1. Scientific Question

A foundational question in cognitive science is whether linguistic knowledge is fundamentally categorical (Sprouse, 2007 [9], Fong et al. 2013 [4]) or probabilistic (Abney, 1995 [1], 2011 [2]; Manning, 2003 [7]) in nature. Grammatical judgments present a problem for probabilistic models in that probabilities cannot be mapped directly to grammaticality, because of the influence of sentence length and lexical frequency. In this paper we look at the problem of predicting grammaticality judgments using probabilistic models. We tested four models. We tested the relative contribution of each model, and each feature subset through which speakers represent their syntactic knowledge may diverge significantly from classical formal theories of syntactic structure.

2. Data and Methodology

Lau et al. [2014] report the results of an experiment in which 500 sentences from the British National Corpus (BNC) are translated into four languages, and then back into English using Google Translate. This produces a test set of 2500 English sentences exhibiting various degrees of syntactic and lexical infelicity, as well as a significant subset of well-formed sentences. We annotated this test set using Amazon Mechanical Turk (AMT) crowd sourcing to obtain a large collection of individual and manual native speaker judgments. We employed three modes of presentation for judgement. These included binary, four way, and a sentence with an underlying range of 100 points. We found a high Pearson coefficient correlation of judgements in pairwise comparisons among these modes of presentation. In general, the judgements for the test set display a substantial amount of gradience. This pattern was confirmed in a subsequent AMT experiment on 100 randomly chosen "linguists examples" (50 good sentences and 50 starred ones) from a text book on syntactic theory.

3. Unsupervised Language Models

In recent work we have constructed enriched language models to predict speakers' grammaticality judgements. Building on the results of Clark et al. [2013] we devise various forms of normalisation to control for a variety of largely unsupervised learning models, we show high correlations between the predictions of our models and manual native speaker judgments. These results suggest that probabilistic models are, in principle, capable of accounting for observed grammaticality judgments. We are primarily motivated by the question of how speakers represent syntactic knowledge. However, there are also significant engineering applications for a system that can successfully predict speakers' grammaticality judgements. These include language generation, machine translation, and text summarisation systems. Such a system could also contribute to automatic essay scoring, and to second language learning.

4. Grammaticality Measures

We apply the following grammaticality measures to map logprob-values into relative grammaticality scores.

- Mean Logprob
- Norm Logprob
- SOR
- Minimum


5. Results

We used the Pearson correlation coefficient to test the predictions of each model against mean speakers' judgements for our test set. The results for the best grammaticality measures are summarised below.

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependency Grammar</td>
<td>0.22</td>
</tr>
<tr>
<td>Lexical 2-gram</td>
<td>0.37</td>
</tr>
<tr>
<td>Lexical 3-gram</td>
<td>0.42</td>
</tr>
<tr>
<td>Lexical 4-gram</td>
<td>0.43</td>
</tr>
<tr>
<td>One-Tier BHMM</td>
<td>0.45</td>
</tr>
<tr>
<td>Two-Tier BHMM</td>
<td>0.50</td>
</tr>
<tr>
<td>Lexical 4-grams</td>
<td>0.43</td>
</tr>
<tr>
<td>BHMMs</td>
<td>0.59</td>
</tr>
<tr>
<td>All Models</td>
<td>0.62</td>
</tr>
</tbody>
</table>

We tested the relative contribution of each model, and each class of models, with feature ablation.

6. Comparison with Current Work

Heilman et al. [2014] [5] present a supervised system for predicting grammaticality judgements. This system uses features from a collection of supervised probabilistic parsers, as well as a spelling feature. They train it on a corpus of English as a second language (ESL) learners' essays, annotated with grammaticality judgements in a four category specification mode of presentation. They test their system on a hold out set from this corpus. They report a Pearson correlation of 0.044 between the predicted scores of their system and the mean judgements of the annotators. For the unsupervised experiment we used our models as trained on the BNC. For SV regression we trained them on their annotated corpus. In both cases we tested the models on their test set. In non-supervised mode our best result is given by a 4-gram model, which approaches 0.5. When we combine all our models we obtain 0.64. Adding spelling, which is central to Heilman et al.'s system, and combining our features optimally (lexical 4-gram BHMM + spelling feature) for our SV regression gives us 0.645.

7. Discussion and Conclusions

We have found that of the models that we tested, our Bayesian HMMs provide the best results for predicting speakers grammaticality judgements. This result has been sustained across two distinct domains, AMT annotations of Google translated BNC sentences, and expert annotations of sentences extracted from ESL essays. Our second-tier BHMM is, in effect, a data driven POS classifier, and our two-tier BHMM is a type of data driven classifier. The fact that these two BHMMs consistently outperform a generative dependency grammar on the task of predicting grammaticality judgements raises the intriguing possibility that the models through which speakers represent their syntactic knowledge may diverge significantly from classical formal theories of syntactic structure.

8. Acknowledgements

The research described in this paper was done in the framework of the Statistical Models of Grammaticality (SMG, http://www.dos.kcl.ac.uk/staff/lappin/smg/), funded by grant ES/I008688/1 from the Economic and Social Research Council of the UK. We are grateful to Douglas Sandy and Garry Smith of the Centre for Integrative Neuroscience and Neurodynamics at the University of Reading for giving us access to the computing cluster at the Centre, and for their generous support. We also thank Shuy Cohen for giving us so freely of his time and expertise in assisting us with the adaptation and parallelisation of the code for Dagecon.

References