UTFPR at SemEval-2020 Task 7: Using Co-Occurrence Frequencies to Capture Unexpectedness

Gustavo H. Paetzold
Universidade Tecnológica Federal do Paraná - Campus Toledo
ghpaetzold@utfpr.edu.br

Abstract

We describe the UTFPR system for SemEval-2020's Task 7: Assessing Humor in Edited News Headlines. Ours is a minimalist unsupervised system that uses word co-occurrence frequencies from large corpora to capture unexpectedness as a mean to capture funniness. Our system placed 22nd on the shared task's Task 2. We found that our approach requires more text than we used to perform reliably, and that unexpectedness alone is not sufficient to gauge funniness for humorous content that targets a diverse target audience.

1 Introduction

It is no secret that the Natural Language Processing community has been developing a growing interest in increasingly challenging text classification and regression tasks in recent years. Nowadays the community is coming together to find effective solutions to tasks that delve ever deeper into human subjectivity, such as identifying offensive language and racist remarks (Zampieri et al., 2019), categorizing complex and nuanced emotions (Klinger et al., 2018), identifying sarcasm (Ghosh and Veale, 2017), quantifying suicidal tendencies (O'dea et al., 2015), just to name a few. It is easy to see how a reliable approach to such tasks could help both businesses and institutions in more reliably and cost-effectively gauging public opinion, monitoring user behavior, driving the creation of new policies, products, services, and so on. Similarly, gauging how humorous a certain piece of content is could also help the entertainment industry, for example, to both evaluate and maybe even automatically create quality content for its users.

Automatic humor recognition has been addressed in many insightful contributions already. Cattle and Ma (2018), for instance, explore the use of many different kinds of word embeddings and neural models in judging whether or not a certain piece of text is humorous. Similarly, Barbieri and Saggion (2014) combine an extensive set of features with tree-based machine learning algorithms to identify humor and irony in text. What these contributions have in common is the fact that, because identifying humor is something inherently subjective and complex, they combine numerous, sometimes hard to obtain resources to perform well. In this contribution, we attempt to take a different approach and try to address this task as minimalistically as possible to establish a foundation for more elaborate strategies.

As it has been discussed in many academic contributions (Goatly, 2017; Borgianni et al., 2017; Potash et al., 2017), one of the most important characteristics of quality humorous content is its unexpectedness, regardless of the means through which it is delivered (text, audio, or video). Because unexpectedness is something that could potentially be captured through Natural Language Processing techniques, we decided to create a system that tries to gauge funniness by using unexpectedness alone. Because of our interest in potentially adapting our approach to under-resourced languages, we also set out to make ours an unsupervised system that uses as few resources as possible.

In this paper, we describe the UTFPR systems submitted to SemEval-2020's Task 7: Assessing Humor in Edited News Headlines. Section 2 describes the task and associated datasets, Section 3 showcases our approach and preliminary results obtained on the dev set, Section 4 reveals our performance on the official shared task, and Section 5 summarizes our contributions and conclusions.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1066–1070
Barcelona, Spain (Online), December 12, 2020.

2 Task Summary

The UTFPR systems were submitted to the SemEval-2020's Task 7: Assessing Humor in Edited News Headlines (Hossain et al., 2020). In this task, participants were challenged to conceive systems that quantify and/or compare the funniness of news headlines that had been modified for humorous effect. The task was divided into two sub-tasks:

1. Assigning a funniness score to each headline in the range of 0.0 (not funny) to 3.0 (funny); and

2. Given two different edited versions of the same headline, judging which one was funniest.

We chose to participate in Task 2, exclusively. The training, dev, and test set for Task 2 were composed of 9381, 2355, and 2960 instances, respectively. Each instance was composed of a news headline with a highlighted word, two replacements for the highlighted word, and a score indicating which one of the replacements made the headline funnier.

As discussed by Hossain et al. (2019), the annotation process of the dataset was quite extensive, involving both academics and Amazon turkers. Both the replacements and the funniness scores were produced by humans subject to monetary compensation. Several measures were taken to prevent annotator's fatigue and ensure annotation quality.

3 The UTFPR System

As was mentioned in Section 1, our system is a very minimalist unsupervised co-occurrence-based model that gauges headlines' funniness by trying to capture how unexpected they are. Given a headline h and a funny replacement w_f to the headline's target word, our system calculates its unexpectedness coefficient $U(h, w_f)$ using Equation 1, where C is a set of text corpora with one sentence per line, and count(c, w, w_f) is the number of sentences in which w and w_f co-occur in a given corpus $c \in C$.

$$U(h, w_f) = \frac{1}{\sum_{c \in C} \sum_{w \in h,\, w \neq w_f} \text{count}(c, w, w_f)} \tag{1}$$

We disregarded the co-occurrence counts of all stop-words in h, such as prepositions, conjunctions, etc. Our set C was composed of three corpora, all of which were made available through the translation shared tasks of WMT 2019 (Barrault et al., 2019):

- **Europarl:** 2,295,044 lines of sentences extracted from European Parliament transcripts.
- **News Commentary:** 545,918 lines of sentences extracted from articles on political and economic commentary.
- **News Week:** 3,309,171 lines of sentences extracted from assorted news articles.

These corpora were chosen because the type of content they contain can be easily obtainable for most languages, even if under-resourced. Parliamentary/governmental transcripts are commonly publicly available records, and news articles are ubiquitous.

It is quite self-evident that, just because two words co-occur very rarely, that does not imply that a sentence that includes both of them is funny. There are numerous reasons why they could co-occur very rarely, such as: They cause the sentence to be ungrammatical, incoherent, or incohesive, they pertain to completely unrelated subjects, one (or both) of these words are rarely used by news outlets because they are not commonly used in day-to-day conversations, etc. However, by inspecting the dataset, we noticed that, because the funny replacements on the dataset were suggested by well-trained humans exclusively, almost none of them would cause the original headline to be ungrammatical, incoherent, or incohesive. Also, because annotators were deliberately instructed to replace the target word with something that would make the headline funnier, most of the replacements suggested are popularly used words.

Our system, named UTFPR, was our official submission to the shared task, but we also tested 3 other variants of our model:

- **UTFPR-Max**: Instead of summing the co-occurrence counts of all non-stop-words in each sentence, we instead calculated the maximum co-occurrence between them and the funny replacement using Equation 2.

$$U(h, w_f) = \frac{1}{\sum_{c \in C} \max_{w \in h,\, w \neq w_f} \text{count}(c, w, w_f)} \tag{2}$$

- **UTFPR-SVM**: A binary SVM classifier model that uses the counts employed by the UTFPR and UTFPR-Max systems as input. It uses a total of 6 features as input: The sum of co-occurrence counts from each corpus (3 features) and the max co-occurence counts from each corpus (3 features). The SVM model uses the default configuration of the scikit-learn library (Pedregosa et al., 2011).

- **UTFPR-DT**: The same as UTFPR-SVM, except it uses a Decision Tree for a model. It uses the default configuration of the scikit-learn library (Pedregosa et al., 2011).

- **UTFPR-LR**: The same as UTFPR-DT, except it uses a Logistic Regression model. It uses the default configuration of the scikit-learn library (Pedregosa et al., 2011).

Table 1 features the Accuracy (A) and macro-averaged F-scores (F) obtained for each of these models on the dev set. As it can be observed, neither UTFPR-Max nor our machine learning models managed to beat our original approach on the dev set.

	A	**F**
UTFPR	**0.565**	**0.562**
UTFPR-Max	0.506	0.355
UTFPR-SVM	0.514	0.481
UTFPR-DT	0.532	0.531
UTFPR-LR	0.533	0.521

Table 1: Accuracy (A) and macro-averaged F-scores (F) obtained for the dev set.

4 Performance on Shared Task

Table 2 showcases the Accuracy obtained by all UTFPR systems on Task 2 along with the top and bottom 3 ranked systems. We placed 22nd overall, missing first place by a 12% Accuracy margin.

Inspecting the output produced for the test set, we observed a few issues with our approach and the dataset that may explain our somewhat subpar performance on the shared task. First, we noticed that almost 8% of the time our model found exactly 0 co-occurrences for both headline edits, which forced it to choose at random. This was caused by both the fact that we used a rather small set of corpora to calculate our unexpectedness scores, and the fact that news headlines are inherently very short, hence providing few opportunities for co-occurrence.

We also found that the assumption that all of the edits would be both coherent and cohesive is false, since a good portion of them weren't. A good example of that would be the headline "Sean Spicer to release **book** next summer", where book could be replaced by either "nudes", which is quite clever and would not cause for coherence/cohesiveness issues, or "read", which arguably would. These problems make it harder for a simple approach like ours to capture unexpectedness.

Finally, we also noticed that, although phenomena such as "millennial humor" (Koltun, 2018) have lead to a sharp increase in the popularity of humorous content that relies much more heavily on senseless unexpectedness than clever word-play, humor in news headlines commonly targets a broader, more diverse target audience that seems to require more than just unexpectedness to be entertained. For instance, one could argue that replacing **book** with "forklifts" in the previous example is more unexpected than replacing it with "nudes", but that would not necessarily make it funnier to a diverse mainstream audience.

	System	Accuracy
1	Hitachi	0.674
2	Amobee	0.660
3	YNU-HPCC	0.659
22	**UTFPR**	**0.569**
-	UTFPR-Max	0.500
-	UTFPR-SVM	0.508
-	UTFPR-DT	0.514
-	UTFPR-LR	0.563
30	heidy	0.419
31	SO	0.329
32	HumorAAC	0.320

Table 2: Accuracy scores obtained by the UTFPR system on the test set along with the top 3 and bottom 3 ranked systems in Task 2.

5 Conclusions

In this paper we introduced the UTFPR systems for the SemEval 2020's Task 7: Assessing the Funniness of Edited News Headlines. Ours is an unsupervised approach that uses co-occurrence frequencies from raw text corpora to capture the unexpectedness of a news headline edit. Due to their minimalist nature, our systems can be easily adapted to even under-resourced languages. Despite being so simple in nature, our best system placed 22nd out of 32 participants, missing the first spot by 12% Accuracy. We found that our systems could have potentially performed better if we had used larger text corpora to calculate co-occurrence scores. We also found that, although some edits in the dataset made it difficult for our systems to capture unexpectedness due to coherence/cohesion problems, we ultimately noticed that unexpectedness is not enough to determine the funniness of a headline.

In the future, we intend to try to use larger corpora, employ more sophisticated co-occurrence metrics that take into consideration the syntactic relations between words, and also use more sophisticated machine learning models.

Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

References

Francesco Barbieri and Horacio Saggion. 2014. Automatic detection of irony and humour in twitter. In *ICCC*, pages 155–162.

Loïc Barrault, Ondřej Bojar, Marta R Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61.

Yuri Borgianni, Gillian Hatcher, et al. 2017. Similarities and differences between humorous and surprising products. In *Proceedings of the 21st International Conference on Engineering Design (ICED 17)*, pages 031–040.

Andrew Cattle and Xiaojuan Ma. 2018. Recognizing humour using word associations and humour anchor extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1849–1858, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Aniruddha Ghosh and Tony Veale. 2017. Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 482–491, Copenhagen, Denmark. Association for Computational Linguistics.

Andrew Goatly. 2017. Lexical priming in humorous discourse. *The European Journal of Humour Research*, 5(1):52–68.

Nabil Hossain, John Krumm, and Michael Gamon. 2019. "president vows to cut <taxes> hair": Dataset and analysis of creative text editing for humorous headlines. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 133–142, Minneapolis, Minnesota. Association for Computational Linguistics.

Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020. Semeval-2020 Task 7: Assessing humor in edited news headlines. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain.

Roman Klinger, Orphée De Clercq, Saif Mohammad, and Alexandra Balahur. 2018. IEST: WASSA-2018 implicit emotions shared task. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 31–42, Brussels, Belgium. Association for Computational Linguistics.

Kim Koltun. 2018. Rick, morty, and absurdism: The millennial allure of dark humor. In *The Forum: Journal of History*, volume 10, page 12.

Bridianne O'dea, Stephen Wan, Philip J Batterham, Alison L Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on twitter. *Internet Interventions*, 2(2):183–188.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. Semeval-2017 task 6:# hashtagwars: Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 49–57.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.