Guidelines and a Corpus for Extracting Biographical Events

Guidelines and a Corpus for Extracting Biographical Events

Marco A. Stranisci*, Enrico Mensa*, Ousmane Diakite*, Daniele P. Radicioni*, Rossana Damiano*

*Department of Computer Science - University of Turin, Turin, Italy

{marcoantonio.stranisci, enrico.mensa, daniele.radicioni, rossana.damiano}@unito.it

ousmane.diakite@edu.unito.it

Abstract

Despite biographies are widely spread within the Semantic Web, resources and approaches to automatically extract biographical events are limited. Such limitation reduces the amount of structured, machine-readable biographical information, especially about people belonging to underrepresented groups. Our work challenges this limitation by providing a set of guidelines for the semantic annotation of life events. The guidelines are designed to be interoperable with existing ISO-standards for semantic annotation: ISO-TimeML (SO-24617-1), and SemAF (ISO-24617-4). Guidelines were tested through an annotation task of Wikipedia biographies of underrepresented writers, namely authors born in non-Western countries, migrants, or belonging to ethnic minorities. $1,000$ sentences were annotated by 4 annotators with an average Inter-Annotator Agreement of 0.825. The resulting corpus was mapped on OntoNotes. Such mapping allowed to to expand our corpus, showing that already existing resources may be exploited for the biographical event extraction task.

Keywords: Event Extraction, Semantic Annotation, Interoperability

1. Introduction

The Semantic Web shift led in few years to a growth of biographical information online. Knowledge Graphs (KG), such as Dbpedia (Auer et al., 2007) and Wikidata (Vrandečić and Krötzsch, 2014), allow the gathering of structured socio-demographic attributes and facts about people. Notwithstanding, many unstructured data conveying biographical information are still not mapped in KGs. Wikipedia pages express more content than their corresponding Wikidata profile: for instance, all the places where a person lived within their life and all their migrations. The enrichment of existing KGs with such information would be crucial in improving several tasks such as community detection (Wang et al., 2018), prosopography (Booth, 2008), and social bias detection (Sun and Peng, 2021).

Although several semantic models have been proposed to formally represent a biographical event (Krieger and Declerck, 2015; Tuominen et al., 2018), computational resources for the automatic extraction of biographical events from text are still missing, and there are no annotated corpora, nor annotation schemes specifically designed for this task.

In this paper, we describe a novel set of annotation guidelines specifically developed for this task, built on two Semantic Annotation Frameworks, ISO 24617-1 (Pustejovsky et al., 2010), and ISO 24617-4 (Bunt and Palmer, 2013). The guidelines have been adopted to annotate a corpus of $1,000$ sentences extracted from Wikipedia pages of under-represented writers, namely writers born in non-Western countries, migrants or belonging to ethnic minorities (Stranisci et al., 2021b). The resource is designed to be interoperable with existing language resources (Pustejovsky et al., 2003; Hovy et al., 2006), in order to augment the corpus with additional data through a systematic mapping. Such data augmentation is crucial for the future implementation of a pipeline for the automatic extraction of biographical events.

The paper is structured as follow. In Section 2, a review of works on biographical encoding and event extraction is provided. Section 3 describes data collection and annotation guidelines design. In Section 4, results of the annotation are presented. Section 5 presents the mapping of the resource with existing corpora. Finally, Section 6 concludes the paper with some insights on future work.

2. Related Work

The extraction of biographical events from text brings into play two main research lines, namely...

Semantic Roles and Events Annotation Frameworks. The annotation of semantic roles has been addressed by a number of approaches with specific focuses (see Petukhova and Bunt (2008)). FrameNet (FN) (Baker et al., 1998) and PropBank (PB) (Kingsbury and Palmer, 2002) are two databases of semantic roles: the former is not syntactically bounded and relies on a detailed taxonomy of semantic roles; the latter is centered on verbs and the classification of arguments is coarse-grained. Other approaches are focused on a general notion of semantic role. VerbNet's (VN) (Schuler, 2005) aim is the classification of English verbs on the basis of semantic-syntactic properties; LIRICS identifies 'relational notions which link a participant to some real or imagined situation ('event')' (Bunt and Romary, 2002). In last years, attempts to unify such resources have been made. The Semantic Annotation Framework (SemAF) (Bunt and Palmer, 2013) provides an unifying framework according to which a semantic annotation relies on a finite set of eventualities (EV) and participants (PT) that form entity structure pairs with

markables, namely tokens to which an EV or PT can be attached. Pairs are then combined in links through link structure. For instance, in the sentence 'she published poetry' three entity structure pairs may be annotated: ϵ_1 = < She, POET >; ϵ_2 = < published, PUBLISH >, and ϵ_3 = < poetry, POEM >. A link structure triple connect ϵ_1 and ϵ_2, assigning to the former the role of agent: L_1 = < ϵ_1, ϵ_1, Agent>.

Frameworks for the annotation of events are heterogeneous, reflecting the high variety of existing event extraction tasks (see Xiang and Wang (2019)). The ACE/ERE initiative (Song et al., 2015) resulted in a series of news corpora in which textual triggers were annotated and labelled by referring to a close set of event types. For instance, the word 'migration' triggers an event of the type 'Movement'. The Topic Detection and Tracking initiative (TDT) (Allan, 2012) led to a corpus in which the story rather is annotated and labelled with reference to actual historical events (eg: Death of Kim Jong Il, Cuban Riot in Panama, etc.) rather than general categories. The ISO-TimeML framework (Pustejovsky et al., 2010) is a standard for the annotation of temporal expressions, events, and temporal relations between events. According to such approach, an instance of the type 'EVENT' must be used to annotate a situation that happens or occurs. Furthermore, events are categorized by some linguistic properties. For instance, the word 'start' triggers an event of the type 'ASPECTUAL', whereas 'say' is a 'REPORTING' event. The Richer Event Description (RED) guidelines (O'Gorman et al., 2016) is a reformulation of ISO-TimeML in which the taxonomy of event properties is simplified, but further annotation layers are defined: entities, causal relations between events, and link between entities.

Our annotation guidelines for biographies take inspiration from two existing frameworks. On one side, they adopt the semantic formalism of SemAF (Bunt and Palmer, 2013), while on the other side they partly inherit the taxonomy of events proposed in ISO-TimeML (Pustejovsky et al., 2010).

Biographical Events Extraction. Despite the existence of several semantic models for biographical events encoding, few works focused on the extraction of biographical information. Russo et al. (Russo et al., 2015) collected 782 biographies of people deported to Nazi concentration camps with the aim of extracting a predefined set of information from both raw text and DBpedia. Then, all information was arranged into a structured representation by using the TimeML framework (Pustejovsky et al., 2010). Menini et al. (2017) defined a set of verbal motion frames and used it to extract migration events from Wikipedia biographies.

Both works adopt a top-down approach. First, a number of information to be retrieved is defined, then an event extraction pipeline is built.

Our guidelines rely on a bottom-up approach: instead of a predetermined classification of event types to be extracted, the focus is on all events in which the entity of the type *writer* is involved as a participant.

3. Data Collection and Annotation Scheme Design

In this section, the data gathering and preprocessing from Wikipedia is described; then, the annotation guidelines are presented.

3.1. Data Gathering

The corpus is a collection of sentences extracted from $8,047$ Wikipedia English pages of under-represented writer, namely authors born in non-Western countries, migrants or ethnic minorities. Specifically, Wikidata properties 'place of birth', 'occupation', and 'ethnic group' were exploited in order to identify all writers born in a former colony or writers belonging to a minority group that were born in a Western country. The data gathering process was performed in four steps: (i) each biography was split in sentences using Stanford Core NLP (Manning et al., 2014); (ii) for each sentence, all the named entities of the type Location or Organization were identified using the same tool; (iii) an automatic semantic role labelling was performed on each sentence, using SRL Bert (Shi and Lin, 2019). The resulting dataset of $218,198$ tuples of predicates and semantic arguments contains at least one Location or one Organization. Below some examples are reported:

- **predicate**:move,**ARG2**:to New York City;
- **predicate**:study,**ARGM-LOC**:in the Convent of Jesus and Mary School in New Delhi;
- **predicate**:confer,**ARG0**:by the municipality of Kautokeino and the Kautokeino Sámi Association.

In the final step (iv), we identified the most frequently occurring combinations of 'predicate,ARG0', 'predicate,ARG1', and 'predicate,ARG2' in order to select a sample representative of the sentences in the data set for annotation.

3.2. Annotation Guidelines

Annotation guidelines were developed in order to annotate all events in which the subject of the biography is a participant in the event. It is important to notice that there is no one-to-one correspondence between a tuple of the type <predicate,argument> and a sentence, since most sentences contain more than one predicate, as it can be observed in the following example:

"In 1974 he left South Africa, living in North America, Europe and the Middle East, before returning in 1986".

Hence, a separate annotation for each relevant subject-predicate pair was made.

The selection of the most significant semantic arguments in biographical events is guided by previous work (Stranisci et al., 2021a) in which a set of combinations of life events and named entities types were

recognized as salient for biographies: locations for migrations; organizations for education and career events. Therefore, our guidelines mainly focus on events in which the subject of the biography is involved with such named entities. Moreover, since time is a crucial feature for biographical narratives, guidelines includes the identification of temporal expressions.

Identification of the entity and their semantic role. The prerequisite for an event to be annotated was that it had to involve the biography subject. This involvement was not always direct, though: an author could be mentioned through their works, as in "Her third novel, Missing in Machu Picchu (2013), was awarded" or through a group they were part of, as in "At the age of nine, her family moved to Ghana". According to the RED guidelines[1], the former case was a BRIDGING relation, while the latter was a SET-MEMBER link. In our guidelines all these types of entity had to be annotated as if they were an instance of the writer, in order to consider important biographical events of the type 'his book win a prize', in which the writer is only indirectly mentioned.

Together with the identification of the writer, annotators had to specify her/his semantic role, in order to classify their participation in the event. Two labels were created for this purpose, both inspired by the Propbank framework: 'writer-ARG0', when the entity plays roles covered by this argument, such as 'Agent' or 'Perceiver', 'writer-ARGx', if they play roles covered by other argument types, like 'Patient'. Even though grouping such arguments slightly reduces the expressiveness of the PropBank framework, it has the advantage of helping the annotators to focus on a more general distinction between events in which writers have an active role and events in which they have not.

Identification of events, and their taxonomy. Events had to be annotated according to the TimeML scheme and were categorized according to a subset of TimeML event types tag: 'ASP-EVENT' to mark all verbs conveying aspectual information, and 'REP-EVENT', for verbs reporting other states and events, 'STATE', 'EVENT' respectively. The last two are mutually exclusive in each annotation. For instance, in the sentence "Then, she traveled to Venezuela, where she worked in linguistics at the Department of Justice of Venezuela" two separate annotations had be provided: one for the pair 'she-traveled', and another one for the pair 'she-worked'. 'ASP-EVENT' and 'REP-EVENT' may occur jointly with another 'STATE' or 'EVENT', in expressions such as 'he started working', which results in the link structure $< started, working, ASP >$ and 'he said he moved', which is encoded as $< said, moved, REP >$

Since some sentences contained nominal utterances and there were semantically empty verbs like the copular *be*, guidelines allowed for the annotation of names as events or states in subordinate clauses like "After a brief time in Toronto", or in nominal predicates such "He was a professor". The annotation of nominal events was supported by NomBank frames (Meyers et al., 2004).

Identification of arguments containing a location or an organization. The third component of the guidelines was aimed at identifying the relation between the writer and some named entities that may signal their migration or their condition of being a migrant in a given place. Annotators were asked to select the entire argument containing a location or an organization, and to mark the latter as 'ARGx-ORG', and the former 'ARGx-LOC'. The focus of this annotation stage was not to identify the specific semantic argument, but to label the cases in which a named entity is part of a semantic role. This allowed to refine clusters of arguments and map them onto existing taxonomies. For instance, in 'He works for $organization', the ARGx-ORG may be mapped onto the VerbNet 'Beneficiary' thematic role.

Identification of temporal arguments. Finally, the guidelines establish the annotation of temporal arguments. Rather than identifying only the token triggering a time expression, the entire argument had to be selected and labelled as 'ARGM-TIME'. For instance, in the example "In 1974 he left South Africa" the entire semantic argument "in 1974" had to be annotated.

A fully annotated example of the sentence below is the following:

"In 1974 he left South Africa, living in North America, Europe and the Middle East, before returning in 1986".

ϵ_1 = <he, WRITER>
ϵ_2 = <left, LEAVE>
ϵ_3 = <South Africa, LOCATION>
ϵ_4 = <living, LIVE>
ϵ_5 = <in South Africa, LOCATION>
ϵ_6 = <in 1974, TIME>
L_1 = <ϵ_1, ϵ_2, writer-ARG0>
L_2 = <ϵ_3, ϵ_2, ARGx-LOC>
L_3 = <ϵ_1, ϵ_4, writer-ARG0>
L_4 = <ϵ_5, ϵ_4, ARGx-LOC>
L_5 = <ϵ_6, ϵ_2, ARGM-TIME>

4. Annotation Task and Results

The annotation task involved 4 annotators who evaluated $1,000$ sentences sampled from $8,047$ Wikipedia English pages of under-represented writers. One of them (ann_01 in Table 1) evaluated all sentences 1000, while the others annotated respectively 200 (ann_02), 100 (ann_03), and 200 (ann_04) sentences. The annotation has been performed on Label Studio[2], an Open Source platform that easily allows to organize chunk annotation tasks. Annotators were asked to provide one

[1] https://github.com/timjogorman/RicherEventDescription

[2] https://labelstud.io/

Table 1: Inter-Annotator Agreement (F-measure).

annotator	Event	State	Writer-ARG0	Writer-ARGx	ARGx-LOC	ARGx-ORG	ARGM-TIME
ann_01 (baseline ann_02)	0.83	0.72	0.90	0.87	0.78	0.75	0.91
ann_01 (baseline ann_03)	0.83	0.76	0.91	0.92	0.38	0.75	0.94
ann_01 (baseline ann_04)	0.84	0.66	0.91	0.90	0.65	0.83	0.85
ann_02 (baseline ann_01)	0.83	0.66	0.91	0.89	0.85	0.94	0.94
ann_03 (baseline ann_01)	0.82	0.64	0.93	0.95	0.91	0.92	0.94
ann_04 (baseline ann_01)	0.84	0.61	0.91	0.89	0.75	0.70	0.87
Average	**0.83**	**0.67**	**0.91**	**0.90**	**0.75**	**0.81**	**0.91**

separate annotation for every EVENT or STATE identified in each sentence. As it is shown in Figure 1, the same sentence has received two separated annotations. The first is the chunk 'jailed' labelled as an EVENT, the second is the chunk 'detained', labelled as a STATE.
The IAA was computed through averaged pairwise F-measure: in this setting, the annotations of one annotator are used as the reference against which the annotations of the other annotator are compared. In order to maximize the agreement between annotators, we did not only consider the exact match between chunk, but also the cases in which one chunk contained the other. Adopting such an approach has allowed to resolve some recurrent inconsistencies. Let us consider the two pairs of annotations:

1. awarded / was awarded
2. the United Nations / to the United Nations

In the first one (1) all the smaller chunk was kept. Conversely, in (2) the larger chunk was kept, in order to preserve the semantic role of the argument containing an entity of the type location.
Table 1 shows the F-measure of the agreement between annotators for each class. Agreement is larger than 0.8 in almost all classes, with the exception of STATE and ARGx-LOC. From a qualitative analysis we observed a mismatch in the recognition of nominal events in proposition such as in (3). Lower agreement in ARGx-ORG identification seems to be caused by the broadness of such a type of entity that results in a variety of irrelevant usages for the annotation task, as in (4).

3. after one year of studies
4. when Sri Lanka banned the burka on 2019, Nasrin took to Twitter to show her support for the decision

The resulting corpus contains $1,489$[3] semantic annotations. Table 2 summarizes the number of ST in the corpus, in which there are 894 events and 695 states. Furthermore, 215 aspectual or reported events were annotated; they occurred in 72 semantic annotations. In 143 cases, they jointly appear with an event or a state (eg: 'he $[started]^{ASP-EVENT}$ $[working]^{STATE}$'). Writers hold the semantic role of agent in $1,205$ annotations, other roles in 445. Arguments containing an organization or a location are $1,203$. More specifically, there are 281 sentences in the corpus in which the presence of a named entity of the type LOCATION or ORGANIZATION was not relevant, despite the corpus to annotate was created by relying on a combination of Named Entity Recognition and Semantic Role Labelling (see Section 3).

Table 2: All the occurrences of Semantic Types in the corpus.

Semantic Type	Occurrences
EVENT	894
STATE	695
ASP-EVENT	114
REP-EVENT	101
writer-ARG0	1,205
writer-ARGx	445
ARGx-LOC	532
ARGx-ORG	671
TIME	525

In Table 3 the 10 most frequently occurring events and states are shown. Some of them are related to the writers' educational journey (eg: graduate, hold, attend, study), others to their career (eg: publish, serve, teach, win, work, write). Finally, there is a set of events framing personal events (eg: bear, die, live, move). From such clusters of predicates, a set of biographical frames may be derived. This is the inverse process of existing works on biographical knowledge extraction from text (Menini et al., 2017; Russo et al., 2015). Rather than selecting a prior number of frames to be used for data gathering, this approach extracts knowledge that must subsequently be aligned to existing resources.

5. Mapping

The annotation guidelines and the corpus presented in this paper constitute a first, yet essential step towards the development of a system for the automatic extraction of biographical events. While such system will be addressed in future work, in this Section we illustrate

[3]The corpus is available at: `https://github.com/marcostranisci/biographicalEvents`

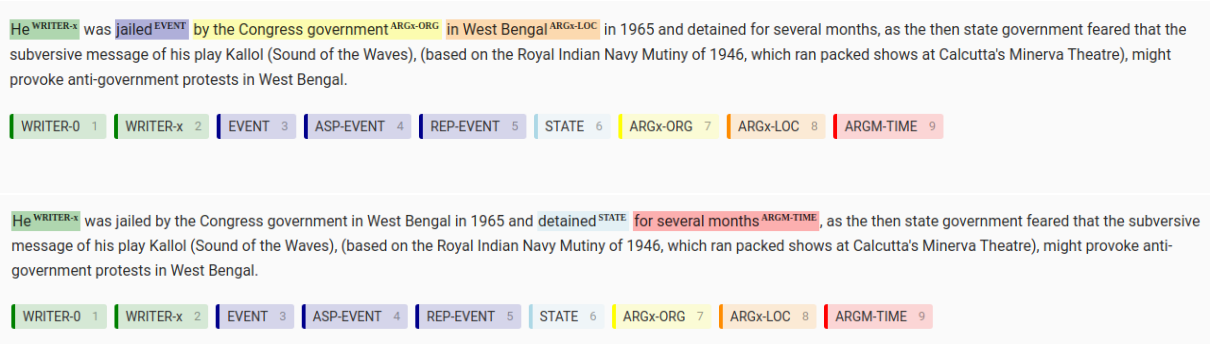

Figure 1: Two examples of annotation in Label Studio.

Table 3: The ten most frequent events and states within the corpus.

Event	occ.	State	occ.
receive	56	work	60
publish	39	write	46
win	36	study	41
award	34	teach	28
write	25	attend	28
move	25	live	21
bear	22	serve	20
graduate	21	hold	15
take	20	spend	14
die	20	writer	13

Table 4: The distribution of arguments containing a Organization (ORG), a Person, or a Geo Political Entity (GPE) for the work.01 PropBank sense in OntoNotes.

argument	n.	ORG	PERSON	GPE
ARG0	996	6.0%	8.1%	3.5%
ARG1	347	7.8%	2.0%	7.2%
ARGM-LOC	248	7.7%	0.4%	18.5%
ARGM-MNR	239	0.4%	0.8%	0.0%
ARGM-TMP	148	1.4%	1.4%	0%
ARGM-DIS	122	0.8%	3.3%	0.0%
ARG2	107	29.0%	8.4%	11.2%
ARG3	99	17.2%	13.1%	8.1%

how the current corpus could be extended to obtain an appropriate training dataset. We show how the data from OntoNotes (Hovy et al., 2006) can be mapped onto our annotation schema, and report some figures regarding this process. Although OntoNotes was selected as the first target for this mapping, the same process could also be applied to other PropBank-like datasets, such as (Kim and Klinger, 2018), for the enrichment of our original corpus.

OntoNotes (Hovy et al., 2006) contains a multi-layer annotation of texts from several domains (e.g., newswires, magazine articles, broadcast news). For each such domain, a PropBank-based semantic annotation and the annotation of named entities is provided. The data set is composed of 99,974 sentences, 249,157 rolesets, and 554,307 semantic arguments. Given a verb, rolesets represent all roles possibly associated to each of its senses according to the PropBank model (Bonial et al., 2014).

In order to align the two corpora, we extracted all verb occurrences and their arguments. Then, we computed the percentage of arguments containing a named entity of the type ORG, GPE, or PERSON. Table 4 shows the 8 most frequently recurring instances for the roleset associated to work.01, which expresses the sense "work, being employed, acts, deeds". As it can be observed, in some of them there is a predominance of GPE and ORG compared to entities of the type PERSON. This enables the identification of some arguments that are more likely to be aligned with our corpus: it is the case of ARG1 and ARG2, which respectively correspond to 'job, project' and 'employer, benefactive'. Let us consider the following examples.

5. <*work*, to improve China's nickel industry's level of technology, technique and equipment, ARG1>

6. <*work*, for the Justice Department, ARG2>

In the former case, the GPE simply adds information about the argument, as in (5). In the latter case, it is directly linked to the verb with the role of 'benefactive', as in (6).

We analyzed the distribution for the 10 most frequently occurring events and states in our corpus (Table 5): they amount to 430 events, covering the 27% of the overall number of instances. Besides the widespread presence of the ARGM-LOC modifier, some patterns emerge. There is a set of events in which an ORG or a GPE has agency on the event: publish.01, award.01. The 60% of the ARG0 linked to publish.01 and the 80% linked to award.01 contain a GPE or a ORG. In fact, many books are published and many prizes are awarded by an organization or a geopolitical entity.

Other patterns may imply the 'benefactive' role: as mentioned before, work.01 is often linked to a bene-

factive as in (6). Conversely, the presence of GPE or ORG in arguments of the type 'benefactive' linked to award.01 seems to be not informative. In some cases, they have an appositive function, as in 'to Waring & LaRosa, New York'. At times, the mapping is less interesting for the specific task, since in some cases organizations are the recipient of a prize, which is not consistent with the biographical domain.

Some arguments are specific to single verbs. For instance, receive.01 always presents an ARG2 associated with the role 'received from', while attend.01 ARG1 always presents instances of type 'thing attended'. Both combinations are common in sentences like 'he attends an institution' and 'he received a degree from an institution'. The distribution confirms such pattern, since the 45.1% of ARG1 linked to attend.01 and the 31.7% paired with receive.02 contain a ORG or a GPE.

Finally, move.02, win.01, and work.01 show a similar behavior when an ARG1 is present. GPE and ORG Entities in this argument are not directly linked to the verb, but rather to further entities within the argument, such as in the example (7):

7. <'win', 'the New York Drama Critics' Circle Award', ARG1>

By definition, the ARG1 of the verb 'win' represents a prize; however, since the organization 'New York Drama Critics' is part of the argument, the entity type ORG is mistakenly considered as a value for the argument in our statistics. Although this behaviour represents an issue when recording descriptive statistics and for the mapping process, such dependency structures should be considered to collect precious and more subtle biographical information that needs further investigation.

Despite the actual limitations, the results of the mapping process is encouraging. In fact, even considering only non-ambiguous argument types, 851 instances may be mapped from OntoNotes to the top ten instances of our corpus, tripling the initial size of the corpus. At the same time, we observed the emergence of patterns helpful to automatically extract and understand events and states from raw text biographies. Further studies may focus on the automatic implementation of such patterns.

Table 5: The most recurring link structures of the type <verb,argument containing ORG or GPE> for the 10 events and states with more occurrences in our corpus.

verb	argument(s)	description
work.01	ARG2 ARG1	benefactive project
write.01	ARG2	benefactive
receive.01	ARG2	received from
publish.01	ARG0	publisher
win.01	ARG1	prize
award.01	ARG0, ARG2	giver, beneficiary
attend.01	ARG1	thing attended
move.01	ARG2	destination
move.02	ARG1	measures
study.01	ARGM-LOC	location
teach.01	ARGM-LOC	location

6. Conclusions and Future Work

In this paper we presented a novel schema for the annotation of biographical events in free text. We have also built a new corpus for this task, containing $1,000$ annotated sentences sampled from $8,047$ Wikipedia English pages pertaining underrepresented writers. Finally, we have shown how existing resources, such as OntoNotes, can be mapped onto our annotation schema in order to increase significantly the size of the corpus.

The developed corpus and the proposed schema are preparatory for the development of an automatic system for the extraction of biographical events from free text, which constitutes the main focus of our future work. Ideally, we could start from existing systems performing semantic role labeling (such as (Shi and Lin, 2019)), and then adapt the results in a manner similar to the one adopted in the mapping process. The mapping process itself also needs to be strengthened with a more thorough evaluation and with the development of specific rules to better detect the entities filling the arguments. Our final focus consists in a study aimed at better understanding and quantifying how the biographical information extracted by the system can be beneficial to tackle other downstream tasks.

7. Bibliographical References

Allan, J. (2012). *Topic detection and tracking: event-based information organization*, volume 12. Springer Science & Business Media.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Bonial, C., Bonn, J., Conger, K., Hwang, J. D., and Palmer, M. (2014). Propbank: Semantics of new predicate types. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3013–3019.

Booth, A. (2008). Orlando: Women's writing in the british isles from the beginnings to the present.

Bunt, H. and Palmer, M. (2013). Conceptual and representational choices in defining an iso standard for semantic role annotation. In *Proceedings Ninth Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation (ISA-9), Potsdam*, pages 41–50.

Bunt, H. and Romary, L. (2002). Requirements on multimodal semantic representations. In *Proceed-*

ings of ISO TC37/SC4 Preliminary Meeting, pages 59–68. KAIST.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Kim, E. and Klinger, R. (2018). Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359.

Kingsbury, P. R. and Palmer, M. (2002). From treebank to propbank. In *LREC*, pages 1989–1993. Citeseer.

Krieger, H.-U. and Declerck, T. (2015). An owl ontology for biographical knowledge. representing time-dependent factual knowledge. In *BD*, pages 101–110.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Menini, S., Sprugnoli, R., Moretti, G., Bignotti, E., Tonelli, S., and Lepri, B. (2017). Ramble on: Tracing movements of popular historical figures. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 77–80.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., and Grishman, R. (2004). The nombank project: An interim report. In *Proceedings of the workshop frontiers in corpus annotation at hlt-naacl 2004*, pages 24–31.

O'Gorman, T., Wright-Bettner, K., and Palmer, M. (2016). Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56.

Petukhova, V. and Bunt, H. (2008). LIRICS semantic role annotation: Design and evaluation of a set of data categories. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Pustejovsky, J., Lee, K., Bunt, H., and Romary, L. (2010). Iso-timeml: An international standard for semantic annotation. In *LREC*, volume 10, pages 394–397.

Russo, I., Caselli, T., and Monachini, M. (2015). Extracting and visualising biographical events from wikipedia. In *BD*, pages 111–115.

Schuler, K. K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Shi, P. and Lin, J. (2019). Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Song, Z., Bies, A., Strassel, S. M., Riese, T., Mott, J., Ellis, J., Wright, J., Kulick, S., Ryant, N., Ma, X., et al. (2015). From light to rich ere: Annotation of entities, relations, and events. In *EVENTS@ HLP-NAACL*, pages 89–98.

Stranisci, M. A., Basile, V., Damiano, R., and Patti, V. (2021a). Mapping biographical events to odps through lexico-semantic patterns? In *12th Workshop on Ontology Design and Patterns, WOP 2021*, volume 3011, pages 1–12. CEUR-WS.

Stranisci, M. A., Patti, V., and Damiano, R. (2021b). Representing the under-represented: A dataset of post-colonial, and migrant writers. In *3rd Conference on Language, Data and Knowledge, LDK 2021*, volume 93, pages 1–14. Schloss Dagstuhl-Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing.

Sun, J. and Peng, N. (2021). Men are elected, women are married: Events gender bias on wikipedia. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 350–360.

Tuominen, J. A., Hyvönen, E. A., Leskinen, P., et al. (2018). Bio crm: A data model for representing biographical data for prosopographical research. In *Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017)*. CEUR Workshop Proceedings.

Vrandečić, D. and Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Wang, R., Yan, Y., Wang, J., Jia, Y., Zhang, Y., Zhang, W., and Wang, X. (2018). Acekg: A large-scale knowledge graph for academic data mining. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 1487–1490.

Xiang, W. and Wang, B. (2019). A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

8. Language Resource References

Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., et al. (2003). The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.