

Brat2Viz: a Tool and Pipeline for Visualizing Narratives from Annotated Texts

Evelin Amorim INESC TEC FCUP-Universidade do Porto evelin.f.amorim@inestec.pt	Alexandre Ribeiro INESC TEC FCUP-Universidade do Porto alexandre.m.ribeiro@inestec.pt	
Inês Cantante INESC TEC FLUP cantante.ines@gmail.com	Alípio Jorge INESC TEC FCUP-Universidade do Porto alipio.jorge@inestec.pt	Brenda Santana Instituto de Informática UFRGS bssantana@inf.ufrgs.br
Sérgio Nunes INESC TEC sergio.nunes@inesctec.pt	Purificação Silvano FLUP CLUP-Universidade do Porto msilvano@letras.up.pt	
António Leal FLUP CLUP-Universidade do Porto jleal@letras.up.pt	Ricardo Campos INESC TEC Polytechnic Institute of Tomar Ci2 - Smart Cities Research Center ricardo.campos@ipt.pt	

Abstract

Narrative Extraction from text is a complex task that starts by identifying a set of narrative elements (actors, events, times), and the semantic links between them (temporal, referential, semantic roles). The outcome is a structure or set of structures which can then be represented graphically, thus opening room for further and alternative exploration of the plot. Such visualization can also be useful during the on-going annotation process. Manual annotation of narratives can be a complex effort and the possibility offered by the Brat annotation tool of annotating directly on the text does not seem sufficiently helpful. In this paper, we propose Brat2Viz, a tool and a pipeline that displays visualization of narrative information annotated in Brat. Brat2Viz reads the annotation file of Brat, produces an intermediate representation in the declarative language DRS (Discourse Representation Structure), and from this obtains the visualization. Currently, we make available two visualization schemes: MSC (Message Sequence Chart) and Knowledge Graphs. The modularity of the pipeline enables the future extension to new annotation sources, different annotation schemes, and alternative visualizations or representations. We illustrate the pipeline using

Copyright © by the paper's authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Finlayson (eds.): Proceedings of the Text2Story'21 Workshop, Online, 1-April-2021, published at <http://ceur-ws.org>

1 Introduction

Narratives texts are often characterized by narrative sequences with features such as the chronological succession of events, causality relations between the events, and the presence of one or more protagonists who suffer a process of transformation throughout the story [Ada92]. These features make narratives both appealing and useful, namely, to aid humans to communicate complex concepts, ideas, realities. Given the huge body of texts containing narratives in cultural heritage and their continuous production, there is a pressing demand for applying and developing Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques for extracting narratives from texts. Therefore, several automatic extraction techniques have been proposed so far [AB02, CJJ20, KBB⁺16, KBM12, MCH⁺16, SPB19, TFNT17, Tou18], the big majority of which require annotated corpora. However, the task of corpus annotation presents some challenges and involves some difficulties.

First, one needs to take into account the specific goals of the project to determine the type of annotation to be made. After choosing the appropriate semantic annotation framework, the process of adapting it to the specificity of the target language can be time-consuming and problematic. Second, a multilayer annotation requires a comprehensive analysis of the annotation framework, which may imply a non-trivial simplification of the tags, attributes, and links, to avoid, for instance, overlapping information and overloaded annotation. At the level of textual analysis, deciding what is relevant to represent from the story may also be troublesome. Finally, there is a lack of suitable tools to label and inspect annotation. Existing annotation tools, like Brat (brat rapid annotation tool) [SPT⁺12] and Prodigy¹, provide multi-purpose interfaces. Despite friendly, they are insufficient for verification of complex annotations.

Using more advanced visualization for annotated narratives has only recently been considered by the research community [HRS⁺20, PPP⁺19]. The visualization of the narratives in the annotation process would enable annotation debugging and therefore a reduction of the enormous human effort required to complete this task.

To tackle the challenge of enabling narrative visualization from annotated texts, we propose the Brat2Viz framework, which complements Brat. Brat2Viz is capable of presenting a narrative visualization from annotated text and can be adapted to different annotation schemes. To deal with different annotation methodologies, we propose to employ the Discourse Representation Structure [GBM20] as a declarative, logic-based, intermediate language. The DRS representation can be converted to different human-readable representations: visual, textual, or other. In this work, we use two existing visualization schemes: MSC (Message Sequence Charts) [HT03] and Knowledge Graphs [EW16].

Our main contributions are as follows: (1) An extensible framework to generate visual representations of narratives from annotated texts; (2) The use of a formal logic-based language, DRS, as an intermediate language to the visual representation, which to the best of our knowledge has not been considered before; (3) A demonstration of our pipeline on a narrative corpus of news stories.

The following sections detail our research.

2 Related Work

One well-known narrative labeled dataset is the ROCStories [MRL⁺17], whose main goals are script learning and story understanding. The proposed task is to choose between two possible ends for each stored story. The kind of labeling adopted by ROCStories, though, disguises the complexity of narratives since their elements remain unidentified. A type of formal representation that is flexible to embody narrative elements and can be used to aid the labeling process are ontologies [AB02, DKK12, EOS⁺14, KBB⁺16]. One advantage of using ontologies is that after they are built, they can be employed in reasoning systems to produce knowledge. Another advantage is the flexibility to represent several aspects of a more complex narrative. Declerck et al. [DKK12], for instance, created an ontology to represent the characters of the folktale “Magic Swan Geese”. Representing narratives using ontologies, nonetheless, demands a huge human effort. In addition to ontologies, other languages can be used to represent narratives, like the declarative language DECLARE [PSVdA07]. Using this language, it is possible to define events and rules that should be followed, which are used to validate the narrative described in the text. However, this language seems more suitable to script narratives. Although it does not embody

¹<https://prodi.gy>

an ontology, our labeling scheme is more specific than ROCStories concerning narrative components and takes advantage of the inference process provided by the ontology. We employ the Discourse Representation Structure (DRS) language [KR93, BBE⁺17]. Therefore, we simplified the process of labeling elements of the narrative, by introducing an intermediate representation language, which can also be used in a reasoning process to infer knowledge. As far as we know, nobody proposed the use of such an intermediate language in this context before.

The visualization of narratives is also studied in the fields of Communication Design and Journalism. Segel and Heer [SH10], for instance, collected visual representations, already built, from news portals, and then identified its main design features. Figueiras [Fig14] analyzed three case studies, also from news portals. The analysis was qualitative and used communication design guidelines. Our main goal, however, is to obtain a visual representation from a manually annotated corpus and employ well-known visual languages and Knowledge Graphs. This visualisation will present the narrative more schematically, allowing then the annotator to verify if the relevant narrative elements have been correctly labeled and if all the necessary links between them have been established.

Similar to our work, Palshikar et al. [PPP⁺19] proposed the adoption of MSC to represent narrative annotations. The authors employed non-supervised approaches for narrative extraction and then used MSC to illustrate the results. Since non-supervised approaches were applied, some simplifications were assumed regarding the narrative elements. Those experiments use corpora with historical narrative text and a Question and Answering dataset. Our work, instead, uses manually annotated data, and we do not consider simplifications regarding narrative elements. In addition to that, we apply our technique in a news story corpus. Hingmire et al. [HRS⁺20] also adopted the MSC to represent historical narrative in the Hindi language. However, the authors proposed some adaptations to deal with specificities of this particular language. Differently, our work proposes the use of DRS as an intermediate between a text in any language and a visual language to prevent possible linguistic ambiguities.

3 The Narrative Annotation Visualization Pipeline

The research presented here stems from our Text2Story project². In this project, we are currently annotating a corpus of news stories written in European Portuguese. A Narrative Annotation Visualization tool, Brat2Viz³, has been developed for supporting the debugging of narrative annotation done with the Brat annotation tool [SPT⁺12]. Brat2Viz implements a pipeline that transforms the annotation into a formal representation (DRS) and from this, to MSC [HT03] and Knowledge Graphs [EW16] visualizations. Next, we detail the annotation scheme and visualization module.

3.1 Annotation Scheme

The annotation of our corpus covered three levels: referential, temporal and semantic role labeling. The annotation scheme used in our project followed the semantic framework from ISO 24617-1/9 [fS07, fS19], for the first two levels, and from Linguistic InfRastructure for Interoperable ResourCes and Systems (LIRICS) for the last one [PB08, SBPSA07]. Some adaptations regarding, for instance, the number of tags and types of attributes were made due to the multilayer annotation and to some properties of the language (European Portuguese) and of the genre of the corpus (news) [Cañ11]. As such, our current annotation scheme has three types of tags: *Actors*, to annotate characters in the story (e.g., “um homem” – “a man”); *Events*, for events (e.g., “assaltou” – “robbed”); and *Timex3*, for temporal expressions. Each one of these tags has attributes (sub-tags) so that we have complete meanings for every component annotated. Figure 1 depicts these two types of tags in one sentence extracted from a news story of our corpus.

In addition to these three general annotation categories, we also have to ensure a proper linkage between the actors; between the events and the actors; and between the events, temporal references and other objects or locations. To accomplish this objective we use three types of links: TLINKS, REF_REL, OBJ_REL. Temporal links (TLINKS) account for the chronological ordering of the events. This type of link allows us to understand if one event happens before another, at the same time, after, etc. It is also used to represent temporal relation between events and temporal expressions. To indicate the relations established between actors, we use Referential Relations (REF_REL) and Objectal Relations (OBJ_REL). The former represents the lexical relations between linguistic units, such as synonymy, antonymy, hyponymy, etc. The latter represents relations between linguistic units, from a discourse point of view. Finally, we adapted, from the concepts of LIRICS, a list of Semantic Roles

²<https://text2story.inesctec.pt/>

³<https://nabu.dcc.fc.up.pt/brat2viz>

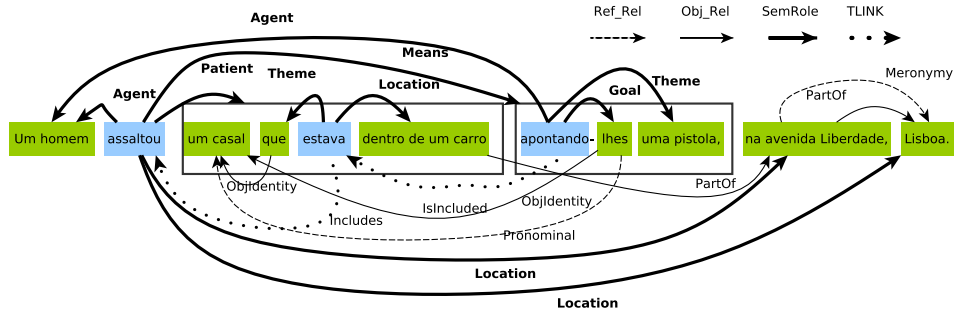


Figure 1: Example of a text labeled according to our annotation framework (“A man robbed a couple who were in a car pointing a gun towards to them, at Liberdade avenue, Lisbon.”).

```

T3 ACTOR 0 8 Um homem
A1 Lexical_Head T3 Noun
A2 Individuation T3 Individual
A3 Actor_Type T3 Per
A4 Involvement T3 1
T9 EVENT 9 17 assaltou
E1 EVENT:T9
A32 Class E1 Occurrence
A33 Tense E1 Past1
A34 Polarity E1 Pos
A35 Time E1 Past

```

```

event(a), eventClass(a,Occurrence), Tense(a,Past1),
Polarity(a,Pos), Time(a,Past), relationRole(agent,T3)

```

(b) Snippet of the DRS.

(a) Sample annotation lines.

Figure 2: Annotation lines (left) and DRS output (right) for the theft example.

that we considered essential, according to our project’s needs, to establish thematic relations. An illustrative representation of these links can be found in Figure 1.

3.2 Visualization

Brat2Viz⁴ consists of two main components: Brat2DRS and DRS2Viz. The Brat2DRS module takes the annotation files generated by Brat, parses them, and creates a DRS representation for each news story. Then, the DRS2Viz module takes as input the DRS representations generated by the previous component, and deploys a web application that produces visualizations of the original news text. In the following, we detail each one of these modules in a more specific manner.

3.2.1 The Brat2DRS module

Following the annotation step, the Brat2DRS module parses the “.ann” file, and builds a dictionary with the linguistic elements found, assigning a symbolic variable to each event. The annotations are then interpreted, and the analyzed structure is converted to a DRS provided by the NLTK library⁵. DRS statements are generated to each expression in a textual format that states the events’ properties, the actors and the time expressions, and relations between them. In Figure 2b we can observe a snippet that declares one event named ‘a’ and some attributes of the event.

This high-level abstract representation of the narrative can now be used by subsequent operations that do not need to go back to the original text of the “.ann” file. Operations may include visualizations, rewriting, and evaluation. Moreover, the DRS representation can be used for inference and reasoning related to this particular narrative. The parser that converts the annotated document into DRS statements is built based on a dictionary that contains the annotation tags. In this way, it is also possible to extend to different annotation formats by going through adjustments in the patterns related to the keys used, i.e., the annotated features and the pattern generated by them in the output “.ann” file.

⁴<https://github.com/LIAAD/brat2viz>

⁵<https://www.nltk.org/howto/drt.html>

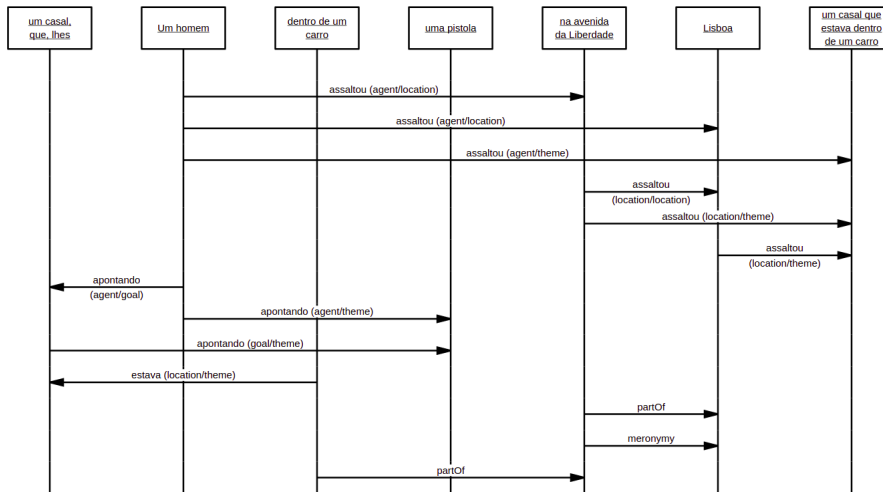


Figure 3: MSC representation from the robbery example.

3.2.2 The DRS2Viz module.

This module consists of a parser component and the visualization engine. It takes the DRS produced in the previous step and generates visualizations in the web browser. As referred above, our currently implemented visual outputs are MSC and Knowledge Graphs. In both visualizations, actors are represented as nodes, and events and relations are represented as links between these nodes. The parser uses the DRS and extracts actors, events and relations into independent data structures. Actors and events are represented in structures that keep track of their identifiers, and the lexical items that represent them in the news article. Using Figure 2a as an example, the actor is represented as $T1$: “Um homem” and the event is represented as $E1$: “assaltou”. Links occur between actors, and thus we have to transform relations between actors and events into relations between pairs of actors, while keeping the references to the events that originated such relations. We must also consider that each actor may be referred in the text through a variety of lexical items. To address redundant actors (e.g., synonymy, object identity and same head), we merge them into single actors while keeping all the lexical items that convey them. Next, we update references to merged actors in the events and relations structures. After parsing the DRS, we are able to generate the visual representations in the browser. The MSC visualization is generated using `mcsngen_js`⁶, a javascript library that renders message sequence charts from MSC strings. Fig 3 shows the MSC output generated from the thief news article example. The Knowledge Graph visual representation is created using `visjs`⁷, a javascript library used to build and display networks. Figure 4 shows the graph output generated from the thief news article example.

This module is extensible to support other annotation labels. However, the user should consider the set of labels used to link redundant actors, which in our case are “synonymy”, “object identity” and “same head”.

4 Conclusion

In this paper, we have described Brat2Viz, a tool for visualizing narratives from annotations produced in Brat. The tool implements a two-step modular pipeline that first transforms narrative annotations into the DRS formal language and visual representations. Currently, we visualize the narratives as MSC and as Knowledge Graphs. The modularity of the pipeline enables its extension and adaptation to other scenarios. Other visualizations from DRS input can be developed, such as timelines. Other types of representations can be used as well (e.g., simplified textual narratives). The annotation scheme can also be adapted to other needs. Narrative extraction algorithms may also be plugged as automatic annotators, resulting in a Narrative Extraction Visualization Pipeline.

Acknowledgments

This work has been carried out as part of the project Text2Story, financed by the ERDF – European Regional Development Fund through the North Portugal Regional Operational Programme (NORTE 2020), under the

⁶https://github.com/sverweij/mcsngen_js

⁷<https://github.com/visjs/vis-network>

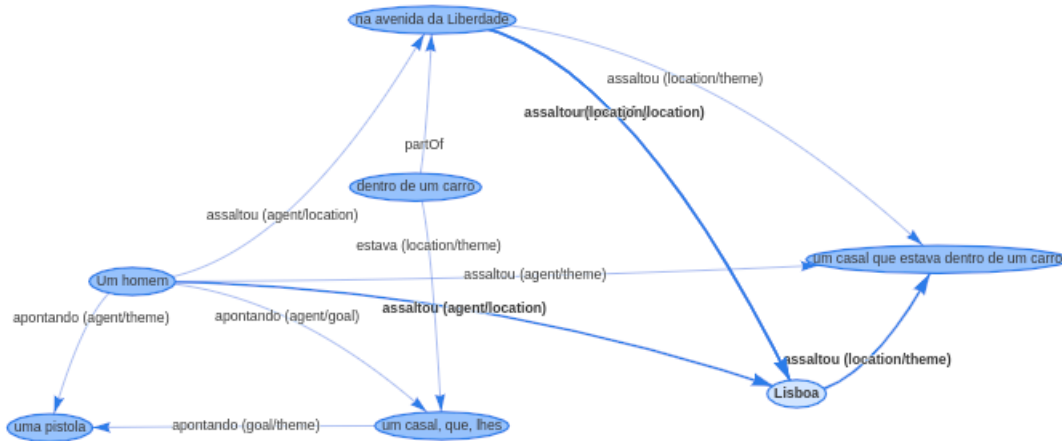


Figure 4: Knowledge Graph representation from the robbery example.

PORTUGAL 2020 and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project PTDC/CCI-COM/31857/2017 (NORTE-01-0145-FEDER-03185).

References

- [AB02] Daniela Alderuccio and Luciana Bordoni. An ontology-based approach in the literary research: two case-studies. In *LREC*, 2002.
- [Ada92] Jean-Michel Adams. Les textes: types et prototypes. *Recit, description, argumentation, explication et dialogue*. France: Nathan, 1992.
- [BBE⁺17] Johan Bos, Valerio Basile, Kilian Evang, Noortje J Venhuizen, and Johannes Bjerva. The groningen meaning bank. In *Handbook of linguistic annotation*, pages 463–496. Springer, 2017.
- [Cañ11] María Teresa Pisa Cañete. La construction discursive de l’ événement rapporté dans les textes des genres informatifs de la presse française. *Çedille. Revista de Estudios Franceses*, (7):272–305, 2011.
- [CJJB20] Ricardo Campos, Alípio Jorge, Adam Jatowt, and Sumit Bhatia. The 3rd International Workshop on Narrative Extraction from Texts: Text2Story 2020. In *European Conference on Information Retrieval*, pages 648–653. Springer, 2020.
- [DKK12] Thierry Declerck, Nikolina Koleva, and Hans-Ulrich Krieger. Ontology-based incremental annotation of characters in folktales. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 30–34, 2012.
- [EOS⁺14] Christian Eisenreich, Jana Ott, Tonio Süßdorf, Christian Willms, and Thierry Declerck. From Tale to Speech: Ontology-based Emotion and Dialogue Annotation of Fairy Tales with a TTS Output. In *International Semantic Web Conference (Posters & Demos)*, pages 153–156, 2014.
- [EW16] Lisa Ehrlinger and Wolfram Wöß. Towards a definition of knowledge graphs. In *SEMANTiCS (Posters, Demos, SuCCESS)*, 2016.
- [Fig14] Ana Figueiras. Narrative visualization: A case study of how to incorporate narrative elements in existing visualizations. In *2014 18th International Conference on Information Visualisation*, pages 46–52. IEEE, 2014.
- [fS07] International Organization for Standardization. ISO/WD 24617-1, Language resource management—semantic annotation framework (semaf)—part 1: Time and events, 2007.
- [fS19] International Organization for Standardization. ISO 24617-9, Language resource management—semantic annotation framework —part 9: Reference annotation framework (RAF), 2019.

- [GBM20] Bart Geurts, David I. Beaver, and Emar Maier. Discourse Representation Theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2020 edition, 2020.
- [HRS⁺20] Swapnil Hingmire, Nitin Ramrakhiyani, Avinash Kumar Singh, Sangameshwar Patil, Girish Palshikar, Pushpak Bhattacharyya, and Vasudeva Varma. Extracting Message Sequence Charts from Hindi Narrative Text. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 87–96, 2020.
- [HT03] David Harel and P. S. Thiagarajan. *Message Sequence Charts*, page 77–105. Kluwer Academic Publishers, USA, 2003.
- [KBB⁺16] Anas Fahad Khan, Andrea Bellandi, Giulia Benotto, Francesca Frontini, Emiliano Giovannetti, and Marianne Reboul. Leveraging a Narrative Ontology to Query a Literary Text. In *7th Workshop on Computational Models of Narrative (CMN 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.
- [KBM12] Oleksandr Kolomiyets, Steven Bethard, and Marie Francine Moens. Extracting narrative timelines as temporal dependency structures. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 88–97, 2012.
- [KR93] Hans Kamp and Uwe Reyle. *Introduction to Model Theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*, volume 42. Springer Netherlands, 1993.
- [MCH⁺16] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, 2016.
- [MRL⁺17] Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51, 2017.
- [PB08] Volha Petukhova and Harry Bunt. LIRICS Semantic Role Annotation: Design and Evaluation of a Set of Data Categories. In *LREC*. Citeseer, 2008.
- [PPP⁺19] Girish Palshikar, Sachin Pawar, Sangameshwar Patil, Swapnil Hingmire, Nitin Ramrakhiyani, Harshimran Bedi, Pushpak Bhattacharyya, and Vasudeva Varma. Extraction of Message Sequence Charts from Narrative History Text. In *Proceedings of the First Workshop on Narrative Understanding*, pages 28–36, 2019.
- [PSVdA07] Maja Pesic, Helen Schonenberg, and Wil MP Van der Aalst. Declare: Full support for loosely-structured processes. In *11th IEEE international enterprise distributed object computing conference (EDOC 2007)*, pages 287–287. IEEE, 2007.
- [SBPSA07] A. Schiffrin, H. Bunt, V. Petukhova, and Susanne Salmon-Al. LIRICS Deliverable D4. 3. Documented compilation of semantic data categories, 2007.
- [SH10] Edward Segel and Jeffrey Heer. Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics*, 16(6):1139–1148, 2010.
- [SPB19] Matthew Sims, Jong Ho Park, and David Bamman. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, 2019.
- [SPT⁺12] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, 2012.

- [TFNT17] Julien Tourille, Olivier Ferret, Aurelie Neveol, and Xavier Tannier. Neural architecture for temporal relation extraction: A bi-lstm approach for detecting narrative containers. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 224–230, 2017.
- [Tou18] Julien Tourille. *Extracting Clinical Event Timelines: Temporal Information Extraction and Coreference Resolution in Electronic Health Records*. PhD thesis, Université Paris-Saclay, 2018.