SMRE: Semi-supervised Medical Relation Extraction

Trugui Sana
Ecole Nationale d'Ingénieurs de Sfax
2 PUBLICATIONS **29** CITATIONS
SEE PROFILE

Ines Boujelben
University of Sfax
16 PUBLICATIONS **54** CITATIONS
SEE PROFILE

Salma Jamoussi
University of Sfax
95 PUBLICATIONS **289** CITATIONS
SEE PROFILE

SMRE: Semi-supervised Medical Relation Extraction

Sana Trigui[1], *Ines Boujelben*[1], *Salma Jamoussi*[2]

[1]Miracl, University of Sfax, Tunisia
[2]Higher Institute of Computer Science and Multimedia of Sfax

truguisana0@gmail.com,Boujelben_ines@yahoo.fr,
jamoussi@gmail.com

Abstract

Relation extraction between named entities presents a useful task for several applications of Natural Languages Processing (NLP) such as a question-answering system [1], automatic summarization [2] and ontology construction [3]. This paper reports our semi-supervised system SMRE to extract relations between medical named entities. Given that semi-supervised learning approach relies on a few labeled data, it requires less human effort and time consuming. Our system is applied to a well-known medical corpus [4] that is built from a Medline medical bibliographic database. The evaluation of our system showed promising results of 83.20% in terms of accuracy without knowing the medical named entities and 100% otherwise.

Index Terms: Named entity, Relation extraction, Semi-supervised learning, Medical domain.

1. Introduction

The task of medical relation extraction serves to discover useful relationships between two medical named entities (NEs). A medical NE is a NE[1] related to the medical field such as diseases (Asthma, Diabetes mellitus), treatments (Antibiotic, analgesic), symptoms (Vomiting, dry mouth), medications (Metformin, Omeprazole) and examinations (Computed tomography, pulmonary x-ray).The medical field is characterized by the complexity and the instability of its vocabulary. This problem, in a sensitive field such as medicine, forces us to develop base of knowledge (medical NE and relations between medical NEs).This base helps us to better understand the text, discover new information and improve the quality of patient care. Given the usefulness of the task of relation extraction between NEs, many researchers have been interested to this task where each of them has tried to find the most optimal solution for solving this problem. Indeed, some works have used linguistic methods (rules, grammars) [5] to solve this problem. Others have chosen to work with automatic classification techniques based on supervised learning [6]. However, these supervised methods are based on large annotated corpora. The annotation of these corpora requires a lot of effort, expertise and time consumption. These limitations have encouraged the introduction of the semi-supervised learning paradigm as a reliable classification tool. Therefore, we have resorted to working with semi-supervised learning methods. The paper is organized as the following: first, we will survey prior semi-supervised studies on relation extraction between medical NEs. The third section is going to illustrate the architecture of our semi-supervised system, in which we detail its main steps. Afterward, we will present the different experiments from which we are going to discuss the reported results. Finally, some conclusions are drawn in order to structure future works.

2. Semi-supervised relation extraction

Semi-supervised learning use both labeled and unlabeled data [7]. It falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). From the labeled data, we can predict labels from the unlabeled data. Several kinds of methods have been developed to carry out the task of semi-supervised learning, among which we can mention Self-training, Co-training, Tri-training. Self-training [8] consists in training a classifier on labeled data (DL). This classifier is then used to label the unlabeled data (DU). Labeled data with a high degree of confidence are then added to the training data (DL). The classifier is re-trained on the DL data and the procedure is repeated until the unlabeled data disappears. Self-training has been applied to several NLP tasks. [9] are based on self-training for the morpho-syntactic categorization of sentences using a Markov model classifier. They obtained a rate of accuracy[2] of 85% when applied to the North American News Corpus NANC. The authors in [10] proposed a method for the classification of sentiments. They showed that the classification rate obtained by their method (84%) is better than the rate obtained by a supervised method (73%). Self-training is a wrapper algorithm. However, this method encounters the problem of the lack of external information; the classifier is supposed to provide additional information from the non-annotated examples based on its own output, including his confidence score. In order to avoid these problems of divergence, the idea here is to use several different classifiers to improve the classification task. This strategy was recognized as a Co-training method. Co-training [11] can be perceived as an extension of the Self-training method. Instead of using a single classifier, Co-training uses two classifiers, and the set of attributes is divided into two independent sets. Its main idea is to train a first classifier with a first set of attributes in order to label the unlabeled data (DU). Then,

[1] NE includes the proper names of persons (e.g. Louis), locations (e.g. New York), and organizations (UNESCO).

[2] Calculated by the percentage of correct classified data

the labeled data with a high degree of confidence are added to the training data (DL) to learn the second classifier. Subsequently, the same procedure is repeated with the second classifier to learn the first classifier. The procedure is repeated until the unlabeled data disappears. An extension of Co-training is recognized as a Tri-training method [12]. This algorithm uses three classifiers. These three classifiers are first trained in labeled data and then used to label unlabeled data. The labeling phase is done by combining the classifiers using the majority vote. Authors in [13] have relied on Co-training for lexical disambiguation of medical words. They have defined two sets of attributes where the first set contains the context of the words around the ambiguous word, and the second set is used to find the different meanings related to that word using the semantic network UMLS[3]. They obtained an accuracy of 85%. The system proposed by [14] is based on a Co-training method for the extraction of two types of relations, one between a disease and symptom, and the other between a symptom and a treatment from Medline. For this purpose, the authors constructed two corpora: the first contains disease-symptom relations and the second contains symptom-treatment relations. They obtained an f-measure of 80% for the second type of relations. To use the Co-training method, the authors [11] defined two conditions: one must have two different views (set of attributes) of the data to be classified, and these two views must be compatibles and independents. A last hypothesis for the proper functioning of Co-training is that each view must be sufficiently consistent to learn a classifier independently. However, the use of this method with the constraints cited above is sometimes impossible because in some cases we cannot have two independent views.

3. Proposed method

Since we seek to rely on a few labeled data, we propose our semi-supervised system "SMRE" for relation extraction between medical NEs. Its general architecture is composed of three general modules of processing, as illustrated in Figure 1: (1) building training data, (2) selection of relevant subsets of attributes, and (3) Semi-supervised training process.

3.1. Building training data

In this module, we construct our training corpus. We used the same corpus of [4] which was also used by [15], [16] and [17].This corpus which is called the Berkeley[4] corpus, is constructed from the titles and abstracts of Medline. It has been annotated with seven types of semantic relationships between the medical entities disease (DIS) and treatment (TREAT). As a first step, we apply a morphological and syntactical analysis using the Stanford[5]tool, which allows segmentation, lemmatization, and morph syntactic categorization. This corpus is used subsequently to extract the different features (attributes) of training in order to build our training base. The different features used in our work are listed in Table 1.

3.2. Selection of relevant subsets of attributes

The process of selecting attributes consists on reducing the number of used attributes by choosing from a large set of attributes the more interesting subset. This reduction can improve the performance of a relation extraction system. We distinguish three approaches to attribute selection: the wrapper approach [18] which uses the classification algorithm to evaluate the subset of generated attributes, the filter-type approach [19] that is completely independent of the used algorithm and uses an evaluation function based on statistical, entropy, consistency and distance to evaluate the subset of attributes. Finally, the embedded approach that combines the two previous approaches [20]. We can notice that the filtering methods are the best in terms of execution time and generality[6]. This is probably the main reason why these methods are the most popular. In order to have the best compromise between the time constraints and the quality of the results, we are interested to filtering methods. However, it is important to note that these filters require choosing the appropriate number of attributes to be selected. For this reason, we try to find an efficient solution to find the right number of attributes to select. The application of a filtering method generates a list of all the attributes sorted in decreasing order according to an evaluation function. In our case, we will use the following functions: dependence measure [21], information gain measure and Principal Component Analysis (PCA) [22].

The idea of selecting the most relevant d attributes from this list is to calculate the accuracy of the first attribute that has the highest score using a training corpus and a test corpus. Then, with every iteration, we add the attribute that succeeds it in the list until to arrive to evaluate all attributes.The selection of *d* attributes is done by selecting a subset of attributes, which presents the maximum accuracy. In order to have *n* subsets of relevant attributes, we follow the same approach while using *n* evaluation criteria (information gain, gain ratio, etc.).

3.3. Semi-supervised training process

The third module presents the core of our method. It takes place in five phases, which are: training phase, selection phase of K unlabeled data, labeling phase, combining phase of the classifiers and finally the phase of updating the corpus.

- **First phase : training phase**

This phase consists in inducing *n* classifiers (same classifier with n different subsets of relevant attributes, the *n* subsets were obtained in the second module) on the training kernel which contains only a small number of labeled data, evaluate these *n* classifiers using a test

[3] https://www.nlm.nih.gov/research/umls/

[4] biotext.berkeley.edu

[5] http://nlp.stanford.edu/software/tagger.shtml

[6] https://en.wikipedia.org/wiki/Feature_selection

corpus. The purpose of using a test corpus is to calculate a confidence score for each classifier. This score is the accuracy measure calculated using the following formula:

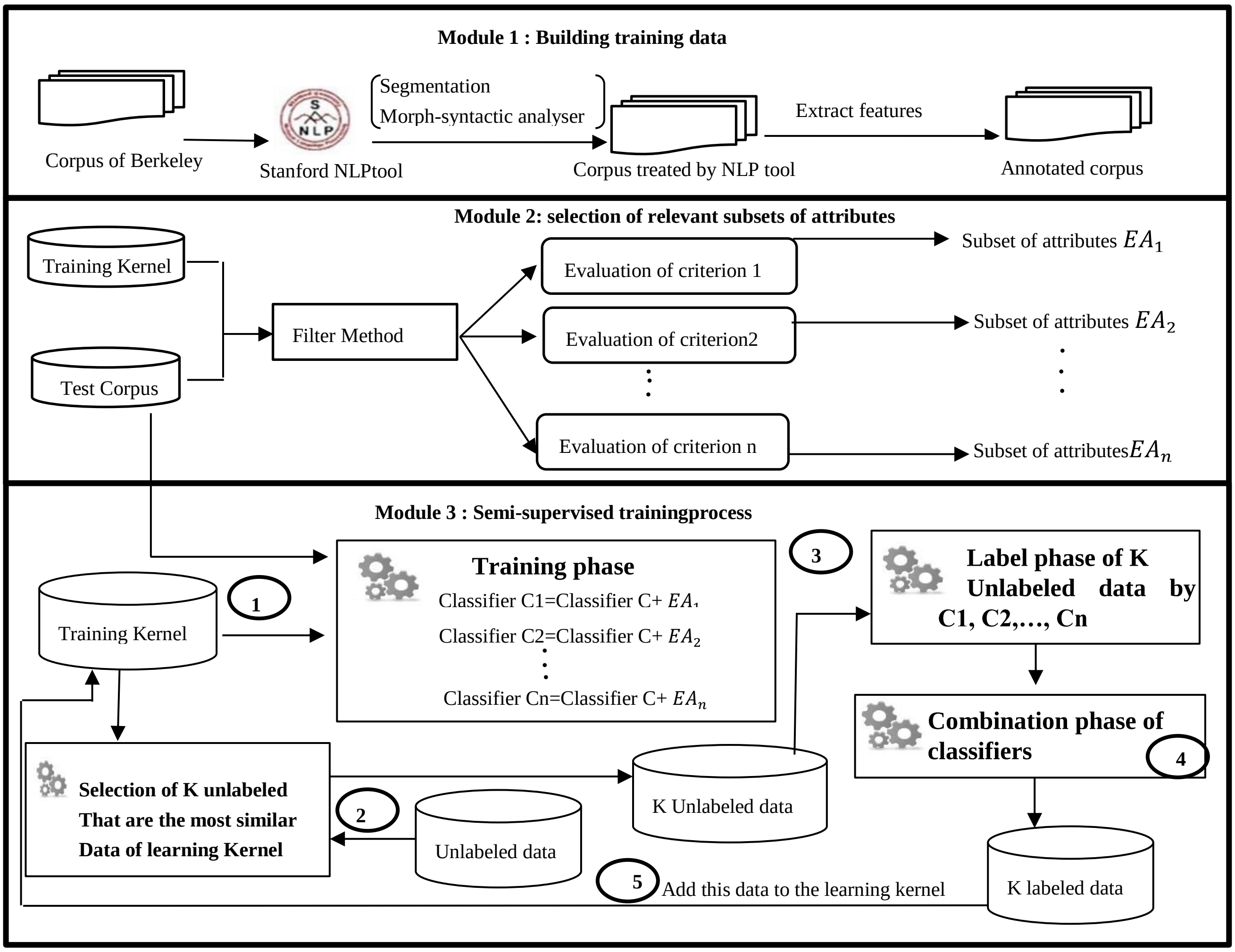
Module 1 : Building training data
Corpus of Berkeley
Stanford NLPtool
Segmentation
Morph-syntactic analyser
Corpus treated by NLP tool
Extract features
Annotated corpus
Module 2: selection of relevant subsets of attributes
Training Kernel
Test Corpus
Filter Method
Evaluation of criterion 1
Evaluation of criterion2
Evaluation of criterion n
Subset of attributes EA_1
Subset of attributes EA_2
Subset of attributesEA_n
Module 3 : Semi-supervised trainingprocess
Training phase
Classifier C1=Classifier C+ EA_1
Classifier C2=Classifier C+ EA_2
Classifier Cn=Classifier C+ EA_n
Training Kernel
1
2
3
4
5
Label phase of K Unlabeled data by C1, C2,…, Cn
Combination phase of classifiers
Selection of K unlabeled That are the most similar Data of learning Kernel
Unlabeled data
K Unlabeled data
K labeled data
Add this data to the learning kernel

Figure 1:*The architecture of our semi-supervised method*

Type	Feature	Description
Lexical	catM1E1	The part of speech of the first word before NE1
	catM2E1	The part of speech of the second word before NE1
	catM1E2	The part of speech of the first word before NE2
	catM2E2	The part of speech of the second word before NE2
	catM1E1E2	The part of speech of the first word between the two NEs
	catM2E1E2	The part of speech of the second word between the two NEs
Semantic	typeEN1	The first NE tag (label)
	typeEN2	The second NE tag
	PaireE1E2	The appearance order of NEs
Syntactic	Type-phrase	Type of sentence(nominal, verbal)
Numeric	NbrM	Number of words in the sentence
	posE1	Position of the first NE
	posE2	Position of the second NE
	nbrMavE1	Number of words before the first NE
	nbrMapE1	Number of words after the first NE
	nbrMavE2	Number of words before the second NE
	nbrMapE2	Number of words after the second NE
	nbrMEnE1E2	Number of words between the two NEs

Table1.*The used features.*

Score(C) = 1- error rate (1)

error rate=number of misclassified data /number total of data (2)

- **Second phase : selecting K unlabeled data**

In this phase, we select the K unlabeled most similar data to the training kernel. The similarity calculation is done by calculating the distance between the unlabeled data and the labeled data. If we have m labeled data and m1 unlabeled data, we will perform m * m1 distance calculations. This requires a lot of time, which will decrease the execution time of our method. To solve this problem, we present the following solution:

1) We start by decomposing the training kernel into M groups with M being the number of classes.

2) For each group, we calculate its barycenter (b) of coordinates $x_1, x_2, \ldots.. x_l$using the following formula:

$$x_i = \frac{\sum_{j=1}^{l} x_{ij}}{l} \qquad (3)$$

Where l is the number of attributes

3) We calculate the distance between the M groups and the unlabeled data using the Euclidean distance:

$$D\ (b, data) = \sum_{i=1}^{l} |x_i - y_i| \qquad (4)$$

4) For each data, it is allocated the smallest distance among all the distances of M groups.

5) The data is sorted increasingly according to distance

6) We select the K data that have the smallest distances where n is the number of attributes, x and y are coordinates of barycenter and the data respectively and i ∈ [1...number of attributes].

- **Third phase : labeling phase**

The K similar data that is obtained from the second phase will be labeled by the *n* classifiers, where each classifier assigns a probability for a data, by knowing each class. This probability is of the form $P\ (data/class_h, C)$:

$$\sum_{h=1}^{numberofclass} P\ (data/class_h, C) = 1 \qquad (5)$$

Where h ∈ [1, number of class].

- **Fourth phase: combination of classifiers**

Our method is based on weighted vote. This is a vote based on weights associated with basic classifiers. All classifiers must label each data. Since each classifier assigns for this data a probability of belonging to each class, this data takes a class h if the probability of belonging to this class is maximum, with respect of the other classes. Then, using this formula (formula 6), we can easily know the class attributed by a classifier c to the instance:

$$P'\ (data/class_h, C_j) = \text{arg-max}\ P\ (data/class_h, C_j) \qquad (6)$$

For each classifier, we multiply its maximum probability of belonging to a class h for a data by its confidence score (according to formula 1) to obtain a calculated weight of a data for its belonging to a class:

$$\text{Weight}\ (data/class_h, C_j) = \text{Score}\ (C_j) * \text{arg-max}\ P'\ (data/class_h, C_j) \qquad (7)$$

For each class, we sum up the weights:

$$S\ (class_h) = \sum \text{Weight}\ (data/class_h, C_j) \qquad (8)$$

The final labeling of an instance is done by assigning the class whose S '(classe$_h$)is maximal.

$$S'\ (class_h) = \text{Arg-max}(S\ (class_h)) \qquad (9)$$

- **Fifth phase: update of all corpus**

As a last step of our method, we add the K labeled data by the n classifiers to the training kernel and remove them from the corpus of unlabeled data.

The third module (Semi supervised training process) is repeated until the unlabeled data disappears. After having presented the architecture of our system, we present the different results obtained when applying our system on our medical corpus.

4. Experimentations and results

The Berkeley corpus has been used by several researchers. Some research studies treated only three relations (Cure, Prevent, SideEffect), others are based on seven relations. For this, we propose to treat these two cases. The statistical study of our corpus reveals the following characteristics[7] as represented in Table 2.

Relations	Number of instances
Cure	783
Only DIS	600
Only TREAT	163
Prevent	58
Vague	45
SideEffect	41
No Cure	4

Table 2.*The statistical characteristics of our medical base*

Our corpus is divided into three sub-corpora: the first sub-corpus is a training kernel, the second is a test sub-corpus, and the third is a sub-corpus of unlabeled data. Table 3 shows the amount of instances allocated for each sub-corpus. We choose to assign 347 instances for the test; it presents the same number of instances used by [4].

Number of instances in the training kernel	229
Number of instances in the test corpus	347
Number of unlabeled data	1080

Table 3.*The distribution of data for each sub-corpus*

We treat the relation extraction task in two cases: the first where the two NEs are not recognized, and the second

[7] The number of sentences available for downloads is not the same as the ones from the original data set published in [Rosario and Hearst, 2004]

when the NEs are recognized. Indeed, the recognition of the types of medical NE facilitates certainly the identification of the type of interacting relationship between these NEs, as illustrated in example 1 and 2:

<DIS_PREV> Measles </DIS_PREV><TREAT_PREV> vaccination </TREAT_PREV> and inflammatory bowel disease (1).

Over half thought <DISONLY> **HIV**</DISONLY> transmission occurred most times or always (2).

As illustrated in example 1, the recognition of the two types of NEs (DIS-PREV and TREAT_PREV) makes it easy to identify the "Prevent" relationship without the need to apply a training algorithm. The same for example 2, when knowing that the NE is annotated by DISONLY, it will be easy to know that the relation is disonly.

We evaluate our method using many classifiers and the best result is obtained by the classifier PART. We use this classifier from WEKA project which is implemented in Java [23]. The parameters used in our system are *n*=3 (number of relevant subsets of attributes), K=10% (number of unlabeled data selected in each iteration).

Afterwards, we evaluate three semi-supervised algorithms Self-training, Co-training and YATSI [24] on the used medical corpus. YATSI contains two steps. It can use any training algorithm (Wrapper algorithm) with the nearest neighbor algorithm. In the first step, an algorithm is trained on the labeled data to construct a training model. Then, this model is used to label unlabeled data. In the second step, the nearest neighbor algorithm is applied using the initial labeled data and the newly labeled data. In addition to the initial classification algorithm, the nearest neighbor algorithm is used to adjust or correct labels assigned to unlabeled data. For the YATSI implementation, we use collective classification package available from MARSDEN project[8] Programs which are written and tested in Java programming language in Eclipse environment. While for Self-training and Co-training, we implement them. To the best of our knowledge, there is no study that has adopted the semi-supervised method using this corpus. Thus, our proposed method represents the first semi-supervised work using the corpus of [4].

As reported in Table 6, we note that our method is more efficient than the three semi-supervised methods considered. Indeed, with the base of 7 relations without recognizing the NEs, we get an accuracy of 83.20%, while with YATSI we get 54%.When applying Self-training method, we get 31.38%. The application of Co-training on the same corpus obtains 57.06% of accuracy. By knowing the types of NEs, the value of accuracy is high for the four algorithms. Here, it is not logical to use NE types as attributes because we have seen previously that the recognition of NEs types facilitates the task of relation extraction and we do not even need to apply a training algorithm.

The originality of our semi-supervised method can be assumed in the contributions acquired, of each phase. For the selection phase of the relevant subset of attributes, we used filtering methods to evaluate the attributes of the data rather than their interactions with a particular classifier. The evaluation of these methods shows more generality (they do not depend on the classifier). Each subset of attributes is the best according to a criteria to be optimized (information gain, ACP,..).The subsets obtained according to the various evaluation criterions may be overlapping or independent. At this level, there are no assumptions or conditions on the relationship between the subsets of attributes found, unlike other works like the work of [11] where they defined a condition that concerns the independence between the subsets of attributes. For the labeling phase of the unlabeled data, we intend to improve it somewhat by selecting in each iteration the most similar data to the labeled data. In addition, our method combines several classifiers. The combination of these classifiers is based on the weighted sum vote. As a conclusion, we can mention that the combination of several classifiers improves the classification results better than the Self-training method that uses only one classifier.

5. Conclusion

In this paper, we described our semi-supervised system SMRE to extract relations between medical NEs .The obtained results are encouraging and promising since the used techniques allowed us to reach a classification rate equal to 100% when we recognize the NEs , and 83.20% without recognizing this information. Using our semi-supervised method, we were able to resolve the labeling problem using semi-labeled corpus. Our proposed method treats only binary relations (between two NEs). As a future works, we intend to treat relations holding between more than two NEs. In addition, we are planning to evaluate our system with other NEs types and different corpora languages and domains.

[8] http://www.cs.waikato.ac.nz/fracpete/projects/collective classification/

Medical corpus	Self-training	Co-training	YATSI	SMRE
Base with 3 relations (NEs are un known)	57.22%	86.67%	82.87%	**90%**
Base with 3 relations (NEs are known)	70%	**100%**	92.72%	**100%**
Base with 7 relations (NEs are unknown)	31.38%	57.06%	54%	**83.20%**
Base with 7 relations (NEs are known)	65.68%	93.95%	**100%**	**100%**

Table 6.*Comparison of our system with other semi-supervised systems in term of accuracy*

References

1. Iftene A. and Balahur-Dobrescu A. ,‘’ NamedEntity Relation Mining Using Wikipedia’’, In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). 28-30 May, Marrakech, Morocco, 2008.
2. Yu X. and Lam W.,’’Jointly IdentifyingEntities and Extracting Relations in Encyclopedia Text via A Graphical Model Approach’’, In Proceedings COLING (Posters), pages 1399–1407, 2010.
3. Nakamura-Delloye Y. and Stern R., ‘’ Extraction de relations et de patrons de relations entre entités nommées en vue de l’enrichissement d’une ontologie’’, Actes de TOTh 2011 (Terminologie & Ontologie : Théories et applications), Annecy, France , 2011.
4. Rosario B. and Hearst M.,’’Classifying semantic relations in bioscience text’’. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), Barcelona. July 2004.
5. Embarek M. and Ferret O.,’’ Une expérience d’extraction de relations sémantiques à partir de textes dans le domaine médical’’, In TALN 2007, pages 37– 46, Toulouse, France, 2007.
6. Kim J., Choe Y. and Mueller K.,’’ Extracting Clinical Relations in Electronic Health Records Using Enriched Parse Trees’’, INNS Conference on Big Data 2015 Program San Francisco, CA, Pages 274-283 , USA 8-10, 2015.
7. X. Zhu,’’Semi-supervised training tutorial’’,Technical report, Department of Computer Sciences University of Wisconsin, Madison, USA, 2007.
8. Agrawala A.,’’Training with a probabilistic teacher”, IEEE Transactions on in- formation Theory, Vol. 16, pp.373-379, 1970.
9. Clark S., Curran J. R. and Osborne M. ,’’Bootstrapping postaggers using unlabelled data’’. In CoNLL, 2003
10. Drury B. and Delopes A. ,’’the identification of indicators of sentiment using a multi-view self-training algorithm’’, Oslo Studies in Language 7(1), 379–395. (ISSN 1890-9639 / ISBN 978-82-91398-12-9), 2015.
11. Blum A. and Mitchell T. ,’’Combining labeled and unlabeled data with co-training’’, COLT: Proceedings of the Workshop on Computational Training Theory, 1999.
12. Zhou Z. and Li M.,’’Tri-training: exploiting unlabeled data using three classiers’’, IEEE Transactions on Knowledge and Data Engineering, 17:15291541, 2005
13. Jimeno Y. A and Aronson A.,’’Self-training and Co-training in biomedical word sense disambiguation’’,Proceedings of BioNLP 2011 Workshop. 2011, Portland, Oregon, USA: Association for Computational Linguistics, 182-183, 2011.
14. Feng Q., Gui Y., Yang Z . and Wang L., Li Y. ,’’Semisupervised Training Based Disease-Symptom and Symptom-Therapeutic Substance Relation Extraction from Biomedical Literature’’,Hindawi Publishing Corporation BioMed Research International Volume 2016, Article ID 3594937, 13 pages, 2016.
15. Dejori M., M. Bundschus, S. Martin, T.Volker and Hans-Peter K.,’’Extraction of semantic biomedical relations from text using conditional random fields’’, BMC Bioinformatics. 9: 207-10.1186/1471-2105-9-207, 2008.
16. Frunza O. and Inkpen D.,’’Extraction of disease-treatment semantic relations from biomedical sentences’’. In : Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, Uppsala, Sweden. Association for Computational Linguistics; pp. 91–8,2010.
17. Ben Abacha A. and Zweigenbaum P. ,’’A Hybrid Approach for the Extraction of Semantic Relations from MEDLINE Abstracts’’, In Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing’11), volume 6608 of Lecture Notes in Computer Science, pages 139-150, Tokyo, Japan,. 2011
18. John G.H., Kohavi R. and Pfleger K.,’’Irrelevant features and the subset selection problem’’. In Proceedings of the Eleventh International Conference on Machine Training, pages 121–129, 1994.
19. Liu H., Yu L.,’’Toward Integrating Feature Selection Algorithms for Classification and Clustering’’, Department of Computer Science and Engineering,Arizona State University, 2005.
20. Navin T., Chappelle O., Weston J. and Elisseeff A.,’’Fearutre extraction, foundations and applications, chapter Embedded Methods’’, Series Studies in Fuzziness and Soft Computing, PhysicaVerlag, Springer, pages 139-167, 2006.
21. Dash M. and Liu H.,’’Feature selection for classification’’, Intelligent Data Analysis, 1 :131-156, 1997.
22. Jolliffe T.,’’Principal Component Analysis’’, Springer-Verlag, 1986.
23. Hall M., Frank E ., Holmes G., Pfahringer B . , Reutemann P .,and Witten IH.,’’The weka data mining software: an update’’. ACM SIGKDD Explorations Newsletter;11(1):10–18, 2009.
24. Driessens K., Reutemann P., Pfahringer B. and Leschi C.,’’Using weighted nearest neighbor to benefit from unlabeled data’’, PAKDD. Mar;60–69,2006