Learning the hyperparameters to learn morphology

Stella Frank
ILCC, School of Informatics
University of Edinburgh
Edinburgh, EH8 9AB, UK
sfrank@inf.ed.ac.uk

Abstract

We perform hyperparameter inference within a model of morphology learning (Goldwater et al., 2011) and find that it affects model behaviour drastically. Changing the model structure successfully avoids the unsegmented solution, but results in oversegmentation instead.

1 Introduction

Bayesian models provide a sound statistical framework in which to explore aspects of language acquisition. Explicitly specifying the causal and computational structure of a model enables the investigation of hypotheses such as the feasibility of learning linguistic structure from the available input (Perfors et al., 2011), or the interaction of different linguistic levels (Johnson, 2008a). However, these models can be sensitive to small changes in (hyper-)parameters settings. Robustness in this respect is important, since positing specific parameter values is cognitively implausible.

In this paper we revisit a model of morphology learning presented by Goldwater and colleagues in Goldwater et al. (2006) and Goldwater et al. (2011) (henceforth GGJ). This model demonstrated the effectiveness of non-parametric stochastic processes, specifically the Pitman-Yor Process, for interpolating between types and tokens. Language learners are exposed to tokens, but many aspects of linguistic structure are lexical; identifying which tokens belong to the same lexical type is crucial. Surface form is not always sufficient, as in the case of ambiguous words. Moreover, morphology in particular is influenced by vocabulary-level type statistics (Bybee, 1995), so it is important for a model to operate on both levels: token statistics from realistic (child-directed) input, and type-level statistics based on the token analyses.

The GGJ model learns successfully given fixed hyperparameter values in the Pitman-Yor Process. However, we show that when these hyperparameters are inferred, it collapses to a token-based model with a trivial morphology. In this paper we discuss the reasons for this problematic behaviour, which are relevant for other models based on Pitman-Yor Processes with discrete base distributions, common in natural language tasks. We investigate some potential solutions, by changing the way morphemes are generated within the model. Our results are mixed; we avoid the hyperparameter problem, but learn overly compact morpheme lexicons.

2 The Pitman-Yor Process

The Pitman-Yor Process $G \sim \text{PYP}(a, b, H_0)$ (Pitman and Yor, 1997; Teh, 2006) generates distributions over the space of the base distribution H_0, with the hyperparameters a and b governing the extent of the shift from H_0. Draws from G have values from H_0, but with probabilities given by the PYP. For example, in a unigram PYP language model with observed words, H_0 may be a uniform distribution over the vocabulary, $U(\frac{1}{T})$. The PYP shifts this distribution to the power-law distribution over tokens found in natural language, allowing words to have much higher (and lower) than uniform probability. We will continue using the language model example in this section, since the subsequent morphology model is effectively a complex unigram language model in which word types correspond to morphological analyses. In our presentation, we pay particular attention to the role of the hyperparameter a, since this value governs the power-law behaviour of the PYP (Buntine and Hutter, 2010).

When G is marginalised out, the result is the PYP Chinese Restaurant Process, which is a useful representation of the distribution of observations (word tokens) to values from H_0 (types). In

Proc. of 5th Workshop on Cognitive Aspects of Computational Language Learning (CogACLL) @ EACL 2014, pages 14–18,
Gothenburg, Sweden, April 26 2014.

this restaurant, customers (tokens) arrive and are seated at one of a potentially infinite number of tables. Each table receives a dish (type) from the base distribution when the first customer is seated there; thereafter all subsequent customers adopt the same dish. The probability of customer z_i being seated at a table k depends on the number of customers already seated at that table n_k. Popular tables will attract more customers, generating a Zipfian distribution over customers at tables.

This Zipfian/power-law behaviour can be similar to that of the natural language data, and is the principal motivation behind using the PYP. However, it is only valid for the distribution of customers to tables. When the base distribution is discrete — as in our language model example and the morphology model — the same dish may be served at multiple tables. In most cases, the distribution of interest is generally that of customers (tokens) to dishes (types), rather than to tables, suggesting a preference for a setting in which each dish appears at few tables. This is dependent on a (constrained to be $0 \leq a < 1$), and to a lesser extent on b: If a is small, each dish will be served at a single table, resulting in the type-token and the table-customer power-laws matching. If a is near 1, however, the probability of more than a single customer being seated at a table is small, and the distribution of dishes being eaten by the customers will match the base distribution, rather than being adapted by the caching mechanism of the PYP.

The expected number of tables K grows as $O(N^a)$ (see Buntine and Hutter (2010) for an exact formulation). The number of word types in the data gives us a minimum number of tables, $K \geq T$. When a is small (less than 0.5), the number of expected tables is significantly less than the number of types in a non-trivial dataset, suggesting a lower bound for values of a.

In our language model, the posterior probability of assigning a word w_i to a table k with dish ℓ_k and n_k previous customers is:

$$p(w_i = k | w_1 \dots w_{i-1}, a, b) \propto \begin{cases} (n_k - a) I(w_i = \ell_k) & \text{if } 1 \leq k \leq K \\ (Ka + b) H_0(w_i) & \text{if } k = K + 1 \end{cases} \quad (1)$$

where $I(w_i = \ell_k)$ returns 1 if the token and the dish match, and 0 otherwise. We see that in order to prefer assigning customers to already occupied tables, we need $H_0(w)(Ka + b) < n_k - a$. Given $K \geq T$, and setting $H_0 = \frac{1}{T}$, we can approximate this with $\frac{1}{T}(Ta + b) < n_k - a$. From this we obtain $a < \frac{1}{2}(n_k - \frac{b}{T})$, which indicates that in order for tables with a single customer ($n_k = 1$) to attract further customers, a must be smaller than 0.5. Thus, there is a tension between the number of tables required by the data and our desire to reuse tables. One solution is to fix a to an arbitrary, sufficiently small value, as GGJ do in their experiments. In contrast, in this paper we infer a and b along with the other parameters, and change the other free variable, the base distribution H_0.

3 Morphology

The morphology model introduced by GGJ has a base distribution that generates not simply word types, as in the language model example, but morphological analyses. These are relatively simple, consisting of stem+suffix segmentation and a cluster membership. The probability of a word is the sum of the probability of all cluster c, stem s, suffix f tuples:

$$H_0(w) = \sum_{(c,s,f)} p(c)p(s|c)p(f|c)I(w = s.f) \quad (2)$$

with the stems and the suffixes being generated from cluster-specific distributions. In the GGJ model, all three distributions (cluster, stem, suffix) are finite conjugate symmetric Dirichlet-Multinomial (DirMult) distributions. We retain the DirMult over clusters, but change the morpheme-generating distributions.

The DirMult is equivalent to a Dirichlet Process prior (DP) with a finite base distribution; we use this representation because it allows us to replace the base distributions flexibly. A DP(α, H_0) is also equivalent to a PYP with $a = 0$, and thus also can be represented with a Chinese Restaurant Process, but in this case we sum over all tables to obtain the predictive probability of a (say) stem:

$$p(s|\alpha_s, H_S) = \frac{m_s + \alpha_s H_S}{\sum_{s'} m_{s'} + \alpha_s} \quad (3)$$

Note that the counts m_s refer to stems generated within the base distribution, not to token counts within the PYP.

The original GGJ model, ORIG, is equivalent to setting H_S for stems to $U(\frac{1}{S})$, and likewise $H_F = U(\frac{1}{F})$, where S and F are the number of possible stems and suffixes in the dataset (i.e., all possible prefix and suffix strings, including a null string).

There are two difficulties with this model. Firstly, it assumes a closed vocabulary and requires setting S and F in advance, by looking at the data. As a cognitive model, this is awkward, since it assumes a fixed, relatively small number of possible morphemes.

Secondly, when the PYP hyperparameters are inferred, a is set to be nearly 1, resulting in a model with as many tokens as tables. This behaviour is due to the interaction between vocabulary size and base distribution probabilities outlined in the previous section: this base distribution assigns relatively high probability to words, so new tables have high probability; as the number of tables increases (from its fairly large minimum), the optimal a for this table configuration also increases, resulting in convergence at the token-based model.

We investigate two alternate base distribution over stems and suffixes, both of which extend the space of possible morphemes, thereby lowering the overall probability of the observed words.

DP-CHAR generates morphemes by first generating a length $l \sim \text{Poisson}(\lambda)$. Characters are then drawn from a uniform distribution, $c_{0...l} \sim U(1/|\text{Chars}|)$. A morpheme's probability decreases exponentially by length, resulting in a strong preference for shorter morphemes.

DP-UNI simply extends the original uniform distribution to s and $f \sim U(1/1e6)$, in effect moving probability mass to a large number of unseen morphemes. It is thus similar to DP-CHAR without the length preference.

4 Inference

We follow the same inference procedure as GGJ, using Gibbs sampling. The sampler iterates between inferring each token's table assignment and resampling the table labels (see GGJ for details).

Within the morphology base distribution, the prior for the DirMult over clusters is set to $\alpha_k = 0.5$. To replicate the original DirMult model[1], we set $\alpha_s = 0.001S$ and $\alpha_f = 0.001F$. In the other models, $\alpha_s = \alpha_f = 1$. Within DP-CHAR, $\lambda = 6$ for stems, 0.5 for suffixes.

[1]In this model, the predictive posterior is defined as $p(s|\alpha, S) = \frac{m_s+\alpha}{m_\cdot+S\alpha}$, using an alternate definition of α.

	Eve (Orth.)		Ornat (Orth.)	
	a	Tables/Type	a	Tables/Type
ORIG	0.96	21.2	0.97	10.64
DP-CHAR	0.46	1.4	0.56	1.17
DP-UNI	0.81	7.3	0.70	2.33

Table 1: Final values for a on the orthographic English and Spanish datasets, as well as the average number of tables for each word type. The 95% confidence interval across three runs is ≤ 0.01. (Phonological Eve is similar to Orthographic Eve.)

4.1 Sampling Hyperparameters

We sample PYP a and b hyperparameters using a slice sampler[2]. Previous work with this model has always fixed these values, generally finding small a to be optimal and b to have little effect. In experiments with fixed hyperparameters, we set $a = b = 0.1$.

To sample the hyperparameters, we place vague priors over them: $a \sim \text{Beta}(1, 1)$ and $b \sim \text{Gamma}(10, 0.1)$. The slice sampler samples a new value for a and b after every 10 iterations of Gibbs sampling.

5 Experiments

5.1 Datasets

Our datasets consist of the adult utterances from two morphologically annotated corpora from CHILDES, an English corpus, Eve (Brown, 1973), and a Spanish corpus, Ornat (Ornat, 1994). Morphology is marked by a grammatical suffix on the stem, e.g. *doggy*-PL. Words marked with irregular morphology are unsegmented.

The two languages, while related, have differing degrees of affixation: the English Eve corpus consists of 63 315 tokens (5% suffixed) and 1 988 types (28% suffixed); the Ornat corpus has 43 796 tokens (23% suffixed) and 3 157 types (50% suffixed). The English corpus has 17 gold suffix types, while Spanish has 72.

We also use the phonologically encoded Eve dataset used by GGJ. This dataset does not exactly correspond to the orthographic version, due to discrepancies in tokenisation, so we are unable to evaluate this dataset quantitatively.

[2]Mark Johnson's implementation, available at http://web.science.mq.edu.au/~mjohnson/Software.htm

	Eve (Orth.)			Ornat (Orth.)			Eve (Phon.)	
	% Seg	$\|L\|$	VM	% Seg	$\|L\|$	VM	% Seg	$\|L\|$
Gold	5			23			(5)	
ORIG Fix	7	1680	46.42(10.8)	14	2488	46.63(2.7)	17	1619
ORIG Inf	1	1893	9.94(1.0)	4	2769	17.80(3.7)	1	1984
DP-CHAR Fix	52	1331	15.33(0.3)	83	1828	35.76(1.1)	47	1289
DP-CHAR Inf	50	1330	16.15(0.4)	85	1824	36.47(0.5)	33	1317
DP-UNI Fix	38	1394	17.28(1.7)	51	1874	39.58(0.8)	36	1392
DP-UNI Inf	15	1574	31.54(3.1)	31	1983	42.48(1.1)	21	1500

Table 2: Final morphology results. 'Fix' refers to models with fixed PYP hyperparameters ($a = b = 0.1$), while 'Inf' models have inferred hyperparameters. % Seg shows the percentage of tokens that have a non-null suffix, while $|L|$ is the size of the morpheme lexicon. VM is shown with 95% confidence intervals.

5.2 Results

For each setting, we report the average over three runs of 1000 iterations of Gibbs sampling without annealing, using the last iteration for evaluation.

Table 1 shows what happens when hyperparameters are inferred: ORIG finds a token-based solution, with as many tables as tokens, while DP-CHAR is the opposite, with a small a allowing for just over one table for each word type. DP-UNI is between these two extremes. b is consistently between 1 and 3, confirming it has little effect.

The effect of the hyperparameters can be seen in the morphology results, shown in Table 2. DP-CHAR is robust across hyperparameter values, finding the same type-based solution with fixed and inferred hyperparameters, while the other models have very different results depending on the hyperparameter settings. ORIG with fixed hyperparameters performs best, with the highest VM score (a clustering measure, Rosenberg and Hirschberg (2007)) and a level of segmentation close to the correct one. However, with inferred hyperparameters, this model severely undersegments: it finds the unsegmented maximum likelihood solution, where all tokens are generated from the stem distribution (Goldwater, 2007).

The models with alternate base distributions go to the other extreme, oversegmenting the corpus. As generating new morphemes becomes less probable, the pressure to find the most compact morpheme lexicon grows. This leads to oversegmentation due to many spurious suffixes. The length penalty in DP-CHAR exacerbates this problem, but it can be seen in the DP-UNI solutions as well, particularly when hyperparameters are fixed to encourage a type-based solution.

6 Conclusion

The base distribution in the original GGJ model assigned a relatively high probability to unseen morphemes, allowing the model to generate new analyses for seen words instead of reusing old analyses and leading to undersegented token-based solutions. The alternative base distributions proposed here were effective in finding type-based solutions. However, these over-segmented solutions clearly do not match the true morphology, indicating that the model structure is inadequate.

One reason may be that the model structure is overly simple. The model is faced with an arguably more difficult task than a human learner, who has access to semantic, syntactic, and phonological cues. Adding these types of information has been shown to help morphology learning in similar models (Johnson, 2008b; Sirts and Goldwater, 2013; Frank et al., 2013).

Similarly, the morphological ambiguity that is captured by a model operating over tokens (and ignored in better-performing models that allow only a single analysis for each word type: Poon et al. (2009); Lee et al. (2011); Sirts and Alumäe (2012)) can often be disambiguated using semantic and syntactic information. A model that generates a single analysis per meaningful (semantically and syntactically distinct) word-form could avoid the potential problems of spurious re-generation seen in the original GGJ model as well as the converse problem of under-generation in our alternatives. Such a model might also map onto the human lexicon (which demonstrably avoids both problems) in a more realistic way.

References

Roger Brown. *A first language: The early stages*. Harvard University Press, Cambridge, MA, 1973.

Wray Buntine and Marcus Hutter. A Bayesian view of the Poisson-Dirichlet process. 2010. URL `arXiv:1007.0296`.

Joan Bybee. Regular morphology and the lexicon. *Language and Cognitive Processes*, 10: 425–455, 1995.

Stella Frank, Frank Keller, and Sharon Goldwater. Exploring the utility of joint morphological and syntactic learning from child-directed speech. In *Proceedings of the 18th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

Sharon Goldwater. *Nonparametric Bayesian Models of Lexical Acquisition*. PhD thesis, Brown University, 2007.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. Interpolating between types and tokens by estimating power-law generators. In *Advances in Neural Information Processing Systems 18*, 2006.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. Producing power-law distributions and damping word frequencies with two-stage language models. *Journal of Machine Learning Research*, 12:2335–2382, 2011.

Mark Johnson. Using Adaptor Grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2008a.

Mark Johnson. Unsupervised word segmentation for Sesotho using Adaptor Grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, pages 20–27, June 2008b.

Yoong Keok Lee, Aria Haghighi, and Regina Barzilay. Modeling syntactic context improves morphological segmentation. In *Proceedings of Fifteenth Conference on Computational Natural Language Learning (CONLL)*, 2011.

S. Lopez Ornat. *La adquisicion de la lengua espagnola*. Siglo XXI, Madrid, 1994.

Amy Perfors, Joshua B. Tenenbaum, and Terry Regier. The learnability of abstract syntactic principles. *Cognition*, 118(3):306 – 338, 2011.

Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25 (2):855–900, 1997.

Hoifung Poon, Colin Cherry, and Kristina Toutanova. Unsupervised morphological segmentation with log-linear models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2009.

Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 12th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2007.

Kairit Sirts and Tanel Alumäe. A hierarchical Dirichlet process model for joint part-of-speech and morphology induction. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2012.

Kairit Sirts and Sharon Goldwater. Minimally-supervised morphological segmentation using adaptor grammars. *Transactions of the Association for Computational Linguistics*, 1:231–242, 2013.

Yee Whye Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL)*, Sydney, 2006.