Injecting Temporal-Aware Knowledge in Historical Named Entity Recognition

Carlos-Emiliano González-Gallardo, Emanuela Boros, Edward Giamphy, Ahmed Hamdi, José G. Moreno and Antoine Doucet

University of La Rochelle, France

CERES - 9 juin 2023

Motivation: “Rise of Digitization”

textual corpora for the Humanities and Social Sciences

mass digitization 1980’s - 2000’s

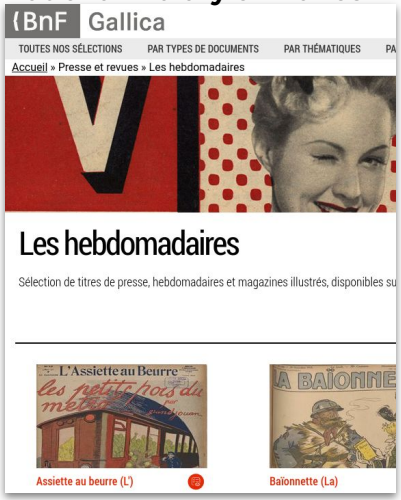
transcripts: manually | OCR | HTR

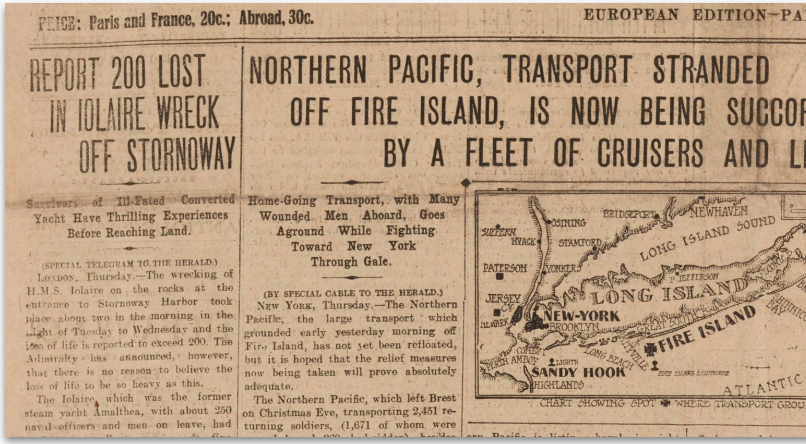
NLP → semantically enriched archives

challenges

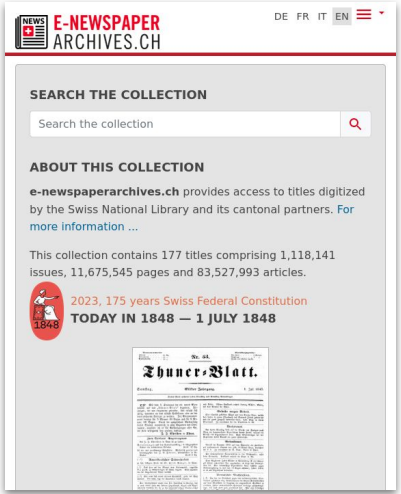
- deteriorated documents
- quality of digitization
- diachrony: language change & evolution

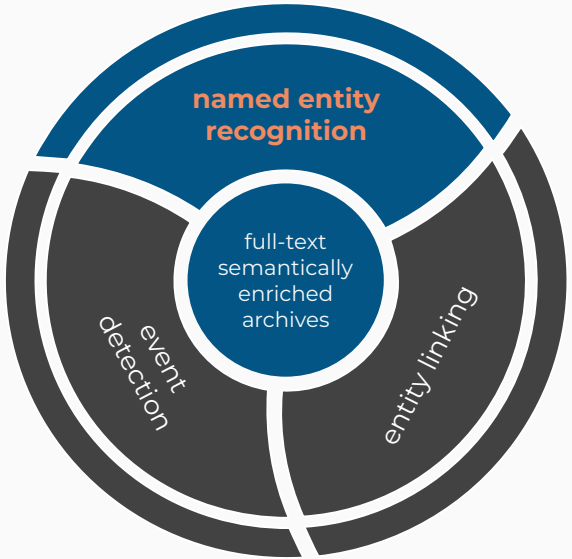
National Library of France





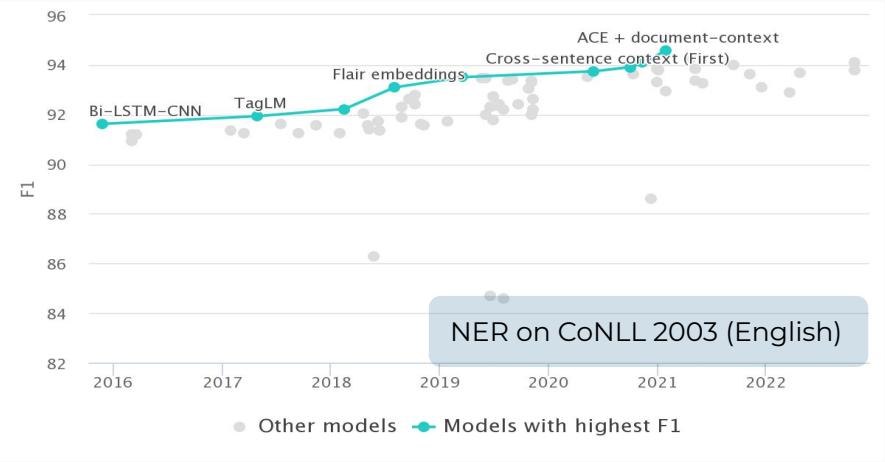
Swiss National Library





Where are we ?

NER in historical and digitized documents less performant & noticeable vs. modern documents



multilingual evaluation campaigns: HIPE@CLEF'20 & '22

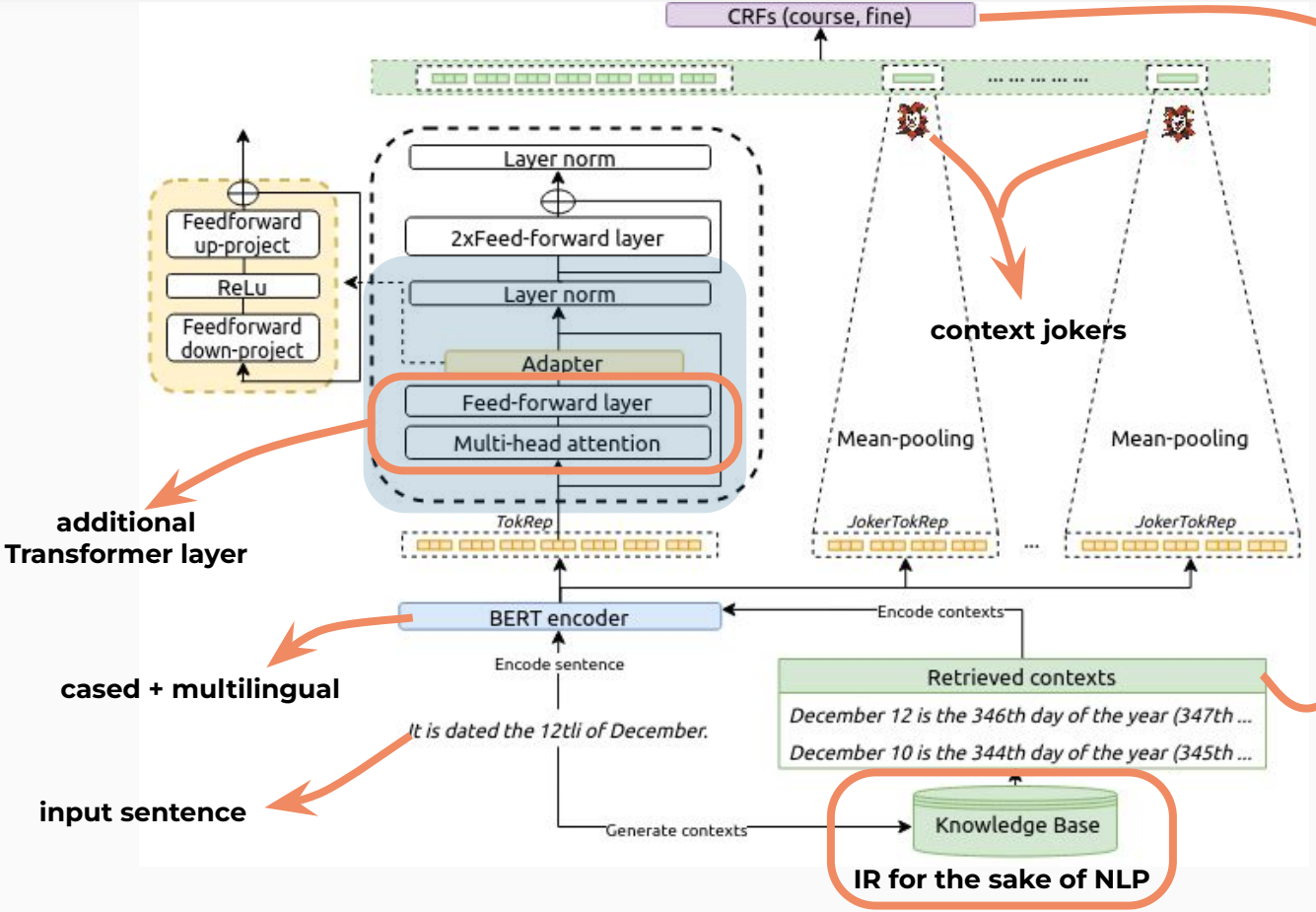
NLP progress

- contextualized embeddings at character level [Bircher, S. 2019] → improves representations of OOVs
- fine-tuning of Transformers ecoders on historical collections [Boros, E. et al., 2020] → alleviates digitization errors
- transformation rules to model diachronic evolution of words [Kogkitsidou, E. et al., 2020] [Díez Platas, M.L, et al., 2021] → recognizes spelling variations
- Historical Multilingual Language Models for Named Entity Recognition (mhBERT) [Schweter, S., et al., 2022] → "historical" language model
- Wenjun THESIS (work in progress) :D

... what about temporality?

Temporal Knowledge-based Contexts for NER

NER Model Architecture



Knowledge Base & temporal information integration

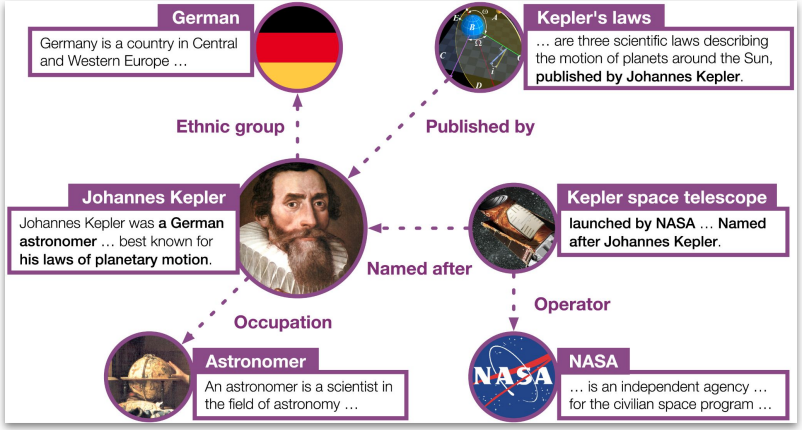
Wikidata5m [Wang X. et al., 2021]

- knowledge graph
- ~ 5M Wikidata entities in the general domain
- aligned to corresponding Wikipedia pages (1st paragraph)

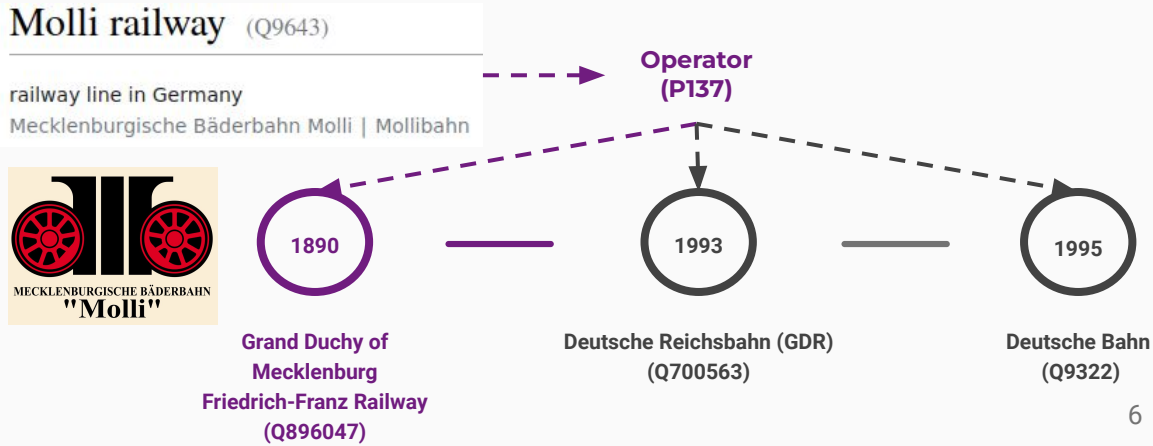
TKG [García-Durán A. et al., 2018]

- > 11k entities
- 150k time-related facts
- 508 - 2017 year scope
- multiple facts per entity → aggregation operator

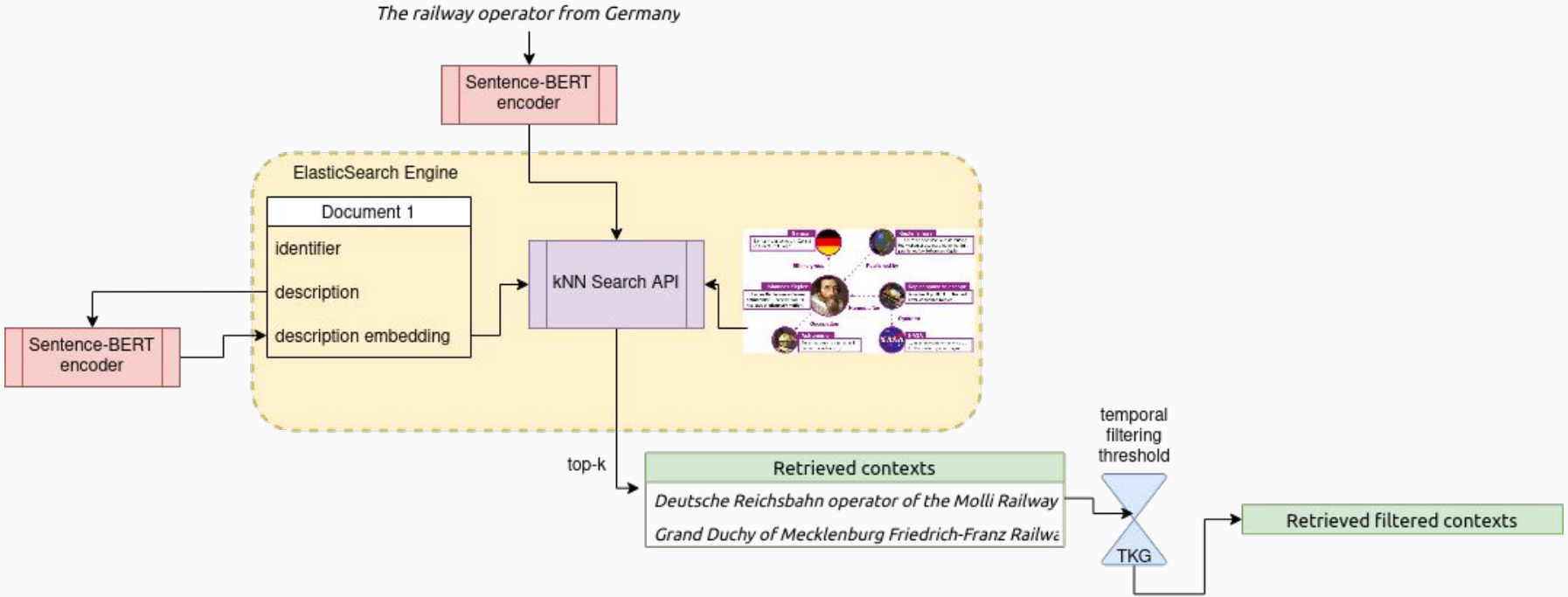
Wikidata5m



TKG



Context Retrieval



Experimental Setup

Historical Collections

newspapers [19C - 20C] (`hipe-2020`)

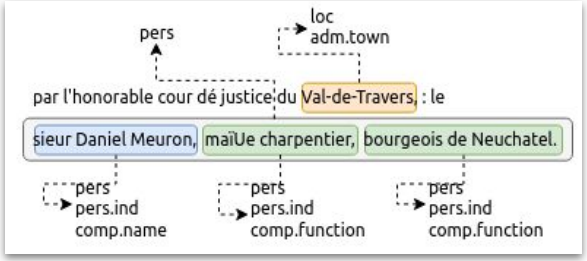
- Swiss, Luxembourgish & American newspapers
- ~ 20k NEs

classical commentaries [19C] (`ajmc`)

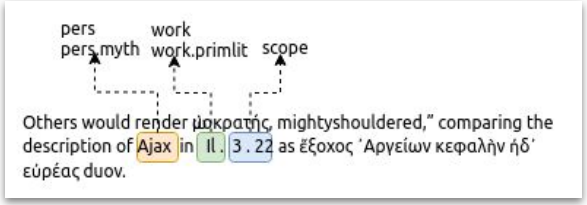
- Ajax Multi-Commentary project
- ~ 7.5K universal & domain-specific NEs

challenges: multilingualism, code-switching, high density NEs

`hipe-2020`



`ajmc`



	`hipe-2020`									`ajmc`								
	French			German			English			French			German			English		
Type	train	dev	test	train	dev	test	train	dev	test	train	dev	test	train	dev	test	train	dev	test
LOC	3,089	774	854	1,740	588	595	–	384	181	15	0	9	31	10	2	39	3	3
ORG	836	159	130	358	164	130	–	118	76	–	–	–	–	–	–	–	–	–
PERS	2,525	679	502	1,166	372	311	–	402	156	577	123	139	620	162	128	618	130	96
PROD	200	49	61	112	49	62	–	33	19	–	–	–	–	–	–	–	–	–
TIME	276	68	53	118	69	49	–	29	17	2	0	3	2	0	0	12	5	3
WORK	–	–	–	–	–	–	–	–	–	378	99	80	321	70	74	467	116	95
OBJECT	–	–	–	–	–	–	–	–	–	10	0	0	6	4	2	3	0	0
SCOPE	–	–	–	–	–	–	–	–	–	639	169	129	758	157	176	684	162	151

Table 1. Overview of the `hipe-2020` and `ajmc` datasets. LOC = Location, ORG = Organization, PERS = Person, PROD = Product, TIME = Time, WORK = human work, OBJECT = physical object, and SCOPE = specific portion of work.

Configurations & evaluation

configurations:

- `no-context`: model with no extra contexts
- `non-temporal`: *context jokers* integration with no time-related information
- `temporal-(10|25|50)`: *context jokers* integration with different year interval thresholds

- micro level precision (**P**), recall (**R**) & F-measure (**F1**)
- strict (***CS***) & fuzzy (***CF***) boundary matching

	French						German						English					
	hipe-2020			ajmc			hipe-2020			ajmc			hipe-2020			ajmc		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
no-context																		
CS	0.755	0.757	0.756	0.829	0.806	0.817	0.754	0.730	0.742	0.910	0.877	0.893	0.604	0.563	0.583	0.789	0.859	0.823
CF	0.857	0.859	0.858	0.883	0.858	0.870	**0.853**	0.826	0.839	0.935	0.901	0.917	0.778	0.726	0.751	0.855	0.931	0.891
non-temporal																		
CS	0.762	**0.767**	**0.765**	0.829	0.783	0.806	0.759	**0.767**	**0.763**	**0.930**	0.898	0.913	0.565	0.601	0.583	0.828	0.871	0.849
CF	0.862	**0.869**	0.866	**0.906**	0.856	0.880	0.847	0.856	0.852	**0.949**	0.916	**0.932**	0.741	0.788	0.764	0.885	0.931	0.908
temporal-50																		
CS	**0.765**	0.765	**0.765**	0.839	0.822	0.830	0.748	0.756	0.752	0.921	**0.911**	**0.916**	**0.643**	0.617	**0.630**	0.855	0.882	0.868
CF	**0.867**	0.867	**0.867**	0.901	0.883	0.892	0.833	0.842	0.838	0.937	**0.927**	**0.932**	**0.794**	0.762	0.777	0.916	**0.945**	0.931
temporal-25																		
CS	0.759	0.756	0.757	**0.848**	**0.839**	**0.844**	0.757	0.743	0.750	0.925	0.903	0.914	0.621	0.630	0.625	0.833	0.876	0.854
CF	0.863	0.859	0.861	0.902	**0.892**	**0.897**	0.852	0.835	0.843	0.938	0.916	0.927	0.787	0.800	**0.793**	0.893	0.940	0.916
temporal-10																		
CS	0.762	0.764	0.763	**0.848**	**0.839**	**0.844**	**0.760**	0.765	0.762	0.917	0.898	0.907	0.605	**0.646**	0.625	**0.866**	**0.888**	**0.877**
CF	0.863	0.866	0.865	0.902	**0.892**	**0.897**	0.852	**0.857**	**0.854**	0.936	0.916	0.926	0.760	**0.811**	0.784	**0.922**	**0.945**	**0.933**
L3i@HIPE-2022																		
CS	0.782	0.827	0.804	0.810	0.842	0.826	0.780	0.787	0.784	0.946	0.921	0.934	0.624	0.617	0.620	0.824	0.876	0.850
CF	0.883	0.933	0.907	0.856	0.889	0.872	0.870	0.878	0.874	0.965	0.940	0.952	0.793	0.784	0.788	0.868	0.922	0.894

Table 2. Results on French, German and English, for the `hipe-2020` and `ajmc` datasets.

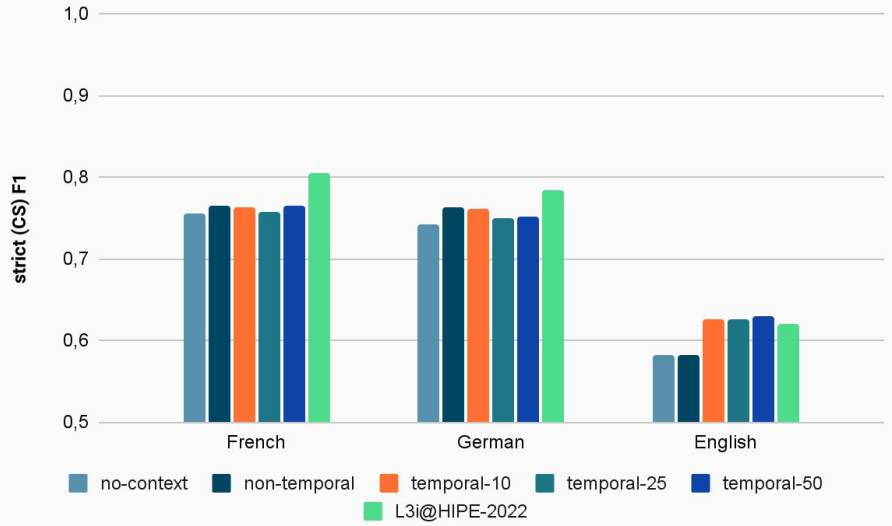
Configurations & evaluation

configurations:

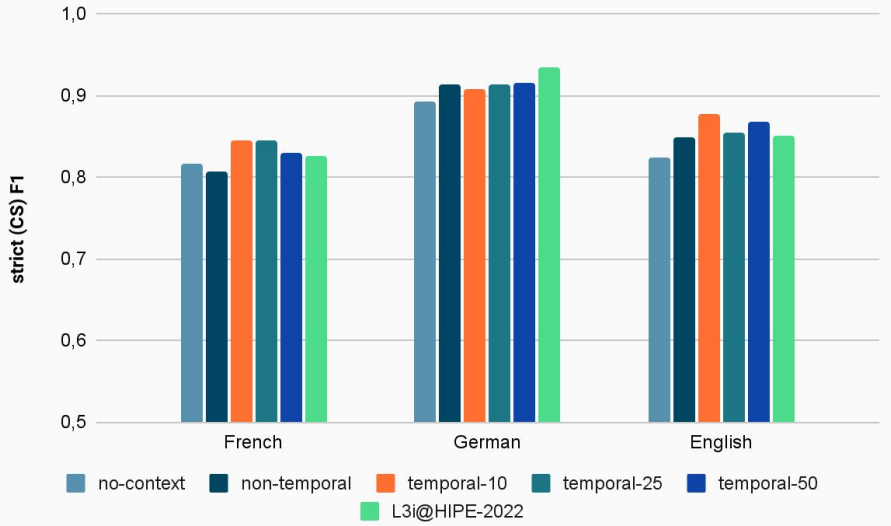
- `no-context`: model with no extra contexts
- `non-temporal`: *context jokers* integration with no time-related information
- `temporal-(10|25|50)`: *context jokers* integration with different year interval thresholds

- micro level precision (**P**), recall (**R**) & F-measure (**F1**)
- strict (***CS***) & fuzzy (***CF***) boundary matching

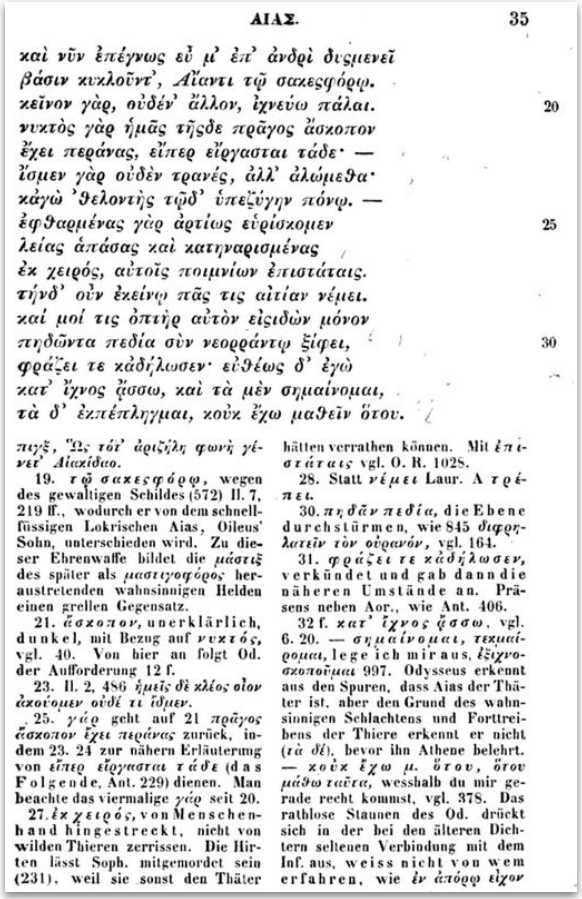
`hipe-2020`



`ajmc`



Impact of Digitization Errors (`ajmc`)



ΑΙΑΣ. 35

καὶ νῦν ἐπέγνως εὖ μ' ἐπ' ἀνδρὶ δυσμενεῖ
βάσιν κυκλοῦντ', Αἴαντι τῷ σακεσφόρῳ.
κεῖνον γὰρ, οὐδέν' ἄλλον, ἰχνεύω πάλαι.
νυκτὸς γὰρ ἡμᾶς τῆσδε πρᾶγος ἄσκοπον
ἔχει περάνας, εἴπερ εἴργασται τάδε· —
ἴσμεν γὰρ οὐδὲν τρανές, ἀλλ' ἀλώμεθα·
κἀγὼ 'θελοντὴς τῷδ' ὑπεξύγην πόνῳ. —
ἐφθαρμένας γὰρ ἀρτίως εὑρίσκομεν
λείας ἁπάσας καὶ κατηριθμένας
ἐκ χειρός, αὐτοῖς ποιμνίων ἐπιστάταις.
τήνδ' οὖν ἐκείνῳ πᾶς τις αἰτίαν νέμει.
καί μοί τις ὀπτὴρ αὐτὸν εἰσιδὼν μόνον
πηδῶντα πεδία σὺν νεορράντῳ ξίφει,
φράζει τε κἀδήλωσεν· εὐθέως δ' ἐγὼ
κατ' ἴχνος ᾖσσω, καὶ τὰ μὲν σημαίνομαι,
τὰ δ' ἐκπέπλημαι, κοὐκ ἔχω μαθεῖν ὅτου.

πιχξ, Ὡς τόϊ' ἀριζήλη φωνὴ γένετ' Αἰακίδαο.

19. τῷ σακεσφόρῳ, wegen des gewaltigen Schildes (572) Il. 7, 219 ff., wodurch er von dem schnellfüssigen Lokrischen Aias, Oileus' Sohn, unterschieden wird. Zu dieser Ehrenwaffe bildet die μάστιξ des später als μαστιγοφόρος heraustretenden wahnsinnigen Helden einen grellen Gegensatz.

21. ἄσκοπον, unerklärlich, dunkel, mit Bezug auf νυκτός, vgl. 40. Von hier an folgt Od. der Aufforderung 12 f.

23. Il. 2, 486 ἡμεῖς δὲ κλέος οἷον ἀκούομεν οὐδέ τι ἴδμεν.

25. γάρ geht auf 21 πρᾶγος ἄσκοπον ἔχει περάνας zurück, indem 23. 24 zur nähern Erläuterung von εἴπερ εἴργασται τάδε (das Folgende, Ant. 229) dienen. Man beachte das viermalige γάρ seit 20.

27. ἐκ χειρός, von Menschenhand hingestreckt, nicht von wilden Thieren zerrissen. Die Hirten lässt Soph. mitgemordet sein (231), weil sie sonst den Thäter hätten verrathen können. Mit ἐπιστάταις vgl. O. R. 1028.

28. Statt νέμει Laur. A τρέπει.

30. πηδᾶν πεδία, die Ebene durchstürmen, wie 845 διαρηλατεῖν τὸν οὐρανόν, vgl. 164.

31. φράζει τε κἀδήλωσεν, verkündet und gab dann die näheren Umstände an. Präsens neben Aor., wie Ant. 406.

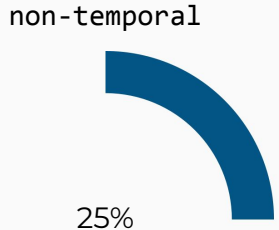
32 f. κατ' ἴχνος ᾖσσω, vgl. 6. 20. — σημαίνομαι, τεκμαίρομαι, lege ich mir aus, ἐξιχνοσκοπῶμαι 997. Odysseus erkennt aus den Spuren, dass Aias der Thäter ist, aber den Grund des wahnsinnigen Schlachtens und Forttreibens der Thiere erkennt er nicht (τὰ δέ), bevor ihn Athene belehrt. — κοὐκ ἔχω μ. ὅτου, ὅτου μάθω ταῦτα, wesshalb du mir gerade recht kommst, vgl. 378. Das rathlose Staunen des Od. drückt sich in der bei den älteren Dichtern seltenen Verbindung mit dem Inf. aus, weiss nicht von wem erfahren, wie ἐν ἀπόρῳ εἶχον

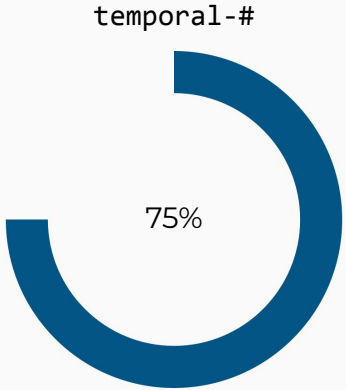
NEs affected by OCR:

- 10% German, English
- 27.5% French

observed improvements:

- clean NEs: `temporal-#` ⬆ 2% vs `non-temporal`
- noisy NEs: `temporal-#` ⬆ 14% vs `non-temporal`
- character error rate <67%:





In Summary...

“Rise of Digitization” + (NLP + IR) → semantically enriched archives for the Humanities and Social Sciences

In a nutshell

NER on historical collections with semantically relevant contexts & temporal information

contexts: mean-pooled representations in a Transformer-based model 

temporality: collection’s metadata & temporal knowledge graphs

Findings & ideas

temporality boosts NER when training data is missing

temporality helps on recognizing entities with digitation errors (to a certain extent)

short time spans → better for collections with restrained entity diversity & narrow year intervals

pertinence of contexts is dependent of time-related metadata & knowledge base... predicting year spans of big knowledge bases?

Merci !

Carlos-Emiliano
González-Gallardo

carlos.gonzalez_gallardo@univ-lr.fr

Eugène Delacroix oil painting of promoting access to historical newspapers in the New Aquitaine region