Computational Aspects of Language Processing

Bogdan Costinel Dumitru

Natural Language Processing (NLP) is a sub-field of computer science concerned with understanding, processing and generating human languages, in order to get computer closer to human language level. A more narrow and formal definition would be: processing of human language in an automatic or semi-automatic way.

Human language understanding is hard because of his complexity in both structure and quantity (e.g., number of languages), but one aspect considerably increase the difficulty, and this is language ambiguity. Human language is not static; it continually grows and evolves in time, shaped by cultural, political and social changes.

Due to language complexity, NLP has quickly grown into a multidisciplinary field. We can see NLP sharing ground with cognitive science, psychology, philosophy and mathematics, where it started with logic, continuing with statistical methods and machine learning.

NLP applications vary from speech recognition, machine translation, question answering, dialog systems, parsing speech understanding, sentiment and emotion analysis, natural language summarization, natural language generation, social computing.

NLP has already a long history with roots back to the 1940s. Warren Weaver started one of the first projects [10] in Machine Translation (MT) based on their World War II experience in breaking enemy codes. A few years later, Weaver's memorandum brought MT research and ideas to a more general audience and inspired many projects that followed.

NLP history and evolution, while a continuous and increasing research field, can be split in various ways like centering on important schools or groups with a long and sustained research history, or time intervals centered on periods of great effervescence. To better follow the description of our own research, periods (waves) are chosen based on the methods and tools used to approach, to study this interdisciplinary field: rationalism, empiricism, and deep learning [8].

In our research, we approach problems from both empiricism and deep learning waves.

1 Contributions

In this thesis, we approach several (open) problems in NLP that spawn over several research areas as sentiment and emotion analysis; text representation, where techniques and research methods from the deep learning wave were used; to more classical problems of similarity and distance, or linguistic problems from empiricism wave.

The main contributions of this thesis can be summarized as follows.

We propose a stylistic analysis of Solomon Marcus' publications, gathered in six volumes, aiming to uncover some of his quantitative and qualitative fingerprints. Moreover, we compare and cluster two distinct periods of time in his writing style: 22 years of the communist regime (1967-1989) and 27 years of democracy (1990-2016). The distributional analysis of Marcus' text reveals that the passing from the communist regime period to democracy is sharply marked by two complementary changes in Marcus' writing: in the pre-democracy period, the communist norms of writing style demanded, on the one hand, long phrases, long words and cliches, and on the other hand, a shortlist of preferred "official" topics; in democracy, the tendency was towards shorter phrases and words, while approaching a broader area of topics. The results are published and presented at RANLP 2018 [2].

We present a deep learning architecture capable of predicting full inflectional paradigms from the uninflected, dictionary form of a word and full orthographic syllabification. We use Romanian as a case study, since its inflectional morphology is rife with stem alter-nations and we show that Sequence to Sequence model receiving the n-grams as input can solve this problem as good as if not better than state of the art. The results are published and presented at CICLing 2018 [5].

We propose a method for semantic difference detection that uses an SVM classifier with features based on co-occurrence counts and shallow semantic parsing. Semantic difference detection attempts to capture whether a word is a discriminative attribute between two other words. For example, the discriminative feature red characterizes the first word from the (apple, banana) pair, but not the second. Modeling semantic difference is essential for language understanding systems, as it provides useful information for identifying particular aspects of word senses. The results are published and presented at SemEval 2018 [7].

We explore a range of deep learning models to predict optimism and pessimism in Twitter at both tweet and user level and show that these models substantially outperform traditional machine learning classifiers used in prior work. Identifying optimistic and pessimistic viewpoints and users from Twitter is useful for providing better social support to those who need such support, for minimizing the negative influence among users and maximizing the spread of positive attitudes and ideas. In addition, we show evidence that a sentiment classifier would not be sufficient for accurately predicting optimism and pessimism on Twitter. Last, we study the verb tense usage as well as the presence of polarity words in optimistic and pessimistic tweets. Parts of these results are published and presented at EMNLP 2018 [1].

Extending the methods used in optimism detection research, we propose a three-layer model with a generic, multi-purpose approach that, without any task-specific optimizations, achieve competitive results in EmoContext competition. We describe our development strategy in detail along with an exposition of our results in a paper published at SemEval 2019 [6].

A new distance between strings, termed Rank Distance, was introduced by Dinu [3]. Since then, the properties of rank distance were studied in several papers. We continue the study of rank distance. More precisely, we tackle three problems that concern the distance between strings:

1. The first problem that we study is *String with Fixed Rank Distance (SFRD)*: given a set of strings S and an integer d decide if there exists a string that is at a distance d from every string in S. For this problem, we provide a polynomial-time exact algorithm.

2. The second problem that we study is named is the *Closest String Problem under Rank Distance (CSRD)*. The input consists of a set of strings S, asks to find the minimum integer d and a string that is at a distance at most d from all strings in S. Since this problem is NP-hard (Dinu and Popa, CPM 2012) it is likely that no polynomial-time algorithm exists. Thus, we propose three different approaches: a heuristic approach and two integer linear programming formulations, one of them using a geometric interpretation of the problem.

3. Finally, we approach the *Farthest String Problem via Rank Distance (FSRD)* that asks to find two strings with the same frequency of characters (i.e., the same Parikh vector) that have the largest possible rank distance. We provide a polynomial-time exact algorithm for this problem [4].

We research how various word embeddings can be retrofitted using available knowledge base databases (e.g., WordNet). By retrofitting, we understand a slight alteration of the word representations using a neural network to capture knowledge derived from WordNet. Results are measured using a rank correlation measure of similarity on canonical datasets and then compared with a baseline from well-known word embeddings [9].

We investigate proto-word reconstruction consisting of recreating the words in an ancient language from related modern languages. Starting from pairs of modern word and proto-word, we perform translation and rotation operations on ancient language word embedding (e.g., Latin) and newly resulted word embeddings are used to detect similar words. Combined with a metric (e.g., Rank Distance) on candidate words, proto-words can be derived.

2 Thesis outline

Chapter 2 introduces all necessary mathematical and computer science apparatus for both major lines of this thesis: neural networks and metrics. We present

few network architectures and methods that form the basis of deep learning for NLP, we present few applications of this methods, and we briefly touch recent breakthroughs that have the potential to greatly change the landscape of NLP in next years greatly. Computational semantics basics are covered next.

In Chapter 3 we explore the idea of retrofitting distributed word representations based on information and structure derived from a knowledge base, WordNet in our case. We build a multidimensional matrix of various similarity measures for all pairs of words available in WordNet and use this matrix to train a neural network to minimize a distance cost between corresponding word representations. Several configurations and methods used are then compared using a ranking metric against widely used test sets that measure similarity and relatedness (e.g., SimLex 999). Another research area where we have relied on word representations was proto-word identification. We try to align word representations of a modern language with word representations of an ancestor language trying to validate the hypothesis that a proto-word and his modern correspondent are not far apart in newly created space.

In Chapter 4, we present research topics from a widely popular area of NLP: sentiment analysis. But first, we start with an authorship attribution type of problem by trying to see if Solomon Marcus' work was influenced by political regimes, more formally, we are looking for stylom variations. On sentiment analysis, we start with optimism and pessimism detection in social media (tweets) using deep neural networks and end the chapter with contextual emotion detection in short conversation. Our proposed goal was to derive a common method to address both problems. We end the chapter with an extension of the original model that incorporates contextualized word representations.

In chapter 5, we present two research topics that fall into the category of computational linguistics. First, we propose a system for capturing discriminative attributes of word triplets, attributes that are used to train an SVM for binary classification. In the next topic, we look at inflection learning on Romanian datasets using a deep neural network architecture.

In Chapter 6, we advance previous research on Rank Distance and propose a solution for three problem: closest rank distance, furthest rank distance, and fixed rank distance. Next, we devise a formula to compute the diameter of the largest sphere that contains a given set of strings.

In final Chapter 7, we draw the final conclusions, reiterate thesis contribution, and briefly outline future research avenues that we intend to pursue or continue and extend those presented in this thesis.

References

[1] Cornelia Caragea, Liviu P. Dinu, and Bogdan Dumitru. "Exploring Optimism and Pessimism in Twitter Using Deep Learning." In: *EMNLP*. Ed. by Ellen Riloff et al. Association for Computational Linguistics, 2018, pp. 652–658. ISBN: 978-1-948087-84-1. URL: http://dblp.uni-trier.de/db/conf/emnlp/emnlp2018.html#CarageaDD18.

[2] Anca Dinu, Liviu P. Dinu, and Bogdan Dumitru. "On the stylistic evolution from communism to democracy: Solomon Marcus study case." In: *RANLP*. Ed. by Ruslan Mitkov and Galia Angelova. INCOMA Ltd., 2017, pp. 201–207. ISBN: 978-954-452-049-6. URL: http://dblp.uni-trier.de/db/conf/ranlp/ranlp2017.html#DinuDD17.

[3] Liviu P. Dinu. "On the Classification and Aggregation of Hierarchies with Different Constitutive Elements." In: *Fundam. Inform.* 55.1 (2003), pp. 39–50. URL: http://dblp.uni-trier.de/db/journals/fuin/fuin55.html#Dinu03.

[4] Liviu P. Dinu, Bogdan Dumitru, and Alexandru Popa. "Algorithms for Closest and Farthest String Problems via Rank Distance". In: *Theory and Applications of Models of Computation - 15th Annual Conference, TAMC 2019, Kitakyushu, Japan, April 13-16, 2019, Proceedings*. 2019, pp. 154–171. DOI: 10.1007/978-3-030-14812-6_10. URL: https://doi.org/10.1007/978-3-030-14812-6%5C_10.

[5] Liviu P. Dinu, Bogdan Dumitru, and Maria Sulea. ""Full Inflection Learning Using Deep Neural Networks"." In: *Computational Linguistics and Intelligent Text Processing - 19th International Conference, CICLing 2017, Hanoi, Vietnam, March 18-24, 2018*. 2018, in press. DOI: 10.1007/978-3-319-77113-7_44. URL: https://doi.org/10.1007/978-3-319-77113-7%5C_44.

[6] Bogdan Dumitru. "GenSMT at SemEval-2019 Task 3: Contextual Emotion Detection in tweets using multi task generic approach." In: *SemEval@NAACL-HLT*. Ed. by Jonathan May et al. Association for Computational Linguistics, 2019, pp. 225–229. ISBN: 978-1-950737-06-2. URL: http://dblp.uni-trier.de/db/conf/semeval/semeval2019.html#Dumitru19.

[7] Bogdan Dumitru, Alina Maria Ciobanu, and Liviu P. Dinu. "ALB at SemEval-2018 Task 10: A System for Capturing Discriminative Attributes." In: *SemEval@NAACL-HLT*. Ed. by Marianna Apidianaki et al. Association for Computational Linguistics, 2018, pp. 963–967. ISBN: 978-1-948087-20-9. URL: http://dblp.uni-trier.de/db/conf/semeval/semeval2018.html#DumitruCD18.

[8] Fernando Pereira. *A (computational) linguistic farce in three acts*. en-US. 2017. URL: https://www.earningmyturns.org/2017/06/a-computational-linguistic-farce-in.html (visited on 01/09/2020).

[9] Marius Popescu and Bogdan Dumitru. "Word Embeddings Retrofitting using Siamese Networks". Preparing for submission. 2020.

[10] Warren Weaver. "Translation". In: *Machine Translation of Languages*. Ed. by William N. Locke and A. Donald Boothe. Reprinted from a memorandum written by Weaver in 1949. Cambridge, MA: MIT Press, 1949/1955, pp. 15–23.