A Closer Look at Recent Results of Verb Selection for Data-to-Text NLG

Guanyi Chen*
Utrecht University
g.chen@uu.nl

Jin-Ge Yao
Microsoft Research Asia
$jinge \cdot yao@microsoft \cdot com$

Abstract

Automatic natural language generation systems need to use the contextually-appropriate verbs when describing different kinds of facts or events, which has triggered research interest on verb selection for data-to-text generation. In this paper, we discuss a few limitations of the current task settings and the evaluation metrics. We also provide two simple, efficient, interpretable baseline approaches for statistical selection of trend verbs, which give a strong performance on both previously used evaluation metrics and our new evaluation.

1 Introduction

The authors of financial headlines often need to use appropriate verbs to describe a percentage change, sometimes also revealing its intensity. For instance, the verb *climb* in the headline *Microsoft's profit climbed 28%* expresses an upward direction, as well as the magnitude, of a percentage change. Likewise, an automatic natural language generation systems for data-to-text generation under similar scenarios should also properly select verbs as well. In earlier systems, neutral verbs such as *increase* and *decrease* that only describes the direction of changes were preferred for simplicity. However, the generated sentences can be more natural if automatic NLG systems could use a more diverse set of verb choices suitable in the context like human writers could do.

Due to the vagueness of word meaning and variations in word usage, data-driven methods are believed to be a reasonable choice for automatic systems (Reiter, 2018). Recently, there indeed exist several studies working towards this direction, with a special focus on statistical or probabilistic methods for selecting verbs to describe trends in percentages. In the most typical task settings, a dataset is extracted from some sentences in a corpora describing percentage changes, with the form of $\mathcal{D} = \{(x_1, v_1), ..., (x_N, v_N)\}$, where x_i is the numeric value of the percentage change (e.g. 12.5%) and v_i is the verb used by human writers in the original sentence to describe the percentage change. Rising trends and falling trends were typically collected separately. The task is to train a verb selector $f(\cdot)$ aiming at mapping a new input percentage x_* to a verb v_*, or a distribution of appropriate verbs from which a verb could be randomly selected to form an utterance.

Thomson Reuters' NLG system (Plachouras et al., 2016) for macro-economic indicator and merger-and-acquisition deals data includes a submodule for trend verb selection (Smiley et al., 2016). For each verb, their system estimates the interquartile range (IQR) of its associated percentage changes from the corpus. Given a new percentage change, their method randomly selects a verb from those verbs whose IQRs could cover the specified percentage value.

One more recent work (Zhang et al., 2018) proposed a Bayesian probabilistic model by estimating the prior probability of each verb as well as the likelihood of seeing the percentage change given the verb. Their evaluation based on the mean reciprocal ranking (MRR) of the human-written verb suggests significant superiority over the Thomson Reuters' approach.

However, we notice that naively outputing the most frequently used verbs for upward / downward direction can perform surprisingly high in terms of a few automatic metrics on the dataset used by Zhang et al. (2018). In this paper, we try to point out a few limitations of currently used evaluation metrics, and provide two simple, efficient and interpretable baseline approaches that achieve results competitive to prior approaches in both previous metrics and our new evaluation strategies.

* Work done during internship at MSRA.

2 Systems in Comparison

The scope of this study is mainly a slightly closer investigation of approaches compared in the recent study by Zhang et al. (2018). For self-containedness, we briefly describe the systems in comparison. The datasets we used to derive or train these systems are those collected and reported by Zhang et al. (2018). [1]

Thomson Reuters: The method adopted by Smiley et al. (2016) as aforementioned.

Neural Networks: A feed-forward neural network with hidden layers and rectified linear unit activations, trained with ℓ_2 regularization. Detailed settings were following Zhang et al. (2018).

Bayesian Models: The method proposed by Zhang et al. (2018), which is a generative model of the posterior $P(v|x)$ inferred by the Bayes rule:

$$P(v|x) \propto P(x|v)P(v), \quad (1)$$

where the likelihood $P(x|v)$ (conditioned on a given verb v) and the verb prior $P(v)$ are estimated from corpus statistics. Zhang et al. (2018) formulated the likelihood model using either kernel density estimation (KDE) or a Beta distribution, while the prior was estimated by frequency ratio with the Jelinek-Mercer smoothing (Jelinek, 1980) on a uniform distribution over all verbs $\mathcal{V}$:

$$P(w) = \lambda \frac{freq(v)}{\sum_{v'} freq(v')} + (1-\lambda)\frac{1}{|\mathcal{V}|}. \quad (2)$$

The choice of λ dictates the trade-off between accuracy and diversity.

In our study, we also introduce two more straightforward baseline approaches that are simpler, more efficient, and more interpretable than non-parametric estimators such as KDE.

The Frequency Baseline: The simplest baseline that directly samples a verb based on the overall frequency distribution was ignored in previous studies when calculating the metrics. In our study, we would like to investigate how different the previous systems perform when compared with this baseline in metrics.

Decision Tree Baseline: One simple improvement of the frequency baseline is segmenting the range of x into groups, and separately calculate frequency distributions within each group. To keep the baseline simple, we only split the range into two groups, which is equivalent to a decision tree with a depth of one. We use information gain as the splitting criterion (Safavian and Landgrebe, 1991). See Figure 1 for a visualization of the split point, which roughly depicts a group of moderate changes versus another group for larger changes.

[1] Retrieved from: https://goo.gl/gkj8Fa ; We only use the WSJ and Reuters subsets in English in this study.

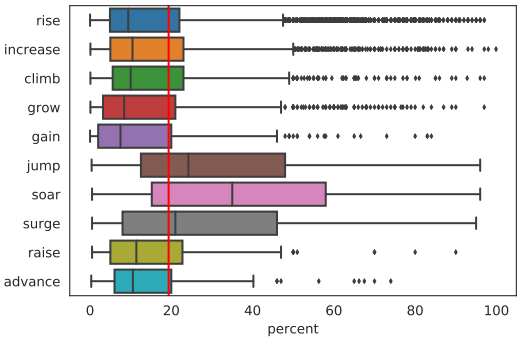

Figure 1: Box plot for the top-10 frequent verbs describing upward trends, with the red line denoting the decision tree split

3 Experiments

3.1 The Original Automatic Metrics

The automatic metrics reported by Zhang et al. (2018) aim at evaluating accuracy and diversity. The accuracy for verb selection is reflected by the Mean Reciprocal Rank (Voorhees et al., 1999; Radev et al., 2002, MRR) of the reference verb:

$$\text{MRR} = \frac{1}{|\mathcal{T}|} \sum_{(x'_i, v'_i) \in \mathcal{T}} \frac{1}{rank(v'_i)}, \quad (3)$$

where T denotes the held-out test set: $\mathcal{T} = \{(x'_1, v'_1), .., (x'_M, v'_M)\}$. For diversity, the number of distinct verbs (*richness*) and the relative abundance (*evenness*) in the selected outputs are reported, along with the well-known Inverse Simpson Index aka Simpsons Reciprocal Index (Simpson, 1949) that takes both aspects into account: $D = (\sum_{i=1}^{R} p_i^2)^{-1}$, where R is the total number of distinct selected verbs. Evenness is calculated as D/R.

Table 1 listed the MRR results along with the difference of top-1 system output with the most frequent baseline. We also calculate the Jensen-Shannon Divergence (Wong and You, 1985; Osterreicher and Vajda, 2003, JSD), which is defined to be the mean of KL divergences in two directions, to measure how different the probability distributions are compared with the frequency distribution. We mostly reproduce the results from (Zhang

corpus	Model	upward verbs			downward verbs		
		MRR(%)	Diff.(%)	JSD	MRR(%)	Diff.(%)	JSD
WSJ	Thomson Reuters	11.92 ± 0.16	-	-	10.55 ± 0.34	-	-
	Neural Network	72.33 ± 1.14	0	0.01	68.48 ± 1.46	0	0.11
	Frequency	72.27 ± 1.18	-	-	68.32 ± 1.66	-	-
	Decision Tree	72.40 ± 1.14	0	0.08	68.56 ± 1.60	0	0.09
	Bayesian ($\lambda = 1$, KDE)	72.41 ± 1.14	0	0.12	68.61 ± 1.56	1.3	0.14
	Bayesian ($\lambda = 1$, Beta)	72.30 ± 1.14	13.20	0.17	68.54 ± 1.53	13	0.18
	Bayesian ($\lambda = 0.05$, KDE)	53.32 ± 1.82	88	0.49	51.57 ± 0.31	84.3	0.51
	Bayesian ($\lambda = 0.05$, Beta)	52.68 ± 1.16	87.6	0.51	53.19 ± 1.08	69.7	0.52
Reuters	Thomson Reuters	37.01 ± 3.33	-	-	33.86 ± 2.27	-	-
	Neural Network	88.91 ± 3.65	0	0.13	88.58 ± 3.52	0	0.09
	Frequency	88.55 ± 3.85	-	-	88.34 ± 3.85	-	-
	Decision Tree	88.76 ± 3.76	0	0.14	88.38 ± 3.52	0	0.18
	Bayesian ($\lambda = 1$, KDE)	88.70 ± 3.80	0	0.15	88.10 ± 3.61	1.9	0.26
	Bayesian ($\lambda = 1$, Beta)	88.67 ± 4.46	15.1	0.19	87.16 ± 3.82	28.9	0.26
	Bayesian ($\lambda = 0.05$, KDE)	72.87 ± 5.97	87.9	0.52	79.91 ± 3.60	73.2	0.54
	Bayesian ($\lambda = 0.05$, Beta)	72.11 ± 6.95	89.1	0.54	69.52 ± 5.36	84.1	0.57

Table 1: MRR along with the top-1 difference and the Jensen-Shannon Divergence with the frequency baseline

et al., 2018) with a difference in the neural network baseline, which appears to be equally competitive against other systems in our experiments when running their implementation. [2] Additionally, we have two more interesting observations:

1. The frequency baseline is not distinguishable from the Bayesian methods in terms of MRR;

2. Systems achieving high MRR in fact yield neglectable difference with the frequency baseline in terms of the top-1 output and JSD.

For automatic metrics measuring diversity, we list the results in Table 2. By correcting an error in the codes from Zhang et al. (2018) where they amplified the differences between two systems by dropping cases that both yield the same verbs, the results are slightly different from what was originally reported. Bayesian models with low smoothing factor (λ) now defeat (not significantly) the Thomson Reuters in the sense of both richness and evenness. Other systems, including the neural network, our two new baselines, and Bayesian models with high smoothing factor has lower diversity, but not goes to zero as stated by Zhang et al. (2018). Almost all the models have exactly the same level of richness. This said, diversity metrics alone cannot reflect quality in selection, as a system could always select every candidate verb with equal probabilities for maximal diversity. We need better ways to simultaneously evaluate appropriateness and comprehensiveness.

3.2 Alternative Evaluation

Our observation of the frequency baseline is reasonable in the sense that it in theory minimises the Bayes risk (Berger, 2013) if the conditional likelihood has tiny subtle differences and is overwhelmed by an overall prior of verb choice, and a number of systems in comparison with high MRR in fact reduce to this baseline by certain degree. More intuitively, the most frequently used verbs in the corpora are *rise* and *fall*, which are often correct but less informative. For each numerical value of percentage, there could be multiple verb choices that are plausible to human writers, while some verbs may not be appropriate. Since neither MRR nor diversity metrics could address this point, we conduct some alternative evaluations by collecting a gold-standard set of multiple plausible candidate verbs on Amazon Mechanical Turk (AMT), using the WSJ subset of the original dataset. Following Smiley et al. (2016), we restricted raters to those located in the United States, with an approval rating above 95% and 1,000 or more HITs approved.

[2] We conjecture that Zhang et al. (2018) encountered a few less likely but substantially low results from one or two trials due to the randomness in training neural networks.

corpus	Model	upward verbs			downward verbs		
		richness	evenness	diversity	richness	eveness	diversity
WSJ	Thomson Reuters	11	0.8555	9.4107	14	0.9485	13.2787
	Neural Network	11	0.2303	2.5330	14	0.1916	2.6819
	Frequency	11	0.2345	2.5799	14	0.2555	3.5770
	Decision Tree	11	0.2461	2.7077	14	0.2361	3.3065
	Bayesian ($\lambda = 1$, KDE)	11	0.2435	2.6789	13	0.2294	2.9822
	Bayesian ($\lambda = 1$, Beta)	11	0.2589	2.8489	14	0.2187	3.0619
	Bayesian ($\lambda = 0.05$, KDE)	11	0.9623	10.8046	14	0.8866	12.4126
	Bayesian ($\lambda = 0.05$, Beta)	11	0.9522	10.5855	14	0.9700	13.5797
Reuters	Thomson Reuters	4	0.9036	3.6142	4	0.8409	3.3635
	Neural Network	4	0.3730	1.4919	4	0.3955	1.5820
	Frequency	4	0.3496	1.3985	4	0.3573	1.4292
	Decision Tree	4	0.3699	1.4797	4	0.3741	1.4962
	Bayesian ($\lambda = 1$, KDE)	4	0.3590	1.4360	4	0.3411	1.3642
	Bayesian ($\lambda = 1$, Beta)	4	0.3668	1.4672	4	0.3623	1.4493
	Bayesian ($\lambda = 0.05$, KDE)	4	0.9129	3.6516	4	0.9262	3.7047
	Bayesian ($\lambda = 0.05$, Beta)	4	0.9705	3.8818	4	0.9175	3.6699

Table 2: The diversity of verb selection measured by the Inverse Simpson Index.

We construct natural language utterances using uniformly sampled percentage values and all the verbs in the dataset (11 verbs describing upward changes and 14 for downward changes in total), and using one consistent subject (*Net Profits*) to reduce the potential variance brought by different subjects which is beyond the scope of this study. AMT workers will try to annotate each sentence with one of the three degrees of appropriateness:

- 3 (Appropriate): The verb is among the most suitable ones to describe the percentage
- 2 (Okay): The verb could be used to describe the percentage, although it might not be one of the most appropriate choices
- 1 (Not Appropriate): The verb is not naturally used to describe the percentage

See supplementary notes for a screenshot of the annotation interface. For either upward and downward verbs, we uniformly random sampled 100 different percentage changes from the interval $[0, 100]$, which results in 12,500 judges in total.

The resulting corpus has the format that each percentage change is paired with a list of candidate verbs, each of which has five judges from five different workers. We treat any verb with more than three judges rated 3 as one appropriate verb. This crowdsourced dataset could be used to evaluate each system on the ability to select multiple plausible verbs, treating verb selection as an ordinary classification task. We randomly leave out 20% of this dataset as the development set, and the rest for the final test set evaluation. For systems giving probability distributions on verbs (as opposed to deterministic selections), it is natural to threshold the accumulated value to decide which verbs to include as the finally selected candidates, with the threshold determined on the development set. We calculate the performance in terms of the precision, recall and F1 measure of all the annotated appropriate verbs, and the results are displayed in Table 3. The values of the cross-entropy loss as we calculated follow similar trends, as included in the supplementary notes.

We can observe that under the new evaluation protocol, the results from the two baseline approaches are more competitive than those reflected by MRR. The frequency baseline is surprisingly strong, indicating that authors of the standard corpora might favor more neutral verbs that could be applicable for almost all percentage values. A slightly more crafted decision tree marginally brings further performance increase. Meanwhile, we can also notice that models achieving higher diversity scores (cf. Table 2) have comparatively lower recall when all of them have precision vary-

Model	upward verbs				downward verbs			
	l	Precision	Recall	F1	l	Precision	Recall	F1
Thomson Reuters	5.67	48.98	51.35	48.22	7.34	39.23	51.04	41.81
Neural Network	9.15	50.10	84.94	61.96	12.29	40.97	89.12	53.45
Frequency	10	47.07	90.91	60.38	13	39.19	**92.86**	52.09
Decision Tree	10	50.58	**92.52**	**62.53**	11.94	45.48	86.59	53.73
Bayesian ($\lambda = 1$, KDE)	7.87	52.13	73.25	57.48	11.56	42.31	85.30	**54.10**
Bayesian ($\lambda = 1$, Beta)	7.62	**52.28**	71.85	56.95	11.01	44.97	80.05	52.64
Bayesian ($\lambda = 0.05$, KDE)	4.29	52.06	43.03	40.85	4.95	48.96	36.51	35.20
Bayesian ($\lambda = 0.05$, Beta)	3.77	44.66	38.38	36.93	3.54	**50.40**	31.08	34.29

Table 3: Results of viewing verb section as a classification task. Precision, recall, and F1 reported here are macro-averaged.; l stands for the average number of verbs each system has selected for each specified percentage.

ing around 50. This suggests that models may have learned rather flat distributions over verbs, with an implication that although these models have higher diversity on selecting verbs, they actually have lower diversity on selecting *appropriate* verbs. Given the current F1 values around 50 to 60, there still exists room for improvements over simple baselines for more precise selection.

4 Discussion

For the task of trend verb selection for data-to-text generation, our observations suggest that the evaluation results for current automatic metrics should be interpreted with caveats. Automatic verb selection systems that achieves good accuracy as reflected by high mean reciprocal ranks could in fact hardly yield real difference compared with just outputting the most frequently used verbs in overall statistics. More complex likelihood modelling using kernel density estimation could not produce more diverse selection of all plausible verbs as it behaves similar to a frequency baseline, while being slightly less interpretable compared with simpler frequency based approaches such as a shallow decision tree. One source of this issue should be the lack of good definition of appropriateness, as the difference in various verbs is often vague (see Figure 1 for an instance, where the used range for a number of verbs could almost span the entire range of the percentages). It also remains an open problem for a more thorough, rigorous, systematic treatment in terms of experimental design and evaluation protocol, given that this work and previous relevant studies have all just temporarily ignored the potential impact of different subjects in a sentence on verb selection for the simplicity of problem settings.

In this study, we focus on a few issues with recent work on trend verb selection in describing various kinds of percentages, with experiments conducted on samples collected from financial news data. We believe that similar caveats should exist in lexical choice problems appeared in other domains as well, such as various kinds of phrases in weather forecasts (Reiter et al., 2005; Ramos-Soto et al., 2014; Li et al., 2016) and in sports match reports (van der Lee et al., 2017; Wiseman et al., 2017; Qin et al., 2018), and hopefully the field will be exploring on some more principled, domain-agnostic approaches in the future.

Acknowledgments

We thank all three anonymous reviewers for helpful comments on our submitted draft.

References

James O Berger. 2013. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Frederick Jelinek. 1980. Interpolated estimation of markov source parameters from sparse data. In *Proc. Workshop on Pattern Recognition in Practice, 1980*.

Chris van der Lee, Emiel Krahmer, and Sander Wubben. 2017. PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 95–104, Santiago de Compostela, Spain. Association for Computational Linguistics.

Xiao Li, Kees van Deemter, and Chenghua Lin. 2016. Statistics-based lexical choice for NLG from quan-

titative information. In *Proceedings of the 9th International Natural Language Generation conference*, pages 104–108, Edinburgh, UK. Association for Computational Linguistics.

Ferdinand Osterreicher and Igor Vajda. 2003. A new class of metric divergences on probability spaces and its applicability in statistics. *Annals of the Institute of Statistical Mathematics*, 55(3):639–653.

Vassilis Plachouras, Charese Smiley, Hiroko Bretz, Ola Taylor, Jochen L Leidner, Dezhao Song, and Frank Schilder. 2016. Interacting with financial data using natural language. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1121–1124. ACM.

Guanghui Qin, Jin-Ge Yao, Xuening Wang, Jinpeng Wang, and Chin-Yew Lin. 2018. Learning latent semantic annotations for grounding natural language to structured data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3761–3771, Brussels, Belgium. Association for Computational Linguistics.

Dragomir R Radev, Hong Qi, Harris Wu, and Weiguo Fan. 2002. Evaluating web-based question answering systems. In *LREC*.

Alejandro Ramos-Soto, Alberto Jose Bugarin, Senén Barro, and Juan Taboada. 2014. Linguistic descriptions for automatic generation of textual short-term weather forecasts on real prediction data. *IEEE Transactions on Fuzzy Systems*, 23(1):44–57.

Ehud Reiter. 2018. Lexical choice needs machine learning!

Ehud Reiter, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167(1-2):137–169.

S Rasoul Safavian and David Landgrebe. 1991. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674.

Edward H Simpson. 1949. Measurement of diversity. *Nature*, 163(4148):688.

Charese Smiley, Vassilis Plachouras, Frank Schilder, Hiroko Bretz, Jochen Leidner, and Dezhao Song. 2016. When to plummet and when to soar: Corpus based verb selection for natural language generation. In *Proceedings of the 9th International Natural Language Generation conference*, pages 36–39, Edinburgh, UK. Association for Computational Linguistics.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82. Citeseer.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Andrew KC Wong and Manlai You. 1985. Entropy and distance of random graphs with application to structural pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (5):599–609.

Dell Zhang, Jiahao Yuan, Xiaoling Wang, and Adam Foster. 2018. Probabilistic verb selection for data-to-text generation. *Transactions of the Association for Computational Linguistics*, 6:511–527.