Does BERT agree? Evaluating knowledge of structure dependence through agreement relations

Abstract

Learning representations that accurately model semantics is an important goal of natural language processing research. Many semantic phenomena depend on syntactic structure. Recent work examines the extent to which state-of-the-art models for pre-training representations, such as BERT, capture such structure-dependent phenomena, but is largely restricted to one phenomenon in English: number agreement between subjects and verbs. We evaluate BERT's sensitivity to four types of structure-dependent agreement relations in a new semi-automatically curated dataset across 26 languages. We show that both the single-language and multilingual BERT models capture syntax-sensitive agreement patterns well in general, but we also highlight the specific linguistic contexts in which their performance degrades.

Introduction

Learning general-purpose sentence representations which accurately model sentential semantic content is a current goal of natural language processing research BIBREF0, BIBREF1, BIBREF2, BIBREF3. A prominent and successful approach is to pre-train neural networks to encode sentences into fixed length vectors BIBREF4, BIBREF5, with common architecture choices based on recurrent neural networks BIBREF6, BIBREF7, convolutional neural networks, or transformers BIBREF8. Many core linguistic phenomena that one would like to model in general-purpose sentence representations depend on syntactic structure BIBREF9, BIBREF10. Despite the fact that none of the aforementioned architectures have explicit syntactic structural representations, there is some evidence that these models can approximate such structure-dependent phenomena under certain conditions BIBREF11, BIBREF12, BIBREF13, BIBREF14, in addition to their widespread success in practical tasks.

The recently introduced BERT model BIBREF15, which is based on transformers, achieves state-of-the-art results on eleven natural language processing tasks. In this work, we assess BERT's ability to learn structure-dependent linguistic phenomena of agreement relations. To test whether BERT is sensitive to agreement relations, we use the cloze test BIBREF16, in which we mask out one of two words in an agreement relation and ask BERT to predict the masked word, one of the two tasks on which BERT is initially trained.

BIBREF17 adapted the experimental setup of BIBREF13, BIBREF11 and BIBREF18 to use the cloze test to assess BERT's sensitivity to number agreement in English subject-verb agreement relations. The results showed that the single-language BERT model performed surprisingly well at this task (above 80% accuracy in all experiments), even when there were multiple “distractors” in the sentence (other nouns that differed from the subject in number). This suggests that BERT is actually learning to approximate structure-dependent computation, and not simply relying on flawed heuristics.

However, English subject-verb agreement is a rather restricted phenomenon, with the majority of verbs having only two inflected forms and only one morphosyntactic feature (number) involved. To what extent does Goldberg's BIBREF17 result hold for subject-verb agreement in other languages, including more morphologically rich ones, as well as for other types of agreement relations? Building on Goldberg's BIBREF17 work, we expand the experiment to 26 languages and four types of agreement relations, which include more challenging examples.

In Section 2, we define what is meant by agreement relations and outline the particular agreement relations under study. Section 3 introduces our newly curated cross-linguistic dataset of agreement relations, while section 4 discusses our experimental setup. We report the results of our experiments in section 5. All data and code are available at https://github.com/geoffbacon/does-bert-agree.

Structure-dependent agreement relations

Agreement phenomena are an important and cross-linguistically common property of natural languages, and as such have been extensively studied in syntax and morphology BIBREF19. Languages often express grammatical features, such as number and gender, through inflectional morphology. An agreement relation is a morphophonologically overt co-variance in feature values between two words in a syntactic relationship BIBREF20. In other words, agreement refers to when the morphosyntactic features of one word are reflected in its syntactic dependents. In this way, agreement relations are overt markers of covert syntactic structure. Thus, evaluating a model's ability to capture agreement relations is also an evaluation of its ability to capture syntactic structure.

Following BIBREF21, we call the syntactically dependent word the “target” of the agreement relation, and the word with which it agrees we call the “controller”. An example of an agreement relation in English is given in (UNKREF2), in which the inflected form of the verb be (are) reflects the plural number of its syntactic head keys. In all examples in this section, the controller and target are given in bold. In this example, keys is the controller and are is the target of the agreement relation.

The keys to the door are on the table.

The agreement relation in (UNKREF2) is between a subject and its verb, but there are other types of agreement relations. In addition to subject-verb agreement, three other types of agreement relations are cross-linguistically common: agreement of noun with i) determiner, ii) attributive adjective and iii) predicate adjective BIBREF22. The latter two types are distinguished by whether the adjective modifies the noun within a noun phrase or whether it is predicated of the subject of a clause. The first two types are sometimes categorized as nominal concord rather than agreement, but for our purposes this is merely a difference in terminology.

The morphosyntactic feature in the agreement relation in (UNKREF2) is number, a feature that is cross-linguistically common in agreement systems. In addition to number, the most commonly involved in agreement relations are gender, case and person BIBREF22.

With its comparatively limited inflectional morphology, English only exhibits subject-verb and determiner agreement (in demonstratives, “this” vs. “these”) and even then only agrees for number. Languages with richer inflectional morphology tend to display more agreement types and involve more features. French, for example, employs all four types of agreement relations. Examples are given in (UNKREF3)-(UNKREF6). The subject and verb in (UNKREF3) agree for number, while the noun and determiner in (UNKREF4), the noun and attributive adjective in (UNKREF5) and the subject and predicated adjective in (UNKREF6) agree for both number and gender.

`The keys to the door are on the table.'

`I can see the keys.'

`I no longer want the completely broken keys.'

`The keys to the door are broken.'

Previous work using agreement relations to assess knowledge of syntactic structure in modern neural networks has focussed on subject-verb agreement in number BIBREF17, BIBREF11, BIBREF13. In our work, we study all four types of agreement relations and all four features discussed above. Moreover, previous work using any method to assess BERT's knowledge of syntactic structure has focussed exclusively on the single-language English model BIBREF23, BIBREF17, BIBREF24, BIBREF25, BIBREF26, BIBREF27. We expand this line of work to 26 languages. Not all languages in our sample

exhibit all four types of agreement nor use all four features examined, but they all exhibit at least one of the agreement types involving at least one of the features.

Data

Our study requires two types of data. First, we need sentences containing agreement relations. We mask out one of the words in the agreement relation and ask BERT to predict the masked word. We are interested in BERT's ability to predict words that respect the agreement relation, that is, words which share the morphosyntactic features of the word with which it agrees. To measure this, we need to know the feature values for each word in BERT's vocabulary. This is our second type of data. Throughout this paper, we refer to the first type of data as the cloze data, and the second as the feature data.

In the design of our datasets, we followed two principles. First, we chose data sources that are available across multiple languages, because we are interested in cross-linguistic generality. The languages in this study are those with sufficiently large data sources that also appear in the multilingual BERT model. Second, we use naturally-occurring data (cf. BIBREF18).

Data ::: Cloze data

We sourced our cloze data from version 2.4 of the Universal Dependencies treebanks BIBREF28. The UD treebanks use a consistent schema across all languages to annotate naturally occurring sentences at the word level with rich grammatical information. We used the part-of-speech and dependency information to identify potential agreement relations. Specifically, we identified all instances of subject-verb, noun-determiner, noun-attributive adjective and subject-predicate adjective word pairs. We then used the morphosyntactic annotations for number, gender, case and person to filter out word pairs that disagree due to errors in the underlying data source (e.g. one is annotated as plural while the other is

singular) or that are not annotated for any of the four features.

This method is language-agnostic, but due to errors in the underlying UD corpora, yielded some false positives (e.g. predicate adjective agreement in English). To correct for this, we consulted reference grammars of each language to note which of the four types of agreement exist in the language. We removed all examples that are of the wrong type for the language (8% of harvested examples). Across the 26 languages, we curated almost one million cloze examples. Their breakdown across agreement type and language is shown in Tables 1 and 2.

In all four types of agreement studied, the controller of the agreement is a noun or pronoun, while the target can be a determiner, adjective or verb. Because of this part-of-speech restriction, we chose to mask out the controller in every cloze example so that BERT is evaluated against the same vocabulary across all four types. This also means that we only need to collect feature data on nouns and pronouns.

Data ::: Feature data

Our feature data comes from both the UD and the UniMorph projects BIBREF29. The UniMorph project also uses a consistent schema across all languages to annotate word types with morphological features. Although this schema is not the same as that used in UD, there is a deterministic mapping between the two BIBREF30.

In this work, a word can take on a particular bundle of feature values (e.g. singular, feminine and third person) if it appears with those features in either UD or UniMorph. The UniMorph data directly specifies what bundles of feature values a word can take on. For the Universal Dependencies data, we say a word can take on a particular bundle if we ever see it with that bundle of feature values in a Universal Dependencies corpus for that language. Both sources individually allow for a word to have multiple

feature bundles (e.g. sheep in English can be singular or plural). In these cases, we keep all possible feature bundles. Finally, we filter out words that do not appear in BERT's vocabulary.

Experiment

Our experiment is designed to measure BERT's ability to model syntactic structure. Our experimental set up is an adaptation of that of BIBREF17. As in previous work, we mask one word involved in an agreement relation and ask BERT to predict it. BIBREF17, following BIBREF13, considered a correct prediction to be one in which the masked word receives a higher probability than other inflected forms of the lemma. For example, when dogs is masked, a correct response gives more probability to dogs than dog. This evaluation leaves open the possibility that selectional restrictions or frequency are responsible for the results rather than sensitivity to syntactic structure BIBREF11. To remove this possibility, we take into account all words of the same part-of-speech as the masked word. Concretely, we consider a correct prediction to be one in which the average probability of all possible correct words is higher than that of all incorrect words. By “correct words”, we mean words with the exact same feature values and the same part of speech as the masked word. By “incorrect words”, we mean words of the same part of speech as the masked word but that differ from the masked word with respect to at least one feature value. We ignore cloze examples in which there are fewer than 10 possible correct and 10 incorrect answers in our feature data. The average example in our cloze data is evaluated using 1,468 words, compared with 2 in BIBREF17.

Following BIBREF17, we use the pre-trained BERT models from the original authors, but through the PyTorch implementation. BIBREF17 showed that in his experiments the base BERT model performed better than the larger model, so we restrict our attention to the base model. For English, we use the model trained only on English data, whereas for all other languages we use the multilingual model.

Results

Overall, BERT performs well on our experimental task, suggesting that it is able to model syntactic structure. BERT was correct in 94.3% of all cloze examples. This high performance is found across all four types of agreement relations. Figure FIGREF13 shows that BERT performed above 90% accuracy in each type. Performance is best on determiner and attributive agreement relations, while worst on subject-verb and predicate adjective.

In figure FIGREF14, we see BERT's performance for each language. BERT performs well for the majority of languages, although some fare much worse than others. It is important to note that it is an unfair comparison because even though the datasets were curated using the same methodology, each language's dataset is different. It is possible, for example, that the examples we have for Basque are simply harder than they are for Portuguese.

Finally, we ask how BERT's performance is affected by distance between the controller and the target, as well as the number of distractors. Figure FIGREF15 shows BERT's performance, aggregated over all languages and types, as a function of the distance involved in the agreement, while figure FIGREF16 shows the same for number of distractors. There is a slight but consistent decrease in performance as the distance and the number of distractors increase. The decline in performance begins later in figure FIGREF16 but drops more rapidly once it does.

Related work

Given the success of large pre-trained language representation models on downstream tasks, it is not surprising that that the field wants to understand the extent of their linguistic knowledge. In our work, we looked exclusively at the predictions BERT makes at the word level. BIBREF24 and BIBREF26 examined

the internal representations of BERT to find that syntactic concepts are learned at lower levels than semantic concepts. BIBREF23 are also interested in syntactic knowledge and propose a method to evaluate whether entire syntax trees are embedded in a linear transformation of a model's word representation space, finding that BERT does capture such information. As a complementary approach, BIBREF27 studied the attention mechanism of BERT, finding clear correlates with interpretable linguistic structures such as direct objects, and suggest that BERT's success is due in part to its syntactic awareness. However, by subjecting it to existing psycholinguistic tasks, BIBREF32 found that BERT fails in its ability to understand negation. In concurrent work, BIBREF33 show that BERT does not consistently outperform LSTM-based models on English subject-verb agreement tasks.

Conclusions & future work

Core linguistic phenomena depend on syntactic structure. Yet current state-of-the-art models in language representations, such as BERT, do not have explicit syntactic structural representations. Previous work by BIBREF17 showed that BERT captures English subject-verb number agreement well despite this lack of explicit structural representation. We replicated this result using a different evaluation methodology that addresses shortcomings in the original methodology and expanded the study to 26 languages. Our study further broadened existing work by considering the most cross-linguistically common agreement types as well as the most common morphosyntactic features. The main result of this expansion into more languages, types and features is that BERT, without explicit syntactic structure, is still able to capture syntax-sensitive agreement patterns well. However, our analysis highlights an important qualification of this result. We showed that BERT's ability to model syntax-sensitive agreement relations decreases slightly as the dependency becomes longer range, and as the number of distractors increases. We release our new curated cross-linguistic datasets and code in the hope that it is useful to future research that may probe why this pattern appears.

The experimental setup we used has some known limitations. First, in certain languages some of the cloze examples we studied contain redundant information. Even when one word from an agreement relation is masked out, other cues remain in the sentence (e.g. when masking out the noun for a French attributive adjective agreement relation, number information is still available from the determiner). To counter this in future work, we plan to run our experiment twice, masking out the controller and then the target. Second, we used a different evaluation scheme than previous work BIBREF17 by averaging BERT's predictions over many word types and plan to compare both schemes in future work.