Modelling Speakers' Grammaticality Judgements

Jey Han Lau, Alex Clark, and Shalom Lappin
King's College London
jeyhan.lau@gmail.com, alexsclark@gmail.com, shalom.lappin@kcl.ac.uk

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 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.

In recent work we have constructed enriched language models to predict speakers' grammaticality judgements. Building on the results of Clark et al. (2013b; 2013a) we use scoring functions to map 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. They can also identify local points of reduction in probability within a string.

We tested four models

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

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 scoring function 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 in Table 1.

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)**

Heilman et al. (2014) 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

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**

Table 1: Results

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.

We trained our models on their corpus and tested them 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.

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.

References

A. Clark, G. Giorgolo, and S. Lappin. 2013a. Statistical representation of grammaticality judgements: the limits of n-gram models. In *Proceedings of the ACL Workshop on Cognitive Modelling and Computational Linguistics*, pages 28–36.

A. Clark, G. Giorgolo, and S. Lappin. 2013b. Towards a statistical model of grammaticality. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pages 2064–2069.

M. Heilman, A. Cahill, N., M. Lopez, M. Mulholland, and J. Tetreault. 2014. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Conference of the ACL*, Baltimore.

J.H. Lau, A. Clark, and S. Lappin. 2014. Measuring gradience in speakers' grammaticality judgements. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, Quebec City, Canada.