When to Plummet and When to Soar: Corpus Based Verb Selection for Natural Language Generation

Charese Smiley[1], **Vassilis Plachouras** [2], **Frank Schilder**[1], **Hiroko Bretz**[1], **Jochen L. Leidner**[2], **Dezhao Song**[1]
[1] Thomson Reuters, Research & Development, 610 Opperman Drive, Eagan, MN, USA
[2] Thomson Reuters, Research & Development, 1 Mark Square, London, EC2A 4EG, UK
firstname.lastname@thomsonreuters.com

Abstract

For data-to-text tasks in Natural Language Generation (NLG), researchers are often faced with choices about the right words to express phenomena seen in the data. One common phenomenon centers around the description of trends between two data points and selecting the appropriate verb to express both the direction and intensity of movement. Our research shows that rather than simply selecting the same verbs again and again, variation and naturalness can be achieved by quantifying writers' patterns of usage around verbs.

1 Introduction

In April 2016, the headline "GoPro's stock rocketed up 19 percent after it poached top Apple designer" was splashed across the top of the Business Insider Tech pages [1]. The authors of stories such as these often use descriptive language such as verbs like *rocketed up* to convey both the direction of motion of a percentage change along with its intensity. Although it is appropriate to use a more neutral verb like *increase* or *decrease* as is the case with most previous research, a more natural sounding text can be generated if we can incorporate the intensity of change.

This paper discusses the use of a large scale news corpus to quantify which verb to use in data-to-speech generation. In this work, we propose that the verb can be collocated to the percentage change such that certain types of trends can be described using a narrow set of verbs while other trends lend themselves to wider variation. We have developed the proposed method in the context of Thomson Reuters Eikon, an NLG system for macro-economic indicator and merger & acquisition deals data (Plachouras et al., 2016). However, the proposed method can be used for other domains with an appropriate corpus. The major contributions of this work are, to the best of our knowledge, the first large scale corpus study of lexical choice for perceptual change verbs with an evaluation using Amazon Mechanical Turk.

This article is structured as follows. Related work is discussed in the next section. Section 3 covers methods. Experiments and Discussion are discussed Section 4. Finally, Section 5 concludes the paper.

2 Related Work

Previous corpus based studies on the relationship between numbers and surrounding context for generation purposes have concentrated on the generation of appropriate numbers for a text in terms of roundness (e.g. 25 vs. 25.9) and format (such as preference for fraction vs. percentages) (Power and Williams, 2012) and hedging and rounding in conjunction with numerical expressions (e.g. less than 25%) (Williams and Power, 2013).

Several studies have explored generation of descriptions of times series data. The TREND system (Boyd, 1998) focuses on the generation of descriptions of historical weather patterns concentrating primarily on the detection of upward and downward trends in the data and using a limited set of verbs (*rose*, *dropped sharply*) to describe the direction and intensity of movement. More recently,

[1] http://uk.businessinsider.com/gopro-stock-rocketed-up-19-percent-2016-4

Proceedings of The 9th International Natural Language Generation conference, pages 36–39,
Edinburgh, UK, September 5-8 2016.

Ramos-Soto et al. (2013) also address the surface realization of weather trend data. They create an "intermediate language" for temperature, wind etc. and then consider 4 different ways to verbalize temperatures based on the minimum, maximum and trend in the time frame considered. In contrast, our method selects the verb based on the trend without hardwiring the mapping at system development, as the associations are learned from a corpus. NLG systems for the visually impaired have also explored the generation of text for trend data (Moraes et al., 2014) around the adaptation of generated descriptions to users' reading levels.

Perhaps the most similar work to ours is that of word choice in SUMTIME-MOUSAM (Reiter et al., 2005). This research conducted an empirical corpus-based study of human-written weather forecasts. One aspect of the research focused on verb selection in weather forecasts. They built a classifier to predict the choice of verb based on type (speed vs. direction), information content (change or transition from one wind state to another) and near-synonym choice. They found that verbs were chosen based upon the most salient semantic information such as whether wind speed, direction, or both constituted the most significant change. After a post-edit analysis where forecasters were asked to edit computer generated texts, they found that lexical choice was highly idiosyncratic based on the individual writer's idiolect. Our research shows that although there is an aspect of variability, writers may be operating within a more limited scope of possible lexical choice depending on factors such as the intensity of change.

3 Methods

For this study, we use the Reuters News Archive, a large corpus of 14 million news articles on a variety of topics collected from the Reuters News Agency[2]. Documents within the corpus were part-of-speech tagged using Stanford Core NLP (Manning et al., 2014). Then phrases that contained an expression of a percentage change in the form (subject, verb, number, percent) were extracted using a simple function in the format shown below:

[GoPro's stock] [rocketed up] [19 percent]

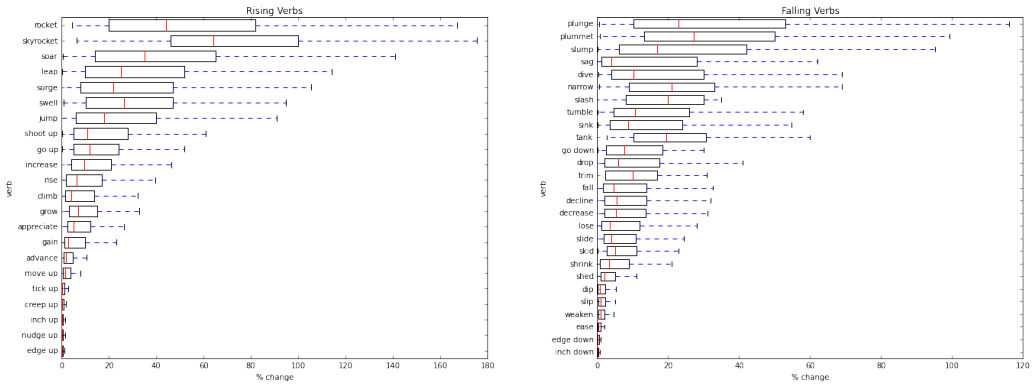
We elected to use percentage changes over absolute numbers as a way of minimizing some of the issues surrounding absolute numbers. An absolute number might be considered relatively small in one instance but large in another. For example, a 10 minute walk might be considered short while a 10 mile walk might be long. On the other hand, a $3 rise in a car priced at $30,000 would be a 0.01% increase whereas a $3 rise in gas priced at $3 would be a 100% increase thus suitably registering the magnitude of the change. Also, when dealing with precise numbers we have to consider the scale on which the number lies (e.g. 24 hours, 7 days, 60 minutes, etc.) (Krifka, 2007). This problem is avoided with percentages. Movements in the form of percentage changes are readily available in our news corpus and can be easily identified and extracted. Also, percentage change can be easily calculated given two data points and then the verb selection algorithm applied making this is useful for data-to-text systems.

After extracting a set of 1.7 million candidate phrases for a total of 5,417 verb types and 182,245 verb tokens, we eliminate rare verbs by removing phrases containing verbs that appear less than 50 times and phrases with noun-verb pairs that occur less than 2 times. We remove all modal and auxiliary verbs and keep only the bare form of the verb. Finally, we manually annotate the motion of the verb as rising or falling removing verbs such as *rebound* which imply a rising motion but have additional meaning of returning from a low to some previous high point. We also remove verbs such as *trade down* which are specific to a particular domain such as the stock market. After preprocessing, we are left with 49 verb types: 22 rising and 27 falling.

For each verb, we calculated the median, standard deviation, and interquartile range (IQR) for all instances of the verb in the corpus. Figure 1 (a) and (b) shows boxplots of the remaining verbs organized along the x-axis in order of ascending IQR with respect to the magnitude of change.

We find that verbs with a small IQR (e.g. *edge up* and *nudge up* are used with very low percentage changes. Verbs with larger IQRs are associated with more extreme changes (e.g. *skyrocket* and *rocket*). This pattern holds for both rising and falling verbs.

[2]A smaller version of this corpus is available at `http://trec.nist.gov/data/reuters/reuters.html`



(a) Boxplot of 22 rising verbs, ordered by interquartile range.

(b) Boxplot of 27 falling verbs, ordered by interquartile range.

Figure 1: Rising and Falling Verbs

4 Experiments

The goal of our evaluation is to test whether our verb generator outperforms a random baseline. That is, if verb selection is truly idiosyncratic, we would expect that raters will have no preference for one verb over another such that their responses cannot be distinguished from chance in the aggregate.

In order to compare our verb selections against human judgements of naturalness, we evaluate using multiple choice questions on Amazon Mechanical Turk (AMT). AMT is a platform which allows requesters to post questions and tasks in order to obtain crowdsourced answers from anonymized workers. Requesters can filter workers on a variety of criteria including location, approval rate and number of Human Intelligence Tasks (HITs) approved. We restricted raters to those located in the United States, with an approval rating above 95% and 1,000 or more HITs approved.

For each question, we asked raters to select the most natural sounding sentence from a pair of sentences that varied only in verb choice. Each question was set up as a HIT (for a total of 2,000 HITs) asking raters to make quick judgements about the naturalness of a sentence. The random baseline is 50% (the chance of arbitrarily choosing either (a) or (b)). An example HIT is shown in Figure 2.

The sentences were generated using 3 topics: *gross domestic product*, *net profits*, and *share prices*, chosen from the most popular subjects in our corpus. We chose 3 noun phrases in the subject position of the sentence in order to reduce the effect of subject on verb selection while somewhat minimizing the repetitiveness of completing multiple HITs. The effect of subject on verb selection will be explored in depth in future research. Percentages were randomly selected from the corpus data. The verbs were generated by randomly selecting a verb where the percentage in question fell within the IQR of the verb. This decision is made to avoid atypical uses of a particular verb. When the percentage change fell within the IQR there were often multiple verbs to choose from. For example, with a 2% increase, our generator would select from among: *move up*, *rise*, *gain*, *advance*, and *climb*. We assume that the specific choice of verb within that range is up to the writer depending on personal preference, writing context, and other factors. To simulate this, we randomly select among the verbs.

The second question was generated by randomly choosing a verb from the list where the percentage did not fall within the IQR. We randomly generated 1,000 question pairs for each of sets of verbs (rising and falling) for a total of 2,000 questions.

For the falling verbs, raters agreed with our selection in 663 / 1,000 instances. For rising verbs, raters agreed in 709 / 1,000 instances. Both findings are statistically significant above the chance baseline of 50% (p <0.0001 two-tail binomial test).

Disagreements between raters and our system were well distributed across all percentages. To keep the task simple for the raters, we did not ask them to justify their rationale for choosing one verb over the other. One limitation of the study, then, is that

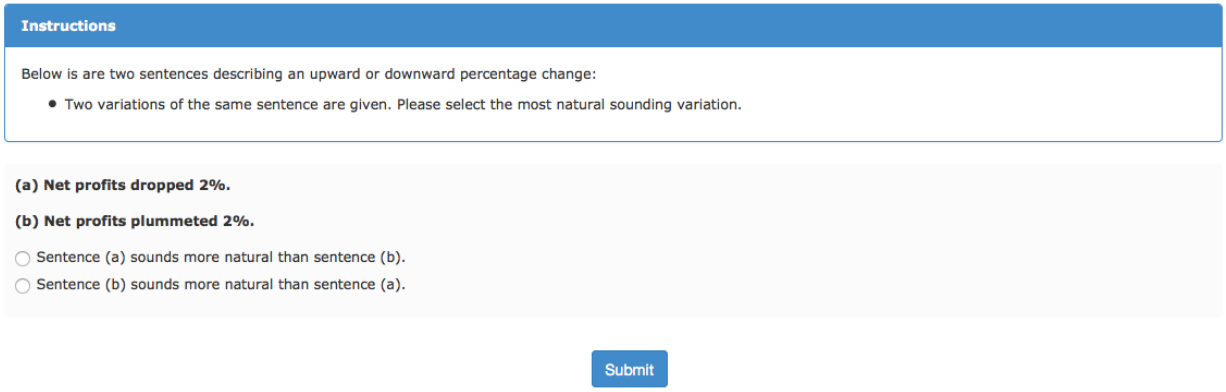


Figure 2: Example Verb Selection HIT on AMT

we cannot reliably distinguish raters who truly disagreed with our system's verb choice and those who are simply chose at random. However, we find it promising that we were able to reach statistical significance in spite of this.

5 Conclusion

We demonstrate verb selection for Thomson Reuters Eikon using a large news corpus. We find that verb selection can be quantified and that the results match our intuitions about which verbs express small and large rates of change. These results are further confirmed using an Amazon Mechanical Turk study of the naturalness of our generated texts.

Acknowledgments

We would like to thank Thomson Reuters F&R including Albert Lojko, Alex Tyrell, Sidd Shenoy, Rohit Mittal, and Jessica Tran as well as Tom Zielund and Khalid Al-Kofahi from Thomson Reuters R&D for their support and discussions. This work received financial support from Thomson Reuters Global Resources.

References

Sarah Boyd. 1998. Trend: a system for generating intelligent descriptions of time series data. In *IEEE International Conference on Intelligent Processing Systems (ICIPS1998)*. Citeseer.

Manfred Krifka. 2007. Approximate interpretation of number words: A case for strategic communication. *Cognitive foundations of interpretation*, pages 111–126.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Priscilla Moraes, Kathleen McCoy, and Sandra Carberry. 2014. Adapting graph summaries to the users? reading levels. *INLG 2014*, page 64.

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.

Richard Power and Sandra Williams. 2012. Generating numerical approximations. *Computational Linguistics*, 38(1):113–134.

Alejandro Ramos-Soto, Alberto Bugarín, Senén Barro, and Juan Taboada. 2013. Automatic generation of textual short-term weather forecasts on real prediction data. In Henrik Legind Larsen, Maria J. Martín-Bautista, M. Amparo Vila, Troels Andreasen, and Henning Christiansen, editors, *Flexible Query Answering Systems - 10th International Conference, FQAS 2013, Granada, Spain, September 18-20, 2013. Proceedings*, volume 8132 of *Lecture Notes in Computer Science*, pages 269–280. Springer.

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

Sandra Williams and Richard Power. 2013. Hedging and rounding in numerical expressions. *Pragmatics & Cognition*, 21(1):193–223.