Asian Journal of Computer Science and Technology
ISSN: 2249-0701 Vol.8 No.1, 2019, pp. 32-35

Comprehensive Review on Effectual Information Retrieval of Semantic Drift using Deep Neural Network

A. Uma Maheswari[1] and N. Revathy[2]
[1]Research Scholar, [2]Associate Professor
[1&2]Department of Computer Science, Hindusthan College of Arts and Science, Coimbatore, Tamil Nadu, India
E-Mail: drnrevathy@gmail.com

***Abstract* - Semantic drift is a common problem in iterative information extraction. Unsupervised bagging and incorporated distributional similarity is used to reduce the difficulty of semantic drift in iterative bootstrapping algorithms, particularly when extracting large semantic lexicons. In this research work, a method to minimize semantic drift by identifying the (Drifting Points) DPs and removing the effect introduced by the DPs is proposed. Previous methods for identifying drifting errors can be roughly divided into two categories: (1) multi-class based, and (2) single-class based, according to the settings of Information Extraction systems that adopt them. Compared to previous approaches which usually incur substantial loss in recall, DP-based cleaning method can effectively clean a large proportion of semantic drift errors while keeping a high recall.**

Keywords: **Semantic Drift, Drifting Points, Deep Neural Network, Information Retrieval**

I. INTRODUCTION

Semantic drift is the evolution of word usage - usually to the point that the modern meaning is completely dissimilar from the original procedure. In diachronic linguistics, semantic drift is a change in one of the meanings of a word. Every word has a variety of senses and suggestions, which can be added, removed, or altered over time, often to the extent that relates across space and time have different meanings. The study of semantic change can be seen as part of etymology, onomasiology, semasiology, and semantics. This is important due to the fact that in many applications of text classification a large number of unlabeled texts are easily accessible, while the receipt of marked texts is quite a difficult task. The paper also shows that the convolution neural network can work better at the level of words, and does not require knowledge of the syntactic or semantic structure of the language. Survey results obtained for text extraction from different sources show that using a Deep Neural Network can also improve the accuracy of the extraction. The concept of semantic drift evolves at streams in data mining. The figure 1 shows various data stream mining tasks that can be carried out in continuously evolving data streams.

II. LITERATURE REVIEW

A. Semantic Drift

Yunfeng Zhu, Fernando De la Torre, Jeffrey F. Cohn [2011] proposed the Dynamic Cascades with Bidirectional Bootstrapping for Action Unit Detection in Spontaneous Facial Behavior.

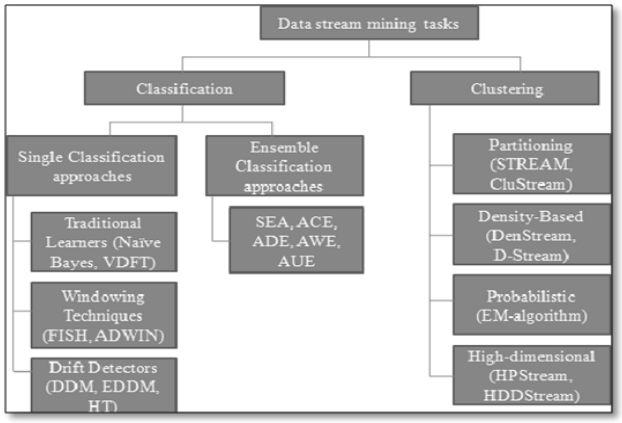

Fig. 1 Data stream mining tasks

They exploit the advantages of feature selection, efficiency, and robustness of Cascade Adaboost. To provide a real-world test, they used the RU-FACS database. Somayeh Kazemi, AyazGhorbani, HamidrezaAmindavar, and Dennis R. Morgan [2016] showed theoretically and experimentally that the bootstrap-based GWT can extract the amplitude and frequency of the two vital-sign components at a range of 3 m in the face of low signal-to-noise ratio and in the presence of phase noise and body motion artifacts, achieving an accuracy that is potentially better than conventional methods can provide. Wing W. Y. Ng,Senior Member, IEEE, Xing Tian, YuemingLv, Daniel S. Yeung [2017] uses a multi hashing to retain knowledge coming from images arriving over time and a weight-based ranking to make the retrieval results adaptive to the new data environment. The current weight-based ranking scheme depends on the performance of each individual hash table and ignores the time at which the hash table has been created. In this paper, they have focused on the semantic image retrieval problems.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q. Zhu [2017] focused on bootstrapping techniques based on syntactic patterns, that is, each iteration finds more syntactic patterns for subsequent extraction. However, syntactic bootstrapping is incapable of resolving the inherent ambiguities in the syntactic patterns. Syntactic patterns as the iteration proceeds, semantic bootstrapping uses a fixed set of patterns. Cheng-Tao Chung, Cheng-Yu Tsai, Chia-

Hsiang Liu, and Lin-Shan Lee [2017] discussed the multiple sets of token labels are then used as the targets of a multi target deep neural network (MDNN) trained on frame-level acoustic features. The results were evaluated using the metrics and corpora defined in the Zero Resource Speech Challenge organized at Inter-speech. The unsupervised tokens were competitive when compared to supervised phoneme recognizer from four other languages on the task of STD.

Peipei Li, Lu He, Haiyan Wang, Xuegang Hu, Yuhong Zhang, Lei Li,andXindong Wu demonstrate that as compared to several well-known concept drifting detection methods in data stream, our approach can detect topic drifts effectively, and it enables handling short text streams effectively while maintaining the efficiency as compared to several state-of-the-art short text classification approaches. Semantic contexts based on the senses of terms hidden in short texts are introduced to make up of the datasparsity and all terms are disambiguated to reduce the noisy impact.Zhixu Li *et al.,* [2018] discussed that most semantic drifts are introduced by a small number of questionable extractions in the earlier rounds of iterations. The experimental results show that the DP cleaning method enables us to clean around 90 percent incorrect instances.

WeichaoShen, Yuwei Wu *et al.,* [2018] explore the use of deep features extracted from the Convolutional Neural Networks (CNNs) to improve the object representation and propose a robust distracter-resistive tracker via learning a multi-component discriminative dictionary. The learned dictionary is more compact and more discriminative, which makes our tracker have better discriminating power to handle appearance changes. Comparisons with 9 state-of-the-art tracking methods on the benchmark dataset have demonstrated that our tracker effectively resists distracters and outperforms existing methods.

B. Deep Neural Network

Muhammad Zain Amin, NomanNadeem .The Classification model presented in this paper is multi-class text classifier. The neural network classifier can be trained on large dataset. They report series of experiments conducted on Convolution Neural Network (CNN) by training it on two different datasets. Neural network model is trained on top of word embedding. Softmax layer is applied to calculate loss and mapping of semantically related words. Gathered results can help justify the fact that proposed hypothetical QAS is feasible.

We further propose a method to integrate Convolutional Neural Network Classifier to an open domain question answering system. The idea of Open domain will be further explained, but the generality of it indicates to the system of domain specific trainable models, thus making it an open domain. The sample neural network comparison between shallow and deep structure is shown in the following figure 2.

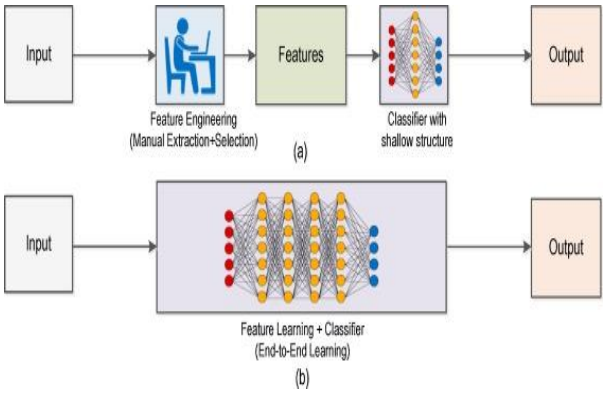

Fig. 2 Neural network

Yu He *et al.,* [2018] proposed that traditional text classification algorithms are based on the assumption that data are independent and identically distributed. However, in most non-stationary scenarios, data may change smoothly due to long-term evolution and short-term fluctuation, which raises new challenges to traditional methods. The authors present the first attempt to explore evolutionary neural network models for time-evolving text classification. They first introduce a simple way to extend arbitrary neural networks to evolutionary learning by using a temporal smoothness framework, and then propose a diachronic propagation framework to incorporate the historical impact into currently learned features through diachronic connections. Experiments on real-world news data demonstrate that our approaches greatly and consistently outperform traditional neural network models in both accuracy and stability.

Du and Huang [2018] research work was mainly based on the classification of keywords and neural network semantic synthesis classification. The former emphasizes the role of keywords, while the latter focuses on the combination of words between roles. The method proposed in this paper considers the advantages of both methods. It uses an attention mechanism to learn weighting for each word. Under the setting, key words will have a higher weight, and common words will have lower weight. Therefore, the representation of texts not only considers all words, but also pays more attention to key words. They feed the feature vector to a soft max classifier. They finally conduct experiments on two news classification datasets published by NLPCC2014and Reuters, respectively. The proposed model achieves F-values by 88.5% and 51.8%on the two datasets. The experimental results show that our method outperforms all the traditional baseline systems.

Siwei Lai, LihengXu, Kang Liu, Jun Zhao [2015] insists that Text classification is a foundational task in many NLP applications. Traditional text classifiers often rely on many human-designed features, such as dictionaries, knowledge bases and special tree kernels. In contrast to traditional methods, they introduce a recurrent convolutional neural network for text classification without human-designed features. In the proposed model, the authors apply a recurrent structure to capture contextual information as far

as possible when learning word representations, which may introduce considerably less noise compared to traditional window-based neural networks. The authors employ a max-pooling layer that automatically judges which words play key roles in text classification to capture the key components in texts. Experiments on four commonly used datasets show that the proposed method out performs the state-of-the-art methods on several datasets, particularly on document-level datasets.

Alexis Conneau *et al.,* [2017] present a new architecture (VDCNN) for text processing which operates directly at the character level and uses only small convolutions and pooling operations. They are able to show that the performance of this model increases with the depth: using up to 29 convolutional layers, we report improvements over the state-of the-art on several public text classification tasks. To the best of our knowledge, this is the first time that very deep convolutional nets have been applied to text processing.

BaoxinWang [2018] Recurrent neural network (RNN) has achieved remarkable performance in text categorization. RNN can model the entire sequence and capture long-term dependencies, but it does not do well in extracting key patterns. In contrast, convolutional neural network (CNN) is good at extracting local and position-invariant features. In this paper, we present a novel model named disconnected recurrent neural network (DRNN), which incorporates position-invariance into RNN. By limiting the distance of information flow in RNN, the hidden state at each time step is restricted to represent words near the current position. The proposed model makes great improvements over RNN and CNN models and achieves the best performance on several benchmark datasets for text categorization**.**

MaazAmajd, Zhanibek Kaimuldenov and Ilia Voronkov[2018] analyze the use of different neural networks for the text classification task. The accuracy of the studied text classifiers can be changed by a small number of previously classified texts. This is important due to the fact that in many applications of text classification a large number of unlabeled texts are easily accessible, while the receipt of marked texts is quite a difficult task. The paper also shows that the convolution neural network can work better at the level of words, and does not require knowledge of the syntactic or semantic structure of the language. On the other hand, a recurrent neural network for the level of data representation in the form of a sequence can effectively classify the text. Experimental results obtained for text corpora from two different sources show that using a vector data representation can also improve the accuracy of the classification.

III. PROPOSED WORK

In Information Extraction (IE), iterative bootstrapping is the most extensively used. One of the biggest issues of iterative information extraction is semantic drift. As iterations proceed, the extractions may shift from the target class to some other classes. State-of-the-art iterative IE methods can be divided into two categories, syntax-based and semantic-based, both of which have the semantic drift problem. The most semantic drifts are introduced by a small number of questionable extractions in the earlier rounds of iterations. These extractions subsequently introduce a large number of questionable results, which lead to the semantic drift phenomenon. These questionable extractions are called as Drifting Points (DPs). If erroneous extractions are the "symptoms" of semantic drift, then DPs are the "causes" of semantic drift. In previous researches, a DP cleaning method has been proposed to minimize semantic drift by identifying the DPs and removing the effect introduced by the DPs. A semi-supervised and multi-task learning based on a small number of automatically labeled training data was used. This method not only leverages unlabeled data for a better understanding of new data, but also exploits the knowledge in related concepts to improve the classifier learning for each concept. However, the computational complexity of this model was high. As a result, in this proposed work, a Deep Neural Network (DNN) model based automatic diagnosing and minimizing the semantic drift in IE. Based on this model, the incorrect extraction by DPs will be automatically identified and minimized.

IV. CONCLUSION

In the proposed research work, a method to minimize semantic drift by identifying the (Drifting Points) DPs and removing the effect introduced by the DPs is proposed. Semantic drift is a problem in information extraction. Several research papers for effective information retrieval using deep neural network model was analyzed in this paper. Based on the review it is concluded that the proposed methodology not only works on the unlabeled data, but also exploits the knowledge in related concepts to improve the classifier learning for each concept. The proposed Deep Neural Network (DNN) model is used to inculcate the automatic identification and minimization of disambiguation. In future the highlighted methodology is to be implemented for recognizing incorrect patterns compared to the methodologies discussed in the literature review.

REFERENCES

[1] [Online] Available: www.wikipedia.com.
[2] Maaz Amajd, Zhanibek Kaimuldenov and Ilia Voronkov, "Text Classification with Deep Neural Networks".
[3] Yunfeng Zhu, Fernando De la Torre, Jeffrey F. Cohn, "Dynamic Cascades with BidirectionalBootstrapping for Action Unit Detection in Spontaneous Facial Behavior", *IEEE*, April-June 2011.
[4] SomayehKazemi, AyazGhorbani, HamidrezaAmindavar, and Dennis R. Morgan, "Vital-Sign Extraction Using Bootstrap-Based Generalized Warblet Transform in Heart and Respiration Monitoring Radar System", *IEEE*, Feb. 2016
[5] Wing W. Y. Ng, Senior Member, IEEE, Xing Tian, YuemingLv, Daniel S. Yeung, "Incremental Hashing for Semantic Image Retrieval in Non-stationary Environments", *IEEE*, Nov. 2017
[6] Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q. Zhu, "Semantic Bootstrapping: A Theoretical Perspective", *IEEE*, Feb. 2017.

[7] Cheng-Tao Chung, Cheng-Yu Tsai, Chia-Hsiang Liu, and Lin-Shan Lee, "Unsupervised Iterative Deep Learning of Speech Features and Acoustic Tokens with Applications to Spoken Term Detection", *IEEE*, Oct. 2017.

[8] Peipei Li, Lu He, Haiyan Wang, Xuegang Hu, Yuhong Zhang and Lei Li,andXindong Wu, Learning From Short Text Streams with Topic Drifts", *IEEE*, September 2018.

[9] Zhixu Li, Ying He, BinbinGu, An Liu, Hongsong Li, Haixun Wang, and Xiaofang Zhou, "Diagnosing and Minimizing Semantic Drift in Iterative Bootstrapping Extraction", *IEEE*, May 2018.

[10] WeichaoShen, Yuwei Wu*, Junsong Yuan, Lingyu Duan, Jian Zhang Senior and YundeJia, "Robust Distracter- Resistive Tracker via Learning a Multi-Component Discriminative Dictionary", *IEEE*, 2018.

[11] Muhammad Zain Amin, Noman Nadeem, "Convolutional Neural Network: Text Classification Model for Open Domain Question Answering System".

[12] Yu He, Jianxin Li, Yangqiu Song, Mutian He, HaoPeng, "Time-evolving Text Classification with Deep Neural Networks", *IJCAI*, 2018.

[13] C. Du and L. Huang, "Text Classification Research with Attention-based Recurrent Neural Networks", International Journal of Computers Communications & Control, February 2018

[14] Siwei Lai, LihengXu, Kang Liu and Jun Zhao, Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[15] Alexis Conneau, HolgerSchwenk, Yann Le Cun, "Very Deep Convolutional Networksfor Text Classification", *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Vol. 1, Long Papers, pp. 1107–1116, Valencia, Spain, April 3-7, 2017.

[16] BaoxinWang, "Disconnected Recurrent Neural Networks for Text Categorization", *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, (Long Papers), pages 2311–2320 Melbourne, Australia, July 15 - 20, 2018.

[17] MaazAmajd, ZhanibekKaimuldenov and Ilia Voronkov, "Text Classification with Deep Neural Networks", ceur-ws.org/Vol-1989, 2018.