SNLP: Statistical Natural Language Processing

Special Session along with ICIMP 2020
The Fifteenth International Conference on Internet Monitoring and Protection
September 27, 2020 to October 01, 2020 - Lisbon, Portugal

https://www.iaria.org/conferences2020/ICIMP20.html

Sebastião Pais
Computer Science Department
NOVA LINCS and UBI
Covilhã, Portugal
Email: sebastiao@di.ubi.pt

Irfan Khan Tanoli
Computer Science Department
University of Beira Interior
Covilhã, Portugal
Email: irfan.khan.tanoli@ubi.pt

Abstract—**Statistical natural language processing (SNLP) is a field lying in the intersection of natural language processing and machine learning. SNLP differs from traditional natural language processing in that instead of having a linguist manually construct some model of a given linguistic phenomenon, that model is instead (semi-) automatically constructed from linguistically annotated resources, unsupervised and language independent. Methods for assigning part-of-speech tags to words, categories to texts, parse trees to sentences, and so on, are (semi-) automatically acquired using machine learning techniques.**

Keywords–Sentiment Analysis; Extreme Sentiment Analysis; Social Media; Natural Language Processing; Extremism; Collective Radicalization; Controlled Natural Language; Social Network Policies; Natural Language

I. Introduction

The field of natural language processing (NLP) has seen a dramatic shift in both research direction and methodology in the past several years. In the past, most work in computational linguistics tended to focus on purely symbolic methods. Recently, more and more work is shifting toward hybrid methods that combine new empirical corpus-based methods, including the use of probabilistic and informationtheoretic techniques, with traditional symbolic methods. This work is made possible by the recent availability of linguistic databases that add rich linguistic annotation to corpora of natural language text. Already, these methods have led to a dramatic improvement in the performance of a variety of NLP systems with similar improvement likely in the coming years [1].

Statistical NLP as we define it compises all quantitative approaches to automated language processing, including probabilistic modeling, information theory, and linear algebra. While probability theory is the foundation for formal statistical reasoning, we take the basic meaning of the term 'statistics' as being broader, encompassing all quantitative approaches to data (a definition which one can quickly confirm in almost any dictionary). Although there is thus some potential for ambiguity, Statistical NLP has been the most widely used term to refer to non-symbolic and non-logical work on NLP over the past decade, and we have decided to keep with this term [2].

Statistical Natural Language Processing (SNLP) is a special track in ICIMP 2020 to invite the researchers to submit their work related to SNLP.

For this special session, we aim the following topics related to SNLP that expects to address the issues, challenges and potential solutions the paradigms covering the following aspects: Statistical Inference; Part-of-Speech Tagging, Parsing, Text Classification, Question Answering, Text Summarization, Conversational agents, Information Extraction , Narrative Science, Lexical semantics, Word sense disambiguation, Speech recognition, text-to-speech and spoken language understanding, Computational Social Science and Social Media, Dialogue and Interactive Systems, Discourse and Pragmatics. Further the track also coers wide range topics of SNLP that includes Information Extraction for NLP, Information Retrieval NLP and Text Mining, Linguistic Theories, Cognitive Modeling and Psycholinguistics, Machine Learning for NLP and its application, Phonology, Morphology and Word Segmentation, Resources and Evaluation, Sentiment Analysis, Stylistic Analysis, and Argument Mining, Speech and Multimodality.

II. Submissions to Statistical Natural Language Processing

In total, three papers are submitted to SNLP Track:

- *'A Lexicon Based Approach to Detect Extreme Sentiments'*
- *'Language-Independent Approach to Detect Extremism and Collective Radicalization Online'*,
- *'Modeling Natural Language Policies into Controlled Natural Language: A Twitter Case Study'*.

In the first paper entitled 'A Lexicon Based Approach to Detect Extreme Sentiments' [3], the authors present an unsupervised lexicon-based approach that detects the extreme sentiments from social media. The authors propose an unsupervised approach for automatic detection of people's extreme sentiments on social networks. For this, the first task was automatically to build a standard lexicon consisting of extreme sentiments terms having high extreme positive and extreme negative polarity. With this new lexicon of extreme sentiments, the final task is to validate this lexicon, for which the authors developed an unsupervised approach for automatic detection of extreme sentiments, and evaluated the performance on five different social networks and media datasets. This final task shows that, in these datasets, posts classified with negative sentiments, there are posts of extremely negative sentiments. On the other hand, in posts classified with positive sentiments, there are posts of extremely positive sentiments. The results are promising, the authors achieved quite good results in the identification of extremely positive posts; however, in the identification of extremely negative posts, we did not achieve good results. The authors provide a standard lexicon that can also be useful for other researchers to exploit it for sentiment analysis studies as well as for anti-extremism authorities, allowing them to identify and prevent violent extremism early.

The second paper, 'Language-Independent Approaches to Detect Extremism and Collective Radicalization Online' [4], the authors present a study about language-independent approaches to detect extremism and collective radicalization online. The objective is understanding and detecting extremism and collective radicalism on the online world has a connection to sentiment analysis and opinion mining. There are many barriers to the understanding of extremism and collective radicalisation; one of them is to differentiate between who is on that process and who is just talking about the theme. The main focus of this work is to find the best ways to identify extremism and collective radicalisation on the internet using sentiment analysis focusing on statistical and probabilistic methods, in order to create an unsupervised and language-independent approach. The area of extremism and/or radicalisation online does not have much previous work. However, there is plenty of information on sentiment analysis in the computer science area; there is also information on the roots of radicalisation and extremism.

The final paper entitled as 'Modeling Natural Language Policies into Controlled Natural Language: A Twitter Case Study' [5], the authors study some policy-oriented controlled natural languages, they adopt them as source languages for translating sample Twitter policies. To move in the direction of managing and enforce access policies automatically, in this work, the authors consider a selection of different machine-oriented, English-based Controlled Natural Languages (CNLs), designed initially within different contexts. They investigate their effectiveness in expressing data policies as specified on a popular SN site. The authors consider samples of real Twitter data policies for translating from their original form in natural language to each of the selected controlled languages. The translations are evaluated concerning key properties defined in the so-called PENS (*Precision, Expressiveness, Naturalness, Simplicity*) classification scheme, having one new property, namely *policy enforcement*. The evaluation will help users to select the most appropriate CNL and to automatically process the terms and conditions under which users' data are accessed, stored, and used for machine readability.

III. Conclusions

The SNLP special track includes a wide range of topics and subtopics related Natural Language Processing (NLP). It covers both academic and industry studies to provide solution and ideas to face the challenges and issues in NLP.

Acknowledgment

The authors woul like to thank the organizers for accepting SNLP 2020 as a special track and organize the SNLP session in ICIMP 2020,

References

[1] M. Marcus, "New trends in natural language processing: statistical natural language processing," Proceedings of the National Academy of Sciences, vol. 92, no. 22, 1995, pp. 10 052–10 059.

[2] C. Manning and H. Schutze, Foundations of statistical natural language processing. MIT press, 1999.

[3] S. Pais, I. Tanoli, M. Albardeiro, and J. Cordeiro, "A lexicon based approach to detect extreme sentiments," Proceedings in special track: SNLP along with ICIMP, 2020.

[4] S. Pais, I. Tanoli, M. Abardeiro, and J. Cordeiro, "Language-independent approaches to detect extremism and collective radicalization online," Proceedings in special track: SNLP along with ICIMP, 2020.

[5] I. Tanoli and S. Pais, "Modeling natural language policies into controlled natural language: A twitter case study," Proceedings in special track: SNLP along with ICIMP, 2020.