# Survey of Information Extraction Techniques for Pharmacovigilance

**Tanay Jayant Kayastha and Pushpak Bhattacharyya**
Department of Computer Science & Engineering
Indian Institute of Technology Bombay
Mumbai, Maharashtra, India - 400076
{tanayjk, pb}@cse.iitb.ac.in

## Abstract

Information Extraction (IE) is concerned primarily with the extraction of entities and relationships from text. This study primarily focuses on various approaches to Information Extraction from textual resources, and their immense application to Pharmacovigilance IE. We explore various approaches to entity extraction, relation extraction including rule-based, classical machine learning-based, and deep learning-based approaches. This paper also presents the auxiliary tools and techniques often used in conjunction with IE models to enhance their performance.

## 1 Introduction

Adverse Drug Event (ADE) is an unfavourable, unintended symptom temporarily associated with the use of the medicinal products[1]. Pharmacovigilance refers to monitoring drug safety by identifying associated ADEs during clinical trials and after its approval. The prime necessity for pharmacovigilance is that the complete ADE profile for a drug is unknown during its approval (Ahmad, 2003) owing to the clinical trials done under a controlled environment that differ from actual drug usage conditions. ADEs are among the top 10 leading causes of death, costing around $75 billion annually in the United States (Ahmad, 2003). Challenges to ADE include delayed/under-reporting of adverse events by pharmaceutical companies (Ahmad, 2003), informal language by reporters, and non-standard medical codes (Friedman, 2009). For example, awake for a long time vs. Unified Medical Language System (UMLS)[1] insomnia. At the same time, there is an increase in the number of patients sharing their drug usage experiences over publicly accessible social media such as twitter (Freifeld et al., 2014) or medical fora like Drugs.com[2] or WebMD[3]. This paper explores the existing techniques for automating the task of Pharmacovigilance.

[1] https://www.ncbi.nlm.nih.gov
[2] https://www.drugs.com/
[3] https://www.webmd.com/

## 2 ADE Extraction Task

We now present the formulation of ADE extraction task as an IE task.

### 2.1 Background Terminology

Before formally describing an ADE task, here is a glossary of frequently encountered terminology in pharmacovigilance[4]: **Indication**: It is an authentic reason to prescribe or perform a specific test, medication or surgery. An indication is visible to the patient, doctors, and others e.g. increased heartbeats or body temperature. **Symptom**: It is evidence for a disease experienced by the patient e.g. nausea, stomach pain, or fatigue. We refer to indication and symptom phrases collectively as treatment events or positive events.

### 2.2 ADE Extraction as IE

Consider an example, *"the doctor prescribed **linaclotide** for **constipation**. I have been feeling **stomach pain** since last 15 hours due to **linaclotide**"*. Here, we are interested in identifying the drug mention (linaclotide) and the phrases associated with the drug, such as an adverse drug event (stomach pain) and a treatment event (constipation). Hence, ADE extraction is an IE task with the set of entities {*Drug, Event Phrase*} and the set of relations {*has_ade(Drug, Event Phrase), treats(Drug, Event Phrase)*}. In the preceding example we have three entities, namely, {*linaclotide, stomach pain, constipation*} and two relationships namely, *has_ade(linaclotide, stomach pain)* and *treats(linaclotide, constipation)*. It is essential to model *treats* relationship for the ADE extraction task explicitly, as one of the major challenges faced by these systems is the mislabelling of an adverse

[4] https://www.medicinenet.com

event phrase with a treatment event phrase or vice-versa. The ADE extraction system must distinguish an adverse event from a treatment event, since both of them belong to the same event space. For example, the drug *Abilify* is prescribed to treat *depression* (here *depression* is a treatment event), whereas anxiety medications like *Diazepam* may produce *depression* (here *depression* is an adverse event).

## 3 Evolution of IE Approaches

Sarawagi et al. (2008) has traced the crucial advances in paradigms promoted by the community to solve IE tasks. This section discusses the major breakthroughs across that timeline.

Traditionally, IE models were designed using a set of logical rules capturing the domain constraints. These rules were hand-crafted by subject matter experts. Since noise is an inherent component of natural language corpora, the Prominent challenge faced by such models is that they perform poorly against such noisy datasets.

IE approaches saw another leap with the success of the classical machine learning models such as logistic regression and support vector machines. These models were used to learn the rules automatically from the data as opposed to hand-crafted rules. These models were still unsuitable for noisy datasets but were able to generalize more as compared to the hand-crafted rules.

Next improvement phase came along with an advent of graphical models. Generative models such as *Hidden Markov Model(HMM)* formed the basis of IE models. Ingrained problem with HMM is that the model assumes independence among all features; which is, of course, need not be the case in natural language. Towards the end of this phase, the community started working on conditional models such as Maximum Entropy Markov Models (MEMM) introduced by McCallum et al. (2000) for IE and segmentation.

To overcome feature independence assumption in HMMs, conditional formulation termed as Conditional Random Fields (CRFs) were proposed by Lafferty et al. (2001). The advantage of using CRFs is that their capability to model task in terms of features and corresponding weights rather than an application of Bayes theorem to model joint distribution. This makes CRFs get rid of features independence assumption. This era too was marked by the use of hand-crafted features by domain experts and thereby, the performance of a model was limited by the capability of the underlying features.

The recent phase of IE took a major leap with the introduction of deep learning-based architectural building blocks such as recurrent networks, convolutional networks, language token representations, and attention mechanisms (Bahdanau et al., 2014). This phase is getting supplemented by a massive amount of natural language data available digitally. All these blocks align themselves harmoniously with the language models.

## 4 Auxiliary Techniques and Tools for IE

This section presents various helper techniques and tools used widely in conjunction with IE models to enhance the overall performance of the task under consideration.

### 4.1 Embedding: Language token representations

Ultimately, all tokens from a natural language sentence must be represented in the form of numeric or decimal values so that an underlying model can learn the domain. One obvious solution is to make use of hand-crafted features capturing the context e.g. tokens separating the entities in combination with the orthographic word features (*allCap?, allSmall?, numeric?, etc*, and dictionary matching such as checking if a token is in a dictionary of company names). The intrinsic problem with such approach is that the performance of a model is limited by the potential of hand-crafted features, mandating a deep understanding of underlying linguistics and the domain. Another problem in case of hand-crafted features is that noise inherent in features obtained from NLP stages prior to entity and relation extraction is propagated down the pipeline, hitting the performance of a final extraction task. To take on this problem, a clever solution is to use embeddings.

#### 4.1.1 Word2Vec Model

In two consecutive papers published namely, Mikolov et al. (2013a) and Mikolov et al. (2013b), a capability of embeddings to represent words of a language as continuous vectors of much lesser dimension than that of textual corpus (usually in millions) have been showcased in word analogy tasks. This embedding is termed as *Word2Vec* model. Two complimentary thoughts were introduced to learn such embeddings. First paradigm is known as *CBOW (Continuous Bag of Words model*, whose task is to predict a representation for the

current word given the representation for words in its context. The second one is termed as *Skip-gram model* that focuses on learning representations for words in the context that are similar to the representation of current word.

Motivation behind *Word2Vec* is to capture both syntactic and semantic relationships among words through a representation that can encode words with similar meanings closely. Patterns learned using *Skip-gram* model can be represented using algebraic transformations such as addition or subtraction. E.g. Word representations for *India* and *Delhi* should exhibit similar transformation as that of *Germany* and *Berlin*.

Training objective is to learn a model to encode words that maximizes the log-likelihood of predicting nearby words $w_c$ correctly given a current word $w$(refer equation 2). This is a very powerful technique to capture idiomatic phrases by representing the entire phrase as a single vector e.g. "new" "york" as "New York". Learned vectors showcase an ability of mathematical composability i.e. vector(India) + vector(capital) $\approx vector(Delhi)$.

$$p(w_c|w) = \frac{\exp(v_{wc}^{\tau} v_w)}{\sum_{w=1}^{W} \exp(v_{wc}^{\tau} v_w)} \tag{1}$$

$$\frac{1}{T}\sum_{t=1}^{T} \sum_{-c \leq k \leq c, k \neq 0} \log p(w_{t+k}|w_t) \tag{2}$$

The denominator in equation: 1 is highly inefficient to compute since the summation runs over entire corpus (usually millions of words). Mikolov et al. (2013b) proposes two solutions to compute the normalizer efficiently viz, *Hierarchical Softmax* and *Negative Sampling*.

Though *Word2Vec* model performed well on word analogy tasks, it has now way to make use of global statistics (word co-occurrence count) available from the corpus. It simply slides a window across all words considering their immediate local context alone. Embeddings learned from this model thus are insufficient to model the similarities between words that are far apart from each other.

#### 4.1.2 GloVe

Pennington et al. (2014) proposed *GloVe (Global Vectors)* as an approach to represent words in a language in continuous vector space that overcame inherent problem of *Skip-gram* model which could not take global context into consideration. The aim of this model is to learn word representations which capture both local as well as global characteristics of the corpus.

To achieve this goal, authors proposed a combination of global matrix factorization with the local context. The solution makes use of a weighted least squares model where training is performed on word-to-word co-occurrence counts.

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^{\tau} w_j' + b_i + b_j' - \log X_{ij})^2 \tag{3}$$

*GloVe* achieved a significant improvement over *Word2Vec* model. Despite these improvements, the problem with GloVe embeddings was it is a static embedding. This means that once trained over a corpus, a word "bank" will always be represented with same vector irrespective of the context in which it is appearing. E.g. *I enjoy doing yoga near the river* ***bank*** vs *I visit* ***bank*** *to deposit money every week.* Performance of *GloVe* is further hit due to inability to represent unseen words effectively.

#### 4.1.3 ELMo

In order to improve the performance of any natural language model, the representations for words in a language must be of remarkable quality. The notion of high-quality compel representations to learn syntactic and semantic associations along with an ability to handle underlying polysemy. Peters et al. (2018) proposes a solution in the form of deep contextualized embeddings that overcome both difficulties. Word representations are mappings learned from internal states in a bidirectional language model subjected to train on massive text corpus. Hence these representations were termed as Embeddings from Language Models (ELMo). These representations are indeed deep in nature owing to the fact that they are represented using a linear function of all LSTM layers in an already disciplined language model. As against *GloVe*, these representations are immune to words that are not present in the training corpus by learning these representations solely at the character level. This helps model learn morphological properties of words and thereby produce high-quality representations for unseen words.

The problem of static word embeddings is solved by considering the context of an entire sentence before accrediting representation to a word[5]. These context-rich embeddings are the aggregations of hidden states from LSTM stacks. Prime disad-

vantage of this model is that the context vector is formed simply by aggregating right-to-left and left-to-right context state. It does not have the ability to consider both contexts simultaneously.

#### 4.1.4 BERT

*Transformer* introduced by Vaswani et al. (2017) created a remarkable impact in deep leaning due to its ability to capture long range dependencies in a sentence that surpassed LSTMs. At the heart of *Bidirectional Encoder Representations from Transformers (BERT)* (Devlin et al., 2018) lies the idea from *ELMo* and *transformers*. *BERT* combines bidirectional dynamic context from *ELMo* with attention-based long range context capturing capability of *transformers*. Unidirectional limitation is overcome with *Masked Language Model (MLM)*. Fundamental architectural unit in *BERT* is stacked bi-directional layers of transformer encoders. These representations showcases the ability to apply pre-trained representation to a vast number of tasks in NLP pipeline by adding just a single task-specific layer on top of it.

### 4.2 Dependency Parsing

A syntactic structure of sentences in a natural language is defined in terms of its constituent words. Dependencies among these words are represented using labeled directed edges. These labels are derived from a fixed set of grammatical relation types and hence are known as typed dependency structures. The head of an entire sentence is marked with a *root* label. The *root* often approximates the semantic association between its arguments(Jurafsky, 2018). E.g., Dextromethorphan is used as a cough suppressant (refer figure 1.

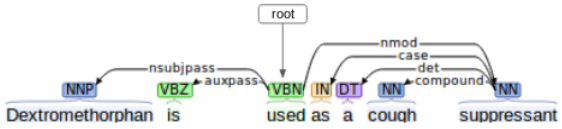


Figure 1: Root label in dependency parse approximating a relationship.

### 4.3 Co-reference resolution

Coreference resolution refers to a task of recognizing all tokens in a text that refers to the same entity[6], often termed as a *canonical entity*. Consider a text extract: ***Metformin*** *is used to treat diabetes.* ***It*** *has been associated with an adverse effect of stomach pain in some cases lately*. In this example, both highlighted expressions refer to the same canonical entity i.e. the drug *Metformin*. Discourse analysis is an integral part of every highly performing IE model which involves considering dynamics of associations among a set of sentences. To support this capability, it is crucial to parse input sentences prior to any downstream task (e.g. relationship extraction) and establish links among all the tokens that refer to the same standard entity. Hence coreference resolution is an auxiliary technique that helps downstream tasks in achieving better performance.

Commonly encountered types of coreference resolution involve *anaphora resolution* and *cataphora resolution*[7].

- Anaphora: When a proform (e.g. pronoun, proverb, etc) occurs after the referred entity. E.g. ***Rimonabant****, an anti-obesity drug was stopped.* ***It*** *has resulted in severe depression among the patients*.

- Cataphora: When a proform occurs prior to the referred entity. E.g. *Although* ***it*** *treats fever, the* ***Paracetamol*** *tablets may result in nausea.*

This report presents two ideas from the literature that are inline with entity and relation extraction task.

#### 4.3.1 Coreference resolution among extracted entities

Gupta et al. (2018) aims at grouping entity mentions in legal judgment documents that refer to the same entity. The proposed solution was based on supervised machine learning approach. For this problem, authors have restricted participants to *Person, Location, or Organization*. Basic entity mentions such as sequence of proper nouns (Mr. Bob), a pronoun (he, she, his, her) or a generic noun phrase (the Judgment) were identified as a target extraction type. Mentions were categorized into two viz, *dependent* and *independent*. If head of a mention is again a participant, it is termed as a dependent mention otherwise it is an independent mention. E.g. Refer figure: 2 for *Mr. Bob, resident*

[5] http://jalammar.github.io/illustrated-bert/

[6] https://nlp.stanford.edu/projects/coref.shtml

[7] https://en.wikipedia.org/wiki/Coreference

*of Mumbai was accused by the firm. Mr. Bob* and *the firm* are independent mentions while *resident* and *Mumbai* are dependent mentions.

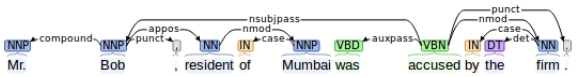


Figure 2: Dependent and independent mention types.

Independent mention can be a basic mention (Lawyer) or can be a composite mention which has dependents such as Mr. Bob. In case of a composite mention, all its dependents are merged recursively. Mention identification was formulated as a sequence labeling task using conditional random fields. The binary classifier was trained to detect candidate coreferences within a contextual window of at maximum 5 sentences. The model used features extracted from both constituency parse and the dependency parse tree. Authors have followed a transfer learning technique due to a scarcity of labeled data in the legal domain. They trained their model over *ACE-2005* dataset and the testing was performed over legal judgment documents. To tackle the problem of transitivity among coreferences, coreference groups were created. Their model was able to achieve the F1 score of 70.20 over prior baseline of 46.5. This result suggests that the performance of a coreference resolution task can be improved for an intended domain by imposing domain characteristics onto the model. E.g. restricting participant mentions to person, location and organization alone.

#### 4.3.2 Marking Participant Alias Links

Patil et al. (2018) identify all alias mentions of participants in a narration. Authors employed linguistic knowledge encoding through Markov Logic Network (MLN) approach to solve the problem. Similar to Gupta et al. (2018), participants were restricted to *person, location or an organization*. Mentions can be dependent of independent (further basic or composite). Composite mentions were formed by recursively merging all dependents of an independent mention (e.g. colleagues of Shiva). In order to make the learned model generalize well, mentions identification was done using WordNet hypernym relation.

First order logic encodes the domain knowledge such as linguistic grammar rules and an MLN based inference helps in linking alias of given mention. Refer figure: 3 for sample input and output format of the proposed system. All input sentences were encoded in Unified Linguistic Denotation Graph (ULDG).

Authors propose a three-phase algorithm to perform alias linking. These are *participant identification, alias identification using MLN inference and composite mention creation*.

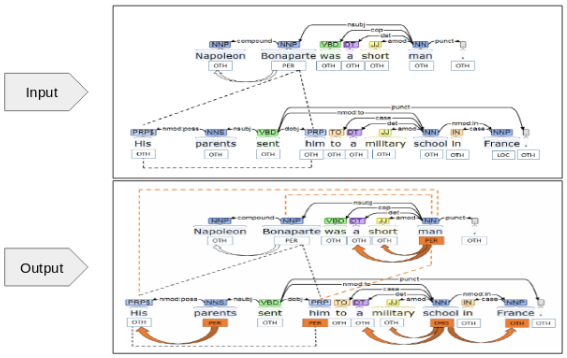


Figure 3: Sample input and output of alias linking. Image source: Patil et al. (2018)

Results from this model beat prior baseline, implying that the use of hypernym relation from WordNet assists in detecting generic noun phrases with improved recall. Overall quality of output results is improved through language knowledge encoded using MLN.

### 4.4 Attention Mechanism

Great success achieved by Bahdanau et al. (2014) in the domain of machine translation led to popularity of attention mechanisms. Attention helps one network focus on different parts of given input with different extents Olah and Carter (2016). Figure: 4 depicts working of an attention technique.

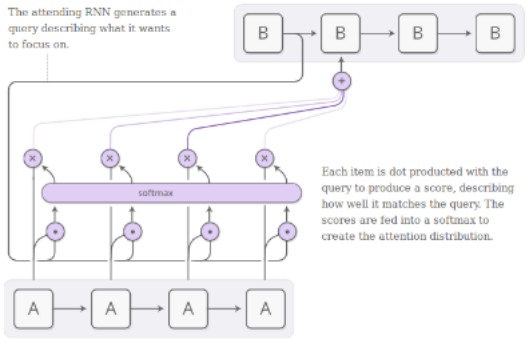


Figure 4: Attention Mechanism in action where network B is attending over network A. Image source: Olah and Carter (2016)

Attention relieves an encoder network from squeezing an entire sentence semantics into a single

vector which forms a bottleneck for a corresponding decoder network. Using attention, an encoder can now send information about each word it sees and meanwhile a decoder network can attend over the relevant words using the unit's hidden state as a query. The idea of attention between two networks can be further extended in which both the attending and the attended network refers to the same network, knows as a *self-attention*. Self-attention is one of the prime components of deep learning based relationship extraction architectures Lin et al. (2016).

Attention mechanism can be a very useful technique to represent an entire paragraph collectively as a single vector, formed by self-attending over a set of sentences in it. Such paragraphs and documents are the usual contenders for input to natural language based systems.

### 4.5 Addressing Labeled Data scarcity

Despite the architectural breakthroughs in models dealing with natural languages, the overall performance is often limited by the availability of robust labeled training data. It is labour intensive to create a good quality text corpus for natural language downstream tasks, as the textual data demands deep linguistic understanding and a scale. Furthermore, with data explosion, it becomes increasingly difficult to manually annotate such massive data. A pharmacovigilance system should capture multi-token entities such as adverse event phrases (e.g., *stomach pain*). If the system is not trained on the domain specific data, it can not perform satisfactorily on such tasks. The pharmacovigilance domain poses a prime concern for data propriety and patient privacy. It is not easy to have direct access to patient adverse event reports submitted to the FDA Adverse Event Reporting System (FAERS), despite the availability of openFDA api[8]. This compels us to explore digital fora to construct a synthetic corpus. The motivation for adopting a synthetic corpus comes from the machine translation area. Lample et al. (2018) validates the usefulness of synthetic data by outperforming the existing state of the art on WMT'14 English-French and WMT'16 German-English benchmarks by an ample margin. This section describes two techniques from the literature adapted to confront the challenge of automatically building high quality labeled text corpus.

[8] https://open.fda.gov/

#### 4.5.1 Distant Supervision

Mintz et al. (2009) introduced the concept of distant supervision according to which any unlabelled textual data can be assigned with labels by superimposing such a raw data over existing knowledge graphs such as ontologies. E.g. If ontology contains two entities *Crocin* and *GSK* associated with a relation of $manufactured\_by$*(Crocin, GSK)*; raw text *Crocin is launched by a pharmaceutical company GSK* will be marked with a relation label $manufactured\_by$.

This means, all the instances from a raw text that mentions entities present in a knowledge graph are considered blindly as positive instances for the exhibited relationship among those entities, which need not always be the case (e.g. *Crocin from GSK has been found to cause muscle ache*. In this example, we are interested more in ADR relation rather than company-product relation at the first place.). Hence, this approach creates an additional challenge of wrong labels being assigned to the input data.

Despite the drawback of noise, distant supervision can be used to as a starting point where you can get labels marked to the raw data. This output can further be refined by SMEs (it is better to have some labels rather than raw data).

#### 4.5.2 Domain adaptation

Ganin and Lempitsky (2014) takes on a challenge of labeled data scarcity through *deep domain adaptation*. Aim is to improve the performance of a deep learning architecture in such a situation. *Domain adaptation* is the process of learning a discriminative classifier by subjecting it to a shift between distribution of training and testing samples. In this case, authors have used a huge amount of labeled source domain data and *unlabeled* target domain data, making it an unsupervised domain adaptation.

The approach is to use a combination of domain adaptation with deep feature learning that results in features that are *discriminative* across samples in the source domain and *invariant* to the domain shift. Within a domain, classifier's task is to minimize prediction loss which harnesses the capability of discriminativeness. Feature predictor used in the model aims at minimizing label prediction loss and maximizing domain classifier loss through a technique called *gradient reversal layer* which delivers domain invariance.

Source domain *S* is assumed to be shifted from the target domain *T* by some shift such as *(MNIST,*

*MNIST-M)*[9]. The architecture of the proposed model can be seen in figure: 5.

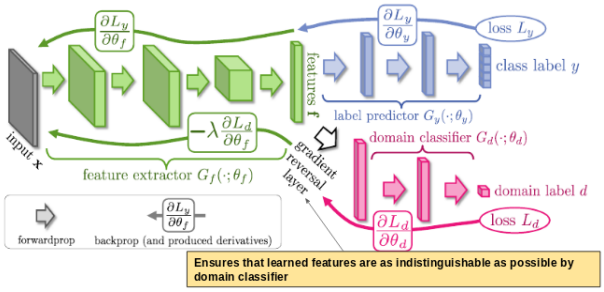


Figure 5: Domain adaptation through backpropagation and gradient reversal layer. Image source: Ganin and Lempitsky (2014)

The proposed model was able to achieve an F1 score of 0.8149 for MNIST, MNIST-M pair which was a prominent gain over baseline that was trained over source domain alone with no target domain data (0.5749).

### 4.6 Generalized Cross Entropy Loss for Noisy Labels

Deep Neural Networks are often vulnerable to overfitting the noisy data owing to their high capacity to memorize. Synthetically constructed corpora suffer from the noisy labels. This problem necessitates a noise-robust loss function. The categorical cross entropy (CCE) loss emphasizes the difficult samples during training. It is faster to converge but overfits the noise present in the dataset. On the contrary, Mean Absolute Error (MAE) gives equal importance to all the samples in the dataset. Hence, it is robust to the noise but takes more time to converge. The solution to this problem is to utilize the advantages of both CCE and MAE loss functions.

The solution, $L_q$ loss proposed by Zhang and Sabuncu (2018) is as follows:

$$L_q(f(x), e_j) = \frac{1 - f_j(x)^q}{q} \tag{4}$$

Where, $q \epsilon (0, 1)$ and $f_j \epsilon [0, 1]$.

It should be noted that the above loss function places relatively less emphasis on difficult samples than in CCE and relatively more attention to difficult ones when compared with the MAE.

### 4.7 BioBERT

BioBERT (Lee et al., 2020) is a BERT-based (Devlin et al., 2018) model trained on $\tilde{1}$ million PubMed abstracts. The model claims the F1 score of 0.88 on the NCBI disease dataset. The intuition behind the BioBERT was that the same token could get a different label under different contexts, and thereby contextual embedding should help capture such differences. Consider an example, *Ibuprofen is prescribed to treat the headache [PE]* and *Naproxen has caused headache [AE] in certain cases*. Here, the same token *headache* is labeled as either an adverse event or a treatment event based on the context.

## 5 Existing approaches to IE in Pharmacovigilance

This section presents notable approaches used in Pharmacovigilance to extract ADEs. These approaches make use of different forms of input text corpus, that can be classified into two categories: *Clinical Health Records* and *Social and Digital Media*. Clinical health records in electronic form represent a set of documents from spontaneous adverse event reporting systems (SAERs) whereas blogposts, news articles, medical articles in literature, posts from social media like twitter constitute Social and Digital media corpus. Following sections present an overview of approaches to ADR extraction from both categories.

### 5.1 Extraction of Adverse Drug Effects from Clinical Health Records

Aramaki et al. (2010) aims at estimating the amount of adverse event information contained in medical records through a manual approach followed by designing a system that can automatically extract AE related information from these records. Authors investigate the overall accuracy of the automatic approach over a massive amount of patient discharge summaries.

Corpus was comprised of 3012 records representing one month worth of patient discharge summaries from a Hospital. The task of term identification (drug and the symptom expression) was performed using CRF-NER tagger and the task of association identification was performed using two methods: *pattern based* and *machine learning based* relation identification.

- Pattern-based relation identification: This is a heuristic rule based approach with set of keywords identifying an association restricted to: {*stopped, cause, side effect, adverse effect, changed*}. Regular expression based pat-

[9] http://yann.lecun.com/exdb/mnist/

terns with all possible combinations of *drug, keyword and symptom* were constructed using wild cards to extract adverse relationships. E.g. *drug * keyword * symptom* → *paracetamol seemed to cause a dizziness*.

This approach has a serious problem: does the extracted term pair drug:symptom pair really have an adverse effect associated with them? E.g. *Dextromethorphan : cough-expectorant* vs *Dextromethorphan : vomiting*

- Machine learning based relation extraction: Authors employed a classical machine learning model known as Support Vector Machine (SVM). Dataset was labeled using term tag (drug expression and symptom expression) and relation tag (adverse effect: drug → symptom). From labeled corpus, feature set {*drug term, symptom term, surface token chain, distance*} was constructed for training SVM model.

Based on author's findings, 7.7% records had AE information. 59% of these records with AE were extracted using their proposed model. Based on the outcome, authors claimed that relation extraction task is tougher than that of entity extraction. Pattern based model (F-score: 0.65) outperformed machine learning based model (F-score: 0.59). One possible reason for this could be scarcity of positively labeled data. To improve performance of a machine learning based classifier, more positive instances should be incorporated or techniques that work well with small corpus can be used. This approach did not give any consideration to a prominent aspect of standardizing terms (concepts with the same intention but different forms). E.g. *no sleep, sleeplessness* means *insomnia*.

### 5.2 ADE Extraction using Natural Language Processing from Electronic Health Records (EHR)

Friedman (2009) takes on the problem of automatic detection of ADEs from narrative EHRs instantly using a statistical approach. Authors make use of *Chi-Square* statistic to extract association between disease-symptom, disease-drug leading to the formation of the drug-disease knowledge base.

Medical narratives were parsed using *MedLEE NLP*[10]toolkit to map general terms to corresponding UMLS standardized concepts. Reports in the form of templatized lab test results were standardized to UMLS concepts. Selection layer was used to take care of medication, disease, pathological conditions and filtering to eliminate abnormal test ranges and wrongly ordered events. Frequencies of co-occurring pairs were computed to identify associations using statistical methods. Final filtering layer was used to eliminate samples using medical domain knowledge (e.g. patients with family history for a disease).

Major contribution by the authors was the creation of disease-symptom knowledge base. Additionally, authors were able to automatically extract AEs, thereby leading to improved patient safety at low cost. Despite an attempt to improve quality of ADE extraction, this approach has a limitation of not capturing an indirect association between multiple drug mentions. Dataset was cherry-picked covering only a limited set of medications. Improvement to their approach could be to use publicly available datasets that cover majority of the medications.

### 5.3 Pharmaceutical Product Monitoring in Twitter

Freifeld et al. (2014) tries to evaluate the degree of consistency between Tweeter AE mentions and spontaneous reports submitted to regulatory authorities. The aim is to evaluate the potential of publicly available user generated clinical data for AE extraction. Authors collected 6 months Twitter posts having resemblance to AEs using twitter API. FAERs data was collected using OpenFDA[11] API which contains AE reports over same period. Natural language terms from the data were mapped to corresponding nearest standard concepts using medical knowledge bases such as MedDRA.

Tree-based dictionary matching algorithm (refer figure: 6) as proposed by Freifeld et al. (2008) was utilized to identify product-symptom AE like mentions in twitter posts.

The frequency of product-event AE pairs was compared with FAERS at System Organ Class (SOC) level to generalize well, since matching directly at term level would not yield good results. Authors claim that a Spearman correlation coefficient of 0.75 was observed between twitter AE mentions and corresponding reports submitted to

[10] http://www.medlingmap.org/taxonomy/term/80

[11] https://open.fda.gov/

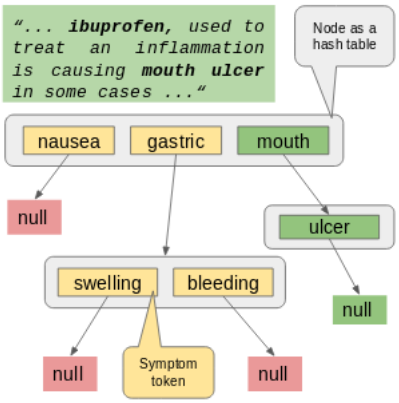


Figure 6: Working of a tree based dictionary matching algorithm.

FAERS for a specified duration at SOC level. Reportedly there were 3 times more AE like mentions in posts thatn FAERS implying that FAERS reports are usually delayed which hampers immediate actions in such cases. Another contribution was to come up with a dictionary that maps internet slang/vernaculars to standardized medical terms cush as MedDRA terms. Hence automatic AE extraction system over social media posts can be seen as a potential hypothesis forming system (since data is cherry-picked i.e. biased towards positive AE mentions and do not represent the real world picture and uncertainties). Performance of this system can be improved further if time series modelling is incorporated into decision making a task which will help in reducing false positives. Dictionary lookup based systems usually fail to generalize as concepts in a dictionary are very rigid and cannot capture variations in morphology and semantics of input tokens. Performance of such a system is usually limited by the quality and robustness ob an underlying dictionary.

### 5.4 Multi-Task Pharmacovigilance

Chowdhury et al. (2018) proposed a neural network based architecture in which multiple tasks from Pharmacovigilance were learned simultaneously to improve the overall accuracy of individual tasks. The aim was to learn a shared representation of these tasks to reduce the chances of mislabelling an indication as an ADR. Authors accomplished above task using a common encoder that was shared among *ADR classification, ADR labeling and indication labeling* along with a separate task-specific decoder for each task. Assumption behind this decision was that the system will eventually be able to capture intricacies involved in all tasks, leading to a distinguishing an ADR from an indication. Authors employed coverage based attention mechanism to detect phrasal mentions. Refer figure: 7 for proposed architecture.

First task is termed as an ADR classification that is a binary classifier which aims at identifying whether a post has AE mention in it or not. Second task is ADR labeling that takes care of finding the most likely sequence of ADR tags given an input post. Third task is indication labeling that aims at producing a most likely sequence of indication labels given a input post. Using simultaneous training, the proposed system was indeed able to achieve a better performance at each of the three individual tasks as each model now has information about others' knowledge about the input post via shared encoder states. Using coverage based attention mechanism helped system capture phrasal AE mentions that span across more than one token (e.g. pain in the stomach).

### 5.5 Identifying Individual Case Safety Reports (ICSRs) from Social and Digital Media

Comfort et al. (2018) aims at detecting AE mentions from social media posts, blogs, literature and investigate if it can be a potential ICSR. Authors harnessed machine learning models to generalize well to casual language text from such source. A narrative is termed as a valid ICSR if it has an identifiable subject (patient), correspondent, drug under suspicion and an AE faced. Authors try to iteratively improve rule based and dictionary matching based system using machine learning model.

Dataset comprised of 300k posts in English language scraped from web. Authors utilized semi-automated filtering and sanitizing to standardize internet vernaculars to nearest MedDRA terms. Idea was to use 4+1 classifiers that correspond to prediction of $\{$*drug label, AE, subject, relation*$\}$ + $\{$*ICSR validator*$\}$. Reporter of the case was assumed as the subject. First iteration was a rule based system in which each classifier was defined using a set of hand-crafted rules built by SMEs and dictionary matching (MedDRA, pronouns, drug generic and brand names, etc). In second iterations, all components except AE classifier were carried forward with addition of ML based AE classifier. In final iteration, ICSR classifier from iteration 2 was replaced by a SVM based ICSR classifier. Set of

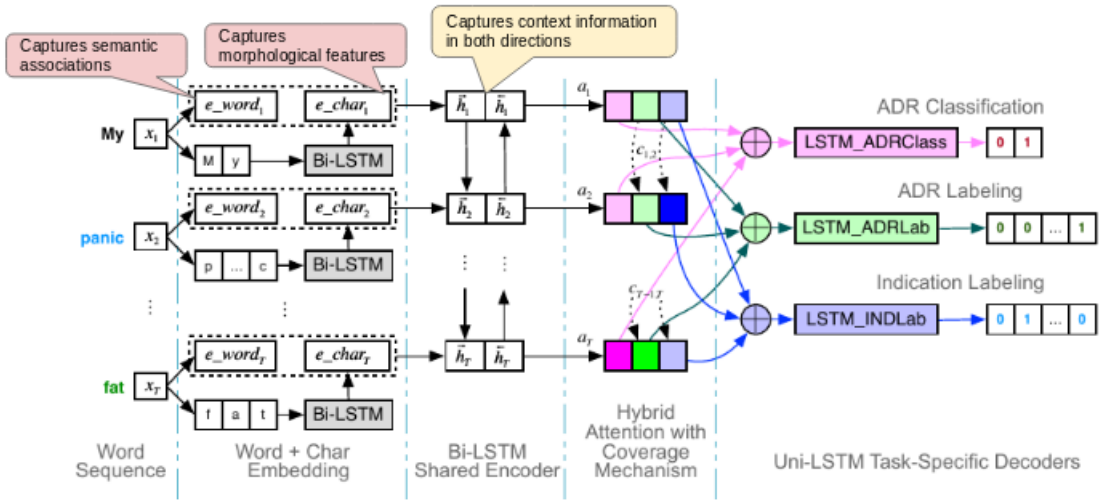


Figure 7: Multitask Pharmacovigilance neural architecture. Image source: Chowdhury et al. (2018)

features utilized for training these ML based models include surface tokens, orthographic features, embeddings and semantic features such as word contexts, features from "gold" reference texts. This iterative approach helped authors to understand the intricacies involve in automatic AE extraction.

Results achieved by model from third iteration (AUC of 0.85) highlights the huge potential of data from social media and blog posts in AE detection task. System serves as a potential hypothesis generation system which gets further verified by SMEs (rather than SMEs starting the whole task from scratch).Authors have claimed that the system could run over the identified dataset in 48h as against estimated 44000h if SMEs were to do this task from scratch. In spite of these achievements, there are few inherent limitations in the system such as the system does not distinguish between cases where there is a drug overdose, a wrong medication and a true AE. Extraction capability is limited to a single sentence alone. There is no consideration to cases involving past history.

## 6 Entity Extraction Approaches

Entity extraction refers to identifying entity mentions and their aspects/characteristics from textual data. Mention may contain a single word (e.g., Crocin) or span across more than two tokens (e.g., skin reaction). **Rule-Based Entity Extraction** (Sarawagi et al., 2008) makes use of rules to identify candidate entity mentions and can be hard-coded or learned from the data. **Statistical Methods for Entity Extraction** (Sarawagi et al., 2008) converts an entity extraction task to an unstructured text decomposition task where the decomposed parts are labeled either jointly or independently. In **Classical Machine Learning-Based Entity Extraction**, Takeuchi and Collier (2005) proposes SVM based named entity tagger for extracting scientific terminologies in the domain of medicines. **Deep Learning-Based Entity Extraction** (Lample et al., 2016) aims at the problem of labeled data scarcity in natural languages and the problem of hand-crafted features that fail to generalize well. The resulting model makes use of Bi-directional LSTM layers followed by a single CRF (Lafferty et al., 2001) layer. Newer models such as *BERT* by (Devlin et al., 2018) achieves the F1 score of 92.80 on the same task.

## 7 Relation Extraction Approaches

Relation extraction refers to a problem of identifying associations between two or more entities from an input source (Sarawagi et al., 2008). **Feature Based Approaches to Relation Extraction** (Sarawagi et al., 2008) use features derived from token in a sentence. The extracted features are passed through machine learning techniques such as logistic regression, SVM, or a decision tree. **Kernel Based approach to Relation Extraction** (Sarawagi et al., 2008) makes use of *kernel functions* such as $K\langle X, X'\rangle$ to capture similarity between two non-flat structures (e.g., graphs) X, X'. **Classical Machine Learning Approach to Relation Extraction** Rink and Harabagiu (2010) propose a semantic relation extraction approach that identifyes appropriate semantic relation and its direction from a set of given relations between la-

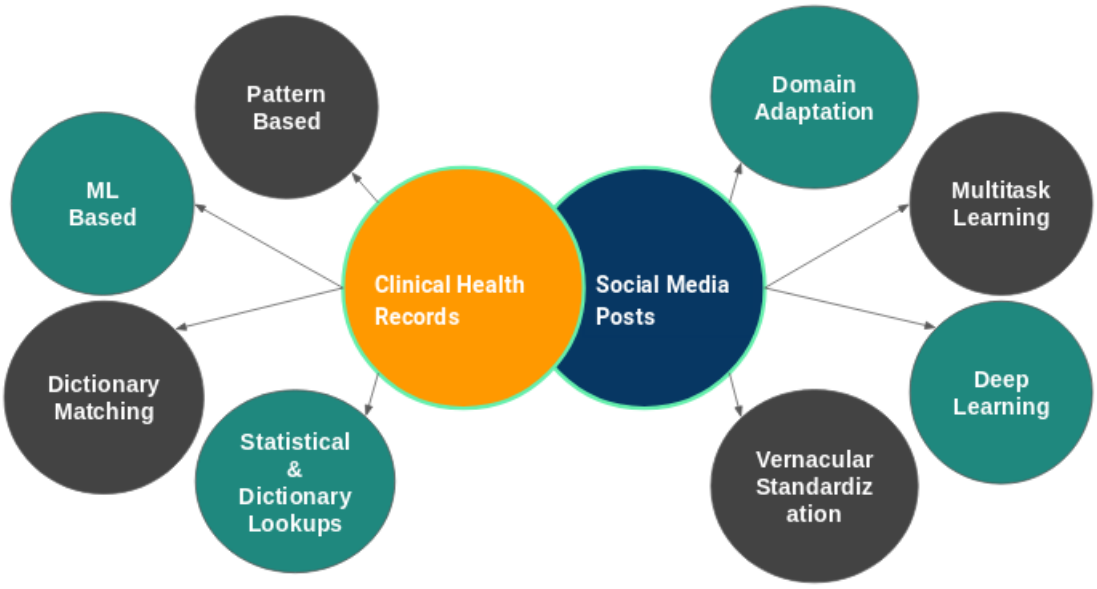


Figure 8: Summary of Existing approaches to Pharmacovigilance IE.

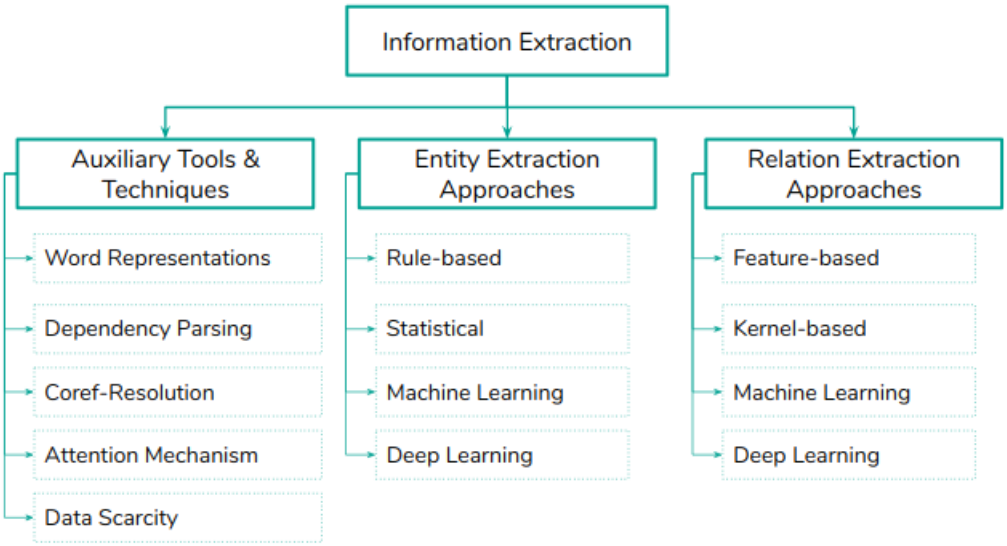


Figure 9: Summary of IE literature survey.

beled entities in a text. **Deep Learning-Based Approaches to Relation Extraction** involve **Attention Based Convolutional Neural Network** (Zeng et al., 2015), a modification to max-pooling operation in conventional Convolutional Neural Network (CNN). It retains semantic information into a corresponding representation called *Piecewise Convolutional Neural Network (PCNN)*. It suffered from noisy labels problem. To tackle these challenges, Lin et al. (2016) proposed a modification to Zeng et al. (2015) approach. The authors introduce an attention-based PCCN relation extractor. **Bidirectional Recurrent Convolutional Neural Network** (Cai et al., 2016) aim at exploiting semantic information present in a dependency graph for relation extraction. Authors have proposed an architecture containing a combination of a convolutional and a recurrent network. **End-to-End Relation Extraction Using Neural and Markov Logic Networks**(Pawar et al., 2017) aim end-to-end relation extraction problem, which involves improving the performance of complete task over ACE-2004[12]dataset.

Authors propose *All Words Pair Neural Network (AWP-NN)* architecture to perform end-to-end relation extraction.

[12]https://catalog.ldc.upenn.edu/LDC2005T09

## 8 Conclusion

In this paper we have presented the various approaches to Pharmacovigilance IE that use either clinical health records or social media text as input source and are summarized in figure 8. The paper also presents auxiliary tools and techniques that work in conjunction with the IE systems. We present a formal definition for Pharmacovigilance use case to extract adverse drug events. The effort also describe various approaches to entity and relation extraction as summarized in figure 9. The important observation from this study was the difficulty involved in gathering an annotated corpus and noise induction through the synthetic corpus creation techniques.

## References


Syed Rizwanuddin Ahmad. 2003. Adverse drug event monitoring at the food and drug administration: your report can make a difference. *Journal of general internal medicine*, 18(1):57–60.

Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Masuichi, Kayo Waki, and Kazuhiko Ohe. 2010. Extraction of adverse drug effects from clinical records. *MedInfo*, 160:739–743.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional recurrent convolutional neural network for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 756–765.

Shaika Chowdhury, Chenwei Zhang, and Philip S Yu. 2018. Multi-task pharmacovigilance mining from social media posts. *arXiv preprint arXiv:1801.06294*.

Shaun Comfort, Sujan Perera, Zoe Hudson, Darren Dorrell, Shawman Meireis, Meenakshi Nagarajan, Cartic Ramakrishnan, and Jennifer Fine. 2018. Sorting through the safety data haystack: using machine learning to identify individual case safety reports in social-digital media. *Drug safety*, 41(6):579–590.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Clark C Freifeld, John S Brownstein, Christopher M Menone, Wenjie Bao, Ross Filice, Taha Kass-Hout, and Nabarun Dasgupta. 2014. Digital drug safety surveillance: monitoring pharmaceutical products in twitter. *Drug safety*, 37(5):343–350.

Clark C Freifeld, Kenneth D Mandl, Ben Y Reis, and John S Brownstein. 2008. Healthmap: global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Informatics Association*, 15(2):150–157.

Carol Friedman. 2009. Discovering novel adverse drug events using natural language processing and mining of the electronic health record. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 1–5. Springer.

Yaroslav Ganin and Victor Lempitsky. 2014. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*.

Ajay Gupta, Devendra Verma, Sachin Pawar, Sangameshwar Patil, Swapnil Hingmire, Girish K Palshikar, and Pushpak Bhattacharyya. 2018. Identifying participant mentions and resolving their coreferences in legal court judgements. In *International Conference on Text, Speech, and Dialogue*, pages 153–162. Springer.

Martin Jurafsky. 2018. *Speech & language processing*. draft.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2124–2133.

Andrew McCallum, Dayne Freitag, and Fernando CN Pereira. 2000. Maximum entropy markov models for information extraction and segmentation. In *Icml*, volume 17, pages 591–598.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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

Chris Olah and Shan Carter. 2016. Attention and augmented recurrent neural networks. *Distill*.

Sangameshwar Patil, Sachin Pawar, Swapnil Hingmire, Girish Palshikar, Vasudeva Varma, and Pushpak Bhattacharyya. 2018. Identification of alias links among participants in narratives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 63–68.

Sachin Pawar, Pushpak Bhattacharyya, and Girish Palshikar. 2017. End-to-end relation extraction using neural networks and markov logic networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 818–827.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Bryan Rink and Sanda Harabagiu. 2010. Utd: Classifying semantic relations by combining lexical and semantic resources. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 256–259. Association for Computational Linguistics.

Sunita Sarawagi et al. 2008. Information extraction. *Foundations and Trends® in Databases*, 1(3):261–377.

Koichi Takeuchi and Nigel Collier. 2005. Bio-medical entity extraction using support vector machines. *Artificial Intelligence in Medicine*, 33(2):125–137.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762.

Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in neural information processing systems*, pages 8778–8788.