A Brother Karamazov: Quantifying Morphology in Topic Modeling of Literary Russian

Virginia Partridge
University of Massachusetts Amherst
vcpartridge@umass.edu

Abstract

TODO

1 Introduction

Latent Dirichlet Analysis (LDA) is a widely adopted approach for unsupervised topic modeling and has been used across disciplines for exploring themes and trends in large document collections. LDA has been applied to explore the ever-growing variety of text from online platforms and social media and to analyze language changes in academic fields over time (Koltsova and Koltsov, 2013; McFarland et al., 2013; Vogel and Jurafsky, 2012; Mitrofanova, 2015). Assuming a bag-of-words approach, LDA produces latent topics as multinomial distributions over words and each topic is viewed as being generated by a mixture of topics (Blei et al., 2003; Steyvers and Griffiths, 2007).

However, what happens when words in this bag-of-words approach are themselves are complex? To this end, we turn to topic modeling on Russian, a flective language with paradigms for nouns, adjectives and verbs (Wade et al., 2020). Russian's inflectional morphology increases the sparsity of words' surface forms in the collection, but it's unclear to what extent this sparsity impacts the interpretability and usefulness of topic models. Stemming and lemmatization treatments are typical text preprocessing steps for topic modeling, even for English, which has relatively little inflectional morphology, but there is a lack of empirical evidence that these treatments improve the models from the perspective of human interpretability or quantitative measures of topic quality (Schofield and Mimno, 2016).

Furthermore, conflation of surface word forms may mask phenomena of interest to researchers. Topic modeling is a popular tool for exploring gender bias in corpora (Vogel and Jurafsky, 2012; Devinney et al., 2020), and many languages, Russian included, have inflectional morphology that marks gender. By normalizing tokens to a single form, topics learned in LDA won't distinguish between Russian's feminine, masculine and neuter word forms, which may or may not be desirable, depending on the domain and researchers' goals.

In this work we explore baseline performance of LDA for topic modeling on a Russian literary corpus and report both quantitatively and qualitatively on the resulting topics. We first establish that topic modeling in Russian behaves similarly to English in terms of correlating with corpus metadata, regardless of stemming or lemmatisation. These observations cast doubt on whether inflectional morphology needs to be addressed at all. By investigating topic models produced with no morphological preprocessing step, we demonstrate that a topic can be dominated by particular morphological features and propose ways quantify this relationship. Finally, we compare various stemming and lemmatisation treatments as preprocessing the corpus and contrast this with post-processing the keywords learned by models.

2 Related Work

3 Methods

3.1 Framework for discussing morphological complexity

We will first clarify terms for discussing Russian's morphological paradigms, following frameworks for quantifying morphological complexity used in linguistics and computational linguistics (Baerman et al., 2015; Cotterell et al., 2019). We draw a distinction between *derivational* morphology, the process by which new words are formed through changing meaning or part of speech, and *inflectional* morphology, which can be simplistically understood as verb paradigms to capture subject-verb agreement or noun declensions for case and grammatical gender. For our purposes here, we are primarily interested in the equivalence classes

formed by normalizing inflectional morphology, for example conflating "respond" and "responds", rather than "respond" and "responsiveness", although more aggressive stemming methods will do both conflations.

In the word-based morphology framework, inflection is captured by triples consisting of the surface form (also called wordform) w, a lexeme signifying the pairing of the surface for with a meaning and a slot σ, which can be understood as a set of "atomic" units of morphological meaning, also called inflectional features (Aronoff, 1976; Sylak-Glassman et al., 2015; Cotterell et al., 2019). A lemma is the surface form used to look up the lexeme in a dictionary, such as the infinitive verb form. Lemmatisation in Russian and set of inflectional features used for Russian is most commonly based on

3.2 Latent Dirichlet Analysis

symmetric vs asymmetric prior mallet gibbs sampling implementation

3.3 Evaluation metrics

4 Corpus

https://github.com/JoannaBy/RussianNovels

4.1 Conflation methods and vocabulary reduction

5 Future Work

Standardization of stop words for more consistent comparison between metrics Shannon-Jenson divergence Stability

Corpora from other domains - Russian National Corpus and OpenCorpus

References

M. Aronoff. 1976. *Word Formation in Generative Grammar*. Linguistic inquiry monographs. MIT Press.

Matthew Baerman, Dunstan Brown, and Greville G Corbett. 2015. Understanding and measuring morphological complexity: An introduction. In Matthew Baerman, Dunstan Brown, and Greville G Corbett, editors, *Understanding and measuring morphological complexity*, chapter 1. Oxford University Press, USA.

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

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7:327–342.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2020. Semi-supervised topic modeling for gender bias discovery in English and Swedish. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 79–92, Barcelona, Spain (Online). Association for Computational Linguistics.

Olessia Koltsova and Sergei Koltsov. 2013. Mapping the public agenda with topic modeling: The case of the russian livejournal. *Policy & Internet*, 5.

Daniel A McFarland, Daniel Ramage, Jason Chuang, Jeffrey Heer, Christopher D Manning, and Daniel Jurafsky. 2013. Differentiating language usage through topic models. *Poetics*, 41(6):607–625.

Olga Mitrofanova. 2015. Probabilistic topic modeling of the russian text corpus on musicology. In *International Workshop on Language, Music, and Computing*, pages 69–76. Springer.

Alexandra Schofield and David Mimno. 2016. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, 4:287–300.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440.

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015. A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 674–680, Beijing, China. Association for Computational Linguistics.

Adam Vogel and Dan Jurafsky. 2012. He said, she said: Gender in the ACL Anthology. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 33–41, Jeju Island, Korea. Association for Computational Linguistics.

T. Wade, D. Gillespie, S. Gural, and M. Korneeva. 2020. *A Comprehensive Russian Grammar*. Blackwell Reference Grammars. Wiley.