

A Probabilistic Approach to Study Features in Opinion Mining using Fuzzy Selection

¹Mrunal Suhas Pendharkar, ²Sweta Kale

¹M.E. Student, ²Head of Dept.

¹Information Technology,

¹RMD Sinhgad College of Engineering, Pune, India

Abstract— Opinion mining (also called as sentiment analysis) aims to understand people's thinking towards entities such as products like mobiles, etc. In opinion mining, an opinion word, or feature in short, extracts a thing or a feature of an entity on which users state their views. The proposed model is a new approach to the recognition of such features from unstructured textual reviews. The opinion features are assigned weights or frequency which is used to state the nature of reviews. It is, thus, important to extract the specific opinionated features from text reviews and associate them to opinions. Using Fuzzy selection techniques, the extracted words are set under a range of 0 to 1 which improves accuracy in sentiment analysis. Finally, after finding an association set of opinion words and product words, the web based application which crawls reviews from websites, posts the classification of reviews.

Index Terms—Information search and retrieval, opinion mining, sentiment analysis, natural language processing, fuzzy selection.

I. INTRODUCTION

Opinion mining also called as sentiment analysis helps to analyse people's views, thoughts, and attitudes towards products, services, and their attributes [1] [2]. Sentiments or views expressed in textual reviews are typically analysed at various levels. For example, sentence-level opinion mining helps to identify the overall bias or sentiment expressed on an entity (e.g., mobile or hotel) in a review document, but it does not correlate opinions with specific aspect (e.g., display, battery) of the entity. In opinion mining, an opinion word, or feature indicates a thing or an attribute of an entity on which user's state in their opinions.

The proposed model is a novel approach to the identification of such features from unstructured textual reviews. Supervised learning model may be tuned to work well in a given domain, but the model must be analyzed if it is applied to different domains [6], [5]. Unsupervised natural language processing (NLP) approaches [9], [7], [4] identify opinion features by defining domain-independent syntactic templates or rules that capture the dependence roles and local context of the feature terms. One key finding is that the distributional structure of an opinion feature in a given domain-dependent review corpus, for example, mobile reviews, is different from that in a domain-independent corpus.

The proposed method is summarized as follows: Firstly, the reviews are extracted and stored in a database and pre-processed, for example, mobile or hotel reviews. Later, using TF-IDF, frequently occurring words are extracted and assigned weights to each word for deciding priority. Then, for each predictable feature candidate, its domain relevance score with respect to the domain-specific and domain independent databases is calculated. Finally, candidate features with low TF-IDF scores and high TF-IDF scores are eliminated. Evaluations observed on two real-world review domains give the effectiveness of the proposed IEDR approach in identifying opinion features.

In a typical fuzzy classifier system, the classification is explicitly described by a number of fuzzy If-Then rules. A fuzzy rule may look like IF X is *Small* AND Y is *Large* then Class1, where X and Y are features and *Small* and *Large* are fuzzy sets. In each classification rule each feature may be described by different fuzzy sets such as *SMALL/LOW*, *MEDIUM*, *LARGE/HIGH*, etc. The proposed method by which optimal fuzzy sets can be selected automatically by using conventional search techniques and a representative labelled data set is used [8].

For example, consider a review of a mobile by a customer, "The battery and features of the iPhone are admirable with advanced technology but the price is too costly". Here the overall review is positive but the aspects 'price' expresses a negative opinion. Hence, a fine grained approach is required to extract the appropriate nature of review.

II. NEED

One key need is that the users should find convenience in decision making while buying any product or visiting any hotel. Hence to classify the reviews posted online, it should be necessary to build a system which classifies the summary of reviews. This leads to propose a method to identify opinion features by exploiting their distribution disparities across different corpora. The purpose of opinion mining is to sort the overall bias or sentiment expressed in an individual review document. On any internet website, reviews by users or manufactures are randomly shown; hence, for any passionate user it is difficult to understand exact review of the product. Using opinion mining, it is easy to distinguish reviews into positive and negative review.

III. RELATED WORK

Opinions and sentiments expressed in text reviews can be generally analyzed at the document, sentence, or even phrase (word) levels. The purpose phrase-level opinion mining is to classify the overall bias or sentiment expressed in an individual review document (sentence). To prevent a sentiment classifier from considering irrelevant or even potentially misleading text, Pang and Lee [12] proposed to first employ sentence-level subjectivity detector to identify the sentences in a document as either subjective or objective, and later discarding the objective ones. Then the resulting subjectivity extract was applied via sentiment classifier, with improved results. Review rating assessment is a much more complicated problem compared to binary sentiment organization. Generally, opinions are expressed in a different way in different domains. The sentiment classification methods mentioned above can be tuned to work very well on a given domain; however, they may fail to classify opinions in a different domain.

Bollegala [2] proposed a fully regular method to create a thesaurus that is responsive to the sentiment of words expressed in different domains. It utilizes all data which can be labelled or unlabeled available for the source domains and unlabeled data from the target domain. A basic problem when applying a sentiment classifier trained on a particular domain to classify reviews on a different domain is that words (hence features) that appear in the reviews in the target domain do not always appear in the trained model. To prevail over this feature difference problem, the author constructed a sentiment aware thesaurus that stores the relatedness of words as used in different domains of databases.

Bing Liu [7] studied the problem of generating feature-based summaries of customer reviews of products sold online. Here, features broadly mean product features (or attributes) and functions. Given a set of user reviews of a particular product, the task involves three subtasks: identify features of the product that customers have expressed their opinions on (called product features); for each feature, classifying review sentences that give positive or negative opinions; and producing a summary using the discovered information. B. Liu also proposed an association rule mining (ARM) approach to mine frequent item sets as probable opinion features, which are nouns and noun phrases with high sentence-level frequency (or support). However, ARM, which relies on the frequency of item sets, has the following limitations for the task of feature identification, frequent but invalid features are extracted incorrectly, and rare but valid features may be overlooked.

IV. METHODOLOGY

An opinion feature such as “battery” in mobile reviews is typically domain-specific. That is, the feature appears often in the given review domain, and infrequently outside the domain such as in a domain-independent corpus. As such, domain-specific opinion features will be mentioned more often in the domain databases of reviews, compared to a domain-independent corpus.

Figure.1 shows the workflow of our proposed method. Given a domain-dependent review database and a domain independent database, first a list of candidate features from the review corpus are extracted by manually defined syntactic rules defined in TABLE I. For each extracted candidate feature, its IDR is calculated, which represents the statistical relation of the candidate to the given domain corpus, and its extrinsic relevance, which gives the statistical relation of the candidate to the domain-independent database.

Domain relevance states how much a term is related to a particular corpus (i.e., a domain) based on two kinds of statistics, namely, dispersion and deviation. Dispersion states how significantly a term is mentioned across all documents by measuring the distributional significance of the term across different documents in the entire corpus (horizontal significance).

Identify Opinion Features

Deviation reflects how frequently a term is mentioned in a particular document by measuring its distributional significance in the document (vertical significance). Both dispersion and deviation are calculated using the well-known term frequency-inverse document frequency (TF-IDF) term weights. Each term T_i has a term frequency TF_{ij} in a document D_j , and a global document frequency DF_i . The weight w_{ij} of term T_i in document D_j is then calculated as follows:

$$w_{ij} = \{(1 + \log TF_{ij}) * \log N / DF_i \text{ if } TF_{ij} > 0\}, \quad (1)$$

$$\text{otherwise } 0$$

where i are total number of terms and j are total number of documents in the corpus.

Dispersion thus measures the normalized average weight of term T_i . It is high for terms that appear frequently across a large number of documents in the entire corpus. The dispersion $disp_i$ of each term T_i in the corpus is defined as follows:

$$disp_i = w_i / s_i \quad (2)$$

where s_i is standard deviation.

Deviation dev_{ij} indicates the degree in which the weight w_{ij} of the term T_i deviates from the average w_{ij} in the document D_j . The deviation thus characterizes how significantly a term is mentioned in each particular document in the corpus. The deviation dev_{ij} of term T_i in document D_j is given by

$$dev_{ij} = w_{ij} - w_j \quad (3)$$

where the average weight w_j in the document D_j is calculated over all M terms as follows:

$$w_j = 1/M \sum w_{ij}$$

The domain relevance dr_i for term T_i in the corpus is finally defined as follows:

$$dr_i = disp_i * \sum dev_{ij} \tag{4}$$

Clearly, the domain relevance dr_i incorporates both horizontal (dispersion $disp_i$) and vertical (deviation dev_{ij}) distributional significance of term T_i in the corpus which reflects the ranking and distributional characteristics of a term in the entire corpus

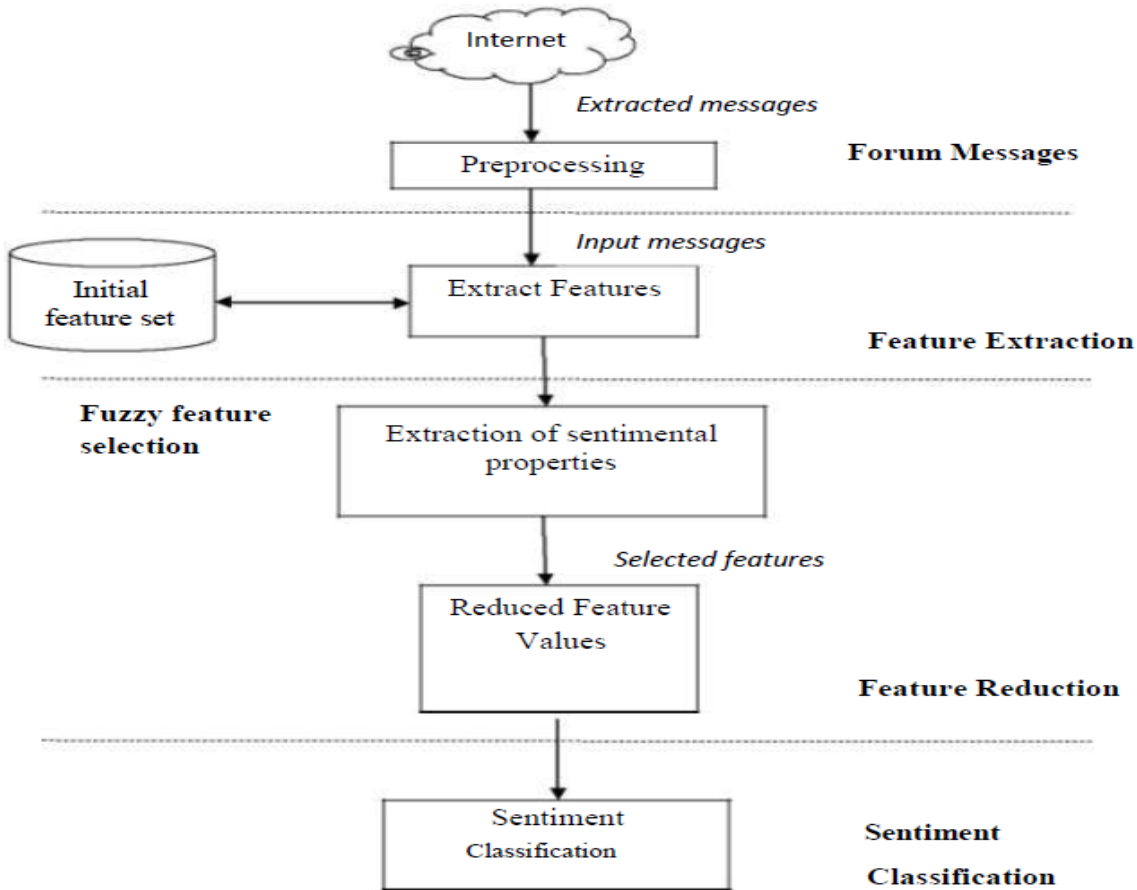


Figure 1. The Proposed Architecture

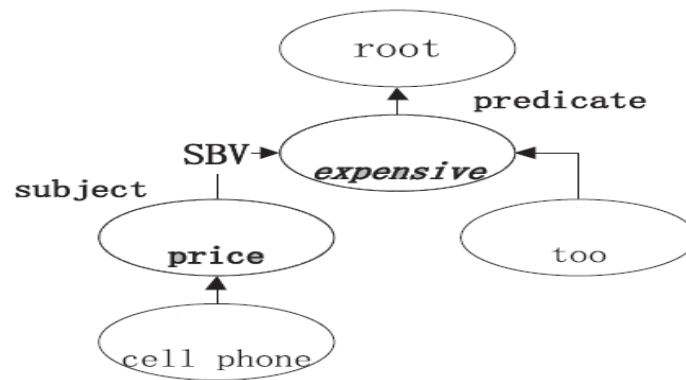
Extracting Opinion Features

The reviews are extracted from any website using web crawler and opinion features are to be identified. Opinion features are generally nouns or noun phrases, which normally appear as the subject or object of a review sentence. In the case of dependence grammar [14], the subject opinion feature has a syntactic relationship of type subject-verb (SBV) with the predicate. The object opinion feature has a dependence association of verb-object (VOB) on the predicate. In addition, it also has a dependence association of preposition-object (POB) on the prepositional word in the sentence.

Table I. Syntactic Rules

Rules	Interpretation
NN+SBV→CF	Identify NN as a CF, if NN has a SBV dependency relation
NN+VOB→CF	Identify NN as a CF, if NN has a VOB dependency relation
NN+POB→CF	Identify NN as a CF, if NN has a POB dependency relation

Given example illustrates the corresponding dependence tree in Figure. 2. As shown in example, the opinion feature “price” (underline), which is associated with the adjective “expensive” (*italic*), is the subject of the sentence. It has a dependence association of SBV with the adjective predicate. From various dependence relations we can represent three syntactic rules “NN” and “CF” denotes nouns phrases and candidate features.



(The price of the cellphone is too *expensive*.)

Figure 2. Example of dependence tree

The candidate feature extraction process works in the following steps: 1) Dependence parsing (DP) is first in use to identify the syntactic structure of each sentence in the given review database; 2) the three rules are applied to the identified dependence structures, and the equivalent nouns or noun phrases are extracted as candidate features whenever a rule is fired. There could be many invalid features in the extracted candidate feature list; the next step is to prune the list using proposed algorithms.

Aspect Based Mining Fuzzy Algorithm

The procedure for computing the domain relevance is the same regardless of the corpus, as concise in Algorithm 1. When the procedure is applied to the domain-specific review database, the scores are called IDR, otherwise they are called EDR. Candidate features with overly high TF-IDF(IDR) scores or gloomily low TF-IDF(EDR) scores are eliminated using the inter-corpus criterion of IEDR. Algorithm 1 gives the proposed IEDR method, where the minimum IDR threshold i^{th} and maximum EDR threshold e^{th} can be determined experimentally.

Algorithm 1:

Input:

- Domain review corpus R and domain independent corpus D

Output: A validated list of opinion features with the nature of reviews

Method:

- Begin
- Extract candidates from the review corpus R;
- For each candidate feature CF_i do
 - Compute IDR score idr_i using (1), (2), (3), (4) on the review corpus R;
 - Compute EDR score edr_i using (1), (2), (3), (4) on domain independent corpus D;
- If $(idr_i \geq i^{\text{th}})$ AND $(edr_i \leq e^{\text{th}})$ then
 - Confirm candidate CF_i as a feature
- End for
- Return a validated set of opinion features.

Algorithm 2:

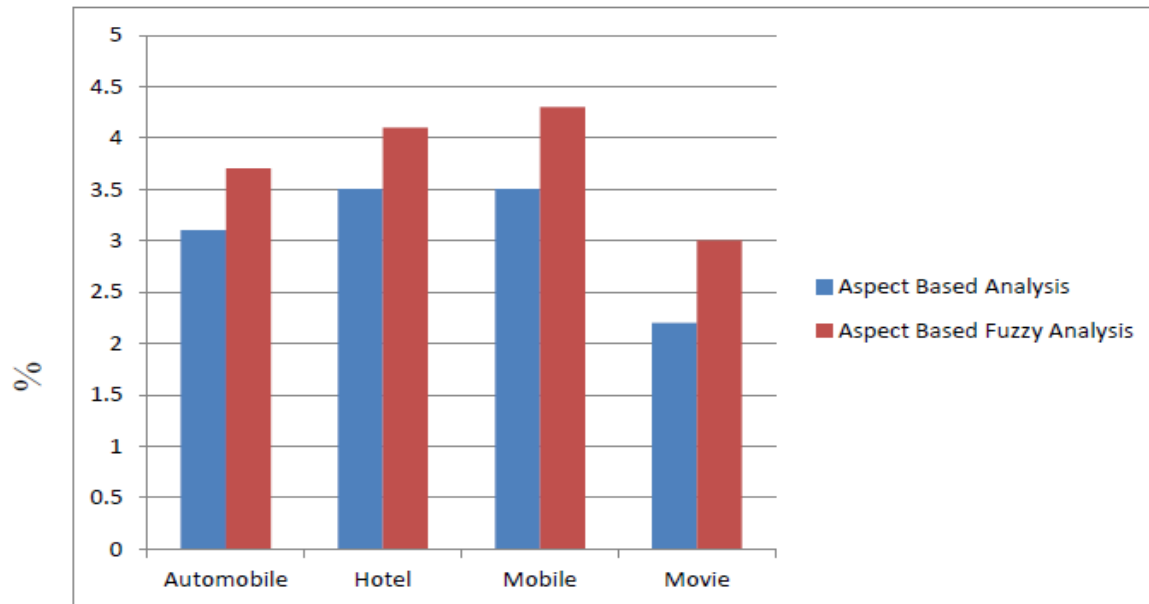
The fuzzy feature selection can be defined by the following pseudo algorithm:

1. Project a labeled data set $X\{x_i/ i=1,2,\dots, n\}$ onto a fuzzy set F_X defined by $U\{u_{ij} | i=1, 2,\dots, p \wedge j=1, 2,\dots, |f_i| \}$. This projection may be defined by linear (e.g. triangular membership function) or nonlinear (e.g. exponential, sigmoid) functions;
2. define a classifier, a criterion function $J(\cdot)$ and $\theta=2^U$;
3. use a feature selection method FS;
4. Find U_{optimal} by using FS in such a way that $J(U_{\text{optimal}})=E(J(U_i)); \forall U_i \in \theta$.

V. EXPERIMENTAL RESULTS

Various experiments are conducted to comprehensively evaluate the proposed system performance on two real-world review domains, mobile and hotel reviews. First reviews are extracted from a website using crawler mainly from Flipkart and Tweeter. Then using sentiment analyzer frequently occurring features are extracted which is pre-processing step. Secondly, the weights of opinion words is calculated using term frequency tool (TF-IDF) which assigns a weight comparing it to a threshold of initial

testing on a database. By using Fuzzy Selection techniques the data set term frequency values are set between the ranges of 0 to 1. Also they are categorized into three fuzzy sets as Low, Medium and High which improves accuracy while classifying sentiment words extracted. These frequencies are required to build a training data set so that it is not necessary to calculate it each time when a sentiment word is added in the database. If any word is already present, comparing with threshold, its priority is increased if frequently occurring in review sentences. Then using Aspect Based Sentiment Analysis algorithm, the nature of reviews is stated. Comparison between Aspect based analysis algorithm and aspect based fuzzy analysis algorithm is represented using graphs shown in fig. 3. The proposed system is a web-based system where user can view their selected product's specifications and reviews posted by other users. There are two modules where one contains an admin part and one is user part. Admin can manage product list, can add, delete product from server database, while users can only post their reviews and view product list. This is for mobile database. In case of hotel database, users can view specification of any hotel, city and reviews of hotel. The proposed system does not require manual annotation as Stanford Parser is used while existing systems require manual annotation.



Data set

Figure 3. Comparison of Existing and Proposed algorithms for various datasets.

Discussion

The opinion feature extraction performance of proposed method (as well as all the competing methods) on the hotel reviews is not as good as that on the mobile reviews. This is because hotel reviews

- are longer and more complex, which makes feature mining much more challenging, and
- Contain large number of noisy domain-unrelated user reviews, which overlap with documents in the domain-independent corpus. In other words, the division between the intrinsic-domain relevance and the extrinsic-domain relevance of a feature is not that clear-cut for the hotel reviews.

As a result, proposed method (and all other competitors) is less successful in identifying features on the hotel domain compared to the mobile domain. It is, thus, very difficult to choose the right domain-independent corpus. According to the experiments carried out, neither corpus size nor topic number in corpus has a large effect on aspect based algorithm feature extraction, but the different nature of the domain-independent corpus/topic from the given review domain makes a big difference.

VI. CONCLUSION

The proposed model is an original inter-related data approach to opinion feature classification based on the IEDR feature-filtering principle, which uses the disparities in distributional description of features across two databases, one related domain and one correlated domain. IEDR identifies candidate description that are specific to the given review domain and yet not overly basic (domain independent). Experimental results reveal that the proposed IEDR not only leads to noticeable improvement over either IDR or EDR, but also overcomes the limitations observed in four conventional methods, namely, LDA, ARM, MRC, and DP, in terms of feature withdrawal performance as well as opinion mining results. Using fuzzy selection techniques, accuracy of algorithm is increased. Moreover, as domain related database is important, it also requires good size data and topic selection on features on which the performance is calculated. Using a domain-independent corpus of a similar size will give way good opinion feature extraction results are found.

VII. ACKNOWLEDGMENT

First and foremost, I would like to express my sincere gratitude to my guide and Head of the Department Prof. Sweta Kale, for her continuous support in completion of my project and helping me to collect reviews from websites. I will always be grateful for her patience, motivation, enthusiasm, and immense knowledge. This work is an extension of the previous research efforts of Zhen Hai, Kuiyu Chang, Jung-Jae Kim and Chrishtopher C. Yang. I am very much indebted to them for their inspiring work which is a boost for future researchers. Finally, I would like to pay my respect and love to my parents, for their encouragement throughout my career.

REFERENCES

- [1] Hai et al., "Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance", IEEE Transactions on Knowledge and Engineering, Vol 26, No. 3, March 2014.
- [2] D. Bollegala, D. Weir, and J. Carroll, "Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus," IEEE Trans. Knowledge and Data Eng., vol. 25, no. 8, pp. 1719-1731, Aug. 2013.
- [3] B. Liu, "Sentiment Analysis and Opinion Mining," Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1-167, May 2012.
- [4] G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion Word Expansion and Target Extraction through Double Propagation," Computational Linguistics, vol. 37, pp. 9-27, 2011.
- [5] N. Jakob and I. Gurevych, "Extracting Opinion Targets in a Single and Cross-Domain Setting with Conditional Random Fields," Proc. Conf. Empirical Methods in Natural Language Processing, pp. 1035-1045, 2010.
- [6] W. Jin and H.H. Ho, "A Novel Lexicalized HMM-Based Learning Framework for Web Opinion Mining," Proc. 26th Ann. Int'l Conf. Machine Learning, pp. 465-472, 2009.
- [7] G. Qiu, C. Wang, J. Bu, K. Liu, and C. Chen, "Incorporate the Syntactic Knowledge in Opinion Mining in User Generated Content," Proc. WWW 2008 Workshop NLP Challenges in the Information Explosion Era, 2008.
- [8] M. Ramze Rezaee, B. Goedhart, B.P.F. Lelieveldt, J.H.C. Reiber, "Fuzzy Feature Selection", Division of Image Processing, Department of Radiology, Netherlands, Pattern Recognition 32, 2008.
- [9] Q. Su, X. Xu, H. Guo, Z. Guo, X. Wu, X. Zhang, B. Swen, and Z. Su, "Hidden Sentiment Association in Chinese Web Opinion Mining," Proc. 17th Int'l Conf. World Wide Web, pp. 959-968, 2008.
- [10] S.-M. Kim and E. Hovy, "Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text," Proc. ACL/COLING Workshop Sentiment and Subjectivity in Text, 2006.
- [11] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 168-177, 2004.
- [12] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, Proc. 42nd Ann. Meeting on Assoc. for Computational Linguistics, 2004.
- [13] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 342-351, 2004.
- [14] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," J. Machine Learning Research, vol. 3, pp. 993-1022, Mar. 2003.
- [15] L. Tesnière, Elements de la syntaxe structurale. Librairie C. Klincksieck, 1959.