Syntactic Topic Models for Language Generation

William Darling
Microsoft
STC Munich
wdarling@microsoft.com

Fei Song
School of Computer Science
University of Guelph
fsong@uoguelph.ca

Abstract

Since topic models' inception as probabilistic *generative* models, it has only been natural to imagine actually applying the generative process to create documents. However, most topic models consist of a generative process that only provides a bag of words which is one critical step short of creating a readable text. With the recent introduction of syntactically sound topic models and structured topic models, we may now be closer than ever to fully statistical natural language generation. In this paper, we present preliminary work and a research agenda for leveraging syntactic topic models for statistical natural language generation along with a discussion on where the research could and should go next.

1 Introduction

Probabilistic topic models are a family of machine learning techniques used principally within natural language processing to uncover the hidden thematic structure of collections of documents [2]. While they have grown in complexity and expressiveness in recent years, all are based on probabilistic generative models that envision in a vastly simplified way how texts are created. Through probabilistic inference, this generative process is reversed and thematic patterns of word use are uncovered.

Though topic models work based on the idealized assumptions that are made with respect to imagined probabilistic document generation, actually applying the generative process for document creation is rarely tried. This makes sense for many reasons including the fact that most topic models in practice utilize a bag-of-words representation as the observed document, and the generative process does not include any mechanism for sensible word ordering.

Nevertheless, more recent approaches have modelled not only the semantic use of words as in LDA, but also the syntactic function of words in combined and more expressive models. HMMLDA [5] and POSLDA [4] use hidden Markov models in the generative process to order words syntactically and help understand the thematic and functional word use patterns in documents. When posterior inference is performed in POSLDA, for example, learned word distributions are consistent across both topics and part-of-speech types, leading to such groups as verbs about health care and nouns about sports.

Another recent advance in topic model research involves adding semantic structure to the word distributions themselves. Factorial LDA (f-LDA) builds on earlier work on multi-faceted topic models and allows word distributions to express multiple latent factors [9]. Whereas LDA's word distributions tend to reflect semantic topics such as sports, economics, or geography, multi-faceted word distributions in f-LDA combine factors such as topic, author viewpoint, and sentiment. For example, a structured word distribution in f-LDA could represent hockey, from the viewpoint of a Montrealer, with positive sentiment. When the latter two factors are integrated out, the distribution would consist of a standard LDA-like word distribution describing hockey.

With these two related advances – syntactically-aware and multi-faceted models – topic models may now represent an excellent tool for weakly supervised statistical natural language generation of original texts. With enough structure to guide the writing style and theme of a document, and with proper syntax encoded into the generative process, topic models are now ready to make direct use of their generative stories. In this paper we outline a vision and some very preliminary experiments that demonstrate the viability of this approach, and a research agenda to guide progress towards the ultimate goal of natural language generation free of templates and canned text. In the next section we discuss some background and motivation for our work. Then, we describe a simple approach using existing and slightly modified syntactic topic models to generate novel texts. Finally, we discuss some preliminary experiments and next steps in this line of work.

2 Background and Motivation

Natural language generation (NLG) has been a long standing goal of AI systems. The dream is to have a system that can generate clear and original natural language text based either on some target output or – more fantastically – on machine intelligence that would allow conversing with machines. Practical applications include legal contract writing based on the laid-out needs of parties, weather reporting based on meteorological data, and abstractive text summarization that can actually paraphrase the important content from an input.

Currently, most NLG work is templated-based [7]. That is, for each target task or domain, a natural language recipe is designed such that it can be filled in with instance-specific data. For example, a very simple weather report could be delivered with a text-template that reads "The weather today is $\{0\}$ and $\{1\}$ degrees" where the first place-holder would be filled with either "sunny", "cloudy", "rainy", or "snowy", and the second place-holder would contain the current temperature, both of which would be provided by electronic instruments. More complex scenarios can be envisaged depending on the task and the sophistication of the creator.

The principal problems with these template-based approaches are that (1) they are limited in original text that can be produced; (2) each template must be built by hand; and (3) they are static. For (1), humans quickly grow tired of repetitive canned text as can be attested by anyone who grew up playing EA Sports video games with "live commentary". For (2), expanding the complexity or coverage of the system is both expensive and time consuming, and for (3), as with (1), the "generated" text is identical each time except for the changing data inputs.

In this work, we are interested in further-looking applications of true novel text generation. We could follow the basic idea of n-gram language models where text is probabilistically generated based on conditional probability distributions learned from unstructured texts [8]. Trying to directly generate text from these types of distributions suffers from similar problems to the canonical topic models; though many text fragments may be syntactically correct if the order of the model is big enough, full sentences will rarely be so because there is no notion of sentence breaks or of syntactic structure. Further, language models are generally topic-agnostic, and there would be little sway in guiding the thematic structure of a text created by following a language model. However, both of these issues might be alleviated by the use of structured syntactic topic models.

Something akin to this approach has been used on a smaller scale in headline generation based on statistical translation [1]. There, headlines for news articles were generated by learning a translation model from a document-like language to a headline-like language, and then realizing the headline by finding the most probable ordering given a language model. Other work also does away with hand-crafted rules and relies on statistical generation of language [8]. There, models are learned to create natural language texts in a statistical manner but the models are learned from aligned corpora using supervised learning. We are interested in the unsupervised framework where the building blocks of our natural language generation system are uncovered through topic modeling-like pattern recognition. Our envisioned approach is in the same spirit of the above work in that we would use structured syntactic topic models to learn fuzzy syntactic rules of writing and also different specific word distributions for topics, viewpoints, writing styles, sentiments, etc. This approach would be essentially unsupervised and would free designers from having to create linguistic rules for specific templates and could also create original text by, for example, combining the writing styles learned from two different sets of input documents.

$\phi^{(t,p,s)} \sim$ Dirichlet(

Topic: *economics*
price
stock
securities
economic
macro

+

Perspective: *liberal*
social
peace
program
government
health

+

Sentiment: *positive*
enjoy
excellent
great
positive
happy

)

Table 1: Drawing a word distribution $\phi^{(t,p,s)}$ for semantic topic "economics", political perspective "liberal", and "positive" sentiment, as in f-LDA [9].

3 Models

In the simplest approach, the POSLDA model could be used to generate a text using posterior distributions learned from some source corpus [4]. Initial experiments show that this method works surprisingly well to generate texts that are plausibly created by humans and are often syntactically correct. We learned a POSLDA model on a large collection of newspaper articles and then set the θ topic distribution to cover a terrorism topic and a geography topic. A "good" example was "The municipality was still identified."; a "bad" example was "The terrorists containing abdel prime slowest aram february egypt ohio a amortization strip."

But much more interesting things can be done. For instance, we would like to structure the entire document such that the narrative can expand, jump around, zero-in, etc. This could be partially addressed with a "sticky" HMM component similar to that in [6] used only in the latter generative step that controls certain factors of the word distributions. For example, in each section (like introduction), draw a variable to determine if we will switch "styles". The probability of changing sections/styles would progressively get higher the longer we stay in that section and could then be reset when a switch actually occurs. Using ideas from f-LDA, word distributions could be drawn that are combinations of semantic topic and section style, depending on the current location of the generated document.

More specifically, f-LDA allows a word distribution to represent a structured "topic" based on K factors. Examples include a 3-factor model based on semantic topic, political perspective, and sentiment of the document. This might result in a word distribution covering (Economics, Liberal, Positive), for example [9]. This is described graphically in Table 1. In fully statistical natural language generation, we could learn to drive content of texts by learning structured distributions from texts with meta-data, and then combining these factors in novel ways for the generation process. This is done by linking the priors for multi-faceted topics that share facets using a log-linear parameterization of the Dirchlet prior for each word distribution.

Then, a document d with manually-set document-specific prior θ_d, sentiment prior $\sigma_d = (0.1, 0.9)$, and author sex $\Delta_d =$ female (for example) could be generated with the following process:

1. For each word token $w_i \in d$:
   (a) Draw style change $\gamma \sim \chi$
   (b) If $\gamma == 1$
      i. $\chi \sim$ Beta(ζ)
   (c) $\chi_{new} \leftarrow \chi_{old} \times p$
   (d) Draw $c_i \sim \pi_{c_{i-1}, i-2, \ldots}$
   (e) If $c_i \notin \mathcal{C}_{\text{CONTENT}}$:
      i. Draw $w_i \sim \phi_{c_i}^{(\text{FUNCTION})}$
   (f) Else:
      i. Draw $z_i \sim \theta_d$
      ii. Draw $s_i \sim \sigma_d$
      iii. Draw $w_i \sim \phi_{c_i, z_i, s_i, \Delta_d}^{(\text{CONTENT})}$

where π is the HMM transition distribution, γ controls whether the current style should change, χ controls how long the current style will remain, p is a constant that lowers the chance of staying in a

Officers suspected armed including engineer condition. In the assembly in some. Guru leader reported declared country want released with million reform to predicted. When largely at some. Their army party to merino. Leaders at with recklessly. The guard. Poor leaders in its southeast that the party bodies after. Radio he were formality. They providing in people leader elections and to arrested. Because patrols promised they to insignificant of asylum of. The agency revas was killed and labor northeast of dispute curricula later supported.

Figure 1: Generated text with $\theta_d = (t_1 = 0.5, t_2 = 0.5, t_3 = 0, ...)$ where $\phi_{t_1} = [military, police, army, ...], \phi_{t_2} = [political, party, national, ...]$.

style the longer it is used, and $\phi^{(k)}$ are the semantic topic word distributions. Importantly, each $\phi^{(k)}$ is drawn from a structured Dirichlet where the prior is a log-linear combination of weight vectors for each of the factors in that distribution (including other styles, not shown in the process above). In this specific example, ϕ is a word distribution that depends on the part-of-speech of the word (c_i), its semantic topic (z_i), the sentiment (s_i), and the author's sex (Δ_d). Further, as in POSLDA, the model distinguishes between "function" word classes that do not change given the multi-faceted topic, and "content" word classes, denoted as $\mathcal{C}_{\text{CONTENT}}$ in the generation process described above, that depend on the semantic topic, sentiment, author perspective, etc. The more factors that are set, the more directed the generated text theoretically could be. A good amount of future research will have to determine how much of the generation process can be left to chance, and how much can be set specifically, to find a workable balance between novel text and existing text re-creation.

The POSLDA model is particularly interesting for this work because style in writing encompasses not just the words that we use in different factors, but also the way that they are combined and ordered [4]. The former can be nicely captured through the f-LDA aspect of the model by collecting annotated collections of texts by different types of authors, discussing different topics, in different kinds of fora, etc. The latter could also be potentially captured by a topic model that includes syntax and word-type. POSLDA has the same aims as the Syntactic Topic Model (STM) [3], but is more flexible in that it is designed to learn the transitions between word types whereas the STM takes previously-parsed sentences as an input and then combines that with the semantic portion of the topic model. While it is indeed far from any kind of understanding of writing style, it could prove to be a useful and interesting way to capture some notions of it.

While we are a ways away from deploying a system that generates understandable text following these techniques, the principle can be demonstrated and experimented with immediately. With the generative process described above, we generated a 100-word document as follows. First, we learned syntactic-semantic word distributions ϕ_k and a transition distribution π using POSLDA with an order-3 HMM, $K = 50$ topics, and $S = 10$ states of which $SS = 5$ were semantic. We then set the document-topic prior distribution θ_d to 0.5 for a learned "politics" distribution and 0.5 for a learned "military" distribution. Then, we followed the generative process outlined above. The output that we obtained is shown in Figure 1. While the output is interesting, it shows that there is clearly much work yet to do.

4 Next Steps

While the basics are set and we even have some initial plausible generated text, there are a number of areas where research will need to be directed to make natural language generation through syntactic topic models a viable approach. First – and most importantly – a means to improve the quality of text needs to be investigated. It could be trivially improved perhaps by going to higher orders in the HMM but more principled approaches will ultimately be required. Second, if this avenue is to be more useful than for simple demos, the preciseness and direction of text must be fully controllable – even if it is by a machine. A weather report generated from an f-LDA-like distribution about weather from a female perspective with a positive sentiment may be interesting but ultimately useless. However, a short text generated from a set of distributions learned on a large collection of texts with certain styles (such as background) removed may be an interesting approach to abstractive automatic text summarization, and this is a task that is already nicely suited to this approach.

We are also interested in the interdisciplinary avenues that this research could lead us to. The possibilities of generating text by combining learned background topic distributions with the styles

of perhaps Hemingway combined with David Foster Wallace told with a liberal political bias all in a neutral sentiment is exciting but also fraught with, *inter alia*, legal questions. If structured syntactic topic models are learned directly from NY Times articles and Jonathan Franzen novels, do those copyright holders have any claim to the intellectual property generated from these techniques? These types of questions can be answered with the help from our colleagues in other disciplines and we look forward to having these sorts of conversations in addition to the ones we will need to have with our fellow machine learning and NLP researchers.

References

[1] Michele Banko, Vibhu O. Mittal, and Michael J. Witbrock. Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 318–325, Stroudsburg, PA, USA, 2000. ACL.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[3] Jordan Boyd-Graber and David M. Blei. Syntactic topic models. In *Advances in Neural Information Processing Systems*, pages 185–192, 2008.

[4] William M. Darling, Michael J. Paul, and Fei Song. Unsupervised part-of-speech tagging in noisy and esoteric domains with a syntactic-semantic bayesian hmm. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 1–9, Stroudsburg, PA, USA, 2012. ACL.

[5] Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. Integrating topics and syntax. In *In Advances in Neural Information Processing Systems 17*, pages 537–544. MIT Press, 2005.

[6] Aria Haghighi and Lucy Vanderwende. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370, Boulder, Colorado, June 2009. ACL.

[7] Irene Langkilde and Kevin Knight. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 704–710, Stroudsburg, PA, USA, 1998. ACL.

[8] François Mairesse, Milica Gašić, Filip Jurčíček, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1552–1561, Stroudsburg, PA, USA, 2010. ACL.

[9] Michael Paul and Mark Dredze. Factorial lda: Sparse multi-dimensional text models. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2591–2599. 2012.