Analysis of Morphology in Topic Modeling

Anonymous EMNLP submission

Abstract

Topic models make strong assumptions about their data. In particular, different words are implicitly assumed to have different meanings: topic models are often used as human-interpretable dimensionality reductions and a proliferation of words with identical meanings would undermine the utility of the top-m word list representation of a topic. Though a number of authors have added preprocessing steps such as lemmatization to better accommodate these assumptions, the effects of such data massaging have not been publicly studied. We make first steps toward elucidating the role of morphology in topic modeling by testing the effect of lemmatization on the interpretability of a latent Dirichlet allocation (LDA) model. Using a word intrusion evaluation, we quantitatively demonstrate that lemmatization provides a significant benefit to the interpretability of a model learned on Wikipedia articles in a morphologically rich language.

1 Introduction

Topic modeling is a standard tool for unsupervised analysis of large text corpora. At the core, almost all topic models pick up on co-occurrence signals between different words in the corpus, that is, words that occur often in the same sentence, are likely to belong to the same latent topic. In languages that exhibit rich inflectional morphology, the signal becomes weaker given the proliferation of unique tokens. In this work, we explore the effect of token-based lemmatization on the performance of topic models.

Syntactic information is not generally considered to exert a strong force on the thematic nature of a document. Indeed, for this reason topic models often make a bag-of-words assumption, discarding the order of words within a document. In morphologically rich languages, however, syntactic information is often encoded in the word form itself. This kind of syntactic information is a nuisance variable in topic modeling and is prone to polluting a topic representation learned from data (Boyd-Graber et al., 2014). For example, consider the Russian name *Putin*; in English, we have a single type that represents in the concept in all syntactic contexts, whereas in Russian Путин appears with various inflections, e.g., Путина, Путину, Путине, and Путином. Which form of the name one uses is fully dependent on the syntactic structure of the sentence. Compare the utterances мы говорим о Путине (*we are speaking about Putin*) and мы говорим Путину (*we are speaking to Putin*): both sentences are thematically centered on Putin, but two different word forms are employed. English stop words like prepositions often end up as inflectional suffixes in Russian, so lemmatization on Russian performs some of the text normalization that stop word filtering performs on English. Topic models are generally sensitive to stop words in English (Wallach et al., 2009a; Blei et al., 2010; Eisenstein et al., 2011), hence we expect them to be sensitive to morphological variation in languages like Russian.

In this study, we show that

- truncated documents, imitating the sparsity seen in social media, reduce interpretability;
- if lemmatization is used, a filtered vocabulary yields more interpretable topics than an informative prior; and
- overall, interpretability is best when the corpus consists of long documents, the vocabulary is filtered, and lemmatization is applied.

2 Morphology and Lemmatization

Morphology concerns itself with the internal structure of individual words. Specifically, we focus on *inflectional morphology*, word internal structure that marks syntactically relevant linguistic properties, e.g., person, number, case and gender on the word form itself. While inflectional morphology is minimal in English and virtually non-existent in Chinese, it occupies a prevalent position in many languages' grammars, e.g., Russian. In fact, Russian will often express relations marked in English with

	Singular	**Plural**
Nominative	пес (*pyos*)	псы (*psy*)
Genitive	пса (*psa*)	псов (*psov*)
Accusative	пса (*psa*)	псов (*psov*)
Dative	псу (*psu*)	псам (*psam*)
Locative	псе (*psye*)	псах (*psax*)
Instrumental	псом (*psom*)	псами (*psami*)

Table 1: A inflectional paradigm for the Russian word пес (*pyos*), meaning "dog". Each of the 12 different entries in the table occurs in a distinct syntactic context. A lemmatizer canonicalizes these forms to single form, which is the nominative singular in, reducing the sparsity present in the corpus.

prepositions, simply through the addition of a suffix, often reducing the number of words in a given sentence. The collection of inflections of the same stem is preferred to as a paradigm. The Russian noun, for example, forms a paradigm with 12 forms. See the sample paradigm in Table 1 for an example.[1] The Russian verb is even more expressive with more than 30 unique forms (Wade, 2010).

In the context of NLP, large paradigms imply an increased token to type ratio, greatly increasing the number of unknown words. One method to combat this issue is to *lemmatize* the sentence. A lemmatizer maps each inflection (an element of the paradigm) to a canonical form known as the lemma, which is typically the form found in dictionaries written in the target language. In this work, we employ the TreeTagger lemmatizer (Schmid, 1994).[2] The parameters were estimated using the Russian corpus described in Sharov and Nivre (2011).

3 Related Work

Though applied in many studies (Deerwester et al., 1990; Hofmann, 1999; Mei et al., 2007; Nallapati et al., 2008; Lin and He, 2009), lemmatization has not been directly explored in the context of topic modeling. An infinite-vocabulary LDA model containing a prior on words similar to an n-gram model has been developed (Zhai and Boyd-Graber, 2013); this prior could be viewed as loosely encoding beliefs of a concatenative morphology, but its effect was not analyzed in isolation.

To measure the effect of lemmatization on topic models we must first define "topic model." In this study, for comparability with other work, we restrict our attention to latent Dirichlet allocation (LDA) (Blei et al., 2003), the canonical Bayesian graphical topic model. We want to measure the performance of a topic model by its interpretability, as topic models are best suited to discovering human-interpretable decompositions of the data (May et al., 2015). We note there are more modern but less widely-used topic models than LDA such as the sparse additive generative (SAGE) topic model, which explicitly models the background word distribution and encourages sparse topics (Eisenstein et al., 2011), or the nested hierarchical Dirichlet process (nHDP) topic model, which represents topics in a hierarchy and automatically infers its effective size (Paisley et al., 2015). These models may render more interpretable results overall. However, we are currently interested in the *relative* impact of lemmatization on a topic model, we are unaware of any direct prior work, and we wish for our results to be widely applicable across research and industry. Thus we leave these alternative topic models as considerations for future work.

While not satisfactorily explored in the topic modeling community, morphology has been actively investigated in the context of word-embeddings. The latent topic vectors that topic models discover have many parallels to continuous embeddings—both are real-valued representations that stand proxy for (some notion of) lexical semantic information. Most notably, Bian et al. (2014) learned embeddings for individual morphemes jointly within the standard WORD2VEC model (Mikolov et al., 2013) and Soricut and Och (2015) used the embeddings themselves to induce morphological analyzers. Character-level embedding approaches have also been explored with the express aim of capturing morphology (Santos and Zadrozny, 2014; Ling et al., 2015).

4 Experiments

For some pre-specified number of topics K and Dirichlet concentration hyperparameters $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$, the LDA topic model represents a vocabulary as a

[1]Note that Table 1 contains several entries that are identical, e.g., the singular genitive is the same as the singular accusative. This is a common phenomenon known as syncretism (Baerman et al., 2005), but it is not universal over all nouns—plenty of other Russian nouns *do* make the distinction between genitive and accusative in the singular.

[2]http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

view	topic
lem	деревня* сельский поселение пункт сельсовет
non	деревня* деревни* деревне* жителей волости
lem	клетка лечение* заболевание† препарат действие
non	лечения* течение лечение* крови заболевания†
lem	японский* япония† корея префектура смотреть
non	считается японии† японский* посёлок японской*
lem	художник* искусство† художественный* картина‡ выставка**
non	искусства† музея картины‡ выставки** выставка**

Table 2: Manually-aligned topic pairs: the first topic in each pair is from the lemmatized model, the second pair is a semantically similar topic in the non-lemmatized model. Within each pair, each of the symbols *, †, ‡, and ** (separately) denotes word forms of a shared lemma. The lemmatized topic representations are more diverse than those of the non-lemmatized topic representations. For example, the non-lemmatized version of the first topic contains three inflections of the Russian word деревня (*village*)—successive inflectional forms add little or no information to the topic.

set of K i.i.d. topics $\boldsymbol{\beta}_k$, represents each document as a an i.i.d. mixture over those topics (with mixture weights $\boldsymbol{\theta}_d$), and specifies that each token in a document is generated by sampling a word type from the document's topic mixture:

$$\begin{aligned}\boldsymbol{\beta}_k &\sim \text{Dirichlet}\,(\boldsymbol{\eta})\\ \boldsymbol{\theta}_d &\sim \text{Dirichlet}\,(\boldsymbol{\alpha})\\ z_{d,n} &\sim \text{Discrete}\,(\boldsymbol{\theta}_d)\\ w_{d,n} &\sim \text{Discrete}\left(\boldsymbol{\beta}_{z_{d,n}}\right)\end{aligned}$$

Meaningful evaluation of topic models is notoriously difficult and has received considerable attention in the literature (Chang et al., 2009; Wallach et al., 2009b; Newman et al., 2010; Mimno et al., 2011). In general we desire an evaluation metric that correlates with a human's ability to use the model to explore or filter a large dataset, hence, the interpretability of the model. In this study we moreover require an evaluation metric that is comparable across different views of the same corpus.

With those concerns in mind we choose a *word intrusion* evaluation: a human expert is shown one topic at a time, represented by its top m words (for some small number m) in random order, as well as an additional word (called the *intruder*) randomly placed among the m topic words (Chang et al., 2009). The intruder is randomly selected from the set of high-probability words from other topics in the model. The expert is tasked with identifying the intruder in each list of $m+1$ words. As in prior work (Chang et al., 2009), we instruct the expert to ignore syntactic and morphological patterns.

If the model is interpretable, the m words from a topic will be internally coherent whereas the intruder word is likely to stand out. Thus a model's interpretability can be quantified by the fraction of topics for which the expert correctly identifies the intruder. We call this value the *detection rate*:

$$\text{DR} = \frac{1}{K}\sum_{k=1}^{K}\delta_{i_k}\left(\omega_k\right)$$

where K is the number of topics in the model, i_k is the index of the intruder in the randomized word list generated from topic k, and ω_k is the index of the word the expert identified as the intruder. We note this is just the mean (over topics) of the *model precision* metric from prior work (Chang et al., 2009) when one expert is used instead of several non-experts.

Our corpus consists of Russian Wikipedia articles from the dump released on 11/02/2015.[3]

We stripped the XML portion of the formatting and then ran the lemmatizer described in Section 2. When the lemmatizer does not recognize a word, we back off to the word form itself.[4]

We consider two preprocessing schemes to account for stop words and other high-frequency terms in the corpus. First, we compute the vocabulary as the top 10,000 words by document frequency,[5] separately for the lemmatized and non-lemmatized data,

[3] The Wikipedia dump is from November 11, 2015.

[4] 11% of the 378 million tokens in the raw corpus were unrecognized by the lemmatizer.

[5] Due to minor implementation concerns the lemmatized and non-lemmatized vocabularies consist of the top 9387 and 9531 words (respectively) by document frequency.

and specify an asymmetric prior on each document's topic proportions $\boldsymbol{\theta}$. We refer to this preprocessing scheme as the *unfiltered-asymmetric* setting. The second modeling scheme we consider uses a vocabulary with high-frequency words filtered out and a uniform prior on the document-wise topic proportions. (We refer to this setting as *filtered-symmetric*.) Specifically, a 10,000 word vocabulary is formed from the lemmatized data by removing the top 100 words by document frequency over the corpus and taking the next 10,000. To determine the non-lemmatized vocabulary, we map the filtered lemmatized vocabulary onto all word forms that produce one of those lemmas in the data. Finally, observing that some of the uninformative high-frequency words reappear in this projection, we remove any of the top 100 words from the lemmatized and non-lemmatized corpora from this list, producing a non-lemmatized vocabulary of 72,641 words. While the large size of this vocabulary slows learning, we do not believe it impacts the results negatively; our priority is retaining the information captured by the lemmatized vocabulary to provide a fair comparison.

In addition to exploring different choices of vocabulary, we also consider truncating the documents to their first 50 tokens.[6] This augmentation simulates data sparsity by reducing the amount of content-bearing signal in each document, so we might expect the truncated documents to more greatly benefit from lemmatization (which can be cast as a dimensionality reduction method).

We learn LDA by stochastic variational inference (Hoffman et al., 2013), initializing the models randomly and using fixed priors.[7] We specify $K = 100$ topics to all models. Uniform priors with $\eta_v = 0.1$ and $\alpha_k = 5/K$ were given to filtered-symmetric models; non-uniform priors with $\eta_v = 0.1$, $\alpha_1 = 5$, and $\alpha_k = 5/(K-1)$ for $k > 1$ were given to unfiltered-asymmetric models. The local hyperparameters $\boldsymbol{\alpha}$ are informed by mean document word usage and document length; in particular, we believe approximately 50% of the word tokens in the corpus are uninformative.

The detection rate for all four configurations

[6] As the vocabulary does not contain rare words, the number of tokens per document seen by the model is less than 50.

[7] In preliminary experiments Gibbs sampling with hyperparameter optimization did not improve interpretability.

			DR		p-val
vocab	prior	docs	non	lem	Δ
unfilt	sym	full	0.54	0.52	0.61
filt	**asym**	**full**	**0.50**	**0.65**	**0.02**
unfilt	sym	trunc	0.37	0.37	0.50
filt	asym	trunc	0.43	0.47	0.28

Table 3: Detection rate for the non-lemmatized (non) and lemmatized (lem) models and p-values for the one-sided detection rate difference tests. (filt and unfilt indicate whether or not the vocabulary is filtered; sym and asym indicate whether the prior is symmetric, trunc and full indicate whether the documents are truncated.) The detection rate benefits significantly from lemmatization on a filtered vocabulary (highlighted in bold).

(filtered-symmetric or unfiltered-asymmetric vocabulary and full-length or truncated documents), and the p-values for one-sided detection rate differences (testing our hypothesis that the lemmatized models yield higher detection rates than the non-lemmatized models), are reported in Table 3. Word intrusion performance benefits significantly from lemmatization on a filtered vocabulary and a symmetric prior. Truncated documents exhibit lower performance overall and are helped less by lemmatization (posing challenges for social media applications). Further, we observe differences between use of an asymmetric prior on an unfiltered vocabulary and use of a symmetric prior on a vocabulary with stop words filtered out.

We find that topics from the unfiltered-asymmetric models often contain stop words despite the first topic receiving half of the prior probability mass. Indeed, many topics consist primarily of stop words, such as the topic и в при с у. Hand-aligned topics from the filtered-symmetric models learned on full-length documents are shown in Table 2. There is significant redundancy (multiple inflected word forms of the same lemma) in the top five words of the non-lemmatized topics; on the other hand, the diversity of words in the lemmatized topics lends to human interpretation.

5 Conclusion

We have demonstrated the impact of lemmatization as a preprocessing step to LDA on Wikipedia articles in Russian. In particular, we have verified the intuition that lemmatization can significantly improve the interpretability of a topic model.

References

Matthew Baerman, Dunstan Brown, and Greville G Corbett. 2005. *The syntax-morphology interface: A study of syncretism*, volume 109. Cambridge University Press.

Jiang Bian, Bin Gao, and Tie-Yan Liu. 2014. Knowledge-powered deep learning for word embedding. In *ECML*, pages 132–148.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, Jan.

David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. 2010. The nested chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):1–30.

Jordan Boyd-Graber, David Mimno, and David Newman. 2014. Care and feeding of topic models: Problems, diagnostics, and improvements. In Edoardo M. Airoldi, David Blei, Elena A. Erosheva, and Stephen E. Fienberg, editors, *Handbook of Mixed Membership Models and Their Applications*. CRC/Chapman Hall.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 288–296. Curran Associates, Inc.

Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society For Information Science*, 41(6):391–407.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1041–1048.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. 2013. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, May.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 289–296.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 375–384, New York, NY, USA. ACM.

Wang Ling, Chris Dyer, Alan W. Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luís Marujo, and Tiago Luís. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *EMNLP*, pages 1520–1530.

Chandler May, Francis Ferraro, Alan McCree, Jonathan Wintrode, Daniel Garcia-Romero, and Benjamin Van Durme. 2015. Topic identification and discovery on text and speech. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Sep.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, pages 171–180. ACM.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, pages 262–272, Jul.

Ramesh M. Nallapati, Amr Ahmed, Eric P. Xing, and William W. Cohen. 2008. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 542–550, New York, NY, USA. ACM.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT)*, pages 100–108, Jun.

J. Paisley, C. Wang, D. M. Blei, and M. I. Jordan. 2015. Nested hierarchical dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):256–270, Feb.

Cicero D Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *ICML*, pages 1818–1826.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *COLING*, volume 12, pages 44–49.

S Sharov and Joakim Nivre. 2011. The proper place of men and machines in language technology. processing russian without any linguistic knowledge. In *Proceedings of the Annual International Conference Dialogue, Computational Linguistics and Intellectual Technologies*, number 10, page 657.

Radu Soricut and Franz Josef Och. 2015. Unsupervised morphology induction using word embeddings. In *NAACL*, pages 1627–1637.

Terence Wade. 2010. *A Comprehensive Russian Grammar*, volume 8. John Wiley & Sons.

Hanna M. Wallach, David Mimno, and Andrew McCallum. 2009a. Rethinking lda: Why priors matter. In *Advances in Neural Information Processing Systems 22 (NIPS)*.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009b. Evaluation methods for topic models. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, Jun.

Ke Zhai and Jordan Boyd-Graber. 2013. Online latent dirichlet allocation with infinite vocabulary. In *Proceedings of the 30th International Conference on Machine Learning*.