Guidelines for the *Vasari* Entity Disambiguation Evaluation Dataset

Cristian Santini, Oleksandra Bruns, Etienne Posthumus, Mary Ann Tan, Tabea Tietz, Harald Sack
FIZ - Information Service Engineering

Introduction

This guidelines are based on the document *Annotation Guidelines for Named Entity Recognition, Entity Linking and Stance Detection* (Hamdi et al., 2021a) for the NewsEye multilingual dataset of Historical Newspapers (Hamdi et al., 2021b).
Our guideline further extends our reference document for what concerns the annotation of common nouns and dates.

1. What is a named entity?

A Named Entity can be generally defined as anything which is mentioned with a proper noun. From the NewsEye guidelines, which inspired this document:

"*Linguistic units considered as named entities must include a proper name, or a definite description having the status of a proper name. Although the definition of a proper name is not straightforward, here are a few characteristics commonly accepted (not valid in all cases nor in all languages): presence of majuscule, non inclusion in lexical but in encyclopedic dictionaries, absence of meaning (the name* George *does not carry - per se - any information about the type of entity that can be called this name, while the noun "table" gives specific information about the type of objects that can be called by it - i.e. having a plateau and feets), and absence of compound meaning (the White House does not refer to any house which is white, la Gare de Lyon is not in Lyon, le Pont Neuf is very old).*"

By this definition, a Named Entity is anything which is mentioned with a proper noun or a definite description, i.e. a noun phrase, having the status of a proper name. This is later elaborated as follows:

"*In front of some definite descriptions, it might be difficult to decide what to do, e.g.* la commission Impériale*,* l'escadre de Nelson*. In such difficult cases, consider the following:*

- definite descriptions which can be considered as named entities tend to have a nominative function (like proper names) rather than a descriptive function. What a definite description says literally about a referent is less important than the nominative aspect.

*- even though some named entities are definite descriptions which are descriptive, e.g. "*Syndicat National de la Magistrature*". In such cases, what makes it a named entity is the referential stability: the entity referred to is always the same.*"

The authors insist on the fact that noun phrases are considered as mentioning named entities when their function is not to describe something literally but rather to mention some external entity. In their words: "what makes it a named entity is the referential stability".
For example, *Republique of France* is a noun phrase which acts as a proper noun, but *pope with the hat* is not considered as a named entity because *a*) does not contain proper nouns, and *b*) the noun phrase has a descriptive rather than nominative function.

Moreover, in order to define what is a named entity we have to take in consideration also the scope of our dataset, which is that of linking entities mentioned in artists biographies to items collected in a KB. For this reason we extended our definition of named entity to also include noun phrases which have a descriptive function, but at the same time appear in the Wikidata encyclopedia as artistic concepts, such as "*Madonna with the Child*".

2. Extending the annotation to common nouns[1] and dates

The creation of this dataset is aimed at the evaluation of an information extraction pipeline for entities described in art history books. An issue that we needed to face since the creation of the dataset is the inclusion of entities which are mentioned in the description of artworks as iconographic motifs. These motifs can be mentioned with common nouns, such as *angel* or *sword*.

As explained in the previous section, common nouns act in the sentence with a descriptive function and they convey literal meanings which can be combined together. This does not occur for proper nouns. Therefore, common nouns are usually annotated separately in our dataset, since noun phrases which do not have a nominative function (as proper nouns) represent an entity from a semantic perspective by aggregating the meanings of the single common nouns, as in *type of flower*. There's only a case when multiple common nouns are annotated together, and it is when they are used in combination to define a single concept present in a Knowledge Base (KB), like in *lead white*, *chiaro e scuro*, *bay horse*, etc.

In order to emphasize the difference between the annotation of proper nouns, which might include definitive descriptions (i.e. noun phrases) and common nouns, we marked each annotation with a boolean value (1 or 0) which expresses if a surface form is a common noun or a named entity.

Another issue which we later faced is the inclusion of dates in our dataset. Dates can be absolute ("June 1964") or relative ("three years later"). They can be either expressed in letters ("nineteen sixtyfour"), numerals (1964) or roman numbers ("MCMLXIV"). In our dataset, we decided to annotate only absolute dates.

id	start_pos	end_pos	surface	type	wb_id	named_entity
1	46	53	boyhood	MISC	Q276258	0
1	62	70	instance	MISC	OOV	0
1	74	83	Ser Piero	PER	Q371916	1
1	94	97	art	MISC	Q735	0
1	103	124	Andrea del Verrocchio	PER	Q183458	1
1	143	156	panel-picture	MISC	Q55439	0
1	160	167	S. John	PER	Q40662	1
1	178	184	Christ	PER	Q302	1
1	191	199	Leonardo	PER	Q762	1
1	211	216	angel	MISC	Q235113	0
1	238	246	garments	MISC	Q11460	0
1	274	277	lad	MISC	Q3010	0
1	279	287	Leonardo	PER	Q762	1

Figure 1: Sample of our Entity Linking annotation.

[1] This paragraph presents the first difference with the annotation guidelines provided by (Hamdi et al., 2021a), which is related to the inclusion of common nouns in the annotation.

3. Nesting and special constructions

a. Nested entities: One problem which is often considered in the annotation of NEL datasets is the one of overlapping entity mentions and nested entities, such as in *the John Fitzgerald Kennedy New York Airport*. The problem of overlapping named entities has been thoroughly investigated in (Jha et al., 2017), (Weichselbraun et al., 2019) and (Brašoveanu et al., 2020). In this last work, the authors generalize three different methods to deal with overlapping named entities. This methods are:

- ØMIN: do not consider nested entities and annotate only the longest surface form (i.e. minimum number of entities)
- ØMAX: annotate each named entity separately with no overlap; ignore surface forms which include multiple named entities as in *JFK NY Airport*.
- OMAX: annotate any entity mentioned in the text by allowing overlaps, annotate also nested entities.

This dataset is annotated by following the ØMIN annotation style. We chose this strategy since it is less prone to split surface forms in multiple entities. Moreover, we decided to not include nested entities to avoid further complexity. Some examples of the ØMIN annotation are reported below:

- <misc>Madonna with the Child<misc>
- <loc>Santa Maria Maggiore in Florence<loc>.
- <per>King Francis of France<per>
- <loc>Church of S. Francesco in Assisi<loc>
- <per>Duke Cosimo of Milan<loc>

b. Surface form boundaries: Surface forms do not include:

- subordinate clauses
- non-restrictive appositives[2]
- incidental clauses
- insertions
- adjectives which do not act in a referential function (*the beautiful* <misc>American Anthem<misc>)

Surface form boundaries include:

- pre-modifiers: only in the form of personal titles (e.g., Saint, Duke, Pope), numerals (e.g., twelve apostles, two thieves) and epithets (e.g., “elder Lorenzo Medici”).
- post-modifiers: regnal numbers, epithets (*S. John the Baptist*), toponymic surnames (*Messer Baldassarre Turini da Pescia*)

c. Coordination: Named entities coordinated based on a coordinate conjunction or a common descriptor are annotated separately.

- <per>Adam<per> *and* <per>Eve<per>
- <per>Santi Pietro<per> *e* <per>Paolo<per>

[2] Non-restrictive appositives were not included due to the fact that a NER algorithm should be capable of recognizing dependency relations between this type of syntagma and the proper noun which is its head, and this task is out of the scope of our evaluation. This is a second difference with respect to (Hamdi et al., 2021a).

4. Entity types

This dataset labels named entities with 5 types:

- Person (PER): any human-like, real or fictional entity, which is referenced with a proper noun.
- Organization (ORG): commercial, educational, government, media, medical-science, non-governmental, religious, sports
- Location (LOC): natural location as well as human-made facilities
- Miscellaneous (MISC): miscellaneous entities, such as events, creative works, products, abstract concepts.
- Date (DATE): absolute dates expressed either with digits, roman numerals or words.

Since they convey generic meanings, we decided to label common nouns as miscellaneous. Moreover, human-like entities which are not mentioned with a common noun are also labeled as miscellaneous. This is the same for human-made facilities and organizations.

- <misc>citizens<misc>
- <misc>king of France<misc>
- <misc>Pope<misc>
- <misc>house<misc>

5. Entity disambiguation, OOV entities and Ambiguities

a. Annotation specificity: Entity disambiguation is carried out by using the Wikidata Knowledge Graph (www.wikidata.org). Wikidata was chosen since it allows to disambiguate both named entities, common nouns and dates, due to its nature of a general-purpose KB which contains named entities as well as common concepts (as a lexicon). For our disambiguation, the goal of our annotation is to link the entity to the most specific resource of the KB, as stated in *A4*.

b. OOV entities: Named entities and common nouns can be *Out Of Vocabulary* OOV if their surface form cannot be mapped to any possible concept in the KB. For proper nouns, this is straightforward and can be verified by checking for the presence of a unique entry for the named entity in the KB. Common nouns represent a special issue, since for a single lemma multiple entities might exist, as in *Figure 2*.

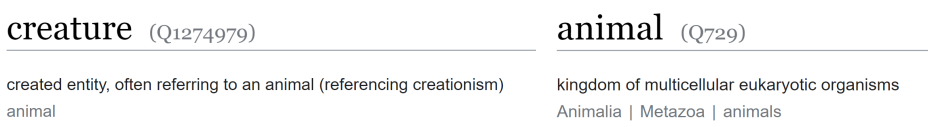

Figure 2: Example of multiple concepts associated with the lemma animal.

In order to identify OOV common nouns, we first verify that 1) the lemma is not present in the KB with a candidate concept which can be used for disambiguation, 2) the KB does not contain synonyms of the lemma which are contained in the KB and are associated to a concept which is an available candidate for the disambiguation of the common noun. The existence of a synonymy relation between the common noun given in the text and the lemma present in the KB is stated by the human-annotator.

c. Special issues - Metonymy: Metonymy occurs when an entity is mentioned with a related concept, as the *White House* is used to refer to the American presidential power. Since our goal is to link the surface form to the most specific entity, entity disambiguation is performed by linking the surface form to the entity which is rhetorically referenced and not to the one which is mentioned literally. However, surface form types can be multiple.

- <misc>Collection<misc> *of* <per>King Francis of France<per> *at* <loc>Fontainebleau<loc>

Here, Fontainebleu *is linked to* wikidata:Palace_of_Fontainebleu *and not to the place where the palace is, since only a palace can contain an art collection.*

- <misc>Abbot<misc> of <org><loc>Vallombrosa<loc><org>

In this sentence, the proper noun Vallombrosa *(a city in Italy) is used to refer to the Vallumbrosan Order (a religious order). Therefore the surface form has both types* org *and* loc *and it is linked to* wikidata:Vallumbrosan_Order.

c. Special issues - Creative Work: Creative works sometimes might be referred to as a definite description, as in *portrait of Monna Lisa*, or a proper noun, as in *The Last Supper*. However, the problem is that most Renaissance paintings are mentioned with the name of their subject, and this may lead to ambiguity in Entity Linking. In order to have a coherent approach, we decided to link creative works to their respective entry in the KB only if they are mentioned directly, i.e. they satisfy the requisite of *referential stability* needed to be considered as a named entity. For this reason, we did not consider all those artwork descriptions which are given in long clauses and which do not refer to the painting directly. For differentiating between named entities and long description with no referential stability, we refer to the discussion in (Jha et al., 2017). An overview of specific cases is given in *Table 1*.

Case	**Example**	**Entity annotated**
Direct mention	*The* <misc>portrait of Monna Lisa<misc>.	wikidata:Mona_Lisa (*painting*)
Long description	*The painting of* <per>S. John<per> *baptizing* <per>Christ<per>.	wikidata:Jesus, wikidata:John_the_Baptist
Long description	*A panel containing the* <misc>Adoration of the Magi<misc>.	wikidata:Adoration_of_the_ Magi (*Christian subject*)
Indirect mention	*A* <misc>Pietà<misc>.	wikidata:Pietà (*artistic theme*)

Table 1: Overview of different artwork mentions and choices for the annotation.

d. Special issues - Coreference: Often, common nouns might refer to specific entities which are earlier or later mentioned in a sentence and they may act as appositives which further describe a named entity. In this case, it might be difficult to state if a common noun should be linked to the same named entity to which it refers, or to the generic concept that the common noun expresses. However, to link a common noun to a named entity implies that we take into account coreference resolution, which is out of the scope of our evaluation.

For this reason, we decided to annotate common nouns only with generic concepts associated with their lemma, and not with named entities. This is also the case for common nouns which indicate temporal roles, such as *King* or *Pope*. For a discussion on temporal role annotation see (Koutraki et al., 2018).

d. Special issues - Dates:
The disambiguation of dates is not always possible, since a specific date (3 February 2020) might not be present yet in the KB. In the case where a date mentioned in the text is not present in Wikidata, we try to link it to the closest date present in the KB.

- <date>28 June 1472<date> → wd:Q6589 (June, year)

Moreover, absolute dates may not be mentioned in a text completely: for example, days or months as well as years can be missing. We decided to annotate also incomplete absolute dates and link them to the corresponding Wikidata entity, if present.

- <date>the seventeenth day of February<date> → wd:Q2341 (February 17, date)

References

Brasoveanu, A. M. P., Weichselbraun, A., & Nixon, L. (2020). In Media Res: A Corpus for Evaluating Named Entity Linking with Creative Works. *Proceedings of the 24th Conference on Computational Natural Language Learning*, 355–364. https://doi.org/10.18653/v1/2020.conll-1.28

Hamdi, A., Pontes, E. L., & Doucet, A. (2021a). *Annotation Guidelines for Named Entity Recognition, Entity Linking and Stance Detection*. https://doi.org/10.5281/zenodo.4574199

Hamdi, A., Linhares Pontes, E., Boros, E., Nguyen, T. T. H., Hackl, G., Moreno, J. G., & Doucet, A. (2021b). A Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2328–2334). Association for Computing Machinery. https://doi.org/10.1145/3404835.3463255

Jha, K., Röder, M., & Ngonga Ngomo, A.-C. (2017). All that Glitters Is Not Gold – Rule-Based Curation of Reference Datasets for Named Entity Recognition and Entity Linking. In E. Blomqvist, D. Maynard, A. Gangemi, R. Hoekstra, P. Hitzler, & O. Hartig (Eds.), *The Semantic Web* (pp. 305–320). Springer International Publishing. https://doi.org/10.1007/978-3-319-58068-5_19

Koutraki, M., Bakhshandegan-Moghaddam, F., & Sack, H. (2018). Temporal Role Annotation for Named Entities. *Procedia Computer Science*, *137*, 223–234. https://doi.org/10.1016/j.procs.2018.09.021

Weichselbraun, A., Brasoveanu, A. M. P., Kuntschik, P., & Nixon, L. J. B. (2019). Improving Named Entity Linking Corpora Quality. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 1328–1337. https://doi.org/10.26615/978-954-452-056-4_152