Metaphor and Figurative Language Detection

Vlad-Rares Furdui
vlad-rares.furdui@s.unibuc.ro

Bogdan Ivan
bogdan.ivan@s.unibuc.ro

Andrei-Alexandru Mihai
andrei-alexandru.mihai@s.unibuc.ro

Abstract

The detection of metaphor and other forms of figurative language presents one of the greatest challenges to computational linguistics. This paper delves into the current practices and advances made in the automatic identification of metaphor in text data. In the following, we compare the actual efficacy of existing techniques for metaphor identification against standard datasets, while newer adaptations are tested against both original and diverse datasets. This is with the aim to try to not only advance the understanding of metaphor detection but also to enhance the accuracy and applicability of these techniques in varied linguistic contexts.

1 Introduction

Metaphor detection and figurative language in general present one of the archetypal challenges within computational linguistics and reflect the intricate interplay of language with meaning. Metaphors color everyday language with richness and expressiveness. At the same time, they pose a significant difficulty for computational systems because they are so very subtle and context-dependent. Our project's methodology is guided by the influential capabilities of the BertForMultipleChoice model; deep contextual understanding of text through a transformer-based architecture, now known as BERT.

One of the first works to overcome traditional weaknesses in metaphor detection, the research uses a very powerful model on a specially created dataset - "Fig-QA". It is a diverse dataset illustrating the usage of figurative language in many different domains and contexts. We try to marry the use of leading technology with a well-curated dataset to make a decisive step forward in the automatic understanding and identification of metaphors in text, pushing the boundary of what computational models can achieve in the interpretation of human language.

2 Related Work

Recent work in the computational detection of metaphors and figurative language draws from a number of perspectives and, within this developing field, has been influenced by advances in natural language processing and machine learning. In the following section, a brief account is given of the main methodologies that have informed our approach, highlighting significant contributions and setting the scene for the use of the BertForMultipleChoice model.

Early works toward the problem of metaphor detection relied on rule-based systems where patterns to catch figurative language were manually defined (Fass, 1991; Martin, 1990). While pioneering, these very early methods suffered in their scale and generalizability to new and unseen data, since they had to involve heavy manual intervention. The variety and complexity in natural language does not allow these kinds of systems to readily accommodate it.

Later on, as computation became more advanced, researchers sought statistical methods that would have the potential to exploit the large corpora in searching for metaphoric expressions, identified through a deviation from the normative linguistic patterns (Gedigian et al., 2006). In this regard, some corpus-based approaches were more scalable and offered automated detection in a more effective and efficient way, while others were highly dependent on surface features, which led to a high rate of false positives.

Many improvements to metaphor detection followed the advent of machine learning models. Techniques like support vector machines and neural networks began to be used, taking advantage of the potential to learn complex patterns from data (Kintsch, 2000; Turney et al., 2011). These models, being rule-based, were much more flexible in adapting to the nuances of language.

The most recent developments include deep learning models, specifically transformer architectures with models like BERT (Devlin et al., 2019). A large leap since it processes language and language understanding at a very deep level, BERT and its task-adapted versions, such as that for multiple-choice questions, have shown startling success not only in detecting figurative language but also in understanding the context in which it was placed, hence very appropriate for our study.

We continue in this work using the BertForMultipleChoice model while developing our approach to the subtle task of metaphor detection. By combining deep learning with a sharply focused technique for multiple-choice question design, this method provided the best results of any method against all benchmarks for metaphor recognition. Such background gives our methodology a profound insight and precise metaphor detection in the text.

3 Method

The methodology can effectively identify metaphor and figurative language, in line with state-of-the-art natural language processing. Our design makes it possible to combine those two advanced machine learning models with comprehensive data handling for robust performance across text genres. This section presents details on the dataset used, the pre-processing steps implemented, and the models employed in our experiments.

3.1 Dataset

The "Fig-QA" dataset therefore forms the core of our research work, metaphor and figurative language detection, based on the fact that it is well-structured and rather close to the requirements of our method and model. This dataset has three different components: the training, development, and testing subsets, which are essential for the all-inclusive training, validation, and evaluation of our computational models.

- **Training Set:** The training set is the largest subset of the dataset. The training set is sourced from train_xl.csv and carries many examples. These examples serve to acquaint the model with a wide range of usages and contexts in which figurative language is used. This can be effected through exposure to complex patterns related to metaphoric expressions. This set, being quite diverse, will allow the model to learn not only the specific instances of metaphor but also how to generalize the pattern regarding the instances over different textual manifestations.
- **Development Set:** The development set is derived from dev.csv, meant mainly for model optimization through the intermittent training process for parameter fine-tuning and interim evaluation. Iterative testing enables any necessary adjustments to be made before the final test on the test set. Adjustments help to calibrate the model performance, especially in areas of balancing overfitting and underfitting.
- **Testing Set:** Finally, this testing set from train_s.csv is done to check the effectiveness of the model to mimic an applied, real-world environment. This subset tests the capability of the model to apply learned experiences to new, previously unseen data and thus gives a trustworthy measure of performance and generalization abilities.

All of these sub-sets are individually processed using pandas in Python, hence maintaining data integrity and consistency. Column standardization is done throughout these subsets, mostly on the "label" column, which is key for supervised learning. Missing labels within the testing set are handled through default assignment, hence managing to evaluate model performance based on complete data. This goes a long way in preparing the dataset properly for training a model that will be effective enough in detecting figurative language.

3.2 Preprocessing

The preprocessing phase of our metaphor detection and figurative language study designed strictly changes the form of raw text data into a structured one that the BertForMultipleChoice model can handle. This, therefore, makes the phase critical to the model, since it will determine how well it learns and performs metaphor detection tasks.

- **Tokenization and Input Configuration:** Basically, the core of the preprocessing routine is BertTokenizer from the transformers library, and it is mostly responsible for breaking down textual data into accessible tokens that are machine-readable by the BERT model. Generally, by using the bert-base-uncased, the tokenizer will cause all text to be in lowercase to ensure uniformity and will also reduce the complexity of the model by normalizing to ensure the model does not learn the same words in both their uppercase and lowercase versions. Generally, each data entry will comprise one "startphrase", which acts as the context or premise, and a couple of "endings" (ending1 and ending2). For each "startphrase", the tokenizer will be used on each corresponding "ending", which will prepare the input for the model in order to evaluate which ending fits the context the best. This step is just a precursor to carrying out the multiple-choice task of metaphor detection.

- **Attention Masks and Input IDs:** The tokenization also generates attention masks and input IDs, which are critical in the process of training the BERT model. It indicates to the model which part of the input it needs to pay attention to and which parts are padding. It forces the model to focus its computations on parts of the text where real data is, hence making it more effective and accurate but also not processing irrelevant information.

- **Preprocessing Function:** The preprocess_function is specifically written to handle the subtlety of multiple-choice questions; it ensures that the choices for a given question are concatenated with it, and together, they are regarded as a single input instance. It will ensure that each "startphrase" is paired with one of the two "endings," so the tokenizer will understand to treat them as two separate situations. It then organizes the tokenized output into structured batches ready for feeding into the model, with input IDs and attention masks aligned appropriately.

This prepares the data for effective learning and optimizes the performance of the model so it can assess and interpret the metaphorical content of the text. Such a careful preparation of input data is very important in order for the predictive power of the model to increase and build a very good foundation in accurate and reliable metaphor detection.

3.3 Models

In our metaphor and figurative language detection project, we employ the BertForMultipleChoice model. It represents a modification of the much-acclaimed BERT model, specifically designed for various tasks where the proper choice must be selected from among several alternatives. The model builds on the robust abilities of BERT, based on transformer architecture, to execute complex language processing tasks very efficiently.

The BertForMultipleChoice is one of the base bricks of our methodology, given that our dataset format, with text entries and multiple possible answers, is very well-suited for it. Each input instance is such that the model is presented with context (the startphrase) and the choices (the ending1 and ending2). Subsequently, it will know which choice properly finishes the phrase correctly in a contextually sensitive way.

To train the model, we set up the environment using the TrainingArguments class from the Hugging Face transformers library. Important parameters are:

- **Learning Rate:** Has been set at the level of 2e-5, which means a relatively low rate for slow and more stable learning, thus avoiding the model to miss subtle data.

- **Batch Size:** Here, the batch size for both training and evaluation is set to be 8 in such a way that the computational efficiency is balanced against memory usage.

- **Epochs:** Training the model is done for 3 epochs, with the number of epochs taken into consideration as a trade-off between underfitting and overfitting, based on the dimensionality and complexity of the data.

- **Evaluation Strategy:** It can be set to evaluate at the end of every epoch, by which it can monitor the model performance and make necessary changes immediately.

Accuracy is the main metric to assess model performance and reflection of the rate of right answers to questions. This metric is important to our project in that it directly quantifies how well the model understands and appropriately uses the linguistic context in order to select appropriate metaphorical or otherwise figurative language.

The model was trained and validated using the Trainer API, which basically has all the components required to facilitate model training on data and, hence, handling device allocation for computational efficiency and evaluating model performance. This configuration of the Trainer API allows not only efficient training but also assurance of the robustness of the model's performance across different subsets of the data.

Our implementation harnesses the power of BertForMultipleChoice to solve one of the challenging tasks: metaphor detection with high accuracy and insight into the subtleties of human language.

4 Conclusion

Our research project on detecting metaphor and figurative language with the BertForMultipleChoice model is reporting the findings and evidence to date of the robustness of the model in complex linguistic tasks. We, therefore, have come up with a system that will not just identify metaphorical language but also increase our understanding of how the use of language that is metaphorical can be computationally differentiated by the application of advanced machine learning techniques on the carefully created Fig-QA dataset.

One of the main contributions of this study is toward the successful adaption of the BERT model, especially in its BertForMultipleChoice form, for the metaphor detection challenges that textual data pose. The deep transformer network-based model architecture used for this task was quite apt in processing and predicting from the context provided by the 'startphrase' and multiple 'endings'. Our results show that the model was able to produce a high level of accuracy in the choice of the contextually appropriate endings, thereby validating our hypothesis that transformer-based models indeed give significant improvements for metaphor detection tasks.

More preprocessing, tokenization, and input configuration for learning were achieved with the dataset. This is key in letting the model concentrate more on the important parts of the data and, therefore, optimize training and model performance. To this effect, the attention mechanisms within BERT made the model more sensitive to picking up subtle nuances and text intricacies, more so often typical with figurative language.

Practical relevance obviously follows from such findings. This work, through the results presented, is capable of showing that models of deep learning, such as BertForMultipleChoice, can effectively interpret and process metaphorical language. We thus pave the way for much finer applications of natural language processing, ranging from content-filtering systems to interactive AI, such as chatbots and virtual assistants, that are sensitive to the fine nuances in human language and communication.

However, the success included a number of challenges as well, especially in the computational resources and the extent of data preprocessing required. The computational demands for training huge models like BERT point to the significance of having the correct hardware in place, most often being the bottleneck for academic research groups. The preprocessing phase indicated that there was a real need for careful data preparation for high-precision models to be available. This, in most cases, is a highly time-consuming and technically challenging process.

5 Future Work

Building on the foundation laid by our current research in metaphor and figurative language detection using the BertForMultipleChoice model, several promising avenues for future work have been identified. These focus on enhancing model performance, expanding dataset diversity, and exploring new applications in practical scenarios.

- **Enhancing Model Capabilities:** While promising, our results with BERT could still be further boosted for both accuracy and efficiency, which may be achieved by investigating alternative architectures like RoBERTa or ALBERT, allowing some improvement in light of the optimized way of training or the lighter network structures mentioned with great payoffs. Going further, finer-tuning attention mechanisms on the specifics of the nuances of figurative language could prove even more beneficial in helping the model differentiate subtle metaphorical nuances from literal expressions.

- **Dataset Expansion and Diversification:** The current "Fig-QA" dataset is quite robust, but it includes English text almost exclusively. An addition with multilingual examples of metaphorical language would increase both generalizability and relevance to various global linguistic contexts. In addition, the introduction of data from sources such as social media, literature, and oral dialogue will help to further challenge and fine-tune the model's ability so that it stays potent over varied media.

- **Practical Applications:** It will also be very important to substantiate the real-world utility of the metaphor detection system. One of such areas of immediate application is educational technology, where such systems can come in for improved understanding of how the learner uses figurative language. It would also enable the integration of such technology into moderation tools for social platforms, which would decrease the chance of any misinterpretation of the users' communications as harmful or inappropriate.

- **Interdisciplinary Collaboration:** In the future, this will be beneficial for research by computational linguists who would therefore work together with cognitive psychologists and cultural scholars. For example, such collaboration jointly identified the way metaphors are experienced differently across cultures, and individual cognitive differences will therefore lead to a more adaptive and sensitive system of AI.

It is in these directions that future research can extend the capabilities of metaphor detection systems, rendering them accurate, efficient, and applicable across a broad range of contexts and languages.

Limitations

Although this research has contributed important findings on metaphor and the detection of figurative language using the BertForMultipleChoice model, the limitations must be noted to accurately guide the future studies and applications.

- **Model-Specific Constraints:** The first constraint arises from our use of the BertForMultipleChoice model. While it is known that BERT-based models work best for many tasks in NLP, they have some disadvantages. Most importantly, such models require large computational resources for both training and inference, complicating their use and prospects for scaling up, especially for researchers and practitioners who do not have access to high-computing-performance facilities. Additionally, BERT-based models are not very well suited for very large texts, considering the maximum input length is fixed, which could cut off context that would be very significant for more complex metaphorical expressions to be understood.

- **Dataset Limitations:** While the "Fig-QA" dataset is quite strong and well annotated, it holds text that has been mostly pre-selected and preformatted to correspond to the requirements of the multiple-choice metaphor detection task. This very fact might limit the model's capability to generalize for other types of text, which might not be represented within the dataset, such as unstructured or larger text passages commonly seen in novels or conversational speech. The dataset is further focused solely on texts that are written in the English language; examples that are not English are omitted. This adds to the non-robust nature of this application because in multilingual contexts, the metaphors and the respective interpretation can have a very broad variation.

- **Performance Metrics:** The basic performance metric that we used in the study is 'accuracy', which could be very misleading and may not be a good representative of model performance. Accuracy metrics sometimes give false indications, particularly with data sets where the class distribution is imbalanced. This might result in a model that appears performative on standard metrics while still failing to capture finer nuances in metaphor usage.

- **Future Enhancements:** The above limitations can be dealt with by model architecture improvements, along with some improvements to the datasets employed. The future work could further explore more diversified and enlarged datasets, incorporation of alternative metrics—such as F1 score or ROC-AUC for a more balanced performance assessment—and experiment with more computationally efficient models or those capable of dealing with longer text inputs.

By acknowledging these limitations, we will be able to set a more focused research agenda in the future, with the objective to overcome these challenges and push the limits of what computational models can achieve in the intricate task of metaphor detection.

Ethical Statement

We are committed to impeccable ethical standards in pursuing our research on the development of work in the detection of metaphor and figurative language using the BertForMultipleChoice model for responsible AI technology. This research will be executed carefully because of the sensitive nature of the implication and the possible uses in other language processing applications.

- **Data Privacy and Security:** The dataset used, "Fig-QA," does not carry any information relating to personally identifiable information. It carries only textual data that carries information that does not refer to an individual. This will ensure that privacy is fully taken care of and also the fact that high ethical standards in using data are maintained.

- **Bias Mitigation:** We do acknowledge that by nature, pre-trained models like BERT are biased since they are trained on data extracted from the internet. We are aware of the biases that the pre-trained models are exposed to and work on their identification and mitigation by including diverse datasets and continuously evaluating on different linguistic and cultural backgrounds.

- **Transparency and Openness:** We are pledged to being transparent in all its operations by making our research methodology, data processing techniques, and research results available for academic scrutiny and public review in furtherance of openness and ethical accountability.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL-HLT*, pages 4171–4186.

Dan Fass. 1991. Met: A method for discriminating metaphor and metonymy. In *Proceedings of the Workshop on Computational Semantics*, pages 1–16.

Mark Gedigian, John Bryant, Shrikanth Narayanan, and Branimir Ciric. 2006. Catching metaphors. In *Proceedings of the Workshop on Scalable Natural Language Understanding*, pages 41–48.

Walter Kintsch. 2000. Metaphor comprehension: A computational theory. *Psychonomic Bulletin Review*, 7(2):257–266.

James H. Martin. 1990. *A Computational Model of Metaphor Interpretation*. Academic Press, Inc.

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 680–690.