Bi-directional Mechanism for Recursion Algorithms: a Case Study on Gender Identification in MOOCs

Tahani Aljohani, Alexandra I. Cristea and Laila Alrajhi

taljohani7@gmail.com
{alexandra.i.cristea,laila.m.alrajhi}@durham.ac.uk

Abstract. Automatically identifying the learner gender, which serves as this paper's focus, can provide valuable information to personalised learners' experiences in MOOCs. However, extracting the gender from learner-generated data (discussion forum) is a challenging task, which is understudied in literature. Using syntactic features is still the state-of-the-art for gender identification in social media. Instead we propose here a novel approach based on Recursive Neural Networks (RecNN), to learn advanced syntactic knowledge extracted from learners' comments, as an NLP-based predictor for their gender identity. We propose a bi-directional composition function, added to NLP state-of-the-art candidate RecNN models. We evaluate different combinations of semantic level encoding and syntactic level encoding functions, exploring their performances, with respect to the task of learner gender profiling in MOOCs.

1 Prologue

MOOCs content can be personalised based on demographic data, particularly "gender" [6]. The gender parameter has already been shown to influence the success of the learning process. Traditionally, demographic data are extracted from pre-course questionnaires that are filled-in by the learners themselves; however, only about 10% provide their demographic data [1]. To resolve this issue, we research automatic extraction of learner characteristics, here, gender, from the traces learners leave in MOOCs. This falls under the umbrella of a more generic research area called *Author Profiling (AP)*, which promotes the use of automatic tools – developed based on Natural Language Processing (NLP) – for the purpose of identifying authors' demographics and characteristics, mainly based on their writing, and only basic types of syntactic representations of text, such as Part of Speech (POS), have been considered in previous studies for gender profiling [3]. The umbrella research question in this paper is: *Can advanced textual features extracted from MOOC discussion forums be used to classify a learner's gender?*. The main contributions of this paper are as follows: examining advanced syntactic features for the learner gender profiling; exploring state-of-the-art recursive models (tree-structured LSTM, SATA, and SPINN models), and applying them, for the first time, to author-profiling (here, for learners gender profiling), and then improving the current recursive models by introducing a novel bi-directional strategy.

2 Methodology

Approximately 322,310 comments from female and male learners in MOOCs
For text normalisation, simple NLP cleaning steps were applied. We made sure that these steps should not harm the learners' writing style. For *semantic representation*, GloVe was used for word input embeddings. These initial inputs are fed to the RecNN models to provide semantic vectors for each word for the leaf nodes or the initial inputs. Sentences are separated based on the full stop (.)" in each comment. This is to reduce the complexity during training time, since samples become shorter, which speeds up the training and generates more samples in the used data. An advanced NLP parse was used (Stanford Probabilistic Context-Free Grammar (PCFG) parser [5]), to convert the phrases at a syntactic level of the text (tree structure) in a binary mode to establish a binary tree. Then, these samples are fed to TreeNNs models for classification (see Figure 1).

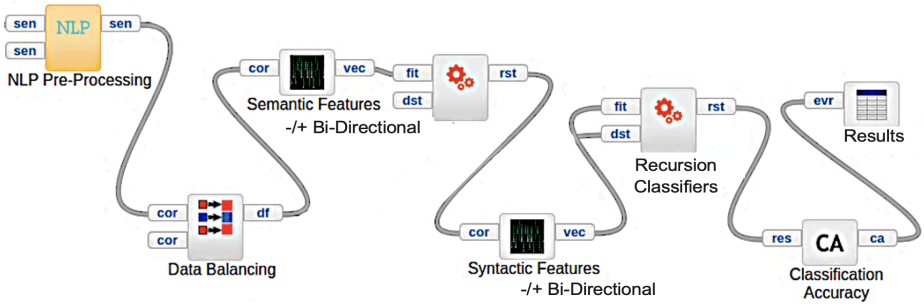
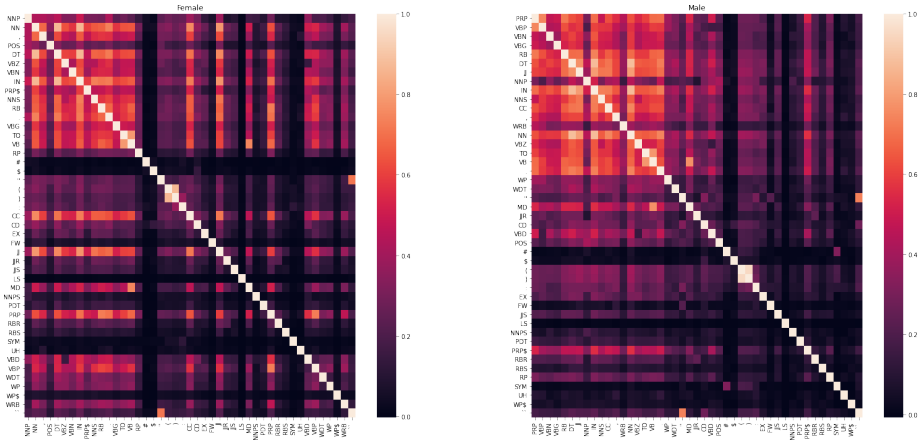


Fig. 1: General Workflow of Gender Identification

As an initial step, a heatmap approach was used to understand the correlation between the POS (in its simple form) and discover a statistical measure linearly. Since there are many POS variables, the aim was to examine how dependent they are on each other, which may be shown in a 2D matrix called a correlation matrix. In the Figures 2a and 2b, the lighter the colour between two variables, the stronger the correlation (and vice versa).

3 Findings

The distribution of POS patterns based on the mean, as shown in the heatmaps figures, does not reveal differences in the writing styles of the two groups. This means that this chosen approach also failed to capture the differences in syntactic patterns, due to its simplicity. Thus, DL approaches (advanced approaches) have been examined for gender profiling in this paper.
Three types of syntactic learning models were applied in this study: the original TreeLSTM [7], the Stack-Augmented Parser-Interpreter Neural Network (SPINN) [2], and Structure-Aware Tag Augmented (SATA) [4]. These models were chosen, as they are state-of-the-art DL models for such text representations. Additionally, we introduce new versions of these models based on a bi-directional composition function with different combinations.



(a) Correlation between POS in female comments (b) Correlation between POS in male comments

Fig. 2: Correlation between POS in female and male comments

We found that no study had examined the performance of these models by adding the bi-directional learning. Bi-directional learning has already shown its effectiveness in improving the sequential LSTM model and it is well-known that bi-directional LSTM outperforms vanilla LSTM for many NLP tasks. Thus, a hypothesis was made in this study that adding bi-directional TreeLSTM would improve the performance of SPINN and SATA. We investigated propagating the top-down direction of information and the bottom-up direction using bi-directional TreeLSTM. In fact, the uni-directional TreeLSTM by default processes inputs from the bottom-up direction in a bottom-up manner through the tree. So, we included the additional set of hidden state vectors in the top-down direction (from root to comment inputs), which then alters the model to the bi-directional paradigm. This is technically another TreeLSTM model, where the final hidden state is the final state vectors of the two LSTMs.

The syntactic learning in a TreeLSTM-based architecture in general consists of the following two steps: word-level encoding with a feedforward neural network or LSTM neural network; and sentence-level encoding with a tree-structured LSTM composition function. While previous literature has recommended using LSTM for word-level encoding, there is no such work to introduce bi-directional LSTM for the word-level encoding. Thus, this research also contributes to fill this gap, by adding the bi-directional LSTM at the word level as well. The motivation for this supplementary bi-directional technique is to increase the high-level representation of tree nodes during the recursive propagation across many branches.

TreeLSTM, SPINN,and SATA in their original structure are considered as baselines. The proposed versions of the bi-directional strategy that were applied for all 18 models (combinations of bidirectional and/or unidirectional of semantic and syntactic learning in each of the three classifiers). All of them were more

effective than the baseline models. They achieved competitive performance in relation to each other. In general, all models achieved high performance in identifying the gender class (80.60% or above). This could promote the idea that the use of phrase-level representation is robust for learner gender classification. Every two versions of each model are very similar, but the bi-directional composition function models achieved slightly better results. The highest observed outcome in this paper was 82.62%. This was achieved by the newly proposed model based on the simple Forward Neural Network combined with the SPINN model.

4 Epilogue

This study indicates that bi-directional learning is promising in terms of improving the gender classification accuracy. It also shows the importance of the extra information that the model obtains during the training, which does not have to be limited to tags of constituents included in the SATA model. Furthermore, it is evident that using only a simple model with fewer parameters for word encoding by the Forward Neural Network (which used linear mapping) still achieves good results. This might be attributable to the fact that using linear mapping better preserves word-level semantics, while the LSTM encoding alters the semantic meaning at the word level, thereby making it harder to structure the sentence from a syntactic perspective. This might be related to task complexity.

References

1. Aljohani, T., Cristea, A.I.: User profiling on the futurelearn platform via deep neural networks, semantic and syntactic representations. Frontiers in Research Metrics and Analytics **6**, 34 (2021)
2. Bowman, S.R., Gauthier, J., Rastogi, A., Gupta, R., Manning, C.D., Potts, C.: A fast unified model for parsing and sentence understanding. arXiv preprint arXiv:1603.06021 (2016)
3. HaCohen-Kerner, Y.: Survey on profiling age and gender of text authors. Expert Systems with Applications p. 117140 (2022)
4. Kim, T., Choi, J., Edmiston, D., Bae, S., Lee, S.g.: Dynamic compositionality in recursive neural networks with structure-aware tag representations. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 6594–6601 (2019)
5. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: Proceedings of the 41st annual meeting of the association for computational linguistics. pp. 423–430 (2003)
6. Shi, L., Cristea, A.I.: Demographic indicators influencing learning activities in moocs: learning analytics of futurelearn courses. *Association for Information Systems* (2018)
7. Tai, K.S., Socher, R., Manning, C.D.: Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075 (2015)