IRISA System for Entity Detection and Linking at CLEF HIPE 2020

Cheikh Brahim El Vaigh, Guillaume Le Noé-Bienvenu, Guillaume Gravier, Pascale Sébillot

IRISA System for Entity Detection and Linking at CLEF HIPE 2020

Cheikh Brahim El Vaigh[1], Guillaume Le Noé-Bienvenu[2], Guillaume Gravier[3], and Pascale Sébillot[4]

[1] INRIA, IRISA, Rennes, France cheikh-brahim.el-vaigh@inria.fr
[2] CNRS, IRISA, Rennes, France guillaume.le-noe-bienvenu@irisa.fr
[3] CNRS, IRISA, Rennes, France guig@irisa.fr
[4] INSA Rennes, IRISA, Rennes, France pascale.sebillot@irisa.fr

Abstract. This note describes IRISA's system for the task of named entity processing on historical newspapers in French. Following a standard entity detection and linking pipeline, our system implements three steps to solve the named entity linking task. Named Entity Recognition (NER) is first performed to identify the entity mentions in a document based on a Conditional Random Fields classifier. Candidate entities from Wikidata are then generated for each mention found, using simple search. Finally, every mention is linked to one of its candidate entities in a so-called linking step leveraging various string metrics and the semantic structure of Wikidata to improve on the linking decisions.

Keywords: Named entity recognition·CRF·Collective entity linking·WRSM entity relatedness measure.

1 Introduction

Entity linking is a core task in textual document processing, which consists in identifying the entities of a knowledge base (KB) that are mentioned in a text. For instance, approaches from the literature implement three stages to solve mention ambiguity in texts. The first stage consists in the detection of named entities within the text and is known as named entities recognition (NER). To further link the mention found in the text, candidate entities are generated for each mention detected in the first stage. Finally, every mention is linked to one of its candidate entities in a so-called linking step. This last step can be performed *independently* for each individual mention, or *collectively* for all mentions at once. In the first case, every mention in a text is assumed to be independent from other mentions and is linked to a candidate entity on sole basis of some similarity between the mention and the candidate entities, so-called *local* scores. By contrast, for *collective entity linking*, entity mentions and the corresponding entities are not assumed independent one from another but somehow semantically related within a (coherent) document, i.e., *mention-to-entity linking decisions are interdependent*. In this case, the local mention-entity scores are complemented with *global* scores reflecting to which extent the candidate entities chosen for the mentions under consideration are

related in the KB, according to a so-called entity *relatedness measure*. The last two stages of the pipeline are also known as named entity linking (NEL).

In the context of the shared task CLEF HIPE 2020 —Identifying Historical People, Places and other Entities—which is a named entity processing on historical newspapers in French, German and English [3], entity linking techniques can be used to retrieve entities from text. CLEF HIPE 2020 is organised as a CLEF 2020 evaluation Lab. However, the historical context makes the linking task harder since texts considered are the results of an optical character recognition (OCR) algorithm which introduces noise. Therefore, we leveraged various features to reduce the impact of the OCR *noise* on named entity processing.

Our system for CLEF HIPE 2020 follows a standard pipeline for entity linking and implements three separate stages:

1. We devise a NER stage on top of the baseline provided by CLEF HIPE 2020 organizers. This system used Conditional Random Fields (CRFs) to detect and classify named entities. We added several features that we found effective for the task of NER.
2. The generation step consists in looking to Wikidata directly when searching entities similar to a given mention. As a lookup in the heavy database (CLEF HIPE 2020 Wikidata dump) is costly in time, we performed automatic searches for the entity mentions using online Wikidata. Note that this search is based on Wikidata indexing algorithm.
3. The linking step is to decide which candidate should be retained for each mention within a document. We tried to link the mentions separately or collectively, training a classifier to predict if a mention is related to one of its candidate entities. The former is based solely on the similarity between a given mention and its candidate entities. The latter which performs the linking collectively for all the mentions at once, beside the previous similarity metrics, makes use of the entity relatedness measure WSRM that we have proposed in [4].
4. The collective linking setup gave the best results and was ranked second for the bundle2 of the shared task CLEF HIPE 2020.

Our source code, datasets and experimental results are made available online for reproducibility purposes[5].

The note is organized as follows. We give the description of our method in Sec. 2. Then we group the experimental results in Sec. 3 before discussing the perspectives and conclude in Sec. 4.

2 System Architecture

This section gives the description of our system. We distinguish two independent tasks for named entity processing, namely the NER and the NEL. Our solutions for the NER and the NEL are described respectively in Sec. 2.1 and Sec. 2.2.

2.1 NER

The NER task aims at detecting the surface forms in a text that correspond to named entities and at classifying those forms as a type (PER, LOC, ORG, TIME, PROD). The NER system that we developed originally came from the NER baseline provided

[5] `https://gitlab.inria.fr/celvaigh/hipe2020`

by the organizing team of the evaluation campaign CLEF HIPE 2020. This system used Conditional Random Fields (CRFs) based on a Python implementation [1] to detect and classify named entities. The features used in this system, as well as the ones we have chosen are described in Table 1.

2.2 NEL

In the NEL stage, we assume that the annotations are known for the mentions (person, organization, location, etc.) for each document. Those annotations are provided by the NER system described in Sec. 2.1 or by an oracle NER. For the candidate generation stage we rely on a simple Wikidata web search. The candidate selection stage accounts for the WSRM entity relatedness measure between candidate entities within the document in an efficient manner, relying here on Wikidata, the KB provided by CLEF HIPE 2020 for named entity processing (see [4] for details on the measure). These different steps are described below.

Candidate Entities Generation To generate candidate entities from the KB for each mention in a document, we chose a simple yet efficient method exploiting the index of Wikidata. For each mention found by the NER phase, we perform online search using Wikidata web pages. We limit ourselves to the top 10 ranked candidate entities. The motivation behind our choice is to speed up the candidate entities generation step as a lookup in the heavy Wikidata dump is costly in time compared to simple web search.

Local Scores The local scores depict the similarity between a mention and its candidate entities. If we assume the mentions to be independent in text, the linking problem can be formalized as

$$\hat{e} = \underset{e_i}{\operatorname{argmax}} \phi(m, e_i) \tag{1}$$

where e_i is a candidate entity, m is an entity mention, and ϕ is the local score function. We tried several metrics for ϕ. Beside the longest contiguous matching sub-sequence, we tried a Levenshtein distance to handle the OCR noise, Wikipedia popularity [2,5] and the cosine similarity based on a word embedding model, similar to the Skip-gram embedding model [6].

Collective Entity Linking In a collective NEL setup, the local score is complemented with a global score accounting for the intricate interrelationships that candidate entities of the different mentions may share. The latter is known as an entity relatedness measure and used to assess entity relationships in the KB, which will allow to estimate the interdependence of the mentions in the text. The CEL problem can be thus formalized as

$$(\hat{e}_1, ..., \hat{e}_n) = \underset{e_{1,...,n}}{\operatorname{argmax}} \left(\sum_{i=1}^{n} \phi(m_i, e_i) + \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \psi(e_i, e_j) \right) \tag{2}$$

where n is the total number of mentions in a text and $\psi(e_i, e_j)$ donates the entity relatedness measure. In the collective linking version of our system, we used the semantic entity relatedness measure WSRM [4] which weights the relation between entities,

where the more relations between the entities, the stronger their relationship. Formally, WSRM is defined between two entities e_i and e_j as

$$WSRM(e_i,e_j) = \frac{|\{r \,|\, (e_i,r,e_j) \in \text{KB}\}|}{\sum_{e' \in E} |\{r' \,|\, (e_i,r',e') \in \text{KB}\}|} \ , \tag{3}$$

where E denotes the set of entities in the KB and $|S|$ the cardinality of the set S.

Because the directions of the relations are somewhat arbitrary in KBs, depending on how the relation vocabulary was designed (e.g., think about the *publishes* and *publishedBy* symmetric RDF properties), we use a symmetric version of WSRM defined as

$$\psi(e_i,e_j) = \frac{1}{2}(\text{WSRM}(e_i,e_j) + \text{WSRM}(e_j,e_i)) \ . \tag{4}$$

Using the NEL Output to Correct the NER Predictions We also exploited the output of the NEL in order to enhance the NER results. First, we used the type (obtained from Wikidata) of the entities retrieved by the NEL and forwarded it to the NER stage, which can be updated accordingly. Then we leveraged WSRM [4] to retrieve, for each entity found by the NEL, a list of potential related entities from the KB. We argue that if an entity is mentioned in the text, its related entities in the KB should be also mentioned in that text. Our aim is to exploit the semantics of the KB for the NER task. Those information provided by the NEL are used as pseudo-labels or features in the CRF to supervise the NER.

3 Experiments

Experimental validation was conducted on the CLEF HIPE 2020 French corpus to assess the quality of our system. The dataset is described in Sec. 3.1. Results for the NER are provided in Sec. 3.2, and in Sec. 3.3 for the NEL.

Due to the high number of results given by the CLEF HIPE 2020 scorer, we decided to focus only on a couple of them, that were given in the produced json file: NE-COARSE-LIT - ALL - strict - F1_micro and NE-COARSE-LIT - ALL - ent_type - F1_micro.

3.1 Dataset

The evaluation corpus is composed of newspaper articles sampled among several Swiss, Luxembourgish and American historical newspapers on a diachronic basis. This corpus is digitised based on an OCR algorithm which hightails the historical context of the evaluation campaign. The time-span of the whole corpus goes from 1798 until 2018. We used only the French version of the corpus composed of a train, a validation and a test sets [6].

3.2 Results of the NER

The NER classifier described in Sec. 2.1 is trained on the CLEF HIPE 2020 dataset. We added several features to the ones of the baseline. We performed a random search to select the best features while controlling the overfitting. We provide the list of the features used in Tab. 1 and the list of the best hyper-parameters in Tab. 2. The system was trained on the train file and then tested on the dev and test files provided by the organizers.

Feature	Was on baseline	Kept
the token in lowercase	yes	yes
the last 3 letters of the token	yes	yes
the last 2 letters of the token	yes	yes
a boolean on whether the token is in uppercase	yes	yes
a boolean on whether the token is in titlecase	yes	yes
a boolean on whether the token is a digit	yes	yes
the correct spelling of the word using an open-source library	no	no
the presence of the token in a list of named entities	no	no
the presence of the token in a list of first names	no	no
if the word is a stop word or not	no	no
if the word is a punctuation mark or not	no	no
the length of the token	no	no
the relative length of the token (small, medium, large)	no	no
the token without redundant letters	no	no
the first 2 characters of the token	no	no
the first 3 characters of the token	no	no
the POS tag of the token	no	no
condition on whether the token matches a date regex	no	no
the token itself, with its diacritical characters converted into their ASCII equivalent	no	yes
the first 100 elements of the vectorial representation of the token using a fastText model provided by the organizers (fr-model-skipgram-300minc20-ws5-maxn-0.vec)	no	yes

Table 1. The list of the features we tested for the NER.

Parameter	Best value found
c1, the coefficient for L1 regularization, between 0 and 1	0.1798
c2, the coefficient for L2 regularization, between 0 and 1	0.0551
min freq, cut-off threshold for occurrence frequency of a feature, between 0 and 1	0
max iterations, the maximum number of iterations for optimization algorithm, between 100 and 1000	192
all possible transitions	false
num memories, the number of limited memories for approximating the inverse Hessian matrix, between 4 and 8	4

Table 2. The best parameters for the CRF.

We compared our NER system with the baseline provided by the CLEF HIPE 2020 organizers on the validation set (dev file). The results are gathered in Tab. 3. We can see that our NER system outperforms the baseline. We believe that its good performance is due to the choice of the selected features, e.g., the use of the tokens present in the text as features for the classifier. The fine tuning of the hyper-parameters of our CRF

[6] Details statistics about the data can be found at `https://impresso.github.io/CLEF-HIPE-2020/datasets.html`

also partly explains the results better than those of the baseline. The results of our system on the test file are gathered in Tab. 3

		Baseline	**Irisa Team**		
	Task	F1	F1	Precision	Recall
Dev file	NERC coarse French strict (literal sense)	0.622	0.716	0.768	0.671
	NERC coarse French fuzzy (literal sense)	0.735	0.821	0.880	0.769
Test file	NERC coarse French strict (literal sense)	-	0.668	0.705	0.634
	NERC coarse French fuzzy (literal sense)	-	0.784	0.828	0.744

Table 3. Scores of NER systems on the test and dev file.

3.3 Results of the NEL

Results of the entity linking process evaluated in terms of micro-averaged F1 classification scores are reported in Tab. 4. The three systems that we submitted to CLEF HIPE 2020 were ranked second (team7 results). We first evaluated the entity linking based on the sole use of the local scores donated by team7_bundle2_fr_1. Second, we added the global score devising a collective entity linking which we named team7_bundle2_fr_2. And finally, we changed the collective linking system to filter the non-linkable mentions (NIL) based on a threshold, meaning we only link a mention to a candidate entity if the prior probability is below a fixed threshold (here 0.5). We can see that the collective linking gave the best results, while the collective linking with a fixed threshold is worse than the non-collective one. These results show the benefit of the collective linking.

3.4 Supervising the NER with the NEL

A few experiments have been carried out to exploit the outputs of the NEL in order to enhance the NER results. The first one consisted in using the types of the entities found by the NEL to change the NER labels; e.g., if the NER detects the entity 'Europe' and classifies it as 'PERS', the NEL links it to 'Q46' and gives the information that the type of 'Q46' is 'LOC'. The second consisted in generating closely related entities

Rank	**Team name**	**System**	**F1**	**Precision**	**Recall**
1	L3i	team10_bundle1_fr_3	0.598	0.594	0.602
2	L3i	team10_bundle1_fr_1	0.597	0.592	0.601
3	L3i	team10_bundle1_fr_2	0.597	0.592	0.602
4	IRISA	team7_bundle2_fr_2	0.421	0.446	0.399
5	IRISA	team7_bundle2_fr_1	0.419	0.450	0.393
6	IRISA	team7_bundle2_fr_3	0.413	0.437	0.391
7	SBB	team33_bundle2_fr_1	0.407	0.594	0.310
8	UvA.ILPS	team31_bundle2_fr_2	0.251	0.352	0.195
9	ERTIM	team16_bundle1_fr_1	0.108	0.150	0.084

Table 4. Linking accuracy (F1 score) on the CLEF HIPE 2020 French dataset for bundle 2.

to the ones found by the NEL. We found that the output of the NEL stage can correct the NER, but can also introduce too much noise. Despite not being able to directly incorporate the output of the NEL with the existing features, we believe that applying a major vote between the different versions of the NER—with and without the NEL output—can lead to an increase of the accuracy of the NER. Nonetheless, our system opened the door to incorporate the semantics of the KB into the NER task.

4 Conclusion

We built an entity processing system based on a CRF classifier for the NER task, and a collective entity linking system for the NEL one, exploiting the WSRM entity relatedness measure that we have proposed in [4]. Our system was evaluated on the CLEF HIPE 2020 French dataset. Though initially expected, we did not succeed in incorporating the output of the NEL to correct the NER step, but we paved the way to fully use the KB semantics in the NER task.

References

1. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., Varoquaux, G.: API design for machine learning software: Experiences from the scikit-learn project. In: European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases Workshop: Languages for Data Mining and Machine Learning. pp. 108–122 (2013)
2. Durrett, G., Klein, D.: A joint model for entity analysis: coreference, typing, and linking. Transactions of the Association for Computational Linguistics **2**, 477–490 (2014)
3. Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S.: Overview of CLEF HIPE 2020: Named entity recognition and linking on historical newspapers. In: CLEF HIPE 2020. pp. 1–25 (2020)
4. El Vaigh, C.B., Goasdoué, F., Gravier, G., Sébillot, P.: Using knowledge base semantics in context-aware entity linking. In: ACM Symposium on Document Engineering 2019. pp. 8:1–8:10. Berlin, Germany (2019)
5. Francis-Landau, M., Durrett, G., Klein, D.: Capturing semantic similarity for entity linking with convolutional neural networks. In: 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1256–1261 (2016)
6. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems. pp. 3111–3119 (2013)