A Note on Topical N-grams

Xuerui Wang and Andrew McCallum
Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003 USA
{xuerui,mccallum}@cs.umass.edu

Technical Report UM-CS-2005-071

December 24, 2005

Abstract

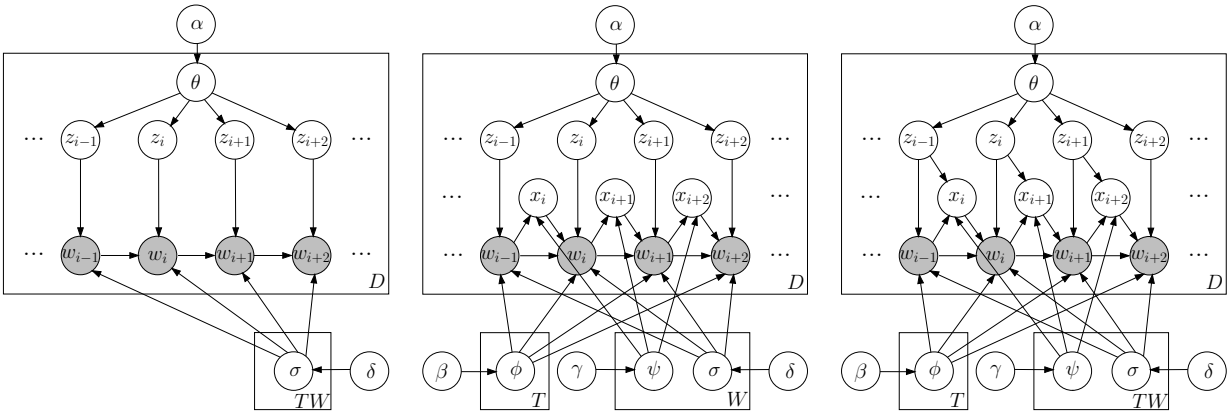
Most of the popular topic models (such as Latent Dirichlet Allocation) have an underlying assumption: bag of words. However, text is indeed a sequence of discrete word tokens, and without considering the order of words (in another word, the nearby context where a word is located), the accurate meaning of language cannot be exactly captured by word co-occurrences only. In this sense, collocations of words (phrases) have to be considered. However, like individual words, phrases sometimes show polysemy as well depending on the context. More noticeably, a composition of two (or more) words is a phrase in some context, but not in other contexts. In this paper, we propose a new probabilistic generative model that automatically determines unigram words and phrases based on context and simultaneously associates them with mixture of topics, and show very interesting results on large text corpora.

1 Introduction

n-gram phrases (or collocations) are fundamentally important in many areas of natural language processing (e.g., parsing, machine translation and information retrieval). Phrase as the whole carries more information than the sum of its individual components, thus it is much more crucial in determining the topics of document collections than individual words. However, most of the topic models (such as Latent Dirichlet Allocation (Blei et al., 2003)) assume that words are generated independently to each other, i.e., under the bag of words assumption. The possible over complicacy caused by introducing phrases makes these topic models completely ignore them. It is true that these models with the bag of words assumption have enjoyed a big success, and attracted a lot of interests from researchers with different backgrounds. We believe that a topic model considering phrases would be more useful in certain applications.

Assume that we conduct topic analysis on a large collection of research papers. Not surprisingly, we will end up with a particular topic on acknowledgment (or funding agency) since many papers have an acknowledgment section (which is not tightly coupled with the content of papers). A topic model with the bag of words assumption ranks very high words like "health" and "science". However, these words have other common meanings and we are not crystal clear why they are ranked so high in acknowledgment topic. A topic model with phrases would associate them with other words to form highly-ranked phrases: "National Institutes of Health" and "National Science Foundation".

Phrases often have specialized meaning, but it is not always the case. For instance, "neural networks" is considered as a phrase because of the frequent use of it as a fixed expression. However, it specifies two distinct concepts: biological neural network in neuroscience and artificial neural networks in modern



(a) Wallach's model (b) LDA-Collocation model (c) Topical n-gram model

Figure 1: Three n-gram models (D: # of documents; T: # of topics; W: # of unique words)

usage. Without consulting the context where the term is located, there is no way to determine its actual meaning. In many situations, topic is very useful to accurately determine the meaning. Also, topic can play a role in phrase discovery. Considering learning English, a beginner usually has difficulty in telling "strong tea" from "powerful tea" (Manning & Schutze, 1999), which are both grammatically correct. The topic associated with "tea" might help to discover the misuse of "powerful".

Is a phrase born to be one? Let us consider another example, in politics topic, "white house" is a proper noun, however, in other topics such as real estate, it does even not mean a phrase, that is, it is not idiomatic at all.

In this paper, we propose a new topical n-gram model that is able to automatically determine unigram words and phrases based on context and simultaneously assign mixture of topics to both individual words and n-gram phrases. The ability to form phrase only where appropriate our model possesses is unique, which distinguish it from the traditional collocation discovery methods discussed in Section 3, in which a *discovered* phrase is treated as a *phrase* no matter what the context is.

2 n-gram Models

Before going to the topical n-gram model, we first describe two related n-gram models in the same flavor. For simplicity, the models discussed in this section take the 1^{st} order Markov assumption, that is, they are actually bigram models. However, all the models have the ability to "model" higher order n-grams (for $n > 2$) by concatenating consecutive bigrams.

2.1 Bigram Topic Model

Wallach (2005) recently developed a Bigram Topic Model on the basis of the Hierarchical Dirichlet Language Model (MacKay & Peto, 1994), by incorporating the concept of topic into bigram models. This model is one of the solutions for the "neural network" example in Section 1. We assume a dummy word w_0 existing at the beginning of each document. The graphical model presentation of this model is shown in Figure 1(a). The generative process of this model can be described as follows:

1. Draw multinomial σ_{zw} from a Dirichlet prior δ;

2. For each document d, draw a multinomial $\theta^{(d)}$ from a Dirichlet prior α; then for each word $w_i^{(d)}$ in document d:
   (a) Draw $z_i^{(d)}$ from multinomial $\theta^{(d)}$;
   (b) Draw $w_i^{(d)}$ from multinomial $\sigma_{z_i^{(d)} w_{i-1}^{(d)}}$.

2.2 LDA Collocation Model

The LDA Collocation Model (Steyvers & Griffiths, 2005) introduces a new set of random variables (for bigram status) $\mathbf{x}$ ($x_i = 1$: w_{i-1} and w_i form a bigram; $x_i = 0$: they do not) which denotes whether a bigram can be formed with the previous word token, in addition to the two sets of random variables $\mathbf{z}$ and $\mathbf{w}$. Thus, it has the power to decide whether to generate a bigram or a unigram. At this aspect, it is more realistic than Wallach's model which always generates bigrams. After all, unigrams are the major components in a document. We assume the status variable x_1 is observed, and only unigram is allowed at the beginning of a document. If we want put more constraints into the model (e.g., no bigram is allowed for sentence/paragraph boundary; only unigram can be considered for the next word after a stop word is removed; etc.), we can assume that the corresponding status variables are observed as well. Although the LDA Collocation model does not generate topic-wise bigrams, a bigram can obtain a topic in a post-hoc way: the first term of a phrase is always generated from the LDA part which carries a topic assignment, and one can take that as the topic of the phrase. This processing does not always assign a reasonable topic to a phrase as everyone can expect. The graphical model presentation of this model is shown in Figure 1(b). The generative process of the LDA Collocation model can be described as follows:

1. Draw multinomial ϕ_z from a Dirichlet prior β;
2. Draw binomial ψ_w from a Beta prior γ;
3. Draw multinomial σ_w from a Dirichlet prior δ;
4. For each document d, draw a multinomial $\theta^{(d)}$ from a Dirichlet prior α; then for each word $w_i^{(d)}$ in document d:
   (a) Draw $x_i^{(d)}$ from binomial $\psi_{w_{i-1}^{(d)}}$;
   (b) Draw $z_i^{(d)}$ from multinomial $\theta^{(d)}$;
   (c) Draw $w_i^{(d)}$ from multinomial $\sigma_{w_{i-1}^{(d)}}$ if $x_i^{(d)} = 1$; else draw $w_i^{(d)}$ from multinomial $\phi_{z_i^{(d)}}$.

2.3 Topical n-gram Model

The topical n-gram (TNG) model is not a pure addition of Wallach's model and LDA Collocation model. It can solve the problem associated with "neural network" example as Wallach's model, and automatically determine whether a composition of two terms is indeed a bigram as in LDA collocation model. However, like other collocation discovery methods discussed in Section 3, a discovered bigram is always a bigram in LDA Collocation model. One of the key contributions of our model is to make it possible to decide whether to form a bigram for the same two consecutive word tokens depending on their nearby context (i.e., co-occurrences). Thus, additionally, our model is a perfect solution for the "white house" example in Section 1. As in LDA collocation model, we may assume some of $\mathbf{x}$ are observed for the same reason. The graphical model presentation of this model is shown in Figure 1(c). Its generative process can be described as follows:

1. Draw multinomial ϕ_z from a Dirichlet prior β;

2. Draw binomial ψ_{zw} from a Beta prior γ;

3. Draw multinomial σ_{zw} from a Dirichlet prior δ;

4. For each document d, draw a multinomial $\theta^{(d)}$ from a Dirichlet prior α; then for each word $w_i^{(d)}$ in document d:

   (a) Draw $x_i^{(d)}$ from binomial $\psi_{z_{i-1}^{(d)} w_{i-1}^{(d)}}$;

   (b) Draw $z_i^{(d)}$ from multinomial $\theta^{(d)}$;

   (c) Draw $w_i^{(d)}$ from multinomial $\sigma_{z_i^{(d)} w_{i-1}^{(d)}}$ if $x_i^{(d)} = 1$; else draw $w_i^{(d)}$ from multinomial $\phi_{z_i^{(d)}}$.

Before discussing the inference problem of our model, let us pause for a brief interlude on topic consistency of terms in a bigram. As shown in the above, the topic assignments for the two terms in a bigram are not required to be identical. Revealing this will surely be enough to cause some readers to stop. However, we are not convinced that this inconsistency is a bad thing from experimental results and our discussion with colleagues. Not like in the LDA Collocation model (the topic of the first term is the topic of the phrase), if a topic of phrase is really needed, we can have the choices to take the topic of the first/last word token or the most common topic in the phrase. In this paper, we will use the topic of the last term as the topic of phrase for simplicity. Furthermore, we could enforce the consistency in the model with ease, by simply adding two more sets of arrows ($z_{i-1} \rightarrow z_i$ and $x_i \rightarrow z_i$). Accordingly, we could substitute Step 4(b) in the above generative process with "Draw $z_i^{(d)}$ from multinomial $\theta^{(d)}$ if $x_i^{(d)} = 1$; else let $z_i^{(d)} = z_{i-1}^{(d)}$;" In this way, a word has the option to inherit a topic assignment from the previous word if they form a bigram phrase. From now on, we will focus on the model shown in Figure 1(c).

Finally we want to emphasize that the topical n-gram model is not only a new method for distilling n-gram phrases depending on nearby context, but also a more sensible topic model than the ones using word co-occurrences alone.

Exact inference like EM on the topical n-gram model in general produces very poor results due to the large number of parameters in the model, thus, many local maxima. We use Gibbs sampling to conduct approximate inference in this paper. To reduce the uncertainty introduced by θ, ϕ, ψ, and σ, we could integrate them out with no trouble because of the conjugate prior setting in our model. Starting from the joint distribution $P(\mathbf{w}, \mathbf{z}, \mathbf{x}|\alpha, \beta, \gamma, \delta)$, we can work out the conditional probabilities $P(z_i, x_i|\mathbf{z}_{-i}, \mathbf{x}_{-i}, \mathbf{w}, \alpha, \beta, \gamma, \delta)$ conveniently[1]using Bayes rule, where $\mathbf{z}_{-i}$ denotes the topic assignments for all word tokens except word w_i, and $\mathbf{x}_{-i}$ represents the bigram status for all tokens except word w_i. During Gibbs sampling, we draw the topic assignment z_i and the bigram status x_i iteratively[2] for each word w_i according to the following conditional probability distribution:

$$P(z_i, x_i|\mathbf{z}_{-i}, \mathbf{x}_{-i}, \mathbf{w}, \alpha, \beta, \gamma, \delta) \propto \frac{\gamma_{x_i} + p_{z_{i-1} w_{i-1} x_i}}{\sum_{k=0}^{1} (\gamma_k + p_{z_{i-1} w_{i-1} k})} (\alpha_{z_i} + q_{dz_i}) \times \begin{cases} \frac{\beta_{w_i} + n_{z_i w_i}}{\sum_{v=1}^{V} (\beta_v + n_{z_i v})} & \text{if } x_i = 0 \\ \frac{\delta_{w_i} + m_{z_i w_{i-1} w_i}}{\sum_{v=1}^{V} (\delta_v + m_{z_i w_{i-1} v})} & \text{if } x_i = 1 \end{cases}$$

where n_{zw} represents how many times word w is assigned into topic z as a unigram, m_{zwv} represents how many times word v is assigned to topic z as the 2^{nd} term of a bigram given the previous word w, p_{zwk} denotes how many times the status variable $x = k$ given the previous word w and the previous word's topic z, and q_{dz} represents how many times a word is assigned to topic z in document d. Note all counts here do not include the assignment of the token being visited.

[1]One could further calculate $P(z_i|\ldots)$ and $P(x_i|\ldots)$ as in a traditional Gibbs sampling procedure.

[2]For some observed x_i, only z_i needs to be drawn.

3 Related Work

Collocation has long been studied by lexicographers and linguists in various ways. Traditional collocation discovery methods range from frequency to variance, to hypothesis testing, to mutual information. The simplest method is counting. Justeson and Katz (1995) combined a small amount of linguistic knowledge (a part-of-speech filter) with frequency and found surprisingly meaningful phrases. Variance based collocation discovery (Smadja, 1993) considered collocations in a more flexible way than fixed phrases. However, high frequency and low variance can be accidental. Hypothesis testing can be used to assess whether or not two words occur together more often than chance. Many statistical tests have been explored, for example, t-test (Church & Hanks, 1989), χ^2 test (Church & Gale, 1991), and likelihood ratio test (Dunning, 1993). More recently, an information-theoretically motivated method for collocation discovery is mutual information (Church et al., 1991; Hodges et al., 1996).

The Hierarchical Dirichlet Language Model (MacKay & Peto, 1994) is closely related to Wallach's model (Wallach, 2005). The probabilistic view of smoothing in language models showed how to take advantage of a bigram model in a Bayesian way.

The main stream of topic modeling has gradually gained a probabilistic flavor as well in the past decade. One of the most popular topic model, Latent Dirichlet Allocation (LDA), which makes the bag of words assumption, has made a big impact in the fields of natural language processing and statistical machine learning (Blei et al., 2003). Three models we discussed in Section 2 all contain an LDA component which is responsible for the topic part.

In our point of view, the HMMLDA model (Griffiths et al., 2005) is the first attack to word dependency in the topic modeling framework. They presented HMMLDA as a generative composite model that takes care of both short-range syntactic dependencies and long-range semantic dependencies between words; its syntactics part is a Hidden Markov Model and the semantic component is a topic model (LDA). Excellent results based on this model are shown on tasks such as part-of-speech tagging and document classification.

4 Experimental Results

We apply the Topical n-gram model to the NIPS proceeding dataset, which consists of the full text of the 13 years of proceedings from 1987 to 1999 Neural Information Processing Systems (NIPS) Conferences. In addition to downcasing and removing stopwords and numbers, we also removed the words appearing less than five times in the corpus—many of them produced by OCR errors. Two-letter words (primarily coming from equations), were removed, except for "ML", "AI", "KL", "BP", "EM" and "IR." The dataset contains 1,740 research papers, 13,649 unique words, and 2,301,375 word tokens in total. Topics found by the Topical n-gram model are shown in Table 1 as anecdotal evidence, with comparison to the corresponding closest (by KL divergence) topics found by LDA.

The "Reinforcement Learning" topic provides an extremely salient summary of the corresponding research area. The LDA topic assembles many common words used in reinforcement learning, but in its word list, there are quite a few generic words (such as "function", "dynamic", "decision") which are common and highly probable in many other topics as well. In TNG, we can find that these generic words are associated with other words to form n-gram phrases (such as "Markov decision process", etc.) which are only highly probable in reinforcement learning. More importantly, by forming n-gram phrases, the unigram word list produced by TNG is also cleaner. For example, because of the prevalence of generic words in LDA, highly related words (such as "Q-learning" and "goal") are not ranked high enough to be shown in the top 20 word list. On the contrary, they are ranked very high in the TNG's unigram word list.

In other three topics, we can find similar phenomena as well. For example, in "Human Receptive System", some generic words (such as "field", "receptive", etc.) are actually the components of the popular

Reinforcement Learning			**Human Receptive System**		
LDA	n-gram (2+)	n-gram (1)	LDA	n-gram (2+)	n-gram (1)
state	reinforcement learning	action	motion	receptive field	motion
learning	optimal policy	policy	visual	spatial frequency	spatial
policy	dynamic programming	reinforcement	field	temporal frequency	visual
action	optimal control	states	position	visual motion	receptive
reinforcement	function approximator	actions	figure	motion energy	response
states	prioritized sweeping	function	direction	tuning curves	direction
time	finite-state controller	optimal	fields	horizontal cells	cells
optimal	learning system	learning	eye	motion detection	figure
actions	reinforcement learning rl	reward	location	preferred direction	stimulus
function	function approximators	control	retina	visual processing	velocity
algorithm	markov decision problems	agent	receptive	area mt	contrast
reward	markov decision processes	q-learning	velocity	visual cortex	tuning
step	local search	goal	vision	light intensity	moving
dynamic	state-action pair	space	moving	directional selectivity	model
control	markov decision process	step	system	high contrast	temporal
sutton	belief states	environment	flow	motion detectors	responses
rl	stochastic policy	system	edge	spatial phase	orientation
decision	action selection	problem	center	moving stimuli	light
algorithms	upright position	steps	light	decision strategy	stimuli
agent	reinforcement learning methods	transition	local	visual stimuli	cell

Speech Recognition			**Support Vector Machines**		
LDA	n-gram (2+)	n-gram (1)	LDA	n-gram (2+)	n-gram (1)
recognition	speech recognition	speech	kernel	support vectors	kernel
system	training data	word	linear	test error	training
word	neural network	training	vector	support vector machines	support
face	error rates	system	support	training error	margin
context	neural net	recognition	set	feature space	svm
character	hidden markov model	hmm	nonlinear	training examples	solution
hmm	feature vectors	speaker	data	decision function	kernels
based	continuous speech	performance	algorithm	cost functions	regularization
frame	training procedure	phoneme	space	test inputs	adaboost
segmentation	continuous speech recognition	acoustic	pca	kkt conditions	test
training	gamma filter	words	function	leave-one-out procedure	data
characters	hidden control	context	problem	soft margin	generalization
set	speech production	systems	margin	bayesian transduction	examples
probabilities	neural nets	frame	vectors	training patterns	cost
features	input representation	trained	solution	training points	convex
faces	output layers	sequence	training	maximum margin	algorithm
words	training algorithm	phonetic	svm	strictly convex	working
frames	test set	speakers	kernels	regularization operators	feature
database	speech frames	mlp	matrix	base classifiers	sv
mlp	speaker dependent	hybrid	machines	convex optimization	functions

Table 1: The four topics from a 50-topic run of TNG on 13 years of NIPS research papers with their closest counterparts from LDA. The **Title** above the word lists of each topic is our own summary of the topic. To better illustrate the difference between TNG and LDA, we list the n-grams ($n > 1$) and unigrams separately for TNG. Each topic is shown with the 20 sorted highest-probability words. The TNG model produces clearer word list for each topic by associating many generic words (such as "set", "field", "function", etc.) with other words to form n-gram phrases.

phrases in this area as shown in the TNG model. "System" is ranked high in LDA, but almost meaningless, and on the other hand, it is not appeared in the top word lists of TNG. Some extremely related words (such as "spatial"), ranked very high in TNG, are absent in LDA's top word list. In "Speech Recognition", the

dominating generic words (such as "context", "based", "set", "probabilities", "database") make the LDA topic less understandable than even just the TNG's unigram word list.

In many situations, a crucially related word might be not mentioned enough to be clearly captured in LDA, on the other hand, it would become very salient as a phrase due to the relatively strong co-occurrence pattern in an extremely sparse setting for phrases. The "Support Vector Machines" topic provides one such example. We can imagine that "kkt" will be mentioned no more than a few times in a typical NIPS paper, and it appears only as a part of the phrase "kkt conditions". The TNG model satisfyingly capture it successfully as a highly probable phrase in the SVM topic.

As we discussed before, higher-order n-grams ($n > 2$) can be approximately modeled by concatenating consecutive bigrams in the TNG model, as shown in Table 1 (such as Markov decision process, hidden Markov model and support vector machines).

To further evaluate the Topical n-gram model against a standard task, we employ the TNG model within language modeling framework to conduct ad-hoc retrieval on TREC collections.

4.1 Ad-hoc Retrieval

Information retrieval performance can be boosted if the similarity between a user query and a document is calculated by common phrases instead of common words (Fagan, 1989; Evans et al., 1991; Strzalkowski, 1995; Mitra et al., 1997). Most research on phrases in information retrieval has employed an independent collocation discovery module, e.g., using the methods described in Section 3. In this way, a phrase can be indexed exactly as an ordinary word. In our topical n-gram model we do not need a separate module for phrase discovery, and everything can be integrated into a language modeling framework. We compare the TNG model with the LDA-based document model recently proposed by Wei and Croft (2006).

5 Conclusions

In this paper, we have presented the Topical n-gram model. The TNG model is able to automatically determine to form a n-gram (and further assign a topic) or not, based on its surrounding context. Examples of topics found by the TNG models are visually better than their LDA counterparts. We also demonstrated how the TNG model can help improve retrieval performance in ad-hoc retrieval tasks on TREC collections.

Unlike some traditional phrase discovery methods, the TNG model provides a systematic way to model (topical) phrases and can be seamlessly integrated with many probabilistic frameworks for various tasks such as ad-hoc retrieval, machine translation and statistical parsing.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and in part by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010, and under contract number HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the sponsor.

References

Blei, D. M., Ng, A. Y., & Jordan, M. J. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Church, K., & Gale, W. (1991). Concordances for parallel text. *In Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research* (pp. 40–62).

Church, K., & Hanks, P. (1989). Word association norms, mutual information and lexicography. *In Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 76–83).

Church, K. W., Gale, W., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. *In Lexical Acquisition: Using On-line Resources to Build a Lexicon* (pp. 115–164). Lawrence Erlbaum.

Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*, 61–74.

Evans, D. A., Ginther-Webster, K., Hart, M., Lefferts, R. G., & Monarch, I. A. (1991). Automatic indexing using selective NLP and first-order thesauri. *In Proceedings of Intelligent Multimedia Information Retrieval Systems and Management (RIAO'91)* (pp. 624–643).

Fagan, J. (1989). The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. *Journal of the American Society for Information Science, 40*, 115–139.

Griffiths, T., Steyvers, M., Blei, D., & Tenenbaum, J. (2005). Integrating topics and syntax. *In Advances in Neural Information Processing Systems 17*.

Hodges, J., Yie, S., Reighart, R., & Boggess, L. (1996). An automated system that assists in the generation of document indexes. *Natural Language Engineering*, *2*, 137–160.

Justeson, J. S., & Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, *1*, 9–27.

MacKay, D. J. C., & Peto, L. (1994). A hierarchical Dirichlet language model. *Natural Language Engineering*, *1*, 1–19.

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Mitra, M., Buckley, C., Singhal, A., & Cardie, C. (1997). An analysis of statistical and syntactic phrases. *Proceedings of RIAO-97, 5th International Conference* (pp. 200–214). Montreal, CA.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, *19*, 143–177.

Steyvers, M., & Griffiths, T. (2005). Matlab topic modeling toolbox 1.3. *http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm*.

Strzalkowski, T. (1995). Natural language information retrieval. *Information Processing & Management*, *31*, 397–417.

Wallach, H. (2005). Topic modeling: beyond bag-of-words. *NIPS 2005 Workshop on Bayesian Methods for Natural Language Processing*.

Wei, X., & Croft, W. B. (2006). LDA-based document models for ad-hoc retrieval. *To appear in Proceedings of the 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval*.