierarchical                          | word embeddings self-trained on Drug subset of PubMed                            | 80.3     |
| Raihani and Laachfoubi [81] | Advice | Drug      | Drug      | DDI     | 826 / 221                          | non-DL: SVM                           | n/a                                                                               | 77.4     |
cessing, reveal the same pattern. However, for text categorization (in the sense of mapping free text to some pre-defined medical category system, such as ICD, SNOMED, or MeSH) this preference is less obvious, since traditional Machine Learning or rule-based models still play an important role here and, more often than for the IE application scenario, show competitive performance against DL approaches. Whether this exception will persist or will be swept away by future research remains an open issue.

Acknowledgments

The first author was partially funded by the German Bundesministerium für Bildung und Forschung (BMBF) within the SMITH project under grant no. 01ZZ1803G. We thank all six reviewers for their insightful and helpful comments.

References

1. Goodfellow IJ, Bengio Y, Courville AC. Deep Learning. MIT Press; 2016.
2. Alom MZ, Taha TM, Yalcopcic C, Westberg S, Sindre P, Nasrin MSet al. A state-of-the-art survey on deep learning theory and architectures. Electronics 2019;8(3):292.
3. Pouyanfar S, Sadiq S, Yan Y, Tian H, Tao Y, Presa Reyes ME, et al. A survey on deep learning: algorithms, techniques, and applications. ACM Computing Surveys 2018;51(5):92(2):1–92:36.
4. Goldberg Y. Neural Network Methods for Natural Language Processing. Number 37 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool; 2017.
5. Belinkov Y, Glass JR. Analysis methods in neural language processing: a survey. Transactions of the Association for Computational Linguistics 2019;7:49–72.
6. Schmidhuber HJ. Deep learning in neural networks: an overview. Neural Networks 2015;61:85–117.
7. Holman FM, Kalig M, Pienta R, Chau DH. Visual analytics in deep learning: an interrogative survey for the next frontiers. IEEE Trans Vis Comput Graph 2019;24(8):2674–93.
8. Nassif AB, Shahin I, Attili I, Azzeh M, Shaalan K. Speech recognition using deep neural networks: a systematic review. IEEE Access 2019;7:19143–65.
9. Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing. IEEE Computational Intelligence Magazine 2018;13(3):55–75.
10. Spasić I, Uzuner Ö, Zhou L. Emerging clinical applications of text analytics. Int J Med Inform 2020;134:103974.
11. Wang Y, Wang L, Rastegar-Mojarad MA, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018;77:34–49.
12. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. J Biomed Inform 2017;73(Supplement C):14–29.
13. Friedman C, Kra P. Rzehetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. J Biomed Inform 2002;35(4):222–35.
14. Newman-Griffis D, Fosler-Lussier E. Writing habits and telltale neighbors: analyzing clinical concept usage patterns with sublanguage embeddings. In: Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis LOUHI@EMNLP-IJCNLP 2019. p. 146–56.
15. Nunez J-J, Carenini G. Comparing the intrinsic performance of clinical concept embeddings by their field of medicine. In: Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis LOUHI@EMNLP-IJCNLP 2019. p. 11–7.
16. Pyysalo S, Ginter F, Moen H, Salakoski T, Ana niadou S. Distributional semantics resources for biomedical text processing. In: Proceedings of the 3rd International Symposium on Languages in Biology and Medicine, LMB 2013. p. 39–43.
17. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Scientific Data 2019;6:52.
18. Chen Q, Peng Y, Lu Z. BioSentVec : creating sentence embeddings for biomedical texts. In: Proceedings of the 7th IEEE International Conference on Healthcare Informatics, ICHI 2019.
19. Beltagy I, Lo K, Cohan A. SciBert : a pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing & 9th International Joint Conference on Natural Language Processing EMNLP-IJCNLP 2019. p. 3615–20.
20. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT : a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–40.
21. Alsentzer E, Murphy JR, William Boag W, Weng W-H, Jin D, Naumann T, et al. Publicly available clinical Bert embeddings. In: Proceedings of the 2nd Workshop on Clinical Natural Language Processing ClinicalNLP @ NAACL-HLT 2019. p. 72–8.
22. Peng Y, Yan S, Zhiyong Lu Z. Transfer learning in biomedical natural language processing: an evaluation of Bert and ELMo on ten benchmarking datasets. In: Proceedings of the 18th SIGBioMed Workshop on Biomedical Natural Language Processing and Shared Task BioNLP @ ACL 2019. p. 58–65.

23. Conway M, Hu M, Chapman WW. Recent advances in using natural language processing to address public health research questions using social media and consumer-generated data. Yearb Med Inform 2019;28:208–17.

24. Gonzalez-Hernandez G, Sarker A, O’Connor K, Savova GK. Capturing the patient’s perspective: a review of advances in natural language processing of health-related text. Yearb Med Inform 2017;26:214–27.

25. Filaninno M, Uzuner Ö. Advancing the state of the art of clinical natural language processing through shared tasks. Yearb Med Inform 2018:27:184–92.

26. Velupillai S, Mowery DL, South BR, Kvist M, Dalianis H. Recent advances in clinical natural language processing in support of semantic analysis. Yearb Med Inform 2015;24:183–93.

27. Meyster SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008;17:128–44.

28. Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley KI, et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database (Oxford) 2016;2016:baw068.

29. Datta S, Bernstein EV, Roberts K. A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. J Biomed Inform 2019;100:103301.

30. Li J, Sun Y, Johnson RJ, Slatky D, Wei C-H, Leaman R, et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database (Oxford) 2016;2016:baw068.

31. DOglan RJ, Robert Leaman R, Lu Z. NCBI Disease Corpus: a resource for disease name recognition and concept normalization. J Biomed Inform 2014;47:1–10.

32. Wang X, Zhang Y, Ren X, Zhang Y, Zitnik M, Shang J, et al. Cross-type biomedical named entity recognition with deep multi-task learning. BMC Bioinformatics 2018;35(10):1745–52.

33. Sachan DS, Xie P, Sachan M, Xing EP. Effective use of bidirectional language modeling for transfer learning in biomedical named entity recognition. In: Proceedings of the 2019 Conference on Machine Learning for Healthcare 2018. p. 383–402.

34. Lou Y, Zhang Y, Qian T, Li F, Xiong S, Ji D. A transition-based joint model for disease named entity recognition and normalization. Bioinformatics 2019;35(1):1235–71.

35. Xu K, Yang Z, Kang P, Wang Q, Liu W. Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition. Comput Biol Med 2019;108:122–32.

36. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 27th Annual Conference on Neural Information Learning Systems NIPS 2013. p. 3111–9.

37. Hong SK, Lee J-G. DTransNER: biomedical named entity recognition with deep learning-based label-label transition model. BMC Bioinformatics 2020;21(1):53.

38. Peters ME, Neumann M, Iyyer M, Gardner M, Clark CT, Lee K, et al. Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018; 1: Long Papers. p. 2227–37.

39. Zhao S, Liu T, Zhao S, Wang F. A neural multitask learning framework to jointly model medical named entity recognition and normalization. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence 2019. p. 817–24.

40. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Fabio Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform 2016;24(2):e53.

41. Collarobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa PP. Natural language processing (almost) from scratch. Journal of Machine Learning Research 2011;12(76):2493–537.

42. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Processing electronic health records. NPJ Digit Med 2018;22(5):1589–604.

43. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, Marlin BW, Rubin DA, et al. Deep learning in electronic health records: a systematic review. J Biomed Semantics 2018;9(1):12.

44. Filaninno M, Uzuner Ö. Using clinical natural language processing to extract clinical cancer phenotypes from electronic medical records. Cancer Res 2019;79(21):5463–70.

45. Zeng D, Sun C, Lin L, Liu B. LSTM-CRF for drug–disease relation extraction. Y earb Med Inform 2017;26:214–27.

46. J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 27th Annual Conference on Neural Information Learning Systems NIPS 2013. p. 3111–9.

47. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the 2015 IEEE International Conference on Computer Vision, ICCV 2015. p. 1026–34.

48. Henry S, Buchan K, Filaninno M, Stubbs A, Uzuner Ö. 2018 n2c2 Shared Task on Adverse Drug Events and Medication Extraction in Electronic Health Records. J Am Med Inform Assoc 2020;27(1):3–12.

49. Uzuner Ö, Solt I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc 2010;17(5):514–8.

50. Johnson AEW, Pollard TJ, Shen L, Lehman L-Wh, Feng M, Ghassemi MM, et al. Mimic-III, a freely accessible critical care database. Scientific Data 2016;3:160035.

51. Jagannatha A, Liu F, Liu W, Yu H. Overview of the First Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (Mado 1.0). Drug Saf 2019;42(1):99–111.

52. Herrero-Zazo M, Segura-Bedmar I, Martinez P, Declerck T. The DDI corpus: an annotated corpus with pharmacological substances and drug–drug interactions. J Biomed Inform 2013;46(5):914–20.

53. Segura-Bedmar I, Martinez P, Herrero-Zazo M, Se-meval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In: Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@ NAACL-HLT 2013. p. 341–50.

54. Wei Q, Li Z, Li Z, Du J, Wang J, Xu J, et al. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. J Am Med Inform Assoc 2019;27(1):13–21.

55. Gligic L, Kornmitzin A, Goldberg P, Nevado-Holgado A. Named entity recognition in electronic health records: a methodological review of deep learning approaches for medication and adverse drug event extraction. J Am Med Inform Assoc 2019;26(1):24–9.

56. Chen T, Himmelstein DS, Beaulieu-Jones BK, Litt MA, Gaffney O, Macdonald EM. 2017. In: Proceedings of the 18th SIGBioMed Workshop on Biomedical Natural Language Processing and Shared Task BioNLP @ ACL 2019. p. 30–47.

57. Li J, Sun A, Han J, Li C. A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering, page [Early Access], 2020. Available at: https://arxiv.org/pdf/1812.09449.pdf.

58. Zhan S, Liu Y, Zhao S, Wang F. A neural multitask learning framework to jointly model medical named entity recognition and normalization. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence 2019. p. 817–24.

59. Dalianis H. Recent advances in clinical natural language processing in support of semantic analysis. Yearb Med Inform 2015;24:183–93.

60. Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley KI, et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database (Oxford) 2016;2016:baw068.

61. Segura-Bedmar I, Martínez P, Herrero-Zazo M, SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In: Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@ NAACL-HLT 2013. p. 341–50.

62. Wei Q, Li Z, Li Z, Du J, Wang J, Xu J, et al. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. J Am Med Inform Assoc 2019;27(1):13–21.

63. Gligic L, Kornmitzin A, Goldberg P, Nevado-Holgado A. Named entity recognition in electronic health records: a methodological review of deep learning approaches for medication and adverse drug event extraction. J Am Med Inform Assoc 2019;26(1):24–9.

64. Zeng D, Sun C, Lin L, Liu B. LSTM-CRF for drug–disease named entity recognition. Entropy 2017;19(6):283.

65. Uzuner Ö, Solt I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc 2010;17(5):514–8.
electronic health records using hierarchical recurrent neural networks with dual-level embedding. Drug Saf 2019;42(1):113–22.
68. Jagannatha AN, Yu H. Bidirectional RNN for medical event detection in electronic health records. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2016. p. 473–82.
69. Jagannatha AN, Yu H. Structured prediction models for RNN based sequence labeling in clinical text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016. p. 856–65.
70. Dandala B, Jooptudi V, Devarakonda MV. Adverse drug event detection in clinical notes by jointly modeling entities and relations using neural networks. Drug Saf 2019;42(1):135–46.
71. Tao C, Filannino M, Uzuner O. Prescription extraction using CRFs and word embeddings. J Biomed Inform 2017;72:60–6.
72. Chapman AB, Peterson KS, Alba PR, DuVall SL, Patterson OV. Detecting adverse drug events with rapidly trained classification models. Drug Saf 2019;42(1):147–56.
73. Yang X, Bian J, Gong Y, Hogan WR, Wu Y. MADEX: a system for detecting medications, adverse drug events, and their relations from clinical notes. Drug Saf 2019;42(1):123–33.
74. Christopoulou F, Tran TT, Sahu SK, Miwa M, Ananiadou S. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. J Am Med Inform Assoc 2020;27(1):39–46.
75. Sun X, Dong K, Ma L, Sutcliffe RFE, He F. An attention-based effective neural model for drug-drug interaction extraction from the literature: a review. Brief Bioinform 2019;bbz087.
76. Zhang Y, Lin H, Yang Z, Wang J, Su Y, Xu B, et al. Neural network-based approaches for biomedical relation classification: a review. J Biomed Inform 2019;99:103294.
77. Vilar S, Friedman C, Hirjipcak GM. Detection of drug-drug interactions through data mining studies using clinical sources, scientific literature and social media. Brief Bioinform 2018;19(5):863–77.
78. Luo Y, Thompson WK, Herr TM, Zeng Z, Berendsen MA, Jommalagadda SR, et al. Natural language processing for EHR-based pharmacovigilance: a structured review. Drug Saf 2017;40(11):1075–89.
79. Xu B, Shi X, Zhao Z, Zheng W. Leveraging biomedical resources in Bi-LSTM for drug-drug interaction extraction. IEEE Access 2018;6:33432–9.
80. Dewi IN, Dong S, Hu J. Drug-drug relation extraction with deep convolutional neural networks. In: Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017. p. 1779–82.
81. Sun X, Ma L, Du X, Feng J, Dong K. Deep convolutional neural networks for drug-drug interaction extraction. In: Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018. p. 1662–8.
82. Grave E, Bojanowski P, Gupta P, Joulin A, Mikolov T. Learning word vectors for 157 languages. In: Proceedings of the 11th International Conference on Language Resources and Evaluation, LREC 2016. p. 3483–7.
83. Mikolov T, Chen K, Corrado GS, Dean J. Efficient estimation of word representations in vector space. In: Proceedings of the 1st International Conference on Learning Representations, ICLR 2013.