Weakly-supervised Symptom Recognition for Rare Diseases in Biomedical Text

Pierre Holat, Nadi Tomeh, Thierry Charnois, Delphine Battistelli,
Marie-Christine Jaulent, Jean-Philippe Metivier

Weakly-supervised Symptom Recognition for Rare Diseases in Biomedical Text

Pierre Holat[1], Nadi Tomeh[1], Thierry Charnois[1],
Delphine Battistelli[2], Marie-Christine Jaulent[3], and Jean-Philippe Métivier[4]

[1] LIPN, University of Paris 13, Sorbonne Paris Cité, Paris, France
[2] MoDyCo, University of Paris Ouest Nanterre La Défense, Paris, France
[3] Inserm, Paris, France
[4] GREYC, University of Caen Basse-Normandie, Caen, France

Abstract. In this paper, we tackle the issue of symptom recognition for rare diseases in biomedical texts. Symptoms typically have more complex and ambiguous structure than other biomedical named entities. Furthermore, existing resources are scarce and incomplete. Therefore, we propose a weakly-supervised framework based on a combination of two approaches: sequential pattern mining under constraints and sequence labeling. We use unannotated biomedical paper abstracts with dictionaries of rare diseases and symptoms to create our training data. Our experiments show that both approaches outperform simple projection of the dictionaries on text, and their combination is beneficial. We also introduce a novel pattern mining constraint based on semantic similarity between words inside patterns.

Keywords: Information extraction, Pattern mining, CRF, Symptoms recognition, Biomedical texts

1 Introduction

Orphanet encyclopedia is the reference portal for information on rare diseases (RD) and orphan drugs. A rare disease is a disease that affects less than 1 over 2,000 people. There are between 6,000 and 8,000 rare diseases and 30 million people are concerned in Europe. The Orphanet initiative aims to improve the diagnosis, care and treatment of patients with such diseases. Among its activities, Orphanet maintains a rare disease database containing expert-authored and peer-reviewed syntheses describing current knowledge about each disease. The syntheses are produced by human specialists following a manual, time-consuming monitoring of the medical literature. The aim of our work is to automatically acquire new knowledge related to rare diseases; we focus on the task of *symptom recognition* in medical publication abstracts.

We use the term *symptom* to refer to features of a disease, as noticed and described by a patient (functional sign), or as observed by a healthcare professional (clinical sign) without distinction. The linguistic structure of symptoms is typically more complex than other biomedical named entities [3] for various reasons

as discussed in [10]. They manifest a considerably larger variability in forms, ranging from simple nouns to whole sentences, and a larger number of syntactic and semantic ambiguities. In the following examples, symptoms as identified by an expert are shown in bold:

- With disease progression patients additionally develop **weakness** and **wasting of the limb and bulbar muscles**.
- Diagnosis is based on clinical presentation, and **glycemia and lactacidemia levels after a meal (hyperglycemia and hypolactacidemia), and after three to four hour fasting (hypoglycemia and hyperlactacidemia)**.

Furthermore, few works have focused on symptom recognition and therefore existing resources are limited and incomplete: to our knowledge, no dataset that is fully annotated with symptoms is available to allow for supervised learning.

To address these issues, we propose a weakly-supervised approach to symptom recognition that combines three independent lexical resources (Sect.2.1): a corpus of unannotated medical paper abstracts and two dictionaries, one for rare diseases and another for symptoms.

We project the dictionaries on the abstracts to create an annotated dataset to train subsequent models. Since the dictionaries are not exhaustive, the annotation is only partial (weak). Given the annotated dataset, we formalize the problem of symptom recognition in two complementary ways: (a) as supervised sequence labeling for which we use Conditional Random Fields (CRF) [8] (Sect.2.3); and (b) as sequential pattern mining under constraints [2] (Sect.2.4). We combine these approaches in a pipeline architecture (Sect.2).

Our contribution is threefold. First, we show experimentally (Sect.3) that sequence labeling and pattern mining are both adequate formalizations of the task, for they outperform a simple projection of the dictionaries on the text; Furthermore, we show that their combination is beneficial since CRFs allow for rich representation of words while pattern mining privileges modeling their context. Second, we introduce a novel pattern mining constraint based on distributional similarities between words (Sect.2.4). Third, we created a gold standard for evaluation (Sect.2.1), manually annotated with symptoms by human experts.

2 A Pipeline Architecture for Symptom Recognition

In this section we describe our iterative pipeline approach to symptom recognition, in the spirit of [11]. Fig.1 depicts the overall architecture of our system.

Input data to our approach contain a collection of unannotated article abstracts and two dictionaries, one for diseases and one for symptoms. First, the dictionaries are projected on the abstracts to obtain partial annotations; the resulting annotated data is used to train a CRF sequence labeler and as an input to a sequential pattern mining algorithm; the learned CRF model and the extracted patterns are used to extract symptoms from the test data, separately or in combination; these symptoms are then compared to manually annotated gold standard for evaluation using F-measure. It should be noted that the learned

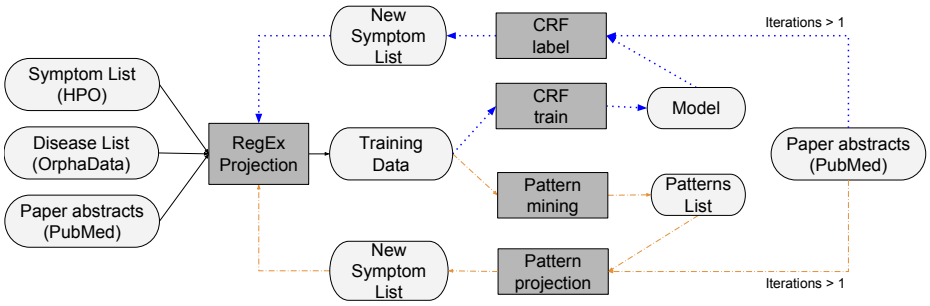

Fig. 1. Overall architecture of the system.

models can be applied on the training data (using cross-validation) to discover new symptoms to be added to the dictionary, and the whole process can be iterated.

2.1 Datasets and Evaluation

The input to our system is composed of three independent online resources:

- The first dataset is a corpus of 10,000 article abstracts extracted from the biomedical literature available on PubMed.[5] To build it, we extracted 100 biomedical paper abstracts for each one of 100 rare diseases selected from OrphaData in advance by an expert;
- The second dataset is a dictionary of 17,469 distinct phenotype anomalies provided by the Human Phenotype Ontology (HPO).[6] A phenotype is all the observable characteristics of a person, such as their morphology, biochemical or physiological properties. It results from the interactions between a genotype (expression of an organism's genes) and its environment. Since many rare diseases are genetic, we follow [11] and consider the above anomalies to be symptoms. This dictionary is not exhaustive;
- The third dataset is a dictionary of 16,576 distinct names of rare disease and their aliases, provided by OrphaData,[7] a comprehensive, high-quality resource related to rare diseases and orphan drugs. This dictionary is not exhaustive.

The testing data are made of 50 biomedical paper abstracts with an average of 184 word by abstract. A first automatic annotation was made on the data, then we ask two medical experts to review the generated symptoms and add missing ones to build a gold standard. Our experts have labeled 407 symptoms and their position in the testing data, so an average of 8,1 symptom by abstract.

[5] `http://www.ncbi.nlm.nih.gov/pubmed`
[6] `http://human-phenotype-ontology.github.io`
[7] `http://www.orphadata.org`

O B I O O B I I O
clinically **silent tumors** often demonstrate **subclinical hormonal activity** .

Fig. 2. Symptom recognition as BIO sequence labeling. Symptoms are bolded.

Performance is evaluated using the standard precision, recall and F-measure.[8]

2.2 Weak Annotation by Projection

Given the input datasets, the first step in our workflow is to project the dictionaries of symptoms and diseases on the abstracts contained in our trainings set. The projection step produce only weak (partial) annotation since Orpha-Data and HPO lists are not exhaustive; they do not contain all symptoms and diseases, nor the various linguistic forms they can take.

The corpus and dictionaries are preprocessed using TreeTagger:[9] texts are tokenized, and each token is lemmatized and part-of-speech (POS) tagged. Each term in the dictionaries (possibly composed of several tokens) is matched against the corpus by comparing using regular expressions. Terms coming from HPO are often generic (e.g. "weakness") and may be supplemented in medical texts with adjectives or object complements (e.g. "severe weakness of the tongue"). Thus, once a term matches, it can be expanded to its nominal phrase using the POS tags assigned to surrounding terms.

The partially annotated corpus resulting from the projection is used as a training set for the subsequent models.

2.3 Symptom Recognition as Sequence Labeling

We formalize the problem of symptom recognition as a supervised sequence labeling problem with BIO notation, for which we use Conditional Random Fields (CRF) [8]. An abstract text is seen as a sequence of words, each of which is labeled with one of three possible labels: `B` (beginning of a symptom), `I` (inside a symptom), and `O` (outside a symptom). Thus, a symptom is a word segment corresponding to a label `B` potentially followed by consecutive `I` labels, as shown in Fig.2.

The main advantage of CRFs is their conditional nature which allows for rich representations of words in a sequence. It is possible to incorporate multiple information sources in the form of feature functions, without having to model their interactions explicitly. On the contrary, applying sequential pattern mining on such rich representations is prohibitively intractable.

We use the same set of features used for named entity recognition in [5] as implemented in Stanford NER recognizer.[10] They include the current word, the

[8] Using the script provided by `http://www.cnts.ua.ac.be/conll2000/chunking/output.html` which take the same input data format (BIO) as our data.

[9] `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger`

[10] `http://nlp.stanford.edu/software/CRF-NER.shtml`

previous and next one, as well as all the words in window of a given size (n-gram features); orthographic features characterizing the form of the words; prefixes and suffixes; and several conjunctions thereof.

2.4 Symptom Recognition as Sequential Pattern Mining

While the sequence labeling CRF approach presented above can easily employ rich representation of words, it cannot efficiently capture rich context. For instance, considering all possible sub-sequences as context for the current word is computationally intractable because their number grows exponentially in the size of the sequence.

To address this issue, we propose to use a sequential pattern mining approach with an emphasis on the context more than the words themselves. Sequential pattern mining allows to take into account the language sequentiality and is one of the most studied and challenging task in data mining. Since its introduction by Agrawal and Srikant [1], the problem has been well formalized:

Let $\mathcal{I} = \{i_1, i_2 \dots i_m\}$ be the finite set of items. An itemset is a non-empty set of items. A sequence S over $\mathcal{I}$ is an ordered list $\langle it_1, \dots, it_k \rangle$, with it_j an itemset over $\mathcal{I}$, $j = 1 \dots k$. A k-sequence is a sequence of k items (i.e., of length k), $|S|$ denotes the length of sequence S. $\mathbb{T}(\mathcal{I})$ will denote the (infinite) set of all possible sequences over $\mathcal{I}$. A *sequence database* $\mathcal{D}$ over $\mathcal{I}$ is a finite set of doubles (SID, T), called transactions, with $SID \in \{1, 2, \dots\}$ an identifier and $T \in \mathbb{T}(\mathcal{I})$ a sequence over $\mathcal{I}$.

Definition 1 (*Inclusion*). *A sequence $S' = \langle is'_1\ is'_2\ \dots\ is'_n \rangle$ is a subsequence of another sequence $S = \langle is_1\ is_2\ \dots\ is_m \rangle$, denoted $S' \preceq S$, if there exist $i_1 < i_2 < \dots\ i_j\ \dots < i_n$ such that $is'_1 \subseteq is_{i_1}$, $is'_2 \subseteq is_{i_2} \dots is'_n \subseteq is_{i_n}$.*

Definition 2 (*Support*). *The support of a sequence S in a transaction database $\mathcal{D}$, denoted $Support(S, \mathcal{D})$, is defined as: $Support(S, \mathcal{D}) = |\{(SID, T) \in \mathcal{D} | S \preceq T\}|$. The frequency of S in $\mathcal{D}$, denoted $freq_S^{\mathcal{D}}$, is $freq_S^{\mathcal{D}} = \frac{Support(S,\mathcal{D})}{|\mathcal{D}|}$.*

Given a user-defined minimal frequency threshold σ, the problem of sequential pattern mining is the extraction of all the sequences S in $\mathcal{D}$ such that $freq_S^{\mathcal{D}} \geq \sigma$. The set of all frequent sequences for a threshold σ in a database $\mathcal{D}$ is denoted $FSeqs(\mathcal{D}, \sigma)$,

$$FSeqs(\mathcal{D}, \sigma) = \{S \mid freq_S^{\mathcal{D}} \geq \sigma\} \tag{1}$$

Our data contains the lemma and the POS of each word. So in our context, let $\mathcal{I}$ be the finite set of all words and part-of-speech tag. An itemset is a non-empty set of the lemma and the part-of-speech of a word. We also add a special item (#symptom#) in $\mathcal{I}$. This item will be used as a placeholder for each annotated symptom in the training data, as shown in the following example:

$< \{we, PP\}\{find, VBD\}\{that, IN\}\{clinically, RB\}\{\textbf{\#symptom\#}\}$
$\{often, RB\}\{demonstrate, VBP\}\{\textbf{\#symptom\#}\}\{., SENT\} >$

Using sequential pattern mining on such sequences allows us to extract linguistic patterns covering symbolic symptoms. However, using only the user-defined minimal frequency threshold σ as a constraint, pattern mining typically yields an exponential number of patterns. Pattern mining under constraints [17] is a powerful paradigm to target relevant patterns [14]. Therefore, we used a pattern mining algorithm under the most used constraints in the literature in addition to σ. The *minimal and maximal gap* constraint imposes a limit on the number of words separating items of a pattern. The *minimal and maximal length* constraint limits the number of items in a pattern. We also used a *belonging constraint* specific to our task, a pattern must contain our specific item #symptom#.

Semantic Similarity Constraint In addition to the above-mentioned constraints, we introduce a new *semantic similarity constraint* based on the distributional properties of the words, estimated from a large unannotated corpus.

We observed during initial experiments a high level of redundancy in extracted patterns, such as a succession of conjunctions or prepositions for instance. We designed a constraint with a limit on the similarity of two adjacent items of a pattern. Therefore, this constraint is designed to discard redundant, uninformative patterns.

To be able to quantify the level of redundancy in the pattern, we used the distributional hypothesis [6]: words that occur in the same context tend to have similar meanings, this hypothesis is the basis for models like *Word2Vec* [12, 13], which learns a low-dimensional continuous vector representation of words from large amount of text. To train the *Word2Vec* model, we extracted 7,031,643 biomedical paper abstract from PubMed, that's 8.7GB of input data for a final model of 1.10GB containing 1,373,138 words in a biomedical context. The learned *Word2Vec* model is loaded by the data mining algorithm, each item i will have an associated vector V_i allowing to measure the cosine distance $D_c(V_i, V_j)$ between two consecutive items.

Definition 3 (Semantic Similarity Constraint). *Given a user-defined maximal similarity threshold ζ. Let i_d the last item of a sequence S, an extension of S by a new item i_n is possible only if :*

$$D_c(V_{i_d}, V_{i_n}) \leqslant \zeta \tag{2}$$

The semantic similarity constraint is anti-monotonic, the Prop.1 allows an efficient pruning of the search space.

Property 1 (Effect of the anti-monotonic semantic similarity constraint). Let S be a sequence. If S does not respect the semantic similarity constraint, it does not exist a sequence S' with $S \preceq S'$ which will respect the constraint. Therefore, the search space of all the extensions of S can be pruned.

3 Experiments and Results

3.1 Iterative Learning Analysis

In this experiment, unlike the next results, the whole annotation process is iterated a number of times in an attempt to discover more symptoms. For the pattern mining module, there is no amelioration because after the first iteration, symptom placeholders cover more parts of the sequence, and the new symptoms are noisy. We almost never find a new symptom after the first iteration. An interesting perspective would be to try this iterative learning with different data set each time, by splitting the data into subsets. It is different for the CRF module, it learns more and more new symptoms at each iterations, but unfortunately it is also mainly noise because the precision tends to decrease.

3.2 Individual Module Results

In this section we compare symptom recognition results, shown in Table 1, for each of the recognition modules in isolation of the other.

Table 1. Details of the best results with tuned parameters for each module

Module	Parameters	Precision	Recall	F-Measure
Dictionary		**57.58**	14.00	22.53
CRF	bag of words	**56.31**	14.25	22.75
CRF	ngrams, ngramLength=6	**56.14**	15.72	24.57
Pattern	σ=0.05%, Gap=0, ζ=0.4	23.12	**38.57**	**28.91**

Dictionary Projection Results. We first created a baseline by projection of the 17,469 symptoms that we gathered in our dictionary on the testing data. Of all our results, that baseline have the best precision but the worst recall. Since we used that, non exhaustive, dictionary to build the training data, it allows us to see if our different modules learn new informations.

CRF Results. The CRF module, successfully learned some context knowledge because there is a 12% raise in recall and a 9% in F-measure for a small decrease of 2.5% in precision.

Pattern Mining Results. Since we replaced each symptom by the item #symptom# in the training data, we only represent the context. A pattern like "(such,JJ) (as,IN) (#symptom#)" can be applied on the testing data and discover symptoms that were not in the training data. Hence, the Pattern mining module has an increase of 175% of recall, but a 60% decline in precision, for a final amelioration of the F-measure by 28%. It was expected that the precision would drop,

but in our context of helping human expert process data, the recall is more important to maximize: missing a potential new symptom is more harmful than producing a large number of false-positives. Figure 3 lists some of the extracted patterns.

```
 (treatment,NN) (IN) (#symptom#) : 62
(development,NN) (of,IN) (#symptom#) : 43
(patient,NNS) (with,IN) (#symptom#) : 295
(diagnosis,NN) (IN) (#symptom#) : 98
(patient,NNS) (IN) (#symptom#) : 306
(case,NN) (of,IN) (#symptom#) : 48
(such,JJ) (as,IN) (#symptom#) : 91
(IN) (patient,NNS) (IN) (#symptom#) : 163
(NNS) (such,JJ) (as,IN) (#symptom#) : 46
(in,IN) (patient,NNS) (IN) (#symptom#) : 89
(in,IN) (patient,NNS) (with,IN) (#symptom#) : 88
(IN) (patient,NNS) (with,IN) (#symptom#) : 161
```

Fig. 3. Examples of extracted patterns with their support.

Table 2 shows the difference between the annotation of each module and the annotation of the human expert on two sentences.

Table 2. Examples of each module annotations.

	diffuse palmoplantar keratoderma and precocious				
Expert Annotation	B	I	I	O	B
Pattern Annotation	B	I	I	I	I
CRF Annotation	B	I	I	O	O

	primary immunodef. disorders with residual cell-mediated immunity						
Expert	B	I	I	O	O	O	O
Pattern	B	I	I	I	I	I	I
CRF	O	B	O	O	O	O	O

3.3 Impact of Training Data Size

Figure 4 shows the variation in precision, recall and F-measure for our modules with increasing size of training data. From 1.000 to 10.000 biomedical abstracts, there is no visible impact on the CRF, even if the best score, in term of F-Measure, is on the maximum size data. Pattern Mining improves its recall with

more data, but the precision tends to drop. The best score is on the middle size data.

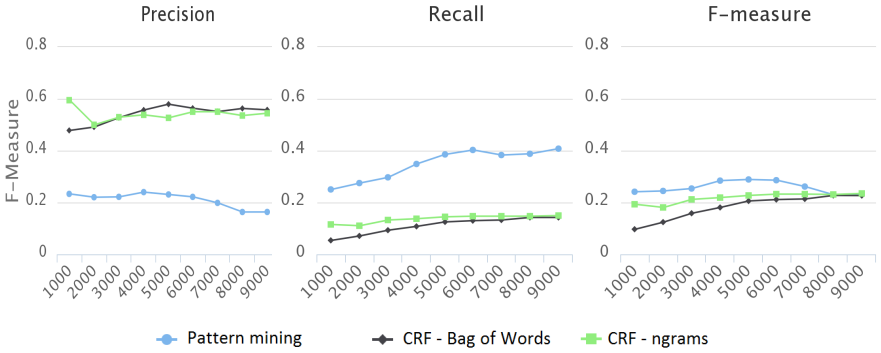

Fig. 4. Impact of the number of abstracts used for training

3.4 Model Combination Analysis

We clearly see on the Fig.4 the difference in the precision/recall ratio of the numerical and the symbolic modules even if they are very close in terms of F-measure. Because the CRF maximize the precision and the pattern mining maximize the recall, we tried to do a combination of the best CRF module result and the best pattern mining module. If the two model made the same decision we keep it, if not we promote the choice of "B". This combination gives better results (cf. Table 3) than each modules separately. In the same logic, we also combine our bests results with the baseline using the projection of the symptom dictionary. If that last enhancement has improved each module result individually, it did not improve the combination of the best CRF and pattern mining result.

Table 3. Details of the best results for each module and their combination.

Module	**Precision**	**Recall**	**F-Measure**
Dictionary (***D***)	**57.58**	14.00	22.53
CRF - ngram	**56.14**	15.72	24.57
Pattern mining	23.12	**38.57**	28.91
Combination	23.46	**39.31**	**29.38**
CRF - ngram + ***D***	**56.90**	16.22	25.24
Pattern mining + ***D***	23.35	**39.07**	29.23
Combination	23.46	**39.31**	**29.38**

3.5 Impact of Semantic Similarity Constraint

The purpose of the semantic similarity constraint is to reduce the number of patterns extracted without jeopardizing the classification accuracy, and like most data mining constraint the results are threshold dependent. Figure 5 shows the success of our constraint, a 30% reduction of extracted patterns without losses in F-Measure. It also shows that a too strong threshold for the constraint lower the performance. With a maximal semantic similarity below 0.4 the constraint tends to produce very few patterns. In terms of cosine distance, a threshold of 0.2 is so low that semantically divergent words would be considered similar.

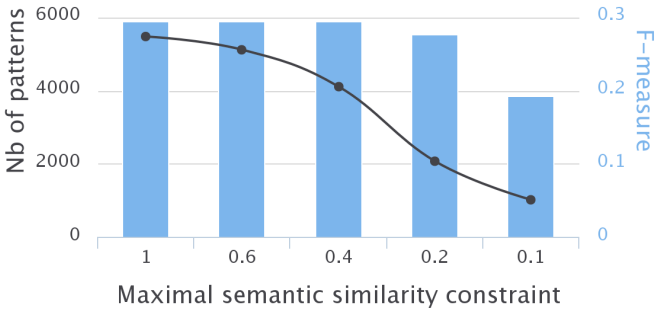

Fig. 5. Impact of the semantic similarity constraint (abscissa) on the number of extracted patterns (black line) and scoring (blue column).

4 Related Work

To our knowledge, there is no annotated dataset which can be used to train a supervised model specific for symptom recognition. Most of the studies are based on clinical reports or narrative corpora without symptom annotation and therefore can not be used in our context for symptom monitoring. Such corpora include the Mayo Clinic corpus [15] and the 2010i2b2/VA Challenge corpus [18]. Other existing biomedical datasets annotate only diseases; they include the NCBI disease corpus [4] which consists of 793 PubMed abstracts with 6,892 disease mentions and 790 unique disease concepts mapped to the Medical Subject Headings (MeSH),[11] and the Arizona Disease Corpus (AZDC) [9] which contains 2,784 sentences from MEDLINE abstracts annotated with disease mentions and mapped to the Unified Medical Language System (UMLS)[12].

Symptom recognition [10] is a relatively new task, often included in more general categories such as *clinical concepts* [19], *medical problems* [18] or *phenotypic information* [16]. Even on these categories, few studies take advantage

[11] `https://www.nlm.nih.gov/mesh/meshhome.html`
[12] `https://www.nlm.nih.gov/research/umls/`

of considering the linguistic context in which symptoms appear, and they are more focused on the linguistic analysis. In [15], the authors notice that most of the symptoms are given in relation with an anatomic location, while in [7] the authors state, after annotating their corpus with MeSH, that 75% of the signs and symptoms co-occur with up to five other signs and symptoms in a sentence. None of the above work was concerned with fully automatic symptom recognition.

5 Conclusion

We have described a system that enable the use of different learning modules for a symptom recognition task in biomedical texts. For the numeric approach we used a CRF module which maximized the precision and for the symbolic approach we used a pattern mining module which maximize the recall. We introduced a new semantic constraint for the pattern mining process which remove redundant patterns without decline in the scoring. Both approach (symbolic and numerical) have been combined to further enhanced the results. A first future direction will be to enhance the combination of the modules. An idea is to use the patterns extracted as features for the CRF. A second future direction is to enhance our similarity constraint to take into account more distant redundancy in a pattern, or to apply the constraint differently in function of the words part-of-speech.

Acknowledgments

This work is supported by the French National Research Agency (ANR) as part of the project Hybride ANR-11-BS02-002 and the "Investissements d'Avenir" program (reference: ANR-10-LABX-0083).

References

1. Agrawal, R., Srikant, R.: Mining Sequential Patterns. In: Proceedings of the Eleventh International Conference on Data Engineering. pp. 3–14 (1995)
2. Béchet, N., Cellier, P., Charnois, T., Crémilleux, B.: Sequence mining under multiple constraints. In: Proceedings of the 30th Annual ACM Symposium on Applied Computing. pp. 908–914 (2015)
3. Cohen, K.B.: BioNLP: biomedical text mining. In: Handbook of Natural Language Processing, Second Edition (2010)
4. Doğan, R.I., Leaman, R., Lu, Z.: Ncbi disease corpus: a resource for disease name recognition and concept normalization. Journal of biomedical informatics 47, 1–10 (2014)
5. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 363–370 (2005)
6. Harris, Z.S.: Distributional structure. Word 10(2-3), 146–162 (1954)

7. Kokkinakis, D.: Developing resources for swedish bio-medical text mining. In: Proceedings of the 2nd International Symposium on Semantic Mining in Biomedicine (SMBM) (2006)
8. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
9. Leaman, R., Miller, C., Gonzalez, G.: Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark. Proceedings of the 2009 Symposium on Languages in Biology and Medicine 82(9) (2009)
10. Martin, L., Battistelli, D., Charnois, T.: Symptom extraction issue. In: Proceedings of BioNLP 2014. pp. 107–111 (June 2014)
11. Métivier, J.P., Serrano, L., Charnois, T., Cuissart, B., Widlöcher, A.: Automatic symptom extraction from texts to enhance knowledge discovery on rare diseases. In: Artificial Intelligence in Medicine, pp. 249–254 (2015)
12. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
14. Pei, J., Han, J., Wang, W.: Constraint-based sequential pattern mining: the pattern-growth methods. Journal of Intelligent Information Systems 28(2), 133–160 (2007)
15. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. Journal of the American Medical Informatics Association 17(5), 507–513 (2010)
16. South, B.R., Shen, S., Jones, M., Garvin, J., Samore, M.H., Chapman, W.W., Gundlapalli, A.V.: Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. BMC bioinformatics 10(9), 1 (2009)
17. Srikant, R., Agrawal, R.: Mining sequential patterns: Generalizations and performance improvements. In: Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology. pp. 3–17 (1996)
18. Uzuner, Ö., South, B.R., Shen, S., DuVall, S.L.: 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association 18(5), 552–556 (2011)
19. Wagholikar, K.B., Torii, M., Jonnalagadda, S.R., Liu, H.: Pooling annotated corpora for clinical concept extraction. Journal of Biomedical Semantics 4(1), 1–10 (2013)