‘ChemXtract’ A System for Extraction of Chemical Events from Patent Documents

Pattabhi RK Rao and Sobha Lalitha Devi
AU-KBC Research Centre,
MIT Campus of Anna University, Chennai, India
sobha@au-kbc.org

Abstract

ChemXtraxt main goal is to extract the chemical events from patent documents. Event extraction requires that we first identify the names of chemical compounds involved in the events. Thus, in this work two extractions are done and they are (a) names of chemical compounds and (b) event that identify the specific involvement of the chemical compounds in a chemical reaction. Extraction of essential elements of a chemical reaction, generally known as Named Entity Recognition (NER), extracts the compounds, condition and yields, their specific role in reaction and assigns a label according to the role it plays within a chemical reaction. Whereas event extraction identifies the chemical event relations between the chemical compounds identified. Here in this work we have used Neural Conditional Random Fields (NCRF), which combines the power of artificial neural network (ANN) and CRFs. Different levels of features that include linguistic, orthographical and lexical clues are used. The results obtained are encouraging.

1 Introduction

Chemical information extraction is a challenging task. Unstructured data in the biomedical domain contain descriptions of chemical entities and the extracting these entities from textual data repositories, in particular from the patents, is becoming increasingly important for researchers and for the industry. Human annotation of patents to generate annotated corpus and populate chemical databases is a tedious task and this can be made easy and fast through the use of automated language processing. The process of automatically extracting the mentions of a particular semantic type in text is known as Information Extraction (IE). IE includes the extraction of names of chemical compounds and assigns a label according to the role it plays within the chemical reaction, popularly known as named entity recognition (NER) and also event relation extraction, where it extracts the chemical event relation that takes place between the chemical compounds. ChemXtract extracts the chemical compound names and its event relation in patent documents.

In this paper we discuss in detail the methods and techniques used in ChemXtract. The extraction identify and label chemical compounds and their specific types, i.e. to assign the label of a chemical compound according to the role which it plays within a chemical reaction, the temperature and reaction time at which the chemical reaction is carried out, the yields obtained for the final chemical product and the label of the reaction. The challenges in extracting the chemical compounds are many and it further increases when it is from patent documents. The language used in patents is very different from the language used in scientific literature. When writing scientific papers, authors strive to make their words as clear and straightforward as possible, whereas patent authors often seek to protect their knowledge from being fully disclosed [34]. Thus the main challenges for natural language processing (NLP) in patent documents arise from its writing style such as long and complex sentences and long list of chemical compounds. As the characteristics of sentences in patent documents bring in challenges in deep syntactic parsing, in this work we have used shallow parsing of the documents. The data used for this work is provided by CheMU, CLEF 2020 [32]. The features and factors used include linguistic, orthographical and lexical clues.

Further the paper is structured as follows, in section 2, a brief overview of the recent published work is given and section 3 details the features

Proceedings of Recent Advances in Natural Language Processing, pages 988–995
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_106

and the methods used in the development of the named Entity recognizer. The Section 4 describes the event extraction, and the evaluation and results are discussed in section 5. The paper ends with the conclusion

2 Literature Review

In recent years Deep Learning is flourishing as a well-known ML methodology for NLP applications. By using the multilayer neural architecture it can learn the hidden patterns from the enormous amount of data and handles the complex problems. In Chemical informatics which is a sub-field of BioNLP the use of Deep Learning for various application related to extraction of information is flourishing as seen in BioIE. Biomedical information extraction (BioIE) automatically extracts relevant structured semantics (e.g. entities, relations and events) from unstructured biomedical text data. BioIE covers a large spectrum of research efforts which includes the tasks such as named entity recognition [6–8], event identification [9–11], and relation extraction [7,12,13]. The domains include medical literature[14], biological literature[15], electronic health records[16], and chemical name extraction[8]. The methodology includes rule-based, knowledge-based, statistics based, learning-based methods and hybrid methods [17–18]. The extraction of information, which uses the natural language processing (NLP) techniques to extract relevant information to understand the underlying mechanisms of disease, is summarized in Gonzalez et al. [19].

Deep learning networks can be roughly categorized into (1) unsupervised/generative, e.g., restricted Boltzmann machines (RBMs)[23], deep belief networks (DBNs)[24]; (2) supervised/discriminative, e.g., deep neural networks (DNNs)[25], convolutional neural networks (CNNs)[26] and recurrent neural networks(RNNs)[27]; and (3) hybrid, e.g., DBNDNN[28] models that combine unsupervised pre-training and supervised fine-tuning.

The identification of chemical entities has to handle with naming variability between and within different chemical subdomains. A chemical entity can be written as a trademark name of a drug, as a short form (abbreviation or acronym), or it can be represented by following the standard naming nomenclature guidelines as provided by the IUPAC. The recent works in this field using deep learning is discussed here. The earlier work on neural network was done by Gallo et.al [1] to classify named entities in ungrammatical text. Their implementation of Multi-Layer Perceptron (MLP) is called as Sliding Window Neural (SwiN) which was specifically developed for grammatically problematic text where the linguistic features could fail. The Deep Neural Framework was developed by Yao et al.[2] to identify the biomedical named entities. They have trained the word representation model on PubMed database with the help of skip-gram model. Yang et al., built a single neural network for identifying multi-level nested entities and non-overlapping NEs. Kuru et al.,[3] used character level representation to identify named entities. They have utilized Bi-LSTMs to predict the tag distribution for each character. Wei et al.,[4] developed a CRF based neural network for identifying the disease names. Along with word embedding the system has also used words, POS information, chunk information and word shape features. Hong et al., [5] developed a deep learning architecture for BioNER which is called as DTranNER. It learns the label to label transition using the contextual information. In this the tag-wise labelling is handled by Unary-Network and the pair-wise network predicts the transition suitability between labels. The networks are then plugged into the CRF of the deep learning framework.

Learning methods used in BioIE falls into three categories: (1) learning from labeled data (i.e. supervised learning); (2) learning from unlabeled data (i.e. semi-supervised and unsupervised learning); (3) Hybrid approach where learning scheme integration to integrate different learning paradigms at outer system level. The approaches used in BioIE are Conditional random fields(CRF)[7] and support vector machines(SSVM)[20] which are supervised learning methods, and deep neural networks[21] which is unsupervised approach and these have been applied to both general domain IE and BioIE. A scalable and reliable approach on IE is the Open information extraction (OpenIE)[22] , which has emerged as a novel information extraction paradigm. OpenIE systems consist of four main components: (1) Automatic Labeling of data using heuristics or distant supervision; (2) Extractor Learning using relation-independent features on noisy self-labeled data; (3) Tuple

Extraction on a large amount of text by the Extractor; (4) Accuracy Assessing by assigning each tuple a probability or confidence score.

3 Extraction of Chemical Entity and its Event Relations

ChemXtract extracts chemical entities and its event relation. It has two components 1) Chemical name identification and 2) event relation Identification. The system follows a pipeline architecture, where the data is first pre-processed to the required format that is needed to train the system. After training the system the NEs are automatically identified from the test set. The overall system architecture is shown in Figure 1. The following section gives in detail the pre-processing required for both the tasks.

3.1 Pre-processing

The data, input to the system, is pre-processed for formatting, where we use a sentence splitter and tokenizer and also it is converted into column format. The formatted data is further annotated for syntactic information which includes the Part-of-speech (POS) and Phrase Chunk (Noun Phrase, Verb phrase) tagging. We have used fnTBL [30], an open source tool for the syntactic analysis of POS and Chunking.

3.2 Named Entity Detection

Identification of chemical compounds from text is a difficult task as it does not follow the common linguistic rules of the language. Hence rule based method do not give expected performance. In ChemXtract, we have used three learning algorithms, one from machine learning CRFs and two from deep Learning, RNN and ANN. The details on all the three algorithms, the feature selection for CRF and the factors incorporated into the layers in RNN and ANN are given in the following sections.

3.2.1 Neural Conditional Random Fields (NCRFs)

Conditional random fields (CRFs) are a probabilistic framework for labeling and segmenting sequential data, based on the conditional approach. Lafferty et al. [33] had first used CRFs for NLP applications. A CRF is a form of undirected graphical model or Markov random field, globally conditioned on X that defines a single log-linear distribution over label sequences given a particular observation sequence.

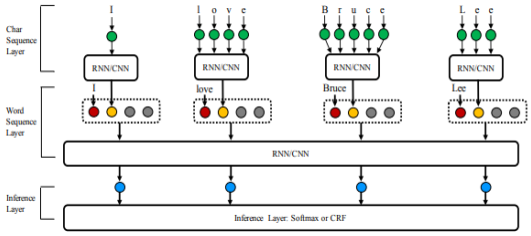

Fig. 1. NCRF architecture for an example sentence. Green, red, yellow and blue circles represent character embeddings, word embeddings, character sequence representations and word sequence representations, respectively. The grey circles represent the embeddings of sparse feature.

Neural CRFs (NCRFs) is designed with three layers: a character sequence layer; a word sequence layer and inference layer. For each input word sequence, words are represented with word embeddings. The character sequence layer can be used to automatically extract word level features by encoding the character sequence within the word. In this we can also incorporate hand crafted features such as capitalization, suffixes etc. Feature selection plays an important role in the performance of any machine learning system. Also, the features selected must be informative and relevant. We have used word, grammatical and functional level terms as features and they are detailed below:

Word level features: Word level features include Orthographical features and Morphological features.

a.Orthographical features contain capitalization, Greek words, combination of digits, symbols.

b.Prefix/suffix of chemical entities are considered as morphological features. Suffixes are the ending sub string of the words for example “acetate”, “mmol”, “dine” etc. Similarly Prefixes are the starting parts of the words (starting sub strings), for example “methyl”, “propyl”. The common sub string parts of the entities are identified which are considered as positive marker for identifying the chemical named entities.

Grammatical features: Grammatical features include words, POS, chunks and combination of words, POS and chunk.

Functional term feature: Functional term helps to identify the chemical named entities and

categorize them to various classes. Example: Alkyl, acid, alkanylene

The NCRF++ tool is used for implementation. It is an open source implementation of NCRFs [31] and is a general purpose tool. The features required for training have been explained above in this section. It learns the patterns of named entities from the tagged corpus and using the model generated using the training data the NEs in the test data can be automatically identified. All the features used are extracted from the training corpus provided by the ChEMU, CLEF Track 2020 and no other external resources have been used.

3.3 Event Extraction

The event and its arguments are extracted for identification of the reaction happening between the chemical compounds. In this work we identify the events and their arguments using NCRFs. The arguments of events are the chemical compounds and entities such as Temperature, Yield_Percent. The main challenges in the event argument extraction are i) Capturing the long range connection between the event trigger and event argument and ii) Identifying the correct role of the event argument with respect to the event type (or the event trigger), and the span of the argument.

Ex. Sentence1:

The crude product was purified by Biotage Isolera™ (3.22 g, 58%).

Ex NEAnnotation1:

The crude <Reaction_Product> product</Reaction_product> was <EventType:Reaction_Step> purified </Event> by Biotage Isolera™ (<Yield_Other>3.22 g</Yield_Other>, <Yield_Percent>58%</Yield_Percent>.

Ex. Event-Argument_Annotation1:

purified --- Arg1 --- product; purified --- ArgM --- 3.22 g; purified --- ArgM --- 58%

In the above example the event trigger is "purified", which is of event type "Reaction_Step". The event arguments for this event are "Reaction_Product", "Yield_Other" and "Yield_Percent".

As discussed earlier the patent document style of writing is a challenge and this is evident from example 2 given below. It is observed that one event trigger has "n" arguments and in the example n=8 i.e., has 8 arguments.

Ex. Sentence 2:

*A microwave vial was **<event>**charged**</event>** with 6-iodo-8-methyl-2-propyl-[1,2,4]triazolo[1,5-a]pyridine (Intermediate 66, 269 mg, 0.89 mmol), methyl 2,2-difluoro-2-(fluorosulfonyl)acetate (0.28 mL, 2.23 mmol), CuI (425 mg, 2.23 mmol), DMPU (0.61 mL, 5.06 mmol), and DMF (5.6 mL).*

In this sentence the event "charged" has one of the event arguments "DMF", which is at far end of the sentence.

The features of POS and Named Entities are used for the identification of Events. The NEs identified in the previous step form the arguments of the event. The motivation behind using the word, POS and NE tags is that it can detect the structures in the input and automatically obtain better feature vectors for classification. Most of the earlier NLP works have used words as input for training.

The POS and NE tags help to add sense and semantic information to the learning. The NE tag will help in identifying whether they are attributes of objects, phenomenon's, events etc. This gives indications on the chemical compounds while learning and thus help in the identification of the chemical events. We have modelled NCRF as pairs of 3-ary observations. The 3-ary consists of word, POS and NE (chemical compound Tag).

These three levels of data in the visible layer (or input layer) are converted to vectors of n-dimension and passed to word sequence layer of NCRF. The word vectors, POS vectors and NE vectors are the vector representations. These are obtained from the word2vec. We make use of the DL4J Word2vec API for this purpose [34].

The output layer uses Support Vector Machine (SVM) for classification. The SVM classifies into two event classes (trigger words): 'WORKUP' or 'REACTION_STEP'. We use the corpus provided by ChEMU 2020 track organizers as data for learning the Word2vec embedding's to convert the data to a 90 dimension of 3-arys for input.

Once the event types are identified we need to identify the arguments of these events. The arguments are identified. The task of identifying the Arguments is modelled as Argument boundary

labelling task. Here this labels "Arg1-Start", "Arg1-End", "ArgM-Start" and "ArgM-End".

The identification of Arg1's two boundaries and ArgM's two boundaries, four language models are built. ArgM-START, Arg1-END, Arg1-START and ArgM-END were identified in series, in that order. The output at each is fed as input to the next model. In other words, in each model, the previously identified boundary is also used as a feature. The choice of the order of identification of bounds was made with the idea that it is easier to first find the boundaries that are in close proximity to the event marker (trigger word) – Arg1-END and ArgM-START. Between these two, ArgM-START was chosen first, based on empirical experiments. The same holds for the choice of Arg1-START to be the third boundary.

4 Evaluation, Results and Discussion

We use the standard evaluation metrics of precision, recall and F measure for evaluating Chemical compounds and Events detection.

4.1 Named Entity Recognition

The results are evaluated and are given in the following table 1. Some examples are given below.

Ex. 1 Sentence:

A solution of hydrogen chloride in diethyl ether (2.0 N, 0.309 mL, 0.618 mmol) was added to a solution of (R)-1-(3-(dimethylamino)piperidin-1-yl)-3-(1-(2,2,2-trifluoroethyl)-1H-imidazol-2-yl)propan-1-one (0.0790 g, 0.238 mmol) in diethyl ether (3.0 mL) at 0° C.

Ex. 1 NE System output:

A solution of <REAGENT_CATALYST>hydrogen chloride</REAGENT_CATALYST> in <OTHER_COMPOUND>diethyl ether</OTHER_COMPUND> (2.0 N, 0.309 mL, 0.618 mmol) was added to a solution of <STARTING_MATERIAL>(R)-1-(3-(dimethylamino)piperidin-1-yl)-3-(1-(2,2,2-trifluoroethyl)-1H-imidazol-2-yl)propan-1-one</STARTING_MATERIAL> (0.0790 g, 0.238 mmol) in diethyl ether (3.0 mL) at <TEMPERATURE>0° C.</TEMPERATURE>

One of biggest challenges in this Chemical domain is that the entity names are alpha-numeric and also consist of parenthesis, comma and hyphens. Also the entity names are lengthy. One of the lengthiest NE had around 1000 characters as single word. Thus use of normal text tokenizer directly is not possible. We did marking of such big entities prior to sending it to the text tokenizer so that these entities are not broken. Identifying them as what type (or class0 of NE is the challenge. We performed linguistic post processing to correct the type of NE recognition and that had improved the NER system.

In Table 1 the evaluation results of CRFs based NER system are provided.

In Table 2 the evaluation results of ANN based NER system are given.

The system based on CRFs had given a very good precision. The recall is low and especially for the entities "YIELD_OTHER" and "YIELD_PERCENT". This could have been improved by using post processing rules.

NE Label	Precision	Recall	F1 Score
EXAMPLE_LABEL	0.9698	0.6932	0.8085
OTHER_COMPOUND	0.9402	0.7566	0.8385
REACTION_PRODUCT	0.9088	0.6338	0.7468
REAGENT_CATALYST	0.8898	0.8098	0.8479
SOLVENT	0.8566	0.8232	0.8395
STARTING_MATERIAL	0.8092	0.9012	0.8527
TEMPERATURE	0.8325	0.8445	0.8384
TIME	0.9521	0.6671	0.7845
YIELD_OTHER	0.9216	0.6452	0.7590
YIELD_PERCENT	0.8998	0.6010	0.7206
Average	**0.8793**	**0.8334**	**0.8037**

Table 1. Results – RNN based NER System

As we can observe from the above table the results are good and are comparable to the state of the art (CHEMU 2020 Track participant's results).

4.2 Event Extraction

The event argument identification module was evaluated with the development data provided in Task 2 CHEMU 2020 CLEF track. The event with its arguments is considered as all correct, if and only if the event marker and all the argument boundaries were correctly identified by the system. The performance of the system was evaluated in terms of precision, recall and f-measure.

Here we have performed two experiments. In the experiment 1 we take the gold tagged data of NEs as given by the CHEMU 2020 CLEF track. In Experiment 2, we take the system output of named entity recognition system as input for Event extraction. This can be said as End-to-End system. Table 2 shows the results of event arguments identification of Experiment 1.

Event Argument – Type	Precision	Recall
ARG1-START	66.67	57.14
ARG1-END	72.95	59.65
ARGM-START	81.54	57.14
ARGM-END	61.54	57.54
ALL 4 Correct	**60.67**	**55.78**

Table 2. Experiment 1- Event Arguments Identification – 10-fold Cross-Validation Results (Average)

For Experiment 2, the output obtained from the NE system as described in section 3.2 is considered, Table 3 shows the results obtained for Experiment 2.

Event Argument – Type	Precision	Recall
ARG1-START	56.67	47.14
ARG1-END	64.25	50.45
ARGM-START	69.43	49.43
ARGM-END	50.65	45.44
ALL 4 Correct	**48.79**	**44.89**

Table 3. Experiment 2 – Event Arguments Identification (End-to-End system) - 10-fold Cross-Validation Results (Average)

From the table 3 we observe that, the final event and event arguments identification results are decreased by 11%. In the NE identification it is observed that the NE types Yield_Other, Yield_Percent and Reaction_Product are not identified properly by the system, the recall of these types is lower, which affects the same in Event extraction.

5 Conclusion

ChemXtract works on extracting names of chemical compounds and event that identify the specific involvement of the chemical compounds in a chemical reaction. We have used Neural Conditional Random Fields (NCRFs) to identify and extract chemical compounds. The patent documents were preprocessed using NLP tools for obtaining syntactic information, Part-of-Speech and Noun/Verb phrases. The relationships between the chemical compounds are based on the chemical reaction events. Again the same Neural Conditional Random Fields (NCRFs) is used to identify the relationships and the relation arguments. The results obtained are encouraging and comparable with the state of the art.

Acknowledgments

We thank the CHEMU 2020 CLEF track organizers [32] for providing us the data.

References

1. Ignazio Gallo, Elisabetta Binaghi, Moreno Carullo, and Nicola Lamberti. (2008). Named entity recognition by neural sliding window. In *2008 The Eighth IAPR International Workshop on Document Analysis Systems* (pp. 567-573). IEEE.
2. Lin Yao, Hong Liu, Yi Liu, Xinxin Li, and Muhammad Waqas Anwar. (2015).Biomedical named entity recognition based on deep neutral network. *Int. J. Hybrid Inf. Technol*, *8*(8), 279-288.
3. Onur Kuru, Ozan Arkan Can, and Deniz Yuret. (2016). Charner: Character-level named entity recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 911-921).
4. Qikang Wei, Tao Chen, Ruifeng Xu, Yulan He, and Lin Gui. (2016). Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks. *Database: The Journal of Biological Databases and Curation. 10.1093/database/baw140.*
5. Hong, S. K., and Jae-Gil Lee. (2020). DTranNER: biomedical named entity recognition with deep learning-based label-label transition model. *BMC bioinformatics*, *21*(1), 53.
6. Larry Smith, Lorraine K. Tanabe, Rie Ando, Cheng Ju Kuo, Fang Chung, Chun Nan Hsu, Yu Shi Lin, Roman Klinger, Christoph M. Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A. Struble, Richard J. Povinelli, Andreas Vlachos, William A. Baumgartner, Lawrence Hunter, Bob Carpenter, Richard Tzong Han TsaiHong Jie Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter Adriaans, Christian Blaschke, Rafael Torres, Mariana Neves, Preslav Nakov, Anna Divoli, Manuel Mana-Lopez, Jacinto Mata, and John Wilbur. (2008). Overview of BioCreative II gene mention recognition. *Genome biology* 2008; 9:S2
7. Ozlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association* 2011; 18(5):552–556
8. Krallinger Martin, Leitner Florian, Rabal Obdulia, Vazquez Miguel, Oyarzabal Julen and Valencia Alfonso. (2013). Overview of the chemical compound and drug name

recognition (CHEMDNER) task. *Proceedings of 4th BioCreative Challenge Evaluation Workshop* 2013; 2:2–33

9. Ananiadou Sophia, Sampo Pyysalo, Junichi Tsujii and Douglas B. Kell. (2010). Event extraction for systems biology by text mining the literature. *Trends in biotechnology 28(7): 381-90*
10. Sofie Van Landeghem, Jari Bjorne, Chih-Hsuan Wei, Kai Hakala, Sampo Pyysalo, Sophia Ananiadou, Hung-Yu Kao, Zhiyong Lu, Tapio Salakoski, Yves Van de Peer, and Filip Ginte.(2013). Large-scale event extraction from literature with multilevel gene normalization. *PLoS ONE 8(4):e55814.*
11. Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. (2013). Overview of BioNLP Shared Task 2013. *Proceedings of the BioNLP Shared Task 2013 Workshop; 1–7*
12. Martin Krallinger, Miguel Vazquez, and Florian Leitner (2011). The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. *BMC bioinformatics 12 Suppl 8(Suppl 8), S3.*
13. Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. (2013). SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013), In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 2:341–350.
14. Kanaka D Shetty, and Siddhartha R Dalal. (2011). Using information mining of the medical literature to improve drug safety. *J Am Med Inform Assoc* 2011; 18:668–674
15. Li Chen, Maria Liakata, and Dietrich Rebholz-Schuhmann. (2013). Biological network extraction from scientific literature: state of the art and challenges. *Briefings in bioinformatics* 2013; 10.1093/bib/bbt006: 1-22
16. Stephane M Meystre, Guergana Savova, Karin Kipper-schuler, and John F. Hurdle. (2008) Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of Medical Informatics,* 17:128–144. PMID: 18660887
17. Jakub Piskorski, and Roman Yangarber (2013) Information Extraction: Past, Present and Future. In: Poibeau, T., Saggion, H., Piskorski, J., Yangarber, R. (eds). Multi-source, Multilingual Information Extraction and Summarization. *Theory and Applications of Natural Language Processing.* Springer, Berlin, Heidelberg. Multilingual Information Extraction and Summarization 2013; 23–49
18. Deyu Zhou, Dayou Zhong, and Yulan He. (2014) Biomedical relation extraction: from binary to complex. *Computational and mathematical methods in medicine*, 298473: 1-18
19. Graciela H Gonzalez, Tasnia Tahsin, Britton C Goodale, Anna C Greene, and Casey S Greene. (2016). Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery. *Briefings in bioinformatics*, 17(1):33–42.
20. Buzhou Tang, Yonghui Wu, Min Jiang, Joshua C. Denny, and Hua Xu. (2013) Recognizing and Encoding Disorder Concepts in Clinical Text using Machine Learning and Vector Space Model. In *Working Notes of CLEF 2013 Conference*, 1179.
21. Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. (2011) Natural language processing (almost) from scratch. The *Journal of Machine Learning Research*, 12:2493–2537
22. Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead and Oren Etzioni. (2007). Open information extraction from the web. In *Proceedings of IJCAI 2007*, 2670–2676
23. Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey E. Hinton. (2007). Restricted Boltzmann Machines for Collaborative Filtering. *Proceedings of the 24th International Conference on Machine Learning 2007*, 791–798
24. Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. (2006). A fast learning algorithm for deep belief nets. *Neural computation* 18(7):1527–1554.
25. Pascal Lamblin, and Yoshua Bengio. (2010). Important gains from supervised fine-tuning of deep architectures on large labeled sets. In *Proceedings of NIPS 2010 Deep Learning and Unsupervised Feature Learning Workshop*, 1-8.
26. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS)*, 1097–1105
27. Richard Socher, Cliff Chiung-Yu Lin, Andrew Y. Ng, and Christopher D. Manning. (2011). Parsing natural scenes and natural language with recursive neural networks. *In Proceedings of the 28th International Conference on International Conference on Machine Learning (ICML'11)*. Omnipress, Madison, WI, USA, 129–136
28. Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent and Samy Bengio. (2010). Why does

unsupervised pre-training help deep learning? *The Journal of Machine Learning Research* 11:625–660

29. Max Valentinuzzi. (2017). Patents and Scientific Papers: Quite Different Concepts. IEEE *Pulse* 8(1): 49 - 53.
30. Grace Ngai and Radu Florian. (2001). Transformation Based Learning in the Fast Lane. In *Proceedings of Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 40–47.
31. Jie Yang and Yue Zhang. (2018). NCRF++: An Open-source Neural Sequence Labeling Toolkit. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, 74–79.
32. He, Jiayuan and Nguyen, Dat Quoc and Akhondi, Saber A… Baldwin, Timothy and Verspoor, Karin. (2020). Overview of ChEMU 2020: Named Entity Recognition and Event Extraction of Chemical Reactions from Patents. In: Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020), LNCS vol. 12260.
33. Lafferty John, Mccallum Andrew and Pereira Fernando. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, 282–289
34. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. (2013). Efficient Estimation of Word Representations in Vector Space. *In ePrint: arXiv:1301.3781 [cs.CL]*.