A Topical Collocation Model

Benjamin Börschinger

09/12/2013

These are some notes on a non-grammar version of Mark Johnson's Topic Collocation Model. While equivalent to the original model, this re-formulation allows for faster inference using a break-point sampler, rather than having to perform actual parsing. In addition, the sampler we describe might be applied to related uses of the Collocation-modeling idea, e.g. in word-segmentation.

1 The Topic Collocation Grammar

Johnson (2010) introduced the Topic Collocation Grammar (TCG) that is a non-parametric extension of LDA. Whereas the HDP (Teh et al., 2006) uses non-parametric priors to allow for an unbounded number of possible topics, the TCG uses non-parametric priors to allow for an unbounded vocabulary, making it possible to learn semantically coherent topical units such as "electrical circuit board".

It's original presentation makes use of Adaptor Grammars, an extension of PCFGs which allows for a simple specification of certain models by defining the generative process as that of generating a tree using context-free rules.

$$
\begin{aligned}
\text{Base} &= \text{WordGen}(\Phi_{Base}, p_{\#}) \\
\theta_i &\sim \text{Dir}(\boldsymbol{\alpha}_\theta) \\
\phi_k &\sim \text{DP}(\alpha_\phi, \text{Base}) \\
z_{i,j} \mid \boldsymbol{\theta} &\sim \text{Disc}(\theta_i) \\
c_{i,j} \mid z_{i,j}, \boldsymbol{\phi} &\sim \phi_{z_{i,j}}
\end{aligned}
$$

In words, we assume a base-distribution which can generate sequences of words. We describe this distribution in more detail below although we simply

assume a Unigram process with fixed emission probabilities Φ_{Base} and a fixed termination-probability $p_{\#}$. Fixing these parameters simplifies implementation of the inference algorithm but is not required. It is easy to put a Dirichlet prior on Φ_{Base} and a Beta on $p_{\#}$.

For each of the N documents, we draw a document-specific mixture over topics, one θ_i for each document $1 \leq i \leq N$. We also generate K topics which are distributions over collocations, that is, sequences of words. This is achieved by drawing each ϕ_k from a DP which takes as input the Base distribution.

Finally, for each of the "units" in each of the documents, we generate a topic indicator from the document specific distribution over topics and then the actual unit by using the corresponding topic distribution over collocations.

2 Inference

Inference in this model is complicated by the fact that we do not actually know the number of units in a given document as the actual observations are not $c_{i,j}$s but the $w_{i,k}$s that make up the collocations. This is illustrated by a toy example using 3 words which, already, are compatible with 4 different hypotheses about the extent of the $c_{i,j}$s, illustrated by bracketing.

- (electric circuit boards)$_1$
- (electric)$_1$(circuit)$_2$(boards)$_3$
- (electric circuit)$_1$(boards)$_2$
- (electric)$_1$(circuit boards)$_2$

In addition to the actual grouping, the assignment to topics is also latent, as in standard topic-models. Assuming a single topic, however, we can immediately make the connection to the Goldwater Unigram model of Word Segmentation model, equating words in the TCG case with characters in the WS case. Goldwater introduced a conceptually simple and easy to implement sampler for her Word Segmentation by realizing that each possible hypothesis about the latent word-variables corresponds to a binary vector of boundary indicators, and that it is easy to derive a Gibbs-Sampler over this representation.

We will apply this idea directly to the TCG in the following paragraph. We will focus on the case of a single document for simplicity.

2.1 State-space

The following latent-variables need to be performed inference for.

- the document specific distribution over topics θ
- the K topic distributions ϕ_k, all of which are draws from (independent) DPs
- the N topic indicators z_i, where N is the (latent) number of collocations
- the actual N collocations c_i, where we constrain these such that their linear concatenation gives rise to the actually observed sequence of words $w_{1:N'}$ of N' words

We integrate out θ and, crucially, all of the K topic distributions, giving rise to the well-known Posterior Predictive Distributions described by the CRP. If the base-distribution for the DPs is fixed, we can ignore the actual seating arrangements as we have a non-hierarchical DP model in this case, simplifying book-keeping. It's straight-forward to derive the added complications a non-fixed base-distribution raises (see Börschinger and Johnson (2011) for details).

The actual state-space of our sampler consists of N' K-valued variables that indicate the presence of a Collocation boundary, as well as the Topic from which this collocation came. This is analogous to the Goldwater (2007) sampler, except that rather than a binary choice a $K+1$-ary choice needs to be made at every possible break-point.

2.1.1 Example

Assume $w_{1:4} =$ red electric circuit boards, and our current indicator vector is $b = \langle 3, 0, 0, 1 \rangle$. This vector encodes the following analysis

$$(\text{red})_3(\text{electric circuit boards})_1$$

That is, there are two collocations, one spanning the single word "red" and coming from Topic 3, and one spanning the words "electric circuit boards" and coming from Topic 1. Let us now resample the value for the first variable. Removing knowledge about it gives the partial vector $b = \langle ?, 0, 0, 1 \rangle$ which, just as in word segmentation, requires us to be non-commital with all affected aspects of the previous analysis. In fact, this affects the entire document as the

next boundary already is the end of the document. Consequently, we would need to

- remove one count for having used Topic 3
- remove onecount for having used Topic 1
- remove one count for having "red" in Topic 3
- remove one count for having "electric circuit boards" in Topic 1
- adjust the overall count of generated Topics / collocations (if there were n_3 uses of Topic 3 before, now there are $n_3 - 1$; if there were $n_{3,red}$ uses of "red" in Topic 3, now there are $n_{3,red} - 1$ such uses

Having made these adjustments, we can calculate the probability of all possible ways of setting ?. Setting it to 0 woudl give rise to the analysis

$$(\text{red electric circuit boards})_1$$

which differs from the original state in having this collocation. Hence, we simply have to calculate

$$P(\text{red electric circuit boards}|\text{Counts}_1)P(\text{topic1} \mid \text{Counts})[P(\text{stop})]$$

where the first factor corresponds to the probability of generating the actual string from the Topic-1-CRP, given its current counts, and the second factor is the probability of generating anything from Topic1 in the current document, given the counts of Topic-uses. The third factor is "optional" — without it, we don't have a fully generative model as we wouldn't be able determine the number of collocations. It's impact is minor in word segmentation but make sense here, it allows weak control over the number of collocations: setting termination probability low encourages analyses with many short collocations, setting it high encourages the use of few collocations.

Let's consider setting ? to any specific topic $x \in \{1..K\}$. This gives rise to

$$(\text{red})_x(\text{electric circuit boards})_1$$

This time, the formula includes four more factors

$P(\text{red} \mid \text{Counts}_x)P(x \mid \text{Counts})(1-P(\text{stop}))P(\text{electric...} \mid \text{Counts}_1+\{\text{red}\})P(1 \mid \text{Counts} +\{x\})P(\text{stop})$

Note that we need to update the counts appropriately as the predictive probabilities for the second collocation are affected by our decision for the first collocation. Also note the additional (1-P(stop)) factor which adds additional cost for positing one more collocation.

That's it.