Controlling Linguistic Style Aspects in Neural Language Generation

Jessica Ficler and Yoav Goldberg

Stylistic Variation Workshop, 2017

Problems Addressed

Model

1. The paper experiments with controlling stylistic aspects of generated text, along with content.
2. Most methods control a single, focused stylistic aspect of the text. This paper achieves finer-grained control over the generated outcome, controlling several stylistic aspects simultaneously.

Proposed Solution

1. The proposed solution is based on the idea of a conditioned RNN language model.
2. The model assumes a set of k parameters $\{p_1, ..., p_k\}$, each parameter p_i with a set of possible values $\{v_1, ..., v_{p_i}\}$.

   Given a specific assignment to these values, the goal is to generate a text that is compatible with the parameters values.
3. The model uses all naturally occurring sentence lengths, and generates text according to:

   a.) Two content-based parameters

   - Sentiment Score (Positive, Neutral, or Negative)
   - Topic (Plot, Acting, Production, Effects, or Other)

   b.) Four stylistic parameters

   - Length ($<$ 11 words, 11 to 20 words, 21 to 40 words, $>$ 40 words)
   - Descriptiveness (Boolean)
   - Whether it is written in personal voice (Boolean)
   - Whether it is written in professional style (Boolean)
4. These parameters are considered using an additional context vector c, which is concatenated to the input word vector of the RNN language model at each time step. c itself is simply a concatenation of embedding vectors considering the above-mentioned textual properties.

Annotating Data

Rotten Tomatoes Movie Reviews Dataset

- Annotations are derived either from the associated metadata or are extracted using one of the following three heuristics:
  - Based on lists of content-words.
  - Based on existence of certain function words.
  - Based on distribution of parts of speech tags.

Metadata based annotation

- A sentence is labelled as *Professional:True* if it is written by either a review that is a professional critic, or by a reviewer that is marked as a "Super Reviewer" on the website.
- For sentiment, the scores are set to a scale of 5, with 0-2 corresponding to negative, 3 corresponding to neutral, and 4-5 corresponding to positive.

  In case of audience reviews, the user rates the movie on a scale of 5.

  In case of critic scores, the score is taken directly from their website and is normalized to a scale of 5.

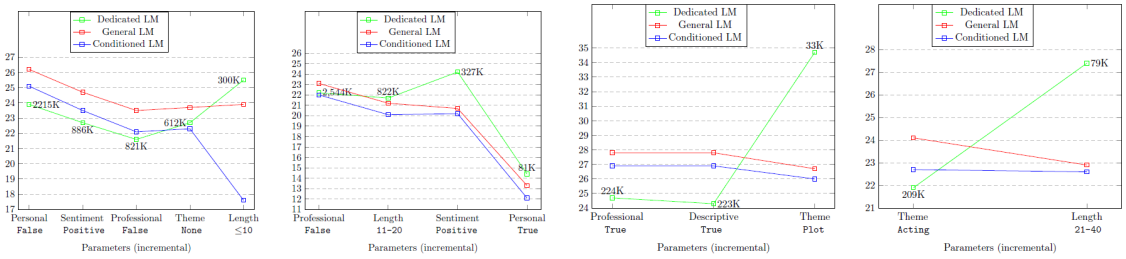
Text based annotation

- Length is counted and placed into one of the 4 bins mentioned above.
- A sentence is said to be written in Personal Voice if it is found to contain any personal pronouns or possessive word.
- Theme is assigned to a sentence by comparing the words in a sentence with the most common words of every category (list of words available in supplementary material).
- A sentence is termed Descriptive if it makes heavy use of adjectives i.e. if at least 35% of all of its POS tags are adjectives.

Experiments

1. The vocabulary is encoded using Byte Pair Encoding (BPE) as it allows representation of an open vocabulary through a fixed size vocabulary by splitting rare words into subword units.

Results



	dev	test
Not-conditioned	25.8	24.4
Conditioned	**24.8**	**23.3**

Table 2: Conditioned and not-conditioned language model perplexities on the development and test sets.

Correct Value	23.3
Replacing Descriptive with non-Descriptive	27.2
Replacing Personal	27.5
Replacing Professional	25
Replacing Sentiment `Pos` with `Neg`	24.3

Table 3: Test-set perplexities when supplying the correct parameter values and when supplying the opposite values.

Quantitative Results

- It is verified that knowing the parameters helps in achieving better language modeling results by comparing the dev-set and test-set perplexities of the conditioned language model to an unconditioned (regular) language model trained on the same data.
- The second baseline to the conditioned LM is to train a separate unconditioned LM on a subset of the data fitting only a few of the characteristics.

  - It is only when just a few conditioning parameters are needed, and the coverage of the parameter combination in the training set is large enough, that the dedicated LM outperforms the conditioned LM. It isn't scalable.
  - As the number of conditioning factors is increased, the amount of available training data to the dedicated model drops, and so does the modeling quality.
  - The conditioned model manages to generalize from sentences with different sets of properties, and is effective also with large number of conditioning factors.
  - Hence, it is verified that the conditioned LM trained on all the data is effective than a dedicated LM, as it is able to generalize across properties-combinations, and share data between the different settings.

- The third baseline is provided by comparing the perplexity when using the correct conditioning values to the perplexity achieved when flipping the parameter value to an incorrect one.
  - There is a major increase in perplexity when flipping the parameter values.
  - The increase is smallest for sentiment, and largest for descriptiveness and personal voice.
  - Hence, it is concluded that the model distinguishes descriptive text and personal voice better than it distinguishes sentiment and professional text.

- When requesting descriptive text, 85.7% of the generated sentences fit the descriptiveness criteria. When requesting nondescriptive text, 96% of the generated sentences are non-descriptive according to our criteria.

- When requesting text in personal voice, 100% of the generated sentences fit the criteria. When requesting non-personal text, 99.85% of the sentences were non-personal.

Requested Length	Avg	Min	Max	Deviation$_{m=2}$
<=10	7.6	1	21	0.2 %
11-20	20.6	5	25	2.6 %
21-40	34	7	49	0.6 %

Table 4: Average, minimum and maximum lengths of the sentences generated according to the correspond length value; as well as deviation percentage with margin (m) of 2.

Requested value	% Plot	% Acting	% Prod	% Effects	% Other
Plot	98.7	0.8	0	0.2	0.3
Acting	2.5	95.3	0	0.6	1.6
Production	0	0	97.4	2.6	0
Effects	0	5.9	0	91.7	2.4
Other	0.04	0.03	0	0.03	99.9

Table 5: Percentage of generated sentences from each theme, when requesting a given theme value.

Qualitative Results

- The professional property of the generated sentences was evaluated manually using Mechanical Turk.
  - For this, 1000 sentences pairs were randomly generated, each with one sentence generated with professional set to true, and the other with the attribute set to false.
  - The annotators were asked to identify the sentence written by a professional critic. Each pair was evaluated by 5 annotators.
  - When taking a majority vote among the annotators, they were able to tell apart the professional from non-professional sentences generated sentences in 72.1% of the cases.

- The sentiment of the generated sentences was also evaluated manually using Mechanical Turk.
  - 300 pairs of generated sentences were randomly created for each of the following settings: positive/negative, positive/neutral and negative/neutral.

- The annotators were asked to mark which of the reviewers liked the movie more than the other.
- Each of the pairs was annotated by 5 different annotators and a majority vote was taken.
- The annotators correctly identified
  a.) 86.3% of the sentence in the Positive/Negative case
  b.) 63% of the sentences in the Positive/Neutral case
  c.) 69.7% of the sentences in the negative/neutral case.

- The ability of the model to generalize was tested by testing whether it could generate sentences for parameter combinations that it had not seen in training.
  - 75,421 sentences which were labeled as theme:plot and personal:true were removed from the training set, and a conditioned LM was re-trained.
  - The trained model saw 336,567 examples of theme:plot and 477,738 examples of personal:true, but has never seen examples where both conditions held together.
  - The trained model was then asked to generate sentences with these parameter values.
    a.) 100% of the generated sentences contained personal pronouns.
    b.) 82.4% of them fit the theme:plot criteria
    c.) In comparison, a conditioned model trained on all the training data managed to fit the theme:plot criteria in 97.8% of the cases.

Bibliography

1. Fadi Abu Sheikha and Diana Inkpen. 2011. Generation of formal and informal sentences. In Proceedings of the 13th European Workshop on Natural Language Generation. Association for Computational Linguistics, Nancy, France, pages 187–193. http://www.aclweb.org/anthology/W11-2826.

2. John A Bateman and Cecile Paris. 1989. Phrasing a text in terms the user can understand. In IJCAI. pages 1511–1517.

3. Douglas Biber and Susan Conrad. 2009. Register, genre, and style. Cambridge University Press.

4. Ondrej Dusek and Filip Jurcicek. 2016a. A context aware natural language generator for dialogue systems. In Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Association for Computational Linguistics, Los Angeles, pages 185–190. http://www.aclweb.org/anthology/W16-3622.

5. Ondrej Dusek and Filip Jurcicek. 2016b. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Berlin, Germany, pages 45–51. http://anthology.aclweb.org/P16-2008.

6. Philip Gage. 1994. A new algorithm for data compression. R & D Publications, Inc., volume 12, pages 23–38.

7. Cong Duy Vu Hoang, Gholamreza Haffari, and Trevor Cohn. 2016. Incorporating side information into recurrent neural network language models. In Proceedings of NAACL-HLT. pages 1250–1255.

8. Sepp Hochreiter and J¨urgen Schmidhuber. 1997. Long short-term memory. MIT Press, volume 9, pages 1735–1780.

9. Eduard Hovy. 1987. Generating natural language under pragmatic constraints. Journal of Pragmatics, volume 11, pages 689–719.

10. Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Controllable text generation. In Proc. of ICML.

11. Chloe Kiddon, Luke Zettlemoyer, and Yejin Choi. 2016. Globally coherent text generation with neural checklist models. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, Texas, pages 329–339. https://aclweb.org/anthology/D16-1032.

12. Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoderdecoders. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, Texas, pages 1328–1338. https://aclweb.org/anthology/D16-1140.

13. Remi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, Texas, pages 1203–1213. https://aclweb.org/anthology/D16- 1128.

14. Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Berlin, Germany, pages 994–1003. http://www.aclweb.org/anthology/P16-1094.

15. Zachary C Lipton, Sharad Vikram, and Julian McAuley. 2015. Capturing meaning in product reviews with character-level generative text models. arXiv preprint arXiv:1511.03683.

16. Franc¸ois Mairesse and Marilyn A Walker. 2011. Controlling user perceptions of linguistic style: Trainable generation of personality traits. MIT Press, volume 37, pages 455–488.

17. David D McDonald and James D Pustejovsky. 1985. A computational theory of prose style for natural language generation. In Proceedings of the second conference on European chapter of the Association for Computational Linguistics. Association for Computational Linguistics, pages 187–193.

18. Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. 2016. What to talk about and how? selective generation using lstms with coarse-to-fine alignment. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, San Diego, California, pages 720–730. http://www.aclweb.org/anthology/N16-1086.

19. Tomas Mikolov, Martin Karafi´at, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In Interspeech. volume 2, page 3.

20. Kate G Niederhoffer and James W Pennebaker. 2002. Linguistic style matching in social interaction. Journal of Language and Social Psychology, volume 21, pages 337–360.

21. Ellie Pavlick and Ani Nenkova. 2015. Inducing lexical style properties for paraphrase and genre differentiation. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Denver, Colorado, pages 218– 224. http://www.aclweb.org/anthology/N15-1023.

22. Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. Transactions of the Association for Computational Linguistics, volume 4, pages 61–74.

23. Richard Power, Donia Scott, and Nadjet Bouayad- Agha. 2003. Generating texts with style. In Proc. of CiCLING. Springer, pages 444–452.

24. Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. arXiv preprint arXiv:1704.01444. Ehud Reiter and Sandra Williams. 2010. Generating texts in different styles. In The Structure of Style, Springer, pages 59–75.

25. Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In Proceedings of NAACL-HLT. pages 35–40.

26. Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics.

27. Jian Tang, Yifan Yang, Sam Carton, Ming Zhang, and Qiaozhu Mei. 2016. Context-aware natural language generation with recurrent neural networks. arXiv preprint arXiv:1611.09900.

28. Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Lisbon, Portugal, pages 1711–1721. http://aclweb.org/anthology/D15-1199.

29. Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In Proceedings of COLING 2012. The COLING 2012 Organizing Committee, Mumbai, India, pages 2899– 2914. http://www.aclweb.org/anthology/C12-1177.

30. Yinfei Yang and Ani Nenkova. 2014. Detecting information-dense texts in multiple news domains. In AAAI. pages 1650–1656.