Statistical Representation of Grammaticality Judgements: the Limits of N-Gram Models

Alexander Clark, Gianluca Giorgolo, and Shalom Lappin
Department of Philosophy, King's College London
firstname.lastname@kcl.ac.uk

Abstract

We use a set of enriched n-gram models to track grammaticality judgements for different sorts of passive sentences in English. We construct these models by specifying scoring functions to map the log probabilities (logprobs) of an n-gram model for a test set of sentences onto scores which depend on properties of the string related to the parameters of the model. We test our models on classification tasks for different kinds of passive sentences. Our experiments indicate that our n-gram models achieve high accuracy in identifying ill-formed passives in which ill-formedness depends on local relations within the n-gram frame, but they are far less successful in detecting non-local relations that produce unacceptability in other types of passive construction. We take these results to indicate some of the strengths and the limitations of word and lexical class n-gram models as candidate representations of speakers' grammatical knowledge.

1 Introduction

Most advocates (Pereira, 2000; Bod et al., 2003) and critics (Chomsky, 1957; Fong et al., 2013) of a probabilistic view of grammatical knowledge have assumed that this view identifies the grammatical status of a sentence directly with the probability of its occurrence. By contrast, we seek to characterize grammatical knowledge statistically, but without reducing grammaticality directly to probability. Instead we specify a set of scoring procedures for mapping the logprob value of a sentence into a relative grammaticality score, on the basis of the properties of the sentence and of the logprobs that an n-gram word model generates for the corpus containing the sentence. A scoring procedure in this set generates scores in terms of which we construct a grammaticality classifier, using a parameterized standard deviation from the mean value. The classifier provides a procedure for testing the accuracy of different scoring criteria in separating grammatical from ungrammatical passive sentences.

We evaluate this approach by applying it to the task of distinguishing well and ill-formed sentences with passive constructions headed by four different sorts of verbs: intransitives (*appear, last*), pseudo-transitives, which take a restricted set of notional objects (*laugh a hearty laugh, weigh 10 kg*), ambiguous transitives, which allow both agentive and thematic subjects (*the jeans / the tailor fitted John*), and robust transitives that passivize freely (*write, move*). Intransitives and pseudo-transitives generally yield ill-formed passives. Passives formed from ambiguous transitives tend to be well-formed only on the agentive reading. Robust transitives, for the most part, yield acceptable passives, even if they are semantically (or pragmatically) odd.

Experimenting with several scoring procedures and alternative values for our standard deviation parameter, we found that our classifier can distinguish pairwise between elements of the first two classes of passives and those of the latter two with a high degree of accuracy. However, its performance is far less reliable in identifying the difference between ambiguous and robust transitive passives. The first classification task relies on local lexical patterns that can be picked up by n-gram models, while the second requires identification of anomalous relations between passivized verbs and *by*-phrases, which are not generally accessible to measurement within the range of an n-gram.

We also observed that as we increased the size of the training corpus, the performance of our enriched models on the classification task also increased. This result suggests that better n-gram language models are more sensitive to the sorts of patterns that our scoring procedures rely on to generate accurate grammaticality classifications.

We note the important difference between

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 28–36,
Sofia, Bulgaria, August 8, 2013. ©2013 Association for Computational Linguistics

grammaticality and acceptability. Following standard assumptions, we take grammaticality to be a theoretical notion, and acceptability to be an empirically testable property. Acceptability is, in part, determined by grammaticality, but also by factors such as sentence length, processing limitations, semantic acceptability and many other elements. Teasing apart these two concepts, and explicating their precise relationship raises a host of subtle methodological issues that we will not address here. Oversimplifying somewhat, we are trying to reconstruct a gradient notion of grammaticality which is derived from probabilistic models, that can serve as a core component of a full model of acceptability.

We distinguish our task from the standard task of error detection in NLP (e.g. Post (2011)), that can be used in various language processing systems, such as machine translation (Pauls and Klein, 2012), language modeling and so on. In error detection, the problem is a supervised learning task. Given a corpus of examples labeled as grammatical or ungrammatical, the problem is to learn a classifier to distinguish them. We use supervised learning as well, but only to measure the upper bound of an unsupervised learning method. We assume that native speakers do not, in general, have access to systematic sets of ungrammatical sentences that they can use to calibrate their judgement of acceptability. Rather ungrammatical sentences are unusual or unlikely. However, we use some ungrammatical sentences to set an optimal threshold for our scoring procedures.

2 Enriched N-Gram Language Models

We assume that we have some high quality language model which defines a probability distribution over whole sentences. As has often been noted, it is not possible to reduce grammaticality directly to a probability of this type, for several reasons. First, if one merely specifies a fixed probability value as a threshold for grammaticality, where strings are deemed to be grammatical if and only if their probability is higher than the threshold, then one is committed to the existence of only a finite number of grammatical sentences. The probabilities of the possible strings of words in a language sum to 1, and so at most $1/\epsilon$ sentences can have a probability of at least ϵ. Second, probability can be affected by factors that do not influence grammaticality. For example, the word 'yak' is rarer (and therefore less probable) than the word 'horse', but this does not affect the relative grammaticality of 'I saw a horse' versus 'I saw a yak'. Third, a short ungrammatical sentence may have a higher probability than a long grammatical sentence with many rare words.

In spite of these arguments against a naive reduction of grammaticality, probabilistic inference does play a role in linguistic judgements, as indicated by the fact that they are often gradient. Probabilistic inference is pervasive throughout all domains of cognition (Chater et al., 2006), and therefore it is plausible to assume that knowledge of language is also probabilistic in nature. Moreover language models do seem to play a crucial role in speech recognition and sentence processing. Without them we would not be able to understand speech in a noisy environment.

We propose to accommodate these different considerations by using a scoring function to map probabilities to grammaticality rankings. This function does not apply directly to probabilities, but rather to the parameters of the language model. The probability of a particular sentence with respect to a log-linear language model will be the product of certain parameters: in log space, the sum. We define scores that operate on this collection of parameters.

2.1 Scores

We have experimented with scores of two different types that correlate with the grammaticality of a sentence. Those of the first type are different implementations of the idea of normalizing the logprob assigned by an n-gram model to a string by eliminating the significance of factors that do not influence the grammatical status of a sentence, such as sentence length and word frequency. Scores of the second type are based on the intuition that the (un)grammaticality of a sentence is largely determined by its problematic components. These scores are functions of the lowest scoring n-grams in the sentence.

Mean logprob (ML) This score is the logprob of the entire sentence divided by the length of the sentence, or equivalently the mean of the logprobs for the single trigrams:
$\mathsf{ML} = \frac{1}{n} \log P_{\text{TRIGRAM}}(\langle w_1, \ldots, w_n \rangle)$
By normalizing the logprob for the entire sentence by its length we eliminate the effect of sentence length on the acceptability score.

Weighted mean logprob (WML) This score is calculated by dividing the logprob of the entire sentence by the sum of the unigram probabilities of the lexical items that compose the sentence:
$\mathsf{WML} = \frac{\log P_{\textsc{trigram}}(\langle w_1,\ldots,w_n\rangle)}{\log P_{\textsc{unigram}}(\langle w_1,\ldots,w_n\rangle)}$
This score eliminates at the same time the effect of the length of the sentence and the lower probability assigned to sentences with rare lexical items.

Synctactic log odds ratio (SLOR) This score was first used by Pauls and Klein (2012) and performs a normalization very similar to WML (we will see below that in fact the two scores are basically equivalent):
$\mathsf{SLOR} =$
$\frac{\log P_{\textsc{trigram}}(\langle w_1,\ldots,w_n\rangle) - \log P_{\textsc{unigram}}(\langle w_1,\ldots,w_n\rangle)}{n}$

Minimum (Min) This score is equal to the lowest logprob assigned by the model to the n-grams of the sentence divided by the unigram logprob of the lexical item heading the n-gram:
$\mathsf{Min} = \min_i \left[\frac{\log P(w_i|w_{i-2}w_{i-1})}{\log P(w_i)}\right]$
In this way, if a single n-gram is assigned a low probability (normalized for the frequency of its head lexical item), then this low score is in some sense propagated to the whole sentence.

Mean of the first quartile (MFQ) This score is a generalization of the Min score. We order the single n-gram logprobs from the lowest to the highest, and we consider the first (lowest) quartile. We then normalize the logprobs for these n-grams by the unigram probability of the head lexical item, and we take the mean of these scores. In this way we obtain a score that is more robust than the simple Min, as, in general, a grammatical anomaly influences the logprob of more than one n-gram.

2.2 N-Gram Models

We are using n-gram models on the understanding that they are fundamentally inadequate for describing natural languages in their full syntactic complexity. In spite of their limitations, they are a good starting point, as they perform well as language models across a wide range of language modeling tasks. They are easy to train, as they do not require annotated training data.

We do not expect that our n-gram based grammaticality scores will be able to idenitfy all of the cases of ungrammaticality that we encounter. Our working hyposthesis is that they can capture cases of ill-formedness that depend on local factors, that can be identified within n-gram frames, as opposed to those which involve non-local relations. If these models can detect local grammaticality violations, then we will have a basis for thinking that richer, more structured language models can recognize non-local as well as local sources of ungrammaticality.

3 Experiments with Passives

Rather than trying to test the performance of these models over all types of ungrammaticality, we limit ourselves to a case study of the passive. By tightly controlling the verb types and grammatical construction to which we apply our models we are better able to study the power and the limits of these models as candidate representations of grammatical knowledge.

3.1 Types of Passives

Our controlled experiments on passives are, in part, inspired by speakers' judgments discussed in Ambridge et al. (2008). Their experimental work measures the acceptability of various passive sentences.

The active-passive alternation in English is exemplified by the pair of sentences

- John broke the window.
- The window was broken by John.

The acceptability of the passive sentence depends largely on lexical properties of the verb. Some verbs do not allow the formation of the passive, as in the case of pure intransitive verbs like *appear*, discussed below, which permit neither the active transitive, nor the passive.

We conducted some prelimiary experiments, not reported here, on modelling the data on passives from recent work in progress that Ben Ambridge and his colleagues are doing, and which he was kind enough to make available to us. We observed that the scores we obtained for our language models did not fully track these judgements, but we did notice that we obtained much better correlation at the low end of the judgment distribution. In Ambridge's current data this judgement range corresponds to passives constructed with intransitive verbs.

The Ambridge data indicates that the capacity of verbs to yield well-formed passive verb phrases forms a continuum. Studying the judgement patterns in this data we identified four reasonably salient points along this hierarchial continuum.

First, at the low end, we have intransitives like *appear*: (**John appeared the book. *The book was appeared*). Next we have what may be described as pseudo-transitives verbs like *laugh*, which permit only notional NP objects and do not easily passivize (*Mary laughed a hearty laugh/*a joke. ?A hearty laugh/*A joke was laughed by Mary*) above them. These are followed by cases of ambiguous transitives like *fit*, which, in active form, carry two distinct readings that correspond to an agentive and a thematic subject, respectively.

- The tailor fitted John for a new suit.
- The jeans fitted John

Only the agentive reading can be passivized.

- John was fitted by the tailor.
- *John was fitted by the jeans.

Finally, the most easily passivized verbs are robust transitives, which take the widest selection of NP subjects in passive form (*John wrote the book. The book was written by John*).

This continuum causes well-formedness in passivization to be a gradient property, as the Ambridge data illustrates. Passives tend to be more or less acceptable along this spectrum. The gradience of acceptability for passives implies the partial overlap of the score distributions for the different types of passives that our experiments show.

The experiments were designed to test our hypothesis that n-gram based language models are capable of detecting ungrammatical patterns only in cases where they do not depend on relations between words that cross the n-word boundary applied in training. Therefore we expect such a model to be capable of detecting the ungrammaticality of a sentence like *A horrible death was died by John*, because the trigrams *death was died*, *was died by* and *died by John* are unlikely to appear in any corpus of English. On the other hand, we do not expect a trigram model to store the information necessary to identify the relative anomaly of a sentence like *Two hundred people were held by the theater*, because all the trigrams (as well as the bigrams and the unigrams) that constitute the sentence are likely to appear with reasonable frequency in a large corpus of English.

The experiments generalize this observation and test the performance of n-gram models on a wider range of verb types. To quantify the performance of the different models we derive simple classifiers using the scores we have defined and testing them in a binary classification task. This task measures the ability of the classifier to distinguish between grammatical sentences, and sentences containing different types of grammatical errors.

The models are trained in an unsupervised manner using only corpus data, which we assume to be uniformly grammatical. In order to evaluate the scoring methods, we use some supervised data to set the optimal value of a simple threshold. This is not however a supervised classification task: we want to see how well the scores *could* be used to separate grammatical and ungrammatical data, and though unorthodox, this seems a more direct way of measuring this conditional property than stipulating some fixed threshold.

3.2 Training data

We used the British National Corpus (BNC) (BNC Consortium, 2007) to obtain our training data. We trained six different language models, using six different subcorpora of the BNC. The first model used the entire collection of written texts annotated in the BNC, for a total of approximately 100 million words. The other models were trained on increasingly smaller portions of the written texts collection: 40 million words, 30 million words, 15 million words, 7.6 million words, and 3.8 million words. We constructed these corpora by randomly sampling an appropriate number of complete sentences.

All models were trained on word sequences. For smoothing the n-gram probability distributions we used Kneser-Ney interpolation, as described in Goodman (2001).

3.3 Test data

We constructed the test data for our hypothesis in a controlled fashion. We first compiled a list of verbs for each of the four verb types that we consider (intransitives, pseudo-transitives, ambiguous transitives, and robust transitives). We selected verbs from the BNC that appeared at least 100 times in their past participle form in the entire corpus in order to ensure a sufficient number of passive uses in the training data.[1] We selected 40 intransitive verbs, 13 pseudo transitives, 23 ambiguous transitives and 40 transitive verbs. To classify the verbs we relied on our intuitions as native speakers of English.

Using these lists we automatically generated four corpora by selecting an agent and a patient from a predefined pool of NPs, randomly selecting a determiner (if necessary) and a number (if the NP allows plurals). The resulting corpora are of the following sizes:

- intransitive verbs – 24480 words, 3240 sentences,
- pseudo transitive verbs – 7956 words, 1053 sentences,
- ambiguous transitive verbs – 14076 words, 1863 sentences,
- robust transitive verbs – 24480 words, 3240 sentences.

Each corpus was evaluated by the six models. We computed our derived scores for each sentence on the basis of the logprobs that the language models assigns.

3.4 Binary classifiers

For each model and for each score we constructed a set of simple binary classifiers on the basis of the results obtained for the transitive verb corpus. We took the mean of each score assigned by the model to the transitive sentences, and we set different thresholds by subtracting from this value a number of standard deviations ranging from 0 to 2.75. The rationale behind these classifiers is that, assuming the passives of the robust transitives to be grammatical, the scores for the other cases should be comparatively lower. Therefore by setting a threshold "to the left" of the mean we should be able to distinguish between grammatical sentences, whose score is to the right of the threshold, and ungrammatical ones, expected to a have a score lower than the threshold. Formally the classifier is defined as follows:

$$c_s(w) = \begin{cases} + & \text{if } s(w) \geq m - S \cdot \sigma \\ - & \text{otherwise} \end{cases} \tag{1}$$

where s is one of our scores, w is the sentence to be classified, $s(w)$ represents the value assigned by the score to sentence w, m is the mean for the score in the transitive condition, σ is the standard deviation for the score again in the transitive condition, and S is a factor by which we move the threshold away from the mean. The classifier assigns the grammatical ($+$) tag only to those sentences that are assigned values higher than the threshold $m - S \cdot \sigma$.

Alternatively in terms of the widely used z-score, defined as $z_s(w) = (s(w) - m)/\sigma$ we can say that w is classified as grammatical iff $z_s(w) \geq -S$.

4 Results

For reasons of space we will limit the presentation of our detailed results to the 100 million word model, as it offers the sharpest effects. We will, however, also report comparisons on the most important metrics for the complete set of models.

In Figure 1 we show the distribution of the five scores for the four different corpora (transitive, ambiguous, pseudo, and intransitive) obtained using the 100 million word model. In all cases we observe the same general pattern: the sentences in the corpus generated with robust transitives are assigned comparatively high scores, and these gradually decrease when we consider the ambiguous, the pseudo and the intransitive conditions. Interestingly, this order reflects the degree of "transitivity" that these verb types exhibit. Notice, however, that the four conditions seem to group into two different macro-distributions. On the right we have the transitive-ambiguous sentences and on the left the pseudo-intransitive cases. This partially confirms our hypothesis that n-gram models have problems recognizing lexical dependencies that determine the felicitousness of passives constructed using ambiguous transitive verbs, as these are, for the most part, non-local. Nevertheless, it is important to note that the overlap of the distributions for these two cases is also due to the fact that many cases in the ambiguous transitive corpus are indeed grammatical.

Figure 2 summarizes the (balanced) accuracies obtained by our classifiers for each comparison, by each model. These results confirm our hypothesis that the classifiers tend to perform better when distinguishing passive sentences constructed with a robust transitive verbs from those headed by

[1] Notice that in most cases the past participle form is the same as the simple past form, and for this reason we set the threshold to such a high value.

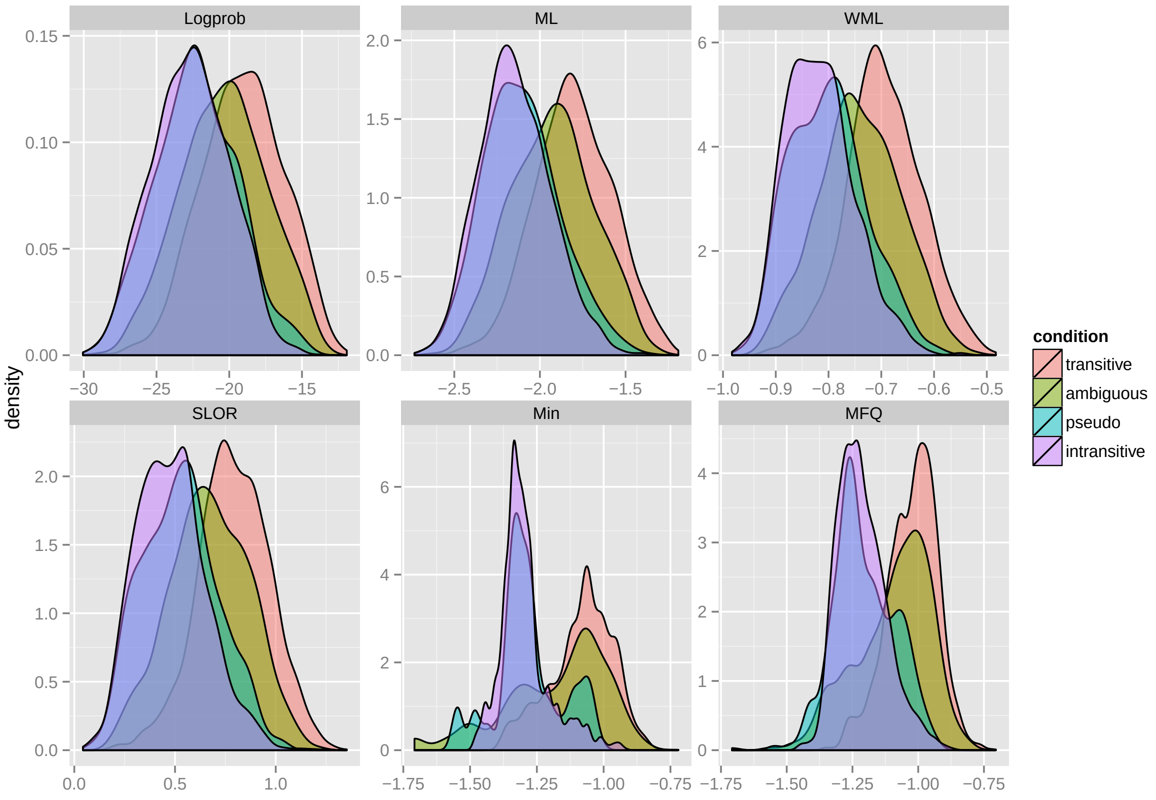

Figure 1: Distributions of the six scores Logprob, ML, WML, SLOR, Min and MFQ for the four different conditions (robust transitive passives, ambiguous transitive passives, pseudo transitive passives and intransitive passives) for the 100 million words language model.

pseudo-transitives and intransitives.

In the comparison between transitive and ambiguous transitive sentences, the classifiers are "stuck" at around 60% accuracy. Using larger training corpora produces only a marginal improvement. This contrasts with what we observe for the transitive/pseudo and transitive/intransitive classification tasks. In the transitive/pseudo task, we already obtain reasonable accuracy with the model trained with the smallest BNC subset. Oddly, the overall best result is achieved with 30 million words, although the result obtained with the model trained on the full BNC corpus is not much lower. For the transitive/intransitive classification task we observe a much steadier and larger growth in accuracy, reaching the overall best result of 85.1%. Table 1 reports the best results for each comparison by each language model. For each condition we report the best accuracy obtained, the corresponding F1 score, the score that achieves the best result, and the best accuracy obtained by just using the logprobs. These results are obtained using different values for the S parameter. However, in general the best results are obtained when the S parameter is set to a value in the interval $[0.5, 1.5]$.

In comparing the performance of the individual scores, we first notice that, while for the transitive/ambiguous comparison all scores perform pretty much at the same level, there is a clear hierarchy between scores for the other comparisons.

We observe that the baseline raw logprob assigned by the n-grams models performs much worse than the scores, resulting in roughly 10% less accuracy than the best performing score in every condition. ML performs slightly better, obtaining around 5% greater accuracy than logprob as a predictor. This shows that even though the length of the sentences in our test data is relatively constant (between 9 and 11 words), there is still an improvement if we take this structural factor into account. The two scores WML and SLOR display the same pattern, showing that they are effectively equivalent. This is not surprising given that they are designed to modify the raw logprob by tak-

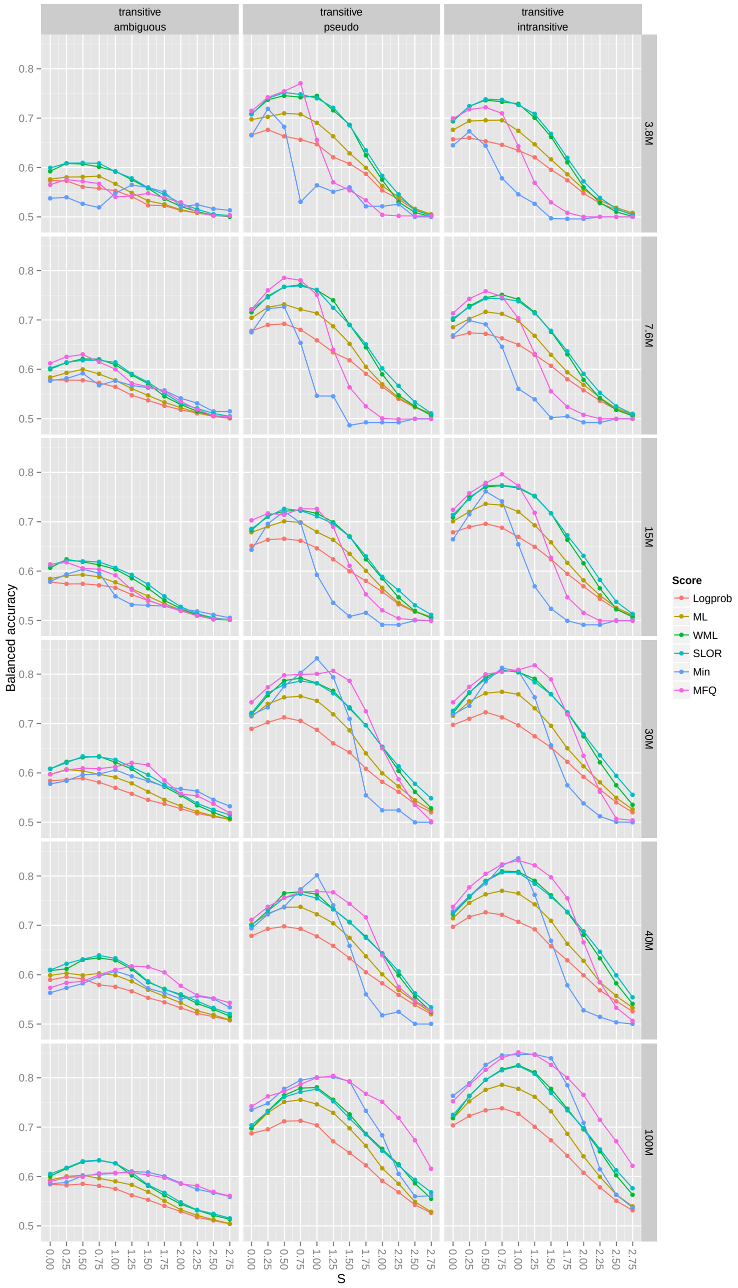

Figure 2: Accuracies for the classifiers for each model. S represents the number of standard deviations "to the left" of the mean of the transitive condition score, used to set the threshold.

ing into account exactly the same factors (length of the sentence and frequency of the unigrams that compose the sentence). These two scores perform generally better in the transitive/ambiguous comparison, and they achieve good performance when the size of the training model is small. However, for the most part, the two scores derived from the logprobs of the least probable n-grams in the sentence, Min and MFQ, get the best results. Min exhibits erratic behavior (mainly due to its non-normal distribution for each condition, as shown in figure 1), and it seems to be more stable only in the presence of a large training set. MFQ has a much more robust contour, as it is significantly less dependent on the choice of S.

5 Conclusions and Future Work

In Clark and Lappin (2011) we propose a model of negative evidence that uses probability of occurrence in primary linguistic data as the basis for estimating non-grammaticality through relatively

Model	Comparison	Best accuracy	F1	Best performing score	Logprob accuracy
3.8M	transitive/ambiguous	60.9%	0.7	SLOR	57.3%
	transitive/pseudo	77%	0.81	MFQ	67.6%
	transitive/intransitive	73.8%	0.72	SLOR	65.6%
7.6M	transitive/ambiguous	62.9%	0.68	MFQ	57.8%
	transitive/pseudo	78.5%	0.76	MFQ	69.1%
	transitive/intransitive	75.8%	0.72	MFQ	67.3%
15M	transitive/ambiguous	62.3%	0.66	WML	57.8%
	transitive/pseudo	72.6%	0.78	SLOR	66.5%
	transitive/intransitive	79.5%	78.3	MFQ	69.5%
30M	transitive/ambiguous	63.3%	0.75	WML	58.9%
	transitive/pseudo	83.1%	0.88	Min	71.2%
	transitive/intransitive	81.8%	0.82	MFQ	72.2%
40M	transitive/ambiguous	63.8%	0.75	SLOR	59.5%
	transitive/pseudo	80.1%	0.86	Min	69.7%
	transitive/intransitive	83.5%	0.83	SLOR	72.6%
100M	transitive/ambiguous	63.3%	0.75	SLOR	58.4%
	transitive/pseudo	80.3%	0.9	MFQ	71.3%
	transitive/intransitive	85.1%	0.85	SLOR	73.8%

Table 1: Best accuracies

low frequency in a sample of this data. Here we follow Clark et al. (2013) in effectively inverting this strategy.

We identify a set of scoring functions based on parameters of probabilistic models that we use to define a grammaticality threshold, which we use to classify strings as grammatical or ill-formed. This model offers a stochastic characterisation of grammaticality without reducing grammaticality to probability.

We expect enriched lexical n-gram models of the kind that we use here to be capable of recognizing the distinction between grammatical and ungrammatical sentences when it depends on local factors within the frame of the n-grams on which they are trained. We further expect them not to be able to identify this distinction when it depends on non-local relations that fall outside of the n-gram frame.

It might be thought that this hypothesis concerning the capacities and limitations of n-gram models is too obvious to require experimental support. In fact, this is not the case. Reali and Christiansen (2005) show that n-gram models can be used to distinguish grammatical from ungrammatical auxiliary fronted polar questions with a high degree of success. More recently Frank et al. (2012) argue for the view that a purely sequential, non-hierarchical view of linguistic structure is adequate to account for most aspects of linguistic knowledge and processing.

We have constructed an experiment with different (pre-identified) passive structures that provides significant support for our hypothesis that lexical n-gram models are very good at capturing local syntactic relations, but cannot handle more distant dependencies.

In future work we will be experimenting with more expressive language models that can represent non-local syntactic relations. We will proceed conservatively by first extending our enriched lexical n-gram models to chunking models, and then to dependency grammar models, using only as much syntactic structure as is required to identify the judgement patterns that we are studying.

To the extent that this research is successful it will provide motivation for the view that syntactic knowledge is inherently probabilistic in nature.

Acknowledgments

The research described in this paper was done in the framework of the Statistical Models of Grammaticality (SMOG) project at King's College London, funded by grant ES/J022969/1 from the Economic and Social Research Council of the UK. We are grateful to Ben Ambridge for providing us with the data from his experiments and for helpful discussion of the issues that we address in this paper. We also thank the three anonymous CMCL 2013 reviewers for useful comments and suggestions, that we have taken account of in preparing the final version of the paper.

References

Ben Ambridge, Julian M Pine, Caroline F Rowland, and Chris R Young. 2008. The effect of verb semantic class and verb frequency (entrenchment) on childrens and adults graded judgements of argument-structure overgeneralization errors. *Cognition*, 106(1):87–129.

BNC Consortium. 2007. The British National Corpus, version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

R. Bod, J. Hay, and S. Jannedy. 2003. *Probabilistic linguistics*. MIT Press.

N. Chater, J.B. Tenenbaum, and A. Yuille. 2006. Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10(7):287–291.

N. Chomsky. 1957. *Syntactic Structures*. Mouton, The Hague.

A. Clark and S. Lappin. 2011. *Linguistic Nativism and the Poverty of the Stimulus*. Wiley-Blackwell, Malden, MA.

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

Sandiway Fong, Igor Malioutov, Beracah Yankama, and Robert C. Berwick. 2013. 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.

Stefan Frank, Rens Bod, and Morten Christiansen. 2012. How hierarchical is language use? In *Proceedings of the Royal Society B*, number doi: 10.1098/rspb.2012.1741.

J.T. Goodman. 2001. A bit of progress in language modeling. *Computer Speech & Language*, 15(4):403–434.

A. Pauls and D. Klein. 2012. 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.

F. Pereira. 2000. Formal grammar and information theory: together again? *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 358(1769):1239–1253.

M. Post. 2011. Judging grammaticality with tree substitution grammar derivations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 217–222.

F. Reali and M.H. Christiansen. 2005. Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science*, 29(6):1007–1028.