Modelling Speakers Grammaticality Judgements

Jey Han Lau, Alexander Clark, and Shalom Lappin

Department of Philosophy, King's College London

`jeyhan.lau@gmail.com, alexsclark@gmail.com, shalom.lappin@kcl.ac.uk`

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. Grammaticality 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 a set of enriched models on a data set of crowd sourced grammaticality judgments for sentences that have had errors introduced through round trip machine translation. Using various normalisation methods, applied to a variety of largely unsupervised learning models, we show high correlations between the predictions of our models and mean 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.

2. Data and Methodology

Lau et al. (2014) [6] 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 mean native speaker judgements. We employed three modes of presentation for judgement. These included binary, four way, and a sliding scale with an underlying range of 100 points. We found a high Pearson coefficient correlation of judgements in pairwise comparisons among these modes of presentation.

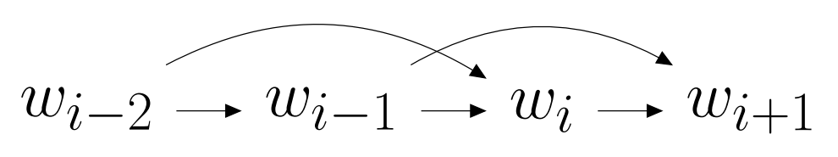
In general, the judgements for the test set display a substantial amount of gradience. This pattern was confirmed in a subsequent AMT experiment on 100 randomly chosen "linguists examples" (50 good sentences and 50 starred ones) from a text book on syntactic theory.

3. Unsupervised Language Models

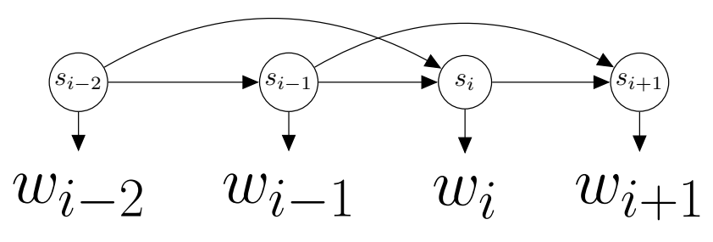
In recent work we have constructed enriched language models to predict speakers' grammaticality judgements. Building on the results of Clark et al. (2013) [3] we devise various forms of normalisation to translate the logprob distributions of a model for a test set to relative acceptability values. These functions modify the logprob values to control for factors like sentence length and word frequency.

We tested four models.

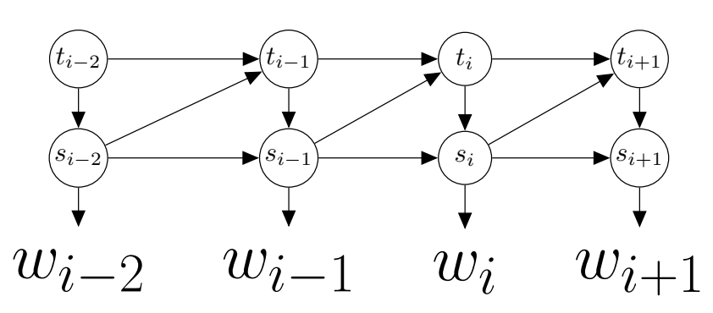
1. Lexical n-gram models (bigram, trigram, and 4-gram)
2. A parallelised implementation of a dependency grammar (Shay Cohen, Dageem, 2008-2011, `http://www.ark.cs.cmu.edu/DAGEEM/`)
3. A second-order Bayesian Hidden Markov Model (BHMM)
4. A two-tier BHMM



Lexical Trigram



Second-order Bayesian HMM



Two-tier Bayesian HMM

Our second-order BHMM can be thought of as a data driven lexical classifier. Our two-tier BHMM can be regarded as a data driven phrasal chunker.

4. Grammaticality Measures

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

$$\text{Mean Logprob} = \frac{\text{Sent Logprob}}{n}$$

$$\text{Norm. Logprob} = \frac{\text{Sent Logprob}}{\text{Sent Unigram Logprob}}$$

$$\text{SLOR} = \frac{\text{Sent Logprob} - \text{Sent Unigram Logprob}}{n}$$

$$\text{Minimum} = \min_i \left(-\frac{\log \theta_{w_i|w_{i-1}w_{i-2}}}{\log \theta_{w_i}} \right)$$

Clark et al. (2013) [3] propose Mean Logprob, Norm. Logprob, and Minimum. Pauls and Klein (2012) [8] suggest SLOR.

Mean Logprob normalises by sentence length, and Norm. Logprob by word frequency. SLOR is effectively equivalent to Norm. Logprob. Minimum normalises by the lowest unigram logprob in a sentence.

Clark et al. (2013) generalise Minimum to Mean of the First Quartile (MFQ) by ordering the single n-gram logprobs from the lowest to the highest, and considering the first (lowest) quartile. MFQ normalizes the logprobs for these n-grams by the unigram probability of the head lexical item, and takes the mean of these scores.

In general Norm. Logprob and SLOR consistently yielded the best results across different models.

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.

Model	Best Correlation
Dependency Grammar	0.32
Lexical 2-gram	0.37
Lexical 3-gram	0.42
Lexical 4-gram	0.43
Bayesian HMM	0.46
Two-Tier BHMM	**0.50**

We employed Support Vector Machine (SVM) regression for supervised learning, to compare the performances of the individual models, and to test their aggregate level of achievement. We obtained the results shown below.

Model(s)	Unsup.	Supervised
Dependency Grammar	0.32	0.34
Lexical 2-gram	0.37	0.43
Lexical 3-gram	0.42	0.48
Lexical 4-gram	0.43	0.50
One-Tier BHMM	0.45	0.55
Two-Tier BHMM	**0.50**	0.57
Lexical N-grams	–	0.51
BHMMs	–	**0.59**
All Models	–	**0.62**

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

Model(s)	Correlation	
All Models	0.62	
– Dependency Grammar	0.62	(±0.00)
– Lexical 2-gram	0.61	(–0.01)
– Lexical 3-gram	0.62	(±0.00)
– Lexical 4-gram	0.62	(±0.00)
– One-Tier BHMM	0.61	(–0.01)
– Two-Tier BHMM	0.59	(–0.03)
– Lexical N-grams	0.59	(–0.03)
– **BHMMs**	**0.52**	**(–0.10)**

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 expert judgements in a four category classification mode of presentation. They test their system on a hold out set from this corpus. They report a Pearson correlation of 0.644 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 with SV regression, we achieve 0.6. Adding spelling, which is central to Heilman et al.'s system, and combining our features optimally (lexical 4-gram + HMM + spelling feature) for our SV regression gives us 0.645.

System	Pearson's r
Heilman et al. (2014)	0.644
Unsupervised Best	0.498
SVR: All Models	0.604
SVR: All Models+Spell	0.623
SVR: 4-gram+BHMM+Spell	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-order BHMM is, in effect, a data driven POS classifier, and our two-tier BHMM is a type of data driven chunker. 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. Acknowlegments

The research described in this paper was done in the framework of the Statistical Models of Grammaticality (SMOG) project at King's College London (`http://www.dcs.kcl.ac.uk/staff/lappin/smog/`), funded by grant ES/J022969/1 from the Economic and Social Research Council of the UK.

We are grateful to Douglas Saddy 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 Shay Cohen for giving so freely of his time and expertise in assisting us with the adaptation and parallelisation of the code for Dageem.

References

[1] Steven Abney. Statistical methods and linguistics. In J. Klavans and P. Resnik, editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pages 1–26. MIT Press, Cambridge, MA, 1996.

[2] Steven Abney. Data–intensive experimental linguistics. *Linguistic Issues in Language Technology*, 6:1–29, 2011.

[3] Alexander Clark, Gianluca Giorgolo, and Shalom Lappin. Statistical representation of grammaticality judgements: The limits of n-gram models. In *Proceedings of the ACL Workshop on Cognitive Modelling and Computational Linguistics*, Sofia, Bulgaria, 2013.

[4] Sandiway Fong, Igor Malioutov, Beracah Yankama, and Robert C. Berwick. Treebank parsing and knowledge of language. In Aline Villavicencio, Thierry Poibeau, Anna Korhonen, and Afra Alishahi, editors, *Cognitive Aspects of Computational Language Acquisition*, Theory and Applications of Natural Language Processing, pages 133–172. Springer Berlin Heidelberg, 2013.

[5] Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), Volume 2: Short Papers*, pages 174–180, Baltimore, Maryland, 2014.

[6] Jey Han Lau, Alexander Clark, and Shalom Lappin. Measuring gradience in speakers' grammaticality judgements. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society (CogSci 2014)*, pages 821–826, Quebec City, Canada, 2014.

[7] Christopher Manning. Probabilistic syntax. In Rens Bod, Jennifer Hay, and Stefanie Jannedy, editors, *Probabilistic Linguistics*. The MIT Press, 2003.

[8] A. Pauls and D. Klein. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 959–968. Jeju, Korea, 2012.

[9] J. Sprouse. Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics*, 1:123–134, 2007.