Modeling German Verb Argument Structures: LSTMs vs. Humans

Abstract

LSTMs have proven very successful at language modeling. However, it remains unclear to what extent they are able to capture complex morphosyntactic structures. In this paper, we examine whether LSTMs are sensitive to verb argument structures. We introduce a German grammaticality dataset in which ungrammatical sentences are constructed by manipulating case assignments (eg substituting nominative by accusative or dative). We find that LSTMs are better than chance in detecting incorrect argument structures and slightly worse than humans tested on the same dataset. Surprisingly, LSTMs are contaminated by heuristics not found in humans like a preference toward nominative noun phrases. In other respects they show human-similar results like biases for particular orders of case assignments.

Introduction

Among neural networks, LSTMs BIBREF0 are commonly used for language modeling. Although new architectures BIBREF1, BIBREF2 challenge this standard, LSTMs remain competitive for language modeling BIBREF3. However, despite the success of LM LSTMs, it is not clear what makes them so effective. In particular, are representations derived through language modeling able to effectively encode syntactic structures and relations? Do they encode them in a reliable and systematic way?

The typical metric used to compare LMs, perplexity, is not adapted to address these questions. Perplexity measures the probability assigned to held-out data from the corpus the LM is trained on. Because the held-out and training data are typically randomly extracted from an initial corpus, they have similar statistics, which is good from a machine learning viewpoint, but bad from the viewpoint of linguistic analysis: perplexity is mostly sensitive to the most common sentence types in the initial corpus and

therefore will not reflect well the behavior of the LM in the tail of the distribution. In addition, the sentences extracted from a natural corpus confound several factors: syntax, semantics, pragmatics, etc. further complicating the interpretation of a good perplexity score.

To circumvent this limitation, recent work has focused on using probing techniques inspired by linguistic and psycholinguistics (for instance, grammaticality or acceptability judgments, or forced choice). In addition, instead of using sentences from the training corpus, studies rely more and more on automatically constructed test sentences, which enable for a removal of the bias in the original corpus and focus on particular linguistic phenomena. Here, we will use acceptability judgments operationalized by the log probability of sentences according to the LM and sets of synthetic sentences generated from template sentences to probe for a challenging linguistic structure: verb argument structure.

Verb argument structure provides languages with a way to link syntactic position in a sentence (subject, direct object, etc) with semantic roles (agent, patient, etc), in other words, to determine who is doing what. It is currently unknown whether neural LMs purely trained from surface statistics are able to capture this kind of structure, or whether additional information from another modality would be needed to provide some semantic grounding.

Verb argument structure is typically correlated to sentence position in many languages like English. But in other languages with relatively free word order, it is indicated by morphological markers. Here, we study German, where the arguments of a verb can occur in any position (when occurring within a relative clause), and is indicated by the case of the noun phrase (nominative, accusative, etc).

We setup a test of argument structure representation by presenting a trained LM with carefully constructed sets of sentences that either have the right set of arguments, or abnormal sentences where one case is missing or duplicated. We use word order permutations to control for unigram and positional

statistics. If the LM is able to track argument structure irrespective of word order, it should assign lower grammaticality scores (log probabilities) to the incorrect sentences as compared to the correct ones.

Since at the level of the sentence, we study a global rather than local syntactic phenomenon, we depart from earlier work BIBREF4, BIBREF5, BIBREF6, BIBREF7 and do not compare pairs of sentences. Rather, we compare a set of valid grammatical variations of the template to a corresponding set of grammatical violations of the template. Thus, for each template, we measure the model's ability to discriminate grammatical sentences from ungrammatical ones using receiver operating characteristic curves, or ROC curves. We also compute the area under the ROC curve, or AUC. In our results, we often report the average AUC over templates as our metric.

We evaluate three LMs on our dataset, the two-layer LSTM of BIBREF8 trained on German Wikipedia text, as well as n-gram baselines using the same corpus. We ask proficient German speakers to annotate our sentences for grammaticality, providing a human comparison. Since some of these sentences are rather implausible because of the permutations, we also collect human meaningfulness scores. We find that our dataset is challenging for both LMs and humans and that LMs lag behind human performance.

Related work

Grammaticality judgments for recurrent networks have been investigated since BIBREF9, who use closely matched pairs of sentences to investigate grammatical correctness. This approach has been adopted recently to assess the abilities of RNNs, and LSTMs in particular, to capture syntactic structures. For instance, BIBREF4 and BIBREF5 use word probes in minimally different pairs of English sentences to study number agreement. To discriminate grammatical sentences from ungrammatical ones, they retrieve the probabilities of the possible morphological forms of a target word, given the probability of the previous words in the sentence. Practically, in the sentence “the boy is sleeping”, the network has detected

number agreement if $\mathbf {P}(w = is) > \mathbf {P}(w = are)$. This methodology has also been adapted by BIBREF10 to models trained with a masked language-modeling objective. Those works find that in the absence of many detractors or complex sentence features, recent language models perform well at the number-agreement problem in English.

More closely related to our work, BIBREF11 use word probes to examine whether LSTMs understand Basque agreement. Like German, Basque is a morpho-syntactically rich language with relatively free word order, thus providing a challenging setting for the LM. In contrast to our work, the LM's ability to understand verb argument structure is tested on number-agreement and on suffix recovery tasks, which involve localized changes rather than whole sentence perturbations and re-orderings.

In parallel to work focusing on word probe probabilities, another closely related line of inquiry has investigated surprisal, the inverse log probability assigned to a specific prediction by a model. For instance, BIBREF12 and BIBREF13 examine many syntactic phenomena, including filler gap dependencies and garden path effects.

We depart from these approaches because our test set encompasses whole sentence variations, such as argument reordering. Word probes are therefore less apt to capture such changes. Instead, we choose to follow BIBREF6 and BIBREF7 in taking the more general approach of comparing whole sentence probabilities as our grammaticality probe. This method, which also corresponds to the sentence-level LogProb acceptability measure of BIBREF14, evaluates whether the model assigns a higher log probability to sentences which are grammatical than to sentences which are not.

In contrast with approaches that seek to probe language models directly, other approaches involve fine-tuning representations to a specific syntactic task using a task-specific supervision signal. For instance, BIBREF15 introduce CoLA, a binary acceptability dataset whose example sentences are taken

from linguistic publications. They train a classifier on top of frozen ELMo BIBREF16 layers to assess performance at acceptability judgments. Later work BIBREF17, BIBREF18 has focused on fine-tuning an entire pre-trained model to the acceptability task, such as is done for BERT BIBREF17. Both of those paradigms do not directly evaluate syntactic ability but rather whether pre-trained representations can be effectively transferred to learn to solve specific syntax problems.

Verb Argument Structure Dataset Construction ::: Templates

Our test sentences were automatically generated from fifty grammatical sentences which we call templates. These templates are all constructed the same way: the main clause “wir wissen, dass...” (“we know that”), followed by a subordinate clause with a subject (nominative case), a verb in the past tense form, a direct object (accusative case) and an indirect object (dative case). For simplicity purposes, we did not use any adjective. In the Template of Figure FIGREF3, “the minister” is the subject, “that bill” the direct object, and “the Senate” the indirect object of “announced”.

We constructed a dataset designed to expose impossible verb argument structures by manipulating the arguments' case assignments. We introduced these changes within subordinate clauses rather than main clauses, because German subordinate clauses have a more flexible noun phrases order than main clauses. This specificity allows us to test whether models are able to capture syntactic dependencies when the arguments' positions vary.

In German, the syntactic role of noun phrases is indicated by the morphological form of its constituents: determiners and nouns take different suffixes, if not completely different forms, according to their case assignment. However, feminine, neutral and all plural noun phrases share common morphological forms. Thus, to avoid sentence duplicates within our dataset, all noun phrases are singular masculine.

Verb Argument Structure Dataset Construction ::: Grammatical Sets

To control for all possible argument orders and words syntactic roles, for each template, we change (i) the positions of the three verb arguments in the subordinate clause and (ii) the case assignments of each noun group. There are three verb arguments, leading to six different position permutations. Similarly, they are three unique case assignments, leading to six possible case assignments. By generating all such permutations, we create $6 \times 6 = 36$ grammatical sentences for each template, yielding 1800 grammatical sentences in total. In Figure FIGREF3, we show an example where only the positions of the subject and the indirect object are switched, which does not alter the meaning. We also show an example where only the case assignments of the subject and the indirect object are switched: “The Senate” becomes the subject and “the minister” the indirect object. The case permutations were done by retrieving the desired case markings (nominative, accusative or dative) from a dictionary mapping the vocabulary's nouns to their morphological forms. Case permutations change sentence meaning. In practice, some of our sentences will be implausible yet grammatical, in contrast with BIBREF6.

Verb Argument Structure Dataset Construction ::: Case Violation Sets

We constructed ungrammatical sentences using the same templates. Briefly, we substituted one of the case assignments by another one already present in the sentence, which creates a grammatical violation: sentences contain three noun phrases and only two case assignments, one being duplicated. In Figure FIGREF3, we show how we apply this to a template sentence to create grammatical violations.

For each case violation, we generated 36 sentences containing a case violation from every template. Thus, from each of our 50 templates, we generated 36 valid grammatical variations and 108 ungrammatical variations. Note also that throughout the number of words in our dataset stays constant (11 words per sentence), so that log probabilities are more comparable. Overall, our dataset comprises

7,200 sentences, of which 1,800 are grammatical and 5,400 are ungrammatical.

Methods ::: Human Evaluations

To generate human results for our dataset, we hire annotators proficient in German on Amazon Mechanical Turk.

Methods ::: Human Evaluations ::: Sentence Grammaticality

We asked Amazon Mechanical turkers to assess the sentence grammaticality on a scale from 1 to 10, where 1 means grammatically incorrect and 10 means grammatically correct. Before the task started, respondents were shown examples of grammatical sentences and ungrammatical sentences. Importantly, it was indicated that grammatical sentences were not necessarily meaningful. As an example, we translated to German Chomsky's famous quote: “Colorless green ideas sleep furiously” BIBREF19. Each respondent graded 50 sentences, with the following constraints: (i) each sentence comes from a different template, to avoid that past sentences impact future ratings; (ii) twenty-five percent of the sentences shown are grammatical, mirroring the construction of the dataset; (iii) sentences selected are randomly chosen among the 144 possibilities for each template, so that each user is exposed to a wide variety of case assignments, argument orders and grammatical violations; (iv) no sentence is annotated twice.

Methods ::: Human Evaluations ::: Sentence Meaningfulness

For grammatical sentences only, we also conduct meaningfulness evaluations. Similarly to our grammaticality experiment, users are asked to grade 50 sentences from 1 to 10, where 1 is meaningless and 10 is meaningful. They were also shown examples of meaningful and meaningless grammaticality correct German sentences before starting the evaluations. Constraints are the same as above, except

that all sentences are grammatical and that there are thus only 36 possibilities per template.

Methods ::: Human Evaluations ::: Ensuring German Proficiency

To ensure that all annotators are proficient in German, we took the following steps: (i) we only accepted annotators from German-speaking countries; (ii) instructions are given in German only; (iii) annotators took a short German grammar test on conjugation and declination knowledge; (iv) filler sentences (easy sentences for which answers are known and obvious to proficient German speakers) are inserted throughout the annotation process to ensure annotators stay focused; (v) we remove annotators who took less than a third of the average time to complete the assignment after checking that they also underperform on our test questions.

Methods ::: Human Evaluations ::: Pairwise Ranking and Individual Grading

As noted, we do not ask humans to compare minimally differing sentences, but rather to grade individual sentences. This setup differs from earlier work such as BIBREF6 who show both sentences simultaneously and ask humans to pick the most grammatical one. This approach prevents humans from using the differences between the sentences to form a judgment on grammaticality; rather they must judge each sentence on its own. In doing so, the human setup is closer to that of language models: when we use log probability scores of LMs, we do not enable them to learn from the differences between the sentences to form a judgment.

In total, we collected 2,750 annotations from 55 annotators for sentence grammaticality (38% of the dataset) and 1,800 annotations from 36 annotators for sentence meaningfulness (100% of grammatical sentences). We do not have grammaticality annotations for all sentences due to a lack of proficient German annotators on Amazon Mechanical Turk. Our human results for grammaticality are computed on

this subset of the dataset.

Methods ::: Language Models

We use the pre-trained word-level language model (German WordNLM) described and trained by BIBREF8. The model is a two-layer LSTM without attention, a hidden dimension of 1,204, and word embeddings of dimension 200 for the 50,000 most frequent words. It was trained on a corpus from German Wikipedia, totalling 819 million words. The 50,000 most-frequent words in this corpus are used as the vocabulary and embedded in 200-dimensional vector space. The model reaches a perplexity of 37.96 on this dataset. We use unigram and bigram language models that use the same corpus with Laplace smoothing as baselines. The probability of test sentences according to the language models is computed using the chain rule:

Each of these log probabilities can be read from the softmax outputs of the LSTM, or directly estimated in the case of the unigram and bigram models. We also tried normalizing for unigram frequency as proposed by BIBREF20 but found like BIBREF6 that it did not improve results for the LSTM.

Results ::: Main Classification Task

Figure FIGREF11 shows the distribution of the log probability scores predicted by the LSTM and the distribution of the grammaticality scores given by humans. Figure FIGREF16 presents the distributions and average of the AUC values computed per template (50 in total), both for the models' log probability scores and the human grammaticality scores. Performances are rather modest, with a mean AUC of 0.56 for the LTSM and of 0.58 for humans, compared to the chance score of 0.5 for the unigram and bigram models. As expected, the n-gram baselines perform exactly at chance, confirming that they do not represent verb argument structures and that LMs need a deeper encoding to be able capture syntax

within sentences. We also notice that AUC varies relatively little across different templates for our models, indicating that the particular choice of template has little impact. For humans, the wider spread in results can be attributed partially to the fact that 55 random permutations out of the 144 permutations were annotated for each template. Therefore, it might have been easier to distinguish grammatical sentences from ungrammatical ones for some templates than others.

Surprisingly, humans performed only slightly better than the LSTM. We believe that this is due two factors. First, we presented the sentences in a scrambled order and asked for an absolute grammaticality judgment. It may be more difficult to put a sentence on a 1 to 10 scale than making pairwise judgments. Second, our sentences may be particularly challenging. The grammatical sentences contained both unusual argument orders and semantically odd situations, thus inciting participants to rate them low. While these factors could be expected to impact the LSTM, it is more surprising that they impact humans, despite precise instructions to rate on grammaticality rather than meaning or frequency. In addition, as can be seen in Figure FIGREF11b, some ungrammatical sentences were rated as highly grammatical by humans. We suspect that these are cases of inattention, as in our test set the distinction between grammatical and ungrammatical rest on a single word, and even a single character (the distinction between 'der' and 'den', for instance).

Results ::: Case Frequency Bias

In Table TABREF18, we further investigate our grammaticality results by segregating them by case violation type (duplicate nominative, accusative or dative). While humans tend to give similar scores for each violation type, models tend to assign higher log probability scores to sentences with doubled nominatives than to grammatical sentences, leading to worse than chance performance on Nominative violations. Conversely, models tend to assign lower log probability scores to sentences with doubled datives, likely because these sentences lack either a nominative or an accusative, both of which are more

frequent than dative. This leads to better than human performance on this case violation. Such behavior is probably due to the fact that German being a non pro-drop language, every verb must have a nominative case, making nominative more frequent than accusative, and that dative even rarer. This frequency bias is worse for models that are directly based on frequency, such as our unigram and bigram models. However, our LSTM is not exempt from it, confirming that RNNs rely in part on frequency cues.

Results ::: Argument Order Preferences

In Figure FIGREF20, we explore the effect of argument order. Despite the fact that all argument orderings should be equally valid from a grammatical perspective, we find that humans tend to favour more 'canonical' orders, with nominative-accusative-dative being the preferred order. Models also assign higher log probability scores to the canonical order compared to others. It is likely that some orders occur more frequently than others in German, thus leading to a frequency bias for both models and humans. Although sentences with shuffled argument order have the same meaning as those without shuffled order, we find a similar bias for the meaningfulness scores.

Interestingly, even though the case orders preferred by the LSTM correlate with those of humans, there are also subtle differences: we also find that models tend to prefer argument orders that start with dative to those that start with accusative, when the opposite is true for human grammaticality scores. The origin of such differences is unclear. Understanding it more fully would require to obtain distributional statistics on the order of such phrases in the original corpus.

Results ::: Animacy Preferences

As mentioned in Section SECREF3, some of our grammatical sentences are semantically implausible though syntactically valid. This is because we create highly unlikely associations of case assignments

and thematic roles when we permute the case assignments from the original sentence template. For instance, one permutation has a bill announcing a minister to the senate. Such unlikely sentences may be rejected by participants as ungrammatical even though they were specifically requested to ignore semantic plausibility. Similarly, they may affect neural models through the distributional correlates of meaningfulness: in any language corpus, a bill being an inanimate object is more likely to be an object (accusative case) than a subject (nominative case).

To check for the existence of such effect, we categorized the nouns in all of our sentences as animate and inanimate, and computed the human and machine scores of our grammatical sentences as a function of the association between case and animacy. Table TABREF22 shows that indeed, both humans and machines are biased by animacy-case associations: all share a preference for animate for nominative (subject) and dative (indirect object). By contrast, negative AUC values for accusative indicate that direct objects are preferred as inanimate.

Results ::: Restricting the Analysis to Plausible Sentences

To see the impact of such biases, we re-analysed the human and machine scores by restricting the AUCs to the non-permuted sentences, i.e, the sentences whose case assignments correspond to that of the original templates. These templates were constructed to be plausible, and indeed the average human plausibility scores for these non-permuted orders of 5.33 is higher than for the permuted ones 3.61. In this analysis, we therefore include the 6 valid grammatical argument order permutations and all 108 grammatical violations for each template sentence.

The results are shown in Table TABREF24. As expected, the human AUC scores are higher in this restricted analysis than in the full dataset shown in Table TABREF18. Note that the model scores are also higher, which suggests that the algorithms are also sensitive to meaningfulness, probably through its

effects on the distribution of case for the different nouns in the training corpus.

Results ::: Correlation between model and human ratings

In Table TABREF26, we show correlations between human judgments of grammaticality, meaningfulness and LSTM log probabilities. Unsurprisingly, all variables are positively correlated, which supports our earlier findings. More surprising is that the LSTM is more correlated with both grammaticality and meaningfulness than meaningfulness is with grammaticality. Note that meaningfulness and grammaticality have been annotated by different annotators, which might help explain this finding.

Conclusions

We set up a well controlled grammaticality test for the processing of argument structure in neural language models and in humans. The results show that LSTMs are better than chance in detecting an abnormal argument structure, despite the fact that the arguments could occur in any position, due to the generally free word order of phrases in German relative clauses. The average performance of models, though, is far from 100% correct and lower than humans, and the error patterns differ markedly. Contrary to humans, neural language models are overly sensitive to frequency distribution of phrase types. For instance, they assign a higher probability to sentences containing multiple nominative phrases than a correct sentence with only one nominative phrase. This frequency bias directly reflects the frequency of nominative, accusative and dative in the language, as the same bias is found in unigram and bigram models. Similar to the conclusion reached by BIBREF21 in their investigation of the error patterns made by RNNs and humans on syntactic agreement, we find that the syntactic representations of humans and LSTMs differ in some respects.

Despite this difference, neural models are able to mimic the pattern of human responses for grammatical

sentences. As has been noted previously, humans are not uniformly considering all grammatical sentences as grammatical, i.e, grammaticality judgments are graded BIBREF22. Humans tend to reject sentences with unusual word orders. For instance, they prefer the canonical Nominative-Accusative-Dative order over all of the others orders. A similar pattern is found in neural models, although the details differ somewhat.

Another point of convergence is found with regards to the association between case and semantic features: humans prefer that nominative phrases are animate, and accusative inanimate, a pattern also found in neural networks. This shows that humans have difficulties in judging grammaticality as separate from other factors like frequency and meaningfulness, especially when sentences are presented independently instead of in minimal pairs. In this respect, humans are quite comparable to neural models.

Overall, the difficulty of neural networks to detect incorrect argument structure as such (especially spectacular in the case of duplicate nominatives), provides us a clue that these models may not be fully able to represent such structures, above and beyond their probability distributions.

Acknowledgments

The team's project is funded by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), Almerys (industrial chair Data Science and Security), and grants from Facebook AI Research (Research Grant), Google (Faculty Research Award), Microsoft Research (Azure Credits and Grant), and Amazon Web Service (AWS Research Credits).