StoryLens: A Multiple Views Corpus for Location and Event Detection

Adrian M.P. Braşoveanu
Swiss Institute for Information Research
University of Applied Sciences Chur
Chur, Switzerland
adrian.brasoveanu@htwchur.ch

Lyndon J.B. Nixon
MODUL Technology GmbH
Vienna, Austria
lyndon.nixon@modultech.eu

Albert Weichselbraun
Swiss Institute for Information Research
University of Applied Sciences Chur
Chur, Switzerland
albert.weichselbraun@htwchur.ch

ABSTRACT

The news media landscape tends to focus on long-running narratives. Correctly processing new information, therefore, requires considering multiple lenses when analyzing media content. Traditionally it would have been considered sufficient to extract the topics or entities contained in a text in order to classify it, but today it is important to also look at more sophisticated annotations related to fine-grained geolocation, events, stories and the relations between them. In order to leverage such lenses we propose a new corpus that offers a diverse set of annotations over texts collected from multiple media sources. We also showcase the framework used for creating the corpus, as well as how the information from the various lenses can be used in order to support different use cases in the EU project InVID for verifying the veracity of online video.

CCS CONCEPTS

• **Information systems** → Incomplete data; Inconsistent data; Extraction, transformation and loading; Data cleaning; Entity resolution; • **Computing methodologies** → Information extraction;

KEYWORDS

Corpus, Named Entity Linking, Geosemantics, Event detection, Information Extraction, Natural Language Processing, Fake news

1 INTRODUCTION

Context should be key when consuming news, regardless of the source, but news media monitoring is full of noise due to massive retweets, rethreads, biases and fake news. Media outlets are biased towards longer narratives, therefore stories can be followed for months or years. While early tweets related to a news event might origin from a particular location, once the event is broadcast, mentions will emerge all over the place, making it difficult to rely on automatically generated location metadata for identifying eyewitnesses or real footage from the scene. News stories need to be understood as a set of events with different actors, locations and relations between them in order to be correctly interpreted, therefore techniques through which to provide additional relational and contextual information are needed.

One of the methods that can be used to improve the NEL processes for detecting locations, events and stories, as well as the relations between them, is the generation of new lenses from existing data in order to test various settings. For example, annotations that take into account entity types, overlaps, stories or even the differences in style and content between media sources (e.g. tweets are shorter and full of abbreviations compared to longer textual news articles), can easily be generated from existing NEL annotation sets by adding some processing rules (e.g., include or do not include overlaps or abbreviations). We call a corpus created through such a method a Multiple Views (or multiple lenses) corpus, as we consider its creation process to be somewhat similar to the one used by photographers when trying lenses with various ranges or focal distances in order to find the right one. This paper presents the process through which such a corpus called StoryLens[1] was created, as well as the various tasks or use cases that can be later performed in order to improve the automated news media monitoring processes.

The remainder of the paper presents the framework used for building the corpus (Section 2), the corpus (Section 3) and several use cases (Section 4). Section 5 discusses related work and Section 6 focuses on the lessons learned.

2 CORPUS CONSTRUCTION FRAMEWORK

In order to build a multiple lenses corpus, a new Python framework has been designed. The framework is highly flexible and contains three components as showcased in Figure 1: (i) *Annotations* - for documents selection, annotation extraction, links mining or clustering; (ii) *Lenses* - focused on creating new types of lenses (e.g., links between the same entities or between events and stories); and (iii) *Evaluation* - for providing statistics on the content of the corpus, as well as on the performance of various NEL tools on this corpus.

Corpus Creation and Annotation Process. During the implementation of the automated processes for differentiating real from fake news, the HORIZON 2020 research project InVID [2] has collected a large number of documents from different media types, including news media and social media (e.g., tweets, YouTube subtitles) with the purpose of evaluating geolocation detection.

A package to select relevant documents in order to create annotations was soon created. The extracted documents were split into three partitions based on the content's provenance: i) news articles; ii) subtitles; iii) tweets. The initial plan was to simply evaluate the three location types included in the TAC-KBP [4] challenges: GPE - Geo-Political Entities (e.g., countries, regions, cities), LOC - Natural Locations (e.g., mountains, rivers, lakes) and FAC - Facilities (e.g., buildings, infrastructure). Due to the serialized nature of news media, there was a need to correctly repair the various errors caused by the location entities embedded in the other entities (e.g., *Grenfell Tower Fire* event title embeds the *Grenfell Tower* FAC entity). As a first step we decided to expand the number of entity types included

[1] https://github.com/modultechnology/storylens
[2] http://www.invid-project.eu/

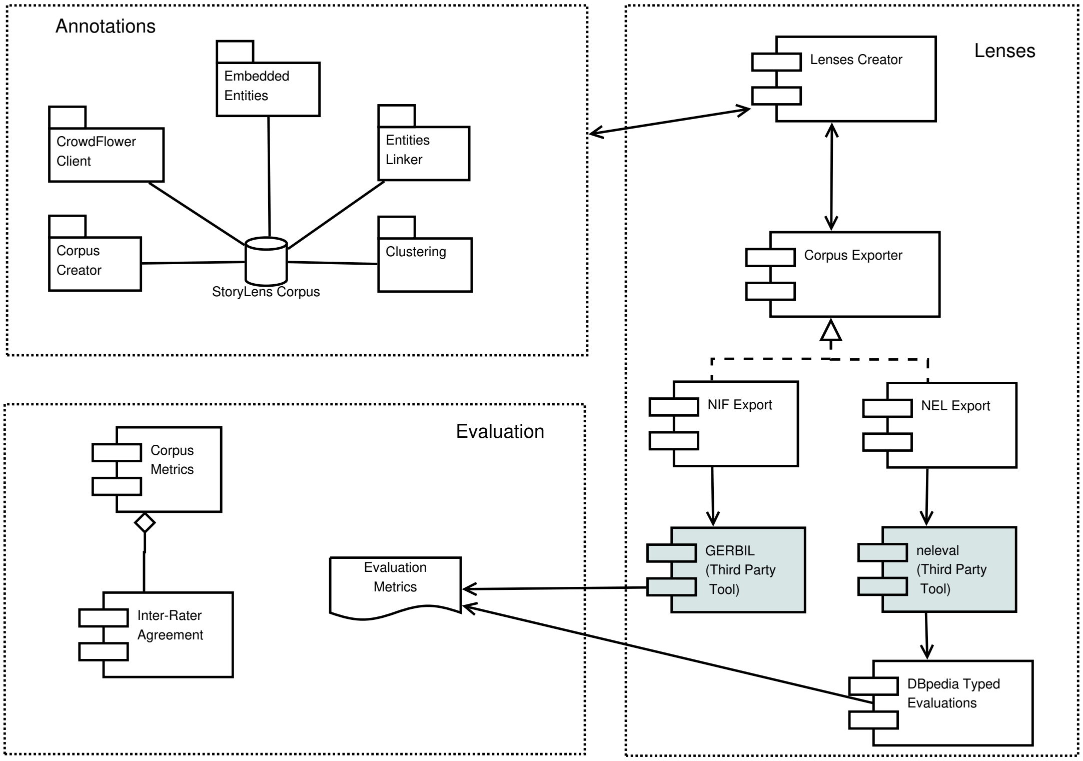

Figure 1: Framework used for corpus construction.

in the corpus and also added PER - Person, ORG - Organisation, EVENT - Event. In order to make the annotations work on several levels (e.g., event or story levels) we decided to add new annotation sets in order to account for correctly nested entities (also known as embedded entities), differentiate between events or stories, or understand the links between entities, therefore building multiple views over the same texts. The annotation rules were included in an Annotation Guideline that was similar to the ones used for TAC-KBP, ACE and related semantic evaluation challenges. For each class we provide a set of examples and a set of rules in order to guide the annotators. The annotators were asked to completely disregard any embedded (nested) entities. The embedded entities were later added in a separate lens automatically by merging all capitalized entities whenever other capitalized entities were found immediately before or after an entity when no punctuation sign was detected. Each document was annotated at least twice: once by a person from our group and once by an external person (via Crowdflower [3]) in order to obtain a somewhat balanced view over the dataset. In both cases the same judge went through all the documents. The annotators were asked to provide information about the surface forms (the text that represents the entity) and entity types. The online Wikipedia version was used for providing the entity links in order to ease the task. The Wikipedia links were later transformed into DBpedia links via SPARQL queries. Where such links were missing, we have constructed them, but added an additional comment in order to explain this provenance. The unlinked entities (usually called NIL) were grouped together using the Clustering package.

While usually in production one would prefer to only use one set of annotations (e.g. with overlaps instead of both with and without overlaps), we felt that such an approach was worth taking especially due to the fact that we included both shorter and longer texts.

Also since stories were considered to be typically composed of sets of related events, an annotation set was dedicated to story annotations. Besides such annotation sets, all the links between the various entities were extracted in order to validate the correctness of the annotation process.

Lenses Creation. In order to define a new lens it is generally required to implement a new package that defines its functionality. Several lenses are described in the next section of the paper. The output of the various lenses is typically exported to the NIF [5], CSV or TAC-KBP [4] formats. Since NIF does not support multiple annotation sets, we generally provide different NIF outputs for the various sets (e.g., with or without embedded entities). For typed evaluations the output only contains the matching entity types (e.g., for geolocation only GPE, FAC and LOC types are taken into account), whereas for TAC-KBP evaluations it is restricted to the five classic types that are typically required (e.g., PER, ORG, GPE, LOC, FAC). A different output file contains all the higher level annotations (e.g., stories). The output for the Twitter partition of the corpus only contains the annotations due to copyright restrictions, but the actual texts of the tweets can be downloaded by ids using

[3] https://www.crowdflower.com/

Type	News	Subs	Tweets
Documents	100	100	300
Entities	975	899	776
Relations	5034	3760	3293

Table 1: Basic Statistics.

free scripts[4]. The annotators have only collected links to the DBpedia [6] Knowledge Base (KB) entries that represent the respective entities, as DBpedia has emerged as the dominant KB for Named Entity Linking (NEL) evaluations.

Agreement. Agreement scores were computed after problematic cases were solved. Only documents with high agreement scores were kept. For each new annotation set our guidelines recommend computing agreements if multiple annotators were involved.

Evaluation. Corpus statistics are computed by Python scripts, whereas agreement is computed with an R script. Evaluation metrics can be computed via Gerbil or TAC-KBP neleval scripts, our metrics package providing a wrapper for these services.

3 STORYLENS CORPUS DESCRIPTION

Some basic statistics can be found in Table 1. A higher number of tweets was included in the corpus simply due to the fact that they are shorter than the other types of documents. As it can be seen from the basic statistics table, the number of entities and relations is balanced across the partitions. There seem to be more relations between the entities from news articles than from the other partitions. The texts were selected from the documents collected by the InVID news monitoring dashboard. Entities were annotated via a manual process, whereas the relational annotations (e.g., links between events or stories, links between entities) were automatically created using a Python script that selected all DBpedia relations between two entities. The unlinked entities were grouped together using a basic NIL Clustering algorithm built with the hierarchical clustering package from scikit learn.

Table 2 depicts some of the stories included in this corpus and the category they belong to. The naming of events and stories has been an issue, as often some narratives have an incidental name at the beginning, which due to lack of information is associated with the city in which the incident took place (e.g., Fire in London, West London Fire), whereas later it can morph into a more precise name that might include even the name of the street (e.g., Grenfell Tower Fire or simply Grenfell Tower). We generally kept the mainstream names in this corpus, but the various bits and pieces of early information are included here as well and can be a good start for anyone interested in how such stories develop over time and various media formats. We chose to stay away from story naming conventions, as we think that understanding how story names change in time is a complex research topic that needs to be developed on its own. The current set of of links between events and stories that has been extracted automatically can aid further research in this direction.

We have created several lenses in order to explore the texts contained in the corpus, as outlined in Table 3. The classic lens is called *Long* as it doesn't include any overlaps between the various entities (e.g., *Trump Hotel DC* would not be annotated two times to account for all the possible cases: *Trump*-PER, *Trump Hotel DC*-ORG), only the longest match for a surface form being considered. The lens that includes the overlaps is called *Embedded*. The *Stories* lens includes the set of documents related to a particular story as we also wanted to capture some information related to long-running narratives. The *Events* lens expands upon the previous lens and collects all the events related to a particular story, offering insights into the development of a story when the various texts are arranged chronologically. The *Links* lens collects all the links between the entities that appear together in a single text. A simplified version of it includes only the counts of relations between each two entities as represented by their links. These links lenses are particularly important for debugging disambiguation methods that rely heavily on KB links (e.g., graph-based algorithms), as one of the first things that needs to be verified when testing them is if such links exists.

4 USE CASES

Our main objective has been to create an initial *benchmark for event detection from different types of media*, as currently most benchmarks are focused on tweets. One major issue has been the lack of DBpedia links, as even though Wikipedia pages are updated daily the same cannot be said about the live DBpedia server yet. Typically DBpedia releases only come in six months increments. Table 2 which depicts the relations between categories and stories is particularly important for this use case.

The classic use case involves *assessing the performance of NEL tools* on this corpus using various settings (e.g., with or without embeddings, classic TAC-KBP entity types, full corpus versus partition performance). The main purpose of these experiments should be identifying requirements for Named Entity Linking tools which allow them to perform well across different text types (e.g., classic news articles, subtitles full of errors, noisy tweets). *Typed evaluations* focused on certain types (e.g., location types like LOC, GPE and FAC) are also important today.

One important use case for providing multiple lenses is the existence of *multiple KBs*. Besides DBpedia, some popular KBs include Wikipedia, Wikidata [10] or LinkedGeoData [8]. As it can be seen from Table 3, lenses for additional KBs have been automatically added. In order to collect strict matches, only the *owl:sameAs* or *skos:exactMatch* have been extracted from the respective KBs. If needed such lenses can also be used in order to provide quick fixes for production for various Knowledge Base and dataset issues [1].

5 RELATED WORK

Jeff Dean argues that the same corpus should be used for solving multiple problems [3] since creating training data is an expensive process. While the practice of creating multiple annotation sets for solving different problems is available in many annotation tools (e.g., GATE [2]), it is rarely used in the field of Named Entity Linking. MEANTIME [7] uses a collection of 120 annotated English texts split into four partitions (Airbus, Apple, GM, Stock) and provides translations into several languages (e.g., Spanish, Italian, Dutch). Since it uses the idea of markables (document-level annotation for a specific task) for events and time annotations and some of these annotations are relational in nature, MEANTIME can be considered

[4] https://gist.github.com/giacbrd/b996cfe2f1d24752f23bd119fdd678f2 - for example

Category	Stories
Politics	Obamacare Repeal, Philando Castile Shooting, Brexit, London Terrorist Attack (May 2017)
Entertainment	Bill Cosby Scandal, Coachella, Wonder Woman
Disaster	Grenfell Tower Fire, USS Fitzgerald Collision, Orlando Pulse Night Club Shooting, WannaCry
Business	Amazon Buys Whole Foods, ExxonMobil XTO Denver Office Closing
Sports	UEFA Champions League Final, Europa League Final, Roland Garros, Wimbledon

Table 2: Example stories included in the corpus.

Category	Description
Long	Longest match without overlaps
Embedded	Entities including overlaps
Wikipedia	Entities linked to Wikipedia
Wikidata	Entities linked to Wikidata
Stories	Stories
Events	Links between events and stories
Links	Links between same text entities

Table 3: Examples of lenses included in the dataset. If not mentioned otherwise (e.g., Wikipedia, Wikidata), DBpedia entity links are provided in the respective lens.

an early example of a multiple lenses corpus. It has to be noted that simply partitioning a corpus or providing multiple annotation sets is not enough in order to consider the respective corpus as a multiple lenses corpus.

In our view the term *Multiple Views* should be used only if there is enough evidence to support that using selected markables or annotation sets, alone or in various combinations, enables researchers to solve new problems. Modern scorers (e.g., Gerbil [9]) allow the use of various evalution types, therefore we offer a NIF version for integration with such tools.

6 CONCLUSION

Providing multiple lenses on the same events helps disambiguate ambiguous entities and sheds light on different media biases. Additional value can be obtained through methods of creating new gold standards from existing ones. Lenses can be used to track the improvements of various KBs in time, especially if the resources are properly versioned (e.g., if each attribute's provenance and last edited properties can be added to track changes), as new attributes discovered through lenses could later be added to KB entities. Another important use case for lenses is for providing links to multiple KBs (e.g., Wikidata, Geonames, LinkedGeoData). Future work will be focused on the use cases described in Section 4, as the corpus will be integrated in the evaluation cycle of the InVID project. We plan to continue publishing new partitions and updates to this corpus.

7 ACKNOWLEDGEMENTS

The research presented here has been conducted as part of the DISCOVER Project (www.htwchur.ch/discover), funded by the Swiss Commission for Technology and Innovation (CTI) and the InVID project funded by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 687786.

REFERENCES

[1] Adrian M.P. Braşoveanu, Giuseppe Rizzo, Philipp Kuntschick, Albert Weichselbraun, and Lyndon J.B. Nixon. 2018. Framing Named Entity Linking Error Types. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (7-12), Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (Eds.). European Language Resources Association (ELRA), Paris, France, 266–271. http://www.lrec-conf.org/proceedings/lrec2018/summaries/612.html

[2] Hamish Cunningham, Diana Maynard, and Kalina Bontcheva. 2011. *Text Processing with GATE.* Gateway Press CA, The University of Sheffield, Department of Computer Science. https://gate.ac.uk/releases/gate-6.1-build3913-ALL/tao.pdf

[3] Jeffrey Dean. 2016. Large-Scale Deep Learning For Building Intelligent Computer Systems. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, February 22-25, 2016*, Paul N. Bennett, Vanja Josifovski, Jennifer Neville, and Filip Radlinski (Eds.). ACM, San Francisco, CA, USA, 1. https://doi.org/10.1145/2835776.2835844

[4] Ben Hachey, Joel Nothman, and Will Radford. 2014. Cheap and easy entity evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers.* ACL, Baltimore, MD, USA, 464–469. https://doi.org/10.3115/v1/P14-2076

[5] Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating NLP Using Linked Data. In *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II*, Harith Alani, Lalana Kagal, Achille Fokoue, Paul T. Groth, Chris Biemann, Josiane Xavier Parreira, Lora Aroyo, Natasha F. Noy, Chris Welty, and Krzysztof Janowicz (Eds.). Lecture Notes in Computer Science, Vol. 8219. Springer, Berlin, Germany, 98–113. https://doi.org/10.1007/978-3-642-41338-4_7

[6] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal* 6, 2 (2015), 103–104. https://doi.org/10.3233/SW-140134

[7] Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. MEANTIME, the NewsReader Multilingual Event and Time Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis (Eds.). ELRA, Paris, France, 4417–4422. http://www.lrec-conf.org/proceedings/lrec2016/summaries/488.html

[8] Claus Stadler, Jens Lehmann, Konrad Höffner, and Sören Auer. 2012. LinkedGeoData: A core for a web of spatial open data. *Semantic Web* 3, 4 (2012), 333–354. https://doi.org/10.3233/SW-2011-0052

[9] Ricardo Usbeck, Michael Röder, Axel-Cyrille Ngonga Ngomo, Ciro Baron, Andreas Both, Martin Brümmer, Diego Ceccarelli, Marco Cornolti, Didier Cherix, Bernd Eickmann, Paolo Ferragina, Christiane Lemke, Andrea Moro, Roberto Navigli, Francesco Piccinno, Giuseppe Rizzo, Harald Sack, René Speck, Raphaël Troncy, Jörg Waitelonis, and Lars Wesemann. 2015. GERBIL: General Entity Annotator Benchmarking Framework. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, Aldo Gangemi, Stefano Leonardi, and Alessandro Panconesi (Eds.). ACM, Florence, Italy, 1133–1143. https://doi.org/10.1145/2736277.2741626

[10] Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85. https://doi.org/10.1145/2629489