TODO

Shantanu Nath
(s4998405)
s.n.nath@student.rug.nl

Abstract

With the advancement of technology, text classification is now being broadly applied to solve many problems in businesses, government, and research and thus has drawn the attention of the research community. However, it is still considered a challenging task due to the unclear meaning of texts and ambiguity in detecting suitable labels. This paper explores Long Short-Term Memory (LSTM) network and Transformer-based architectures such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-Training (GPT) using a publicly available text corpus with 6000 samples in English collected from Twitter. We achieve 90.4% accuracy using Bidirectional LSTM with a popular pretrained word embedding model "GloVe" and 94.4% accuracy using BERT. In addition, we also emphasize experimenting with the traditional machine learning algorithms that yield competitive results in our experiments. We obtain 91.6% accuracy adopting an ensemble learning based approach combining Naive Bayes, Random Forest, and K-Nearest Neighbour (KNN) algorithms. These results demonstrate the effectiveness of BERT compared to the other algorithms even on a smaller corpus.

1 Introduction

Because of having different linguistic properties, language acquisition becomes a challenge in humans. Interestingly, some languages share some linguistic properties, called linguistic universals. According to linguistic universal, quantificational determiners like *all* and *some* always share some structural properties namely monotonicity, quantity, and conservativity (Keenan and Stavi, 1986; Barwise and Cooper, 1981). This means interpreting (1) or (2) only requires examining the restrictor set (birds) and the intersection between the restrictor and scopal set (can fly).

1. All birds can fly.

2. Some birds can fly.

In this paper, I mainly focused on conservative universal determiners, where all simple natural language determiners correspond to conservative quantifiers (Barwise and Cooper, 1981; Keenan and Stavi, 1986). A possible reason for these universal properties lies in our cognition system (Shane Steinert-Threlkeld, Jakub Szymanik). They suggest that linguistic universals arise because expressions satisfying them are easier to learn than those that do not. They measure ease of learning using tools from machine learning and analyze universals in a domain of function words (quantifiers) and content words (color terms). Also, quantifiers having these properties make them simpler (Quantifiers satisfying semantic universals are simpler). Therefore Steinert-Threlkeld and Szymanik (2017) find that neural networks can learn both conservative and non-conservative quantifiers equally well.

On the other hand, Hunter and Lidz [2013] depict that Children aged between 4 to 5 learn conservative quantifiers easier than non-conservative quantifiers. They discovered two new words for conservative and non-conservative, gleeb and gleeb' respectively. They train and test the children with pictures presentation and children show better performance for the conservative quantifiers over non-conservative quantifiers.

In this paper, I want to discuss that there is no difference in learnability between conservative and non-conservative quantifiers. Conservativity have different properties than other quantifiers have. Different experimental setup and neural netowrk models does not support ease of learnability of conservativity. Also varying the distribution of conservative and non-conservative quantifiers in the training data of an LSTM network does not result in a measurable difference between quantifier accuracies.

In Section 2, information about the previous work done on conservativity. In next section 3, I describe fundamental properties of universal quantifiers. In Section 4, I explain experimental setup and evaluation methods that are used to assess model performance. In Section 5, I provide our results, and in Section 6, I discuss our findings and make conclusions on the results that are presented in Section 5.

2 Related Work

Learning a new language is always related to certain linguistics features. Jakub Szymanik proposed that all lexicalized simple quantifiers share common linguistics properties like monotonicity, quantity, and conservativity. They compute minimal description lengths and Lempel- Ziv complexities of each quantifier by using both descriptive statistics and logistic regression. They found quantifier expressions with universal properties are simpler.

Despite learnability and simplicity having a plausible relation, the learning differences between conservative and non-conservative quantifiers remain open. Children learn these conservative quantifiers quickly than non-conservative quantifiers(Hunter and Lidz). They experimented on 20 children aged between 4 to 5 by training them with novel quantifiers. They assigned ***gleeb*** for ***not all*** and ***gleeb'*** for ***not only*** as novel quantifiers and teach children by using a variant of "picky puppet task" (Waxman & Gelman 1986). In their experiment, they reveal a relation between typology in determiners meaning and learnability. They did not find any evidence of learning non-conservative quantifiers meaning compared to conservative quantifiers. They interpret their result as children having a learning bias on conservative quantifiers.

Steinert-Threlkeld and Szymanik tried to explain by computational linguistics with various machine learning models. They proposed a cognition explanation of universal complexity based on learnability. In their study, they train an LSTM network (Hochreiter and Schmidhuber, 1997) to perform a related sequence classification task: the input to the network is a sequence in which each element represents a set-theoretic model and a quantifier, and the output are two possible truth-values given the input. The model is a sequence of entities represented by one-hot encodings of four zones in a Venn diagram: $A \cup B$, $A \setminus B$, $B \setminus A$, and $A \cap B$. For example, an entity in $A \setminus B$ is encoded in a vector [0,1,0,0]. Two encoded quantifiers are used in each experiment: one exhibiting the semantic universal, e.g. the conservative ***not all*** ($A \setminus B = 0$), and one not exhibiting the universal, e.g. the non-conservative not only (***not only*** = 0). Though learning quantifiers with monotonicity and quantity is easier to learn by Recurrent Neural Network computational models, Steinert-Threlkeld and Szymanik [2019] were unable to find a learnability advantage for conservative quantifiers.

Vishwali Mhasawade proposed two reasons for this learning difference between humans and neural networks: either innate or data distribution. They define five pairs of quantifiers including one conservative and one non-conservative in each pair. All quantifiers are asymmetric, and none of them satisfy extensionality which helps them to reduce the quantifier complexity. By using a total of 30000 data points in five experiments, they did not find any learning bias towards conservative quantifiers. All and only are learned to an equal extent and at the same rate across all five experimental conditions. Finally, they conclude that data distribution does not affect learning conservative quantifiers. This learnability bias among children could be innate or representational.

3 Universal properties of quantifier

Quantifier, semantically denoted as Determiners, is a combination of Noun and Verb, Det N VP, like "Some bicycles are Blue". I divided these determiners into simple and complex forms. for example, Some, few, many, assumed as simple and

(e,g. at least 6 or at most 2, an even number of) assumed as complex. Here I mainly focused on simple determiners. This sentence, "Some bicycles are Blue" is an intersection of two different groups Bicycles and Blue things. It means there are some "objects" that are bicycles and Blue is their "property". These two groups of Bicycles, A, and Blue things, B, can be represented as a Model, M = (M, A, B). I will use the Generalized quantifier framework as a collection of these models. Then, a quantifier Q can be represented as a set-theoretic language and we write $Q \in M$ if and only if: M = Q (A, B). For example, we can define all, some, most as follows:

1. 'Every Bicycle is blue' is true $\iff$ $\mid M,A,B \mid : \mid A \cap B \mid \neq 0$
2. 'Some Bicycle is blue' is true $\iff$ $\mid M,A,B \mid : \mid A \cap B \mid > \mid A \setminus B \mid$
3. 'Most Bicycles are blue' is true $\iff$ $\mid M,A,B \mid : \mid A \subseteq B \mid$.

Semantic universals consist of some properties of generalized quantifiers including monotonicity, quantity, and conservativity. A quantifier Q will be upward or downward monotone. If the truth value of a Model does not change after expanding or shrinking its scope, it is called monotonicity. For example, quantifier some is upward monotone and few is downward monotone where exactly five is neither an upward nor downward monotone. The following universal is proposed based on monotonicity:

- All simple (quantified) determiners express monotone quantifiers (Barwise & Cooper, 1981)

Another universal property, all quantifiers describe the quantity. The true value of a model depends on the size of the sets defined as a noun or a verb phrase. For example, "three" can be quantitative. But "first three" can not be quantitative since it is not only addresses the size but also the arrangement of an object. Keenan and Stavi (1986) formulate and defined the following semantic universal:

- All simple determiners are quantitative.

In this sentence, "Every bicycle is blue", here we can ignore external argument, blue, as the bicycle is an internal argument denoted by every. Barwise & Cooper (1981) address this 'living on the internal argument'. Therefore quantifier *every* is called conservative. Also, we can define quantifier, Q as a conservative if and only if the following biconditional is true (Tim Hunter and Jeffrey Lidz):
$R(X)(Y) \iff R(X)(X \cap Y)$
For every, (1) is true only whenever (2) is true:

1. Every bicycle is blue.
2. Every bicycle is a blue bicycle.

The following universal is proposed in terms of conservativity:

- All simple determiners express conservative quantifiers (Barwise & Cooper, 1981)

4 Methods

Here I discuss the following methods using generalized quantifiers in order to find the complexities and learnability in human and neural networks. First of all, Shane Steinert-Threlkeld, Jakub Szymanik proposed a cognition model to compute the complexity in terms of learnability. They claim Minimum expression length, less Kolmogorov complexity make universal quantifier to learn simpler.

4.1 Minimal Expression Length

Quantifier expression length is the summation of the operator in an expression. But quantifiers' expression is not unique. For example, at least 2 can be expressed as $(3 <\mid A \cap B \mid)$ and by $\neg(\mid A \cap B \mid < 2)$. So, the length of these two expressions is 3 and 4 respectively. Minimal expression length is just the shortest length of possible expressions. Shane Steinert-Threlkeld, Jakub Szymanik collected all the expression meanings for each model for all possible sizes. They performed both descriptive statistics and logistic regression to compute the minimal expression length

4.2 Kolmogorov complexity

In one study, Shane Steinert-Threlkeld, Jakub Szymanik used Kolmogorov complexity to compute the differences in quantifier expression complexity. Kolmogorov complexity (K) measures how much an individual sequence of symbols can be compressed. When a sequence contains regularities, these regularities can be exploited to produce a shorter description of that sequence. One of the

drawbacks of using Kolmogorov is that the value of K is uncomputable. For this reason, they use Lempel-Ziv algorithm for lossless data compression (Lempel & Ziv, 1976). Lempel-Ziv complexity LZ(x) is the number of subsequences of a sequence, x. In this way, they compare the differences in the complexity of different quantifiers instead of calculating a fixed complexity value.

4.3 Novel word learning

One of the finest approaches to test learnability is defining a novel word and checking the ability to learn this word. Hunter and Lidz[2013] choose not all and not only as their novel quantifier meanings. They used a method that teaches a contrast indirectly, the "picky puppet" paradigm. This method links a contrast to what a puppet likes and does not like. Large cards, printed in color, displayed various scenes on a beach and a grassy area next to the beach, with boys and girls. Children were introduced to a puppet who likes some cards but dislikes others. For each card the children were told whether the puppet likes the card or not. After this training, participants were presented with five new cards to sort. All training and testing cards were presented in the same fixed order.

Then, Jennifer Spenader and Jill de Villiers replicate the same method with adults and children with a greater number of participants. They also used known and novel quantifiers meaning by using ***flep*** with situation verification with correction. They distributed objects to the two different sets of figurines (police and pirates). Then participants have to describe the situation related to the meaning of ***flep***. Finally, puppet analysis the description whether it is right or wrong.

4.4 Neural Network Model

Shane Steinert-Threlkeld and Jakub Szymanik used Neural Network model to describe learnability based on universal quantifiers. Neural network is a model inspired from human nervous system. Such a network learns to define a function by gradually updating its parameter and tries to find a universal formula to solve the issue. They used Recurrent Neural Network (RNN) architecture which takes two inputs, a Model, M and a quantifier Expression, Q and gives output 1 or 0 if $M \in Q$. or not. The final output was probabilities of True or False. Probabilities close to zero means False and close to 1 means True.

5 Experiments and Results

5.1 Minimum Expression Length and Kolmogorov

Using the described procedures, they generate 24,632 semantically unique quantifiers expressions up to expression length 5. For each quantifier, they tested the property of conservativity, and they computed their complexity scores, for minimal expression length (ML). Both descriptive statistics and Logistic Regression show a negative relationship between complexity and property of the Conservative quantifier. Quantifiers satisfying the universal are simpler than those that do not: they have a shorter minimal description length.

In the Kolmogorov complexity experiment, they took differing pairs of quantifiers in which one satisfies the universal and the other does not. To test the Conservativity universal, they compared the conservative quantifier most, meaning $\mid A \cap B \mid > \mid A \setminus B \mid$, with the non-conservative quantifier M, meaning $\mid A > B \mid$. The descriptive statistics show that for all model sizes and for all model sequences, conservative most has the same complexity as nonconservative M. Second, we compared the conservative quantifier not all, meaning $B \subseteq A$, with the non-conservative quantifier not only, meaning $B \subseteq A$. For model sizes 1 to 10 descriptive statistics show that not all is more complex in 55.9% of the cases and less complex in 40.8% of the cases.

5.2 Human Learnability

Hunter and Lidz have experimented with children to prove that children can learn conservative quantifiers easier than non-conservative quantifiers. In their study, a total of 20 subjects (4-5 years old) were assigned to learn a new quantifier from five training items: either the conservative ***gleeb***, meaning not all, or the nonconservative ***gleeb'***, meaning not only. They found that children had significant success in learning the former, but no subject in the non-conservative ***gleeb'*** condition demonstrated having learned the new quantifier.

Jennifer Spenader and Jill de Villiers replicate the same method with 4 other conditions. First of all, they tried to find the same result with adults as Hunter and Lidz did in their study. They choose 18 english speaking adult participants. They teach conservative ***not all*** to 9 of them and non-

conservative ***not only*** to rest of the 9 participants. In their second experiment, they applied the same method by like Hunter and Lidz extending the test set. In the following experiment, they introduce a novel quantifier without using the negation of know quantifiers. And in the final experiment, they use their novel expression with correction ***flep*** as not all and not only to train the children. After training, children were asked to take over the role of the puppet, first verifying whether or not the situation presented was true or false, given the sentence and question, and then, if the sentence was false, children were asked to change the situation by manipulating the objects to make it true.

5.3 Train Neural Network Model

Shane Steinert-Threlkeld, Jakub Szymanik did 3 experiment for each universal porperties with LSTM model. They choose a pair of quantifiers having a conservative, not all, and non-conservative quantifier, not only, in each pair like Hunter & Lidz. They then train their network with each pair and compare the time it takes to converge for each pair. It will take less time to converge if it is easy to learn. The algorithm has three parameters: the maximum length of a model, the number of data points to generate, and a set of quantifiers. To balance the data distribution, they applied under-sampling. Therefore, each pair have the same number of data points and the learning model did not get biased to certain data points. They also used a combination of conservative 'most' with a non-conservative quantifier with meaning $A \cap B > \mid A \setminus B \mid$. In their first experiment, they applied a paired ttest of the convergence points of quantifiers to determine which quantifier converged earlier than any other quantifier.

6 Discussion and Conclusion

It is observed that all-natural language quantifiers are conservative on language acquisition despite being linguistics differences of different languages.

Though the Minimal Expression Length shows that conservative quantifiers are easier to learn, Kolmogorov complexity of universal quantifiers shows no relation between conservative quantifiers and learnability. Neither complexity nor learnability distinguishes conservative from non-conservative quantifiers. Conservativity may have a different source that differentiates from other quantifier properties.

Hunter and Lidz found in their experiment learning conservative quantifiers have an advantage over non-conservative quantifiers. But it is not true in other experiments. There was no difference in learnability of conservative or non-conservative quantifiers when Spenader and De Villiers replicated this experiment with adults and children. Unlikely, Adults showed a better understanding of non-conservative quantifiers. One of the reasons could be overthinking. Also, Children showed a great performance in learning known quantifiers. Situation verification with a correction maybe help participants to understand the meaning of conservative and non-conservative quantifiers. But children were unable to learn novel quantifiers not all and not only even with situation verification correction. As Negation makes it hard to learn, it might be a reason children have difficulties in learning. In all experiments, they were unable to make a relation between conservativity and learnability.

Similarly, Steinert-Threlkeld and Szymanik found that a conservative quantifier is not easier to learn than a non-conservative one. One of the main differences between the Children and Neural Network model is children learn more from positive aspects whereas an artificial model treats positive and negative sentences equally.

References