Tweet2Story: A framework for automatic extraction of tweets narratives

Vasco Campos[1], Alípio Jorge[1] Ricardo Campos[2], and Inês Cantante[3]

[1] Faculdade de Ciências da Universidade do Porto, Porto, Porto, Portugal
up201908482@up.pt
[2] Instituto Politécnico de Tomar, Tomar, Santarém, Portugal
[3] Faculdade de Letras da Universidade do Porto, Porto, Porto, Portugal

Abstract. Social media platforms have become an important source of news and a stage for discussing current events. The narratives of the discussions on these platforms can help to understand the event and the public opinion about it. Previous works focused on extracting narratives from long documents, or simply relations between entities. However, we propose the automatic extraction of narratives on small sets of tweets about an event. The Tweet2Story frameworks allows us to extract a narrative from a set of tweets into an annotation file and visualize it through a knowledge graph. Lastly, our analysis of this framework shows it can keep up with state-of-the-art tools of Open Information Extraction (OpenIE) and compactly extract the narrative from a set of tweets.

Keywords: narrative extraction, open information extraction, twitter, narrative visualization

1 Introduction

Over the last decade, there was a significant increase on the use of social media platforms. One such platform that saw an exponential growth was *Twitter*. As of Q2 2021[4], *Twitter* has a total of 206 million active accounts and this number keeps growing every year. To a lot of users of this platform, specially for teens, it is also an important source of news [1], which gives *Twitter* a paramount position in today's world. According to a study by Muck Rack [2] about the state of journalism in 2021, **76%** of journalists list *Twitter* as their most valuable social media platform, making them another group of people that is strongly affected by the narratives on the platform.

Oftentimes, current events are discussed on *Twitter* in real time before they become more structured news. Given the short and snappy nature of tweets, as well as the large amounts that are published, it is difficult for journalists to follow up on the different dimensions of stories and opinions revolving around a current event. Thus, in our opinion, it is interesting to have a system that helps interpret these different dimensions of stories and opinions, by automatically extracting a narrative from a set of tweets about a current event.

[4] https://investor.twitterinc.com/home/default.aspx

Ever since its creation in 2006, *Twitter* has been a case study for many authors in the machine learning field, in particular in the Natural Language Processing (NLP) field. Social media platforms are an especially exciting challenge for NLP tasks, due to their colloquial language. Some of the most studied tasks are Sentiment Analysis [3], Keyword Extraction [4], Rumor detection [5] and Open Domain Event Extraction [6]. However, these tasks are usually one dimensional and focus on one aspect of the tweets. To the best of our knowledge, they have yet to tackle the extraction of a complete narrative from tweets.

In this paper, we will focus on our solution to tackle the challenge of automatic narrative extraction from tweets. We start in section 2, by making a study on the related literature about this task, followed by a detailed explanation of our solution for automatic narrative extraction on section 3. Section 4 contains an analysis of the results obtained by our solution, while the final section exposes our conclusions about the narrative extraction tool and its results.

2 Related Work

Metilli et al. [7] have made some steps towards the automation of the extraction of narratives. They identify 8 different steps to successfully extract knowledge from a text and build a narrative. These steps include event detection, named entity recognition and relation extraction, which are task we use in later chapter to build our framework. Despite detailing the pipeline for the extraction of narratives, their work only includes the training and testing of a model for the first two steps - Event Detection and Event Classification.

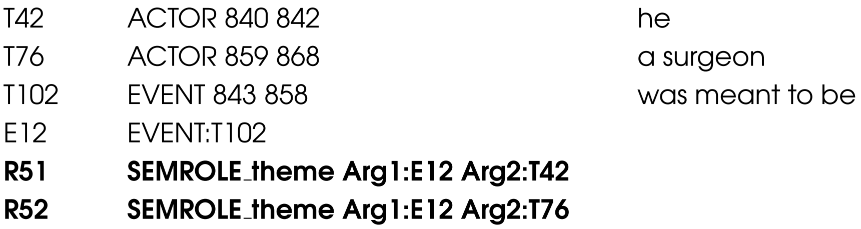
Other works focus on the narrative extraction aspect of long texts. Eisenberg [8] focuses on understanding parts of the narrative structure of a story, such as the narrative point of view and the diegesis. On the other hand, Vargas [9] explores the automatic extraction of narrative information for the task of story generation and to improve the performance of information extraction tools. He uses Russian folktale stories as a corpus to train the models.

Overall, there are not many efforts in the domain of automatic narrative extraction. However, our narrative extraction task can also be seen, in a simpler way, as an entity relation extraction task or as Open Information Extraction (OpenIE). To this extent, there are a few works related to our framework.

Cassirer et al. [10] presented ReVerb, which introduced a syntactic constraint to help extract triples (relations between entities) from a text. A few years later, Del Corro et al. [11] presented ClausIE, which improved OpenIE by simplifying complex sentences with multiple clauses (clause-based). Finally, Angeli et al. [12] presented Stanford OpenIE, which is also a clause-based approach, however it makes use of natural logic inference to shorten the clauses more than the previous systems.

3 Tweet2Story: Automatic extraction of narratives

The primary goal of the Tweet2Story framework, is to extract the narrative behind a set of tweets discussing a certain event. But **how will the narrative be represented?** The answer is through the brat[5] annotation system. Simply put, brat allows us to annotate the entities in a narrative, describe their role and identify the relations between them. Figure 1 shows how the sentence "he was meant to be a surgeon" is annotated.



```
T42    ACTOR 840 842                        he
T76    ACTOR 859 868                        a surgeon
T102   EVENT 843 858                        was meant to be
E12    EVENT:T102
R51    SEMROLE_theme Arg1:E12 Arg2:T42
R52    SEMROLE_theme Arg1:E12 Arg2:T76
```

Fig. 1: brat annotation example

Each row has an identifier (T42), the entity role ("ACTOR"), the character span where the entity can be found in the text (840 842) and the actual entity ("he"). The last two lines represent a relation between entities. In this case, the entity "he" is connected to "was meant to be", which is connected to "a surgeon". This annotation depicts the triple "he - was meant to be - a surgeon" and categorizes the semantic relation as a thematic relation (a general semantic role).

Knowing what the brat annotations are and how they can characterize a narrative, we revise the objective of the Tweet2Story framework. Its goal is to extract the narratives from a set of tweets about an event, by describing them through brat annotations.

In order to achieve this goal, we define an information extraction pipeline with 5 steps to extract text entities and their relations to produce a narrative:

1. ***Named Entity Recognition.*** Retrieve named entities from the text and place them in pre-defined categories. For example, in the sentence "Steve Jobs was the CEO of Apple", the entity "Steve Jobs" fits the category of "person";
2. ***Temporal Entity Extraction.*** Focuses solely on retrieving temporal information and mapping it into context independent representation. For example, the expression "last week" would be parsed as "14-09-2021";
3. ***Co-reference Resolution.*** Find co-references about actors in the entire document and groups them into clusters. Co-references are usually nouns or pronouns that refer to the same entity in a text. For example, in the sentence

[5] https://brat.nlplab.org/

"Sally lives in Paris. She lives in France", both "Sally" and "She" refer to the same entity and, therefore, belong in the same cluster;

4. ***Event Extraction.*** Detects events in the text, typically through verbs and their modifiers. For example, in the sentence "Sally lives in Paris", the event is expressed through the verb **"lives"**;
5. ***Entity Relation Extraction.*** Using the semantic role classification of each word/expression on a sentence, it extracts relations between entities (triples). The relations are always between an actor and an event and are always categorized. For example, "Sally lives in Paris" produces the triple 'Sally - lives - in Paris', which is categorized as a **location** triple.

All five pipeline tasks take advantage of pre-existing technology, usually in the form of pre-trained models, to achieve its objectives. On top of using pre-trained models, for some of them we devise a set of rules, to suit our main goal of extracting a narrative.

First of all, we decide that **verbs** are most likely the cause of events or relations between arguments [13]. Therefore, all events **must** contain a verb. Furthermore, two different verbs that are linked by **one** other argument are still considered part of the same event, as exemplified by table 1. This table shows the identification of the triple "he - was meant to be - a surgeon". The *tag* represents the output of the Semantic Role Labelling model, while the *actor* represents the way we annotate it on the narrative.

Table 1: Semantic roles on the sentence "he was meant to be a surgeon"

word	**tag**	**actor**
he	B-ARG1	T1
was	B-V	EVENT
meant	B-V	EVENT
to	B-ARG1	EVENT
be	B-V	EVENT
a	B-ARG2	T2
surgeon	I-ARG1	T2

Secondly, we decide that the subject should have a specific relation category, therefore all the "ARG0" tags represent an "AGENT" relation. Finally, we also map the modifier arguments ("ARGM") into specific categories. For example "extension" arguments are mapped generically as "theme" relations and "direction" modifiers are mapped as "path" relations.

In conclusion, defining these rules allows the framework to fully extract a narrative from a set of tweets, by annotating it using the brat style. The end product is a ".ann" file with a structure similar to the one on figure 1, but scaled for the whole set of tweets.

4 Analysis of results

To have a comprehensive analysis of the Tweet2Story framework, we measured its results in two different ways: as an Open Information Extraction (OpenIE) tool and as a narrative extractor.

First, as an OpenIE tool, we aim to evaluate the effectiveness of the Tweet2Story framework in the task of extracting entity relations (triples). For this purpose, we test the framework against state-of-the-art tools for OpenIE (e.g. ClausIE) using CaRB [14] - a Crowdsourced Benchmark for OpenIE, which uses a standard OpenIE task and a gold benchmark corpus of triples.

As an OpenIE evaluator, CaRB measures a precision-recall curve, as shown by figure 2. According to the authors, CaRB penalizes tools that extract noisy triples (precision), while benefiting tools that extract triples with relevant information, even if that information is spread across multiple gold triples (recall).

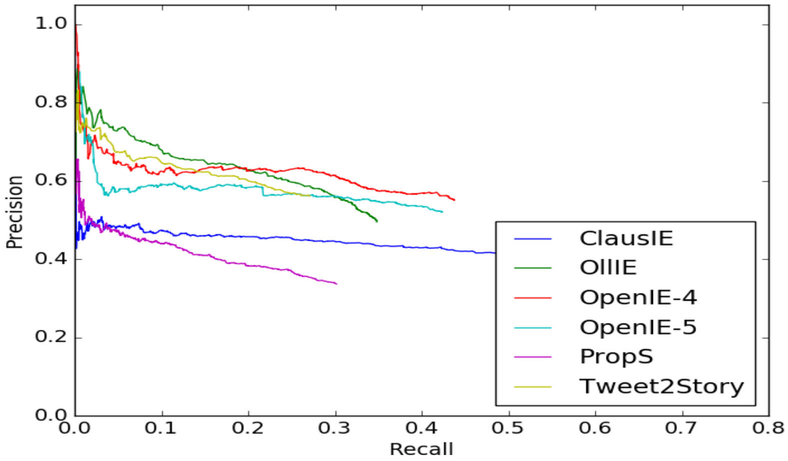


Fig. 2: Tweet2Story vs. state-of-the-art - CaRB evaluation framework

Both figure 2 and table 2 show that Tweet2Story does fairly well against state-of-the-art OpenIE models, despite not being made for OpenIE specifically. In particular, table 2 shows that Tweet2Story has the best precision out of all models, but dips on the recall. This shows that we extract simple triples without noise, but that those triples can sometimes overlook important information. This is on par with how our framework works, since its main purpose is to extract narratives, the triples never have repeated entities. This means that each entity (not including co-references) is only allowed to have one relation (triple), which causes the loss of some information, lowering the recall.

On the other hand, Tweet2Story is meant to be a narrative extraction framework and figure 3 shows an example of a narrative it extracted from a set of tweets about the Grace storm of 2015. Here, we can see the different topics that

Table 2: Results for the optimal threshold with the CaRB benchmark

System	Precision	Recall	F1	AUC
Ollie	0.505	0.346	0.411	0.224
PropS	0.34	0.3	0.319	0.126
OpenIE4	0.553	0.437	**0.488**	**0.272**
OpenIE5	0.521	0.424	0.467	0.245
ClausIE	0.411	**0.496**	0.45	0.224
Tweet2Story	**0.561**	0.271	0.365	0.211

were discussed, such as the speed of the storm, place it might or might not pass through and even predictions about its dissipation.

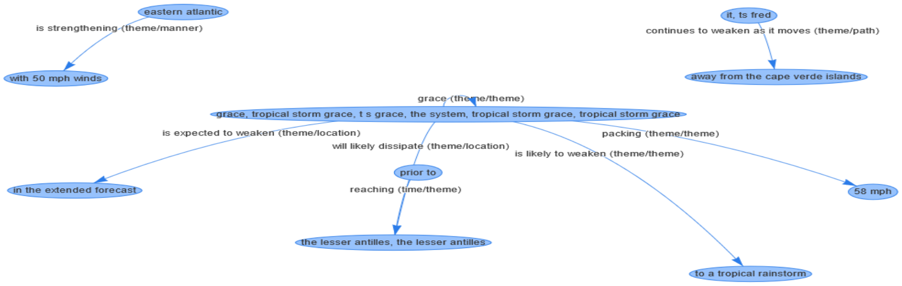

eastern atlantic
is strengthening (theme/manner)
with 50 mph winds
it, ts fred
continues to weaken as it moves (theme/path)
away from the cape verde islands
grace (theme/theme)
grace, tropical storm grace, t s grace, the system, tropical storm grace, tropical storm grace
is expected to weaken (theme/location)
will likely dissipate (theme/location)
packing (theme/theme)
is likely to weaken (theme/theme)
in the extended forecast
prior to
reaching (time/theme)
58 mph
the lesser antilles, the lesser antilles
to a tropical rainstorm

Fig. 3: Knowledge Graph made with annotations from Tweet2Story

All in all, figure 3 shows that Tweet2Story is capable of extracting a narrative from a set of tweets, which is its main goal. But can the tweets reconstruct a news article? The short answer is it depends on whether or not the tweets have enough and accurate information. If we compare figure 3 with figure 4 we can see that the tweets only partially reconstruct the news article. However, most part of the tweets narratives is complementary to the news article and a journalist could possibly use this narrative to understand the public opinion about the event.

5 Conclusion

Despite having room for improvement, we showed the potential of Tweet2Story has a narrative extraction framework. To the best of our knowledge, this is the first framework to extract narratives from small sets of texts, rather than long documents, while also focusing on the relations between entities (triples).

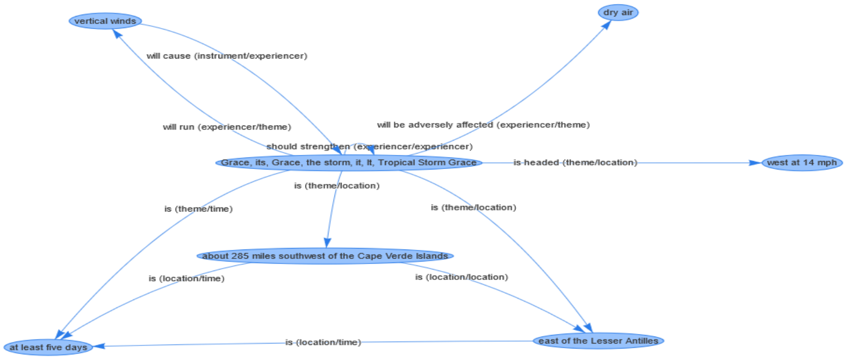

vertical winds
dry air
will cause (instrument/experiencer)
will run (experiencer/theme)
will be adversely affected (experiencer/theme)
should strengthen (experiencer/experiencer)
Grace, its, Grace, the storm, it, It, Tropical Storm Grace
is headed (theme/location)
west at 14 mph
is (theme/location)
is (theme/time)
is (theme/location)
about 285 miles southwest of the Cape Verde Islands
is (location/time)
is (location/location)
at least five days
is (location/time)
east of the Lesser Antilles

Fig. 4: News article Knowledge Graph from annotations made by an expert linguist

The main contributions of this paper are: (i) a gold annotation dataset[6] with 48 news articles annotated in brat format by an expert linguist, which we used to evaluate Tweet2Story; (ii) A framework (Tweet2Story) that extracts the narrative of any short document into a brat style annotation file (".ann"); (iii) A study in which tweets are shown as complementary to the news and where their narratives are seen as helpful in describing public opinion.

As mentioned before, we believe Tweet2Story still has a lot of room for improvement. Retrieving only one triple per verb (event), can sometimes lead to overlooking parts of the narrative. Sometimes, sentence arguments can be neglected simply by not being next to a verb, namely modifier arguments. One future direction could be to start looking at propositions as an extension of an event.

References

1. M. Robb, "Teens and the news: The influencers, celebrities, and platforms they say matter most, 2020," 2020. [Online]. Available: https://www.commonsensemedia.org/research/teens-and-the-news-the-influencers-celebrities-and-platforms-they-say-matter-most-2020
2. MuckRack, "The state of journalism 2021," MUCK RACK Blog, Mar. 15 2021. [Online]. Available: https://muckrack.com/blog/2021/03/15/state-of-journalism-2021
3. Y. Lee, S. Yoon, and K. Jung, "Comparative studies of detecting abusive language on twitter," *CoRR*, vol. abs/1808.10245, 2018. [Online]. Available: http://arxiv.org/abs/1808.10245
4. L. Marujo, W. Ling, I. Trancoso, C. Dyer, A. W. Black, A. Gershman, D. Martins de Matos, J. Neto, and J. Carbonell, "Automatic keyword extraction on Twitter," in *Proceedings of the 53rd Annual Meeting of the Association for*

[6] https://github.com/LIAAD/Tweet2Story

Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 637–643. [Online]. Available: https://aclanthology.org/P15-2105

5. J. Ma, W. Gao, and K.-F. Wong, "Rumor detection on Twitter with tree-structured recursive neural networks," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 1980–1989. [Online]. Available: https://www.aclweb.org/anthology/P18-1184
6. G. Katsios, S. Vakulenko, A. Krithara, and G. Paliouras, "Towards open domain event extraction from twitter: Revealing entity relations," in *DeRiVE@ESWC*, 2015.
7. M. D., B. V., and M. C., "Steps towards a system to extract formal narratives from text," in *Text2Story 2019 - Second Workshop on Narrative Extraction From Texts, pp. 53–61, Cologne, Germany, 14 April 2019*. CEUR-WS.org, Aachen, DEU, 2019.
8. J. D. Eisenberg, "Automatic extraction of narrative structure from long form text," Ph.D. dissertation, Florida International University, 2018.
9. J. V. Vargas, "Narrative information extraction with non-linear natural language processing pipelines," Ph.D. dissertation, Drexel University, 2017.
10. A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, Jul. 2011, pp. 1535–1545. [Online]. Available: https://aclanthology.org/D11-1142
11. L. Del Corro and R. Gemulla, "Clausie: Clause-based open information extraction," in *Proceedings of the 22nd International Conference on World Wide Web*, ser. WWW '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 355–366. [Online]. Available: https://doi.org/10.1145/2488388.2488420
12. G. Angeli, M. J. Johnson Premkumar, and C. D. Manning, "Leveraging linguistic structure for open domain information extraction," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 344–354. [Online]. Available: https://aclanthology.org/P15-1034
13. J. Dutkiewicz, M. Nowak, and C. Jedrzejek, "R2e: Rule-based event extractor," in *Challenge+DC@RuleML*, 2014.
14. S. Bhardwaj, S. Aggarwal, and M. Mausam, "CaRB: A crowdsourced benchmark for open IE," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6263–6268. [Online]. Available: https://www.aclweb.org/anthology/D19-1651