Legal domain named entity recognition

Vasile Păiș

vasile@racai.ro

Research Institute for Artificial Intelligence, Romanian Academy

Overview

- ➢ Introduction
- ➢ Deep Learning Approach
- ➢ Corpus Creation
  - ○ The RELATE platform
  - ○ LegalNERo corpus
  - ○ MicroBloggingNERo corpus
- ➢ Data Augmentation
  - ○ The NL-Augmenter framework
- ➢ Model Training
- ➢ The SAROJ project

Named Entity Recognition

Entity = “something that exists by itself; something that is separate from other things; the existence of a thing as contrasted with its attributes”

(Merriam-Webster dictionary)

Named Entities = *objects (“entities”) from the real world that have a name*

-> persons, locations, organizations

-> time expressions, diseases, chemicals, laws, legal references, emails, bank accounts, currency, …

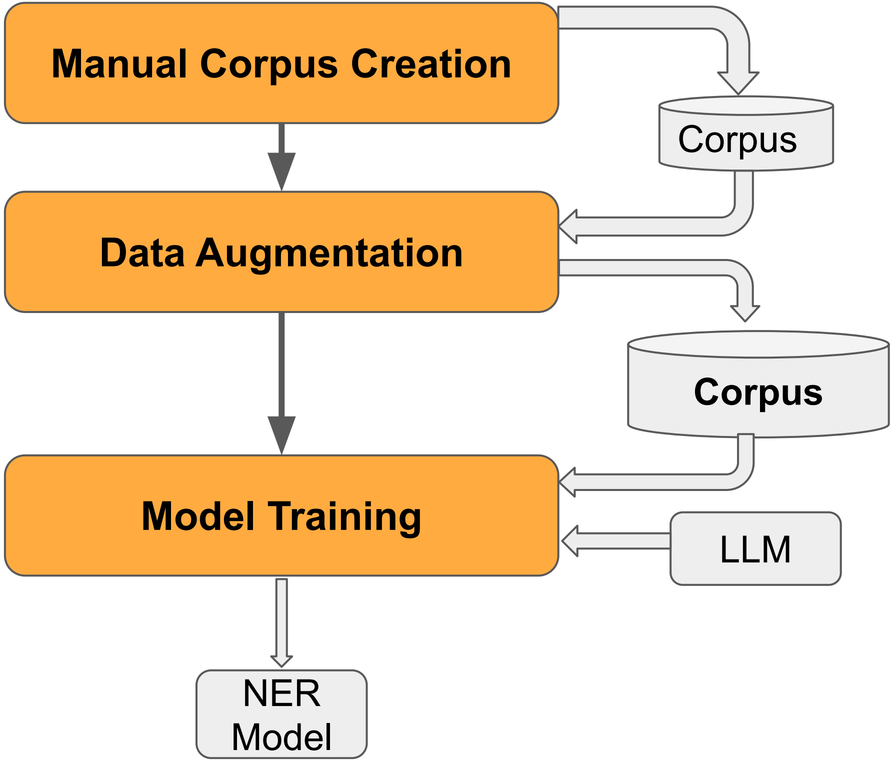
Why is it important ?

- Information extraction
  - Find information relevant to a set of entities
  - Extract text related to a particular product / brand / political figure etc.
- Content recommendation
  - Recommend content with the same NEs
- Customer support
  - Automatically show relevant information from different systems about identified NEs
- Anonymization
  - Anonymize persons, organizations, etc.

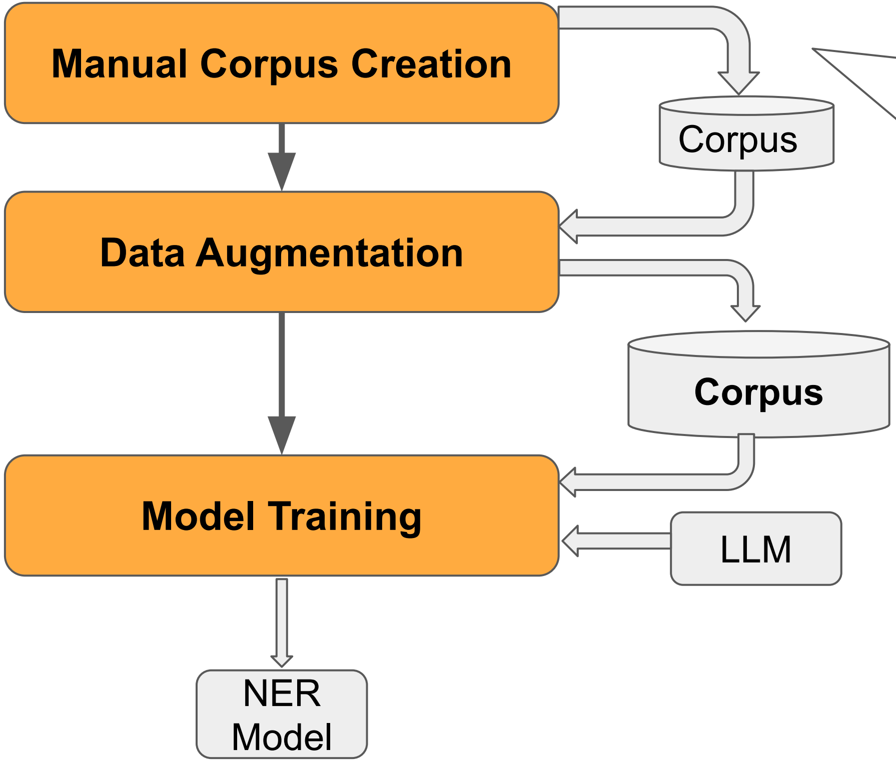
Approaches

- Lists of NEs
- Regular expressions
  - Very good for certain entities: emails/bank accounts/personal id…
- Handcrafted rules
- Conditional Random Fields
- Simple neural networks
- Deep Learning approaches
  - Large language models
  - Large annotated datasets
  - Deep neural networks

Deep Learning Approach

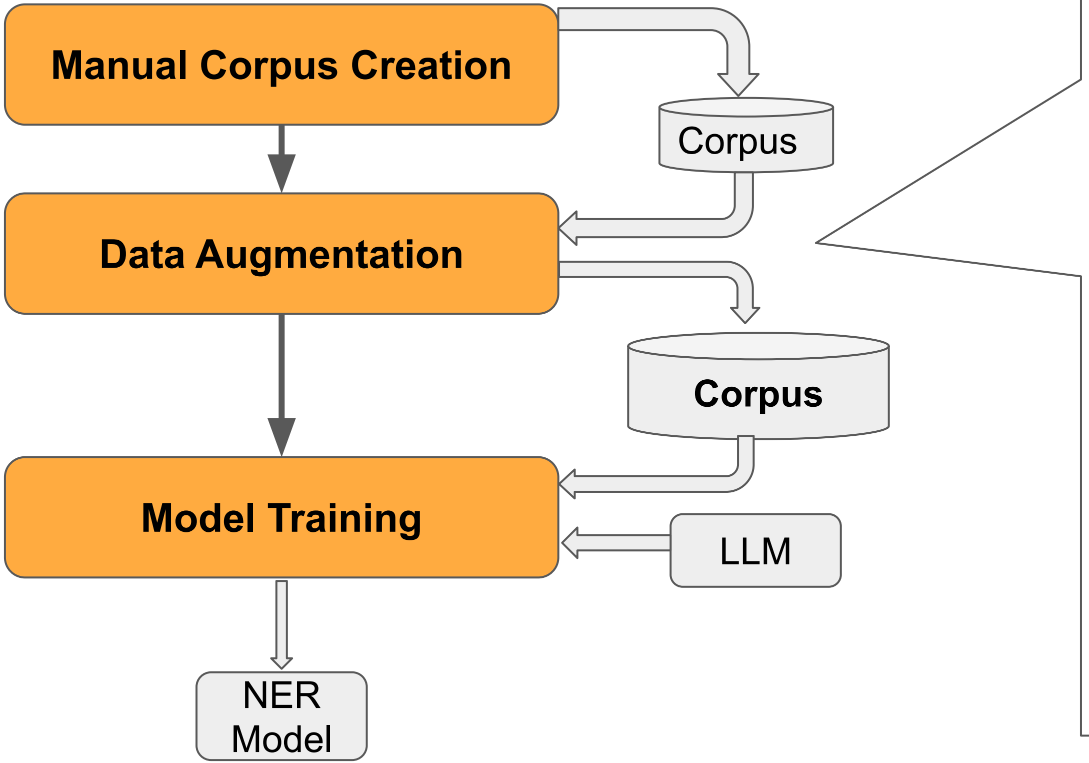


Deep Learning Approach



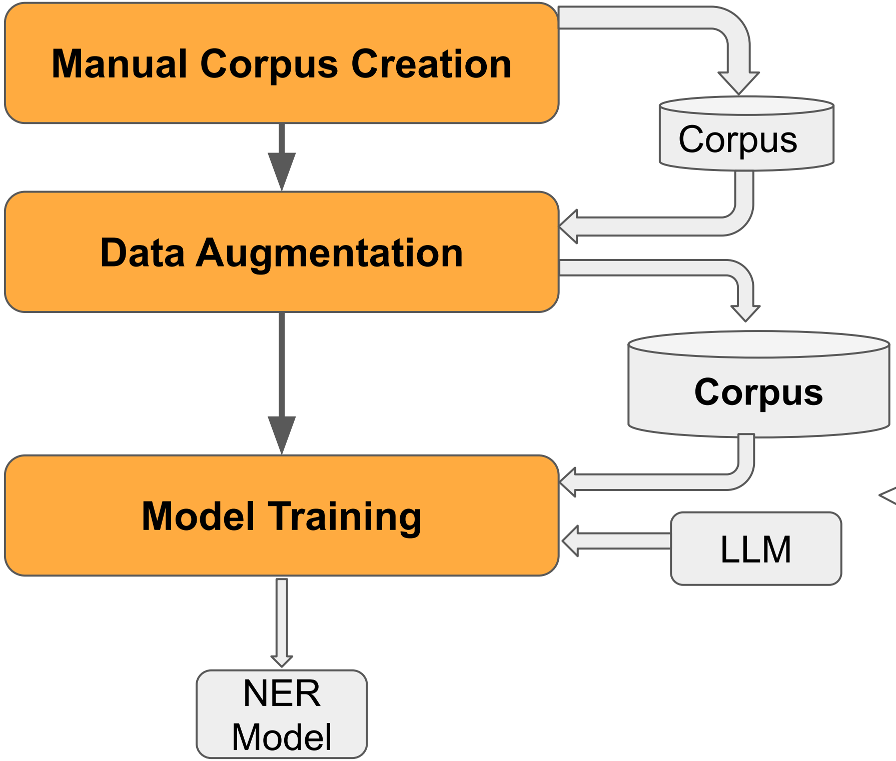
- ➢ Determine the relevant set of entities
- ➢ Manual annotation of text with the relevant entity types
  - ○ Expert annotators (domain experts + annotation experts)
  - ○ Inexperienced annotators (students)
- ➢ Compute corpus statistics, inter-annotator agreement, etc.
- ➢ **The RELATE platform**

Deep Learning Approach



- Construct additional sentences
  - Simple perturbations (typing errors, uppercase/lowercase, etc.)
  - Generative AI methods (paraphrasing, sentence completion, etc.)
- Compute statistics on the resulting corpus
- **The NL-Augmenter framework**

Deep Learning Approach



- ➢ Use the corpus to train a NER model
- ➢ Use a pre-trained LLM to augment the model’s performance

RELATE - **A platform for Romanian language technologies and resources**

Includes technologies developed at ICIA and by Partners in multiple projects:

CoRoLa, ReTeRom, ROBIN, Presidency, MARCELL, CURLICAT,

Enrich4All, ELE, USPDATRO

Follows ELG philosophy:

- web services, REST APIs, dockers when possible
- services can be distributed across different physical nodes
- services can be consumed directly from partners

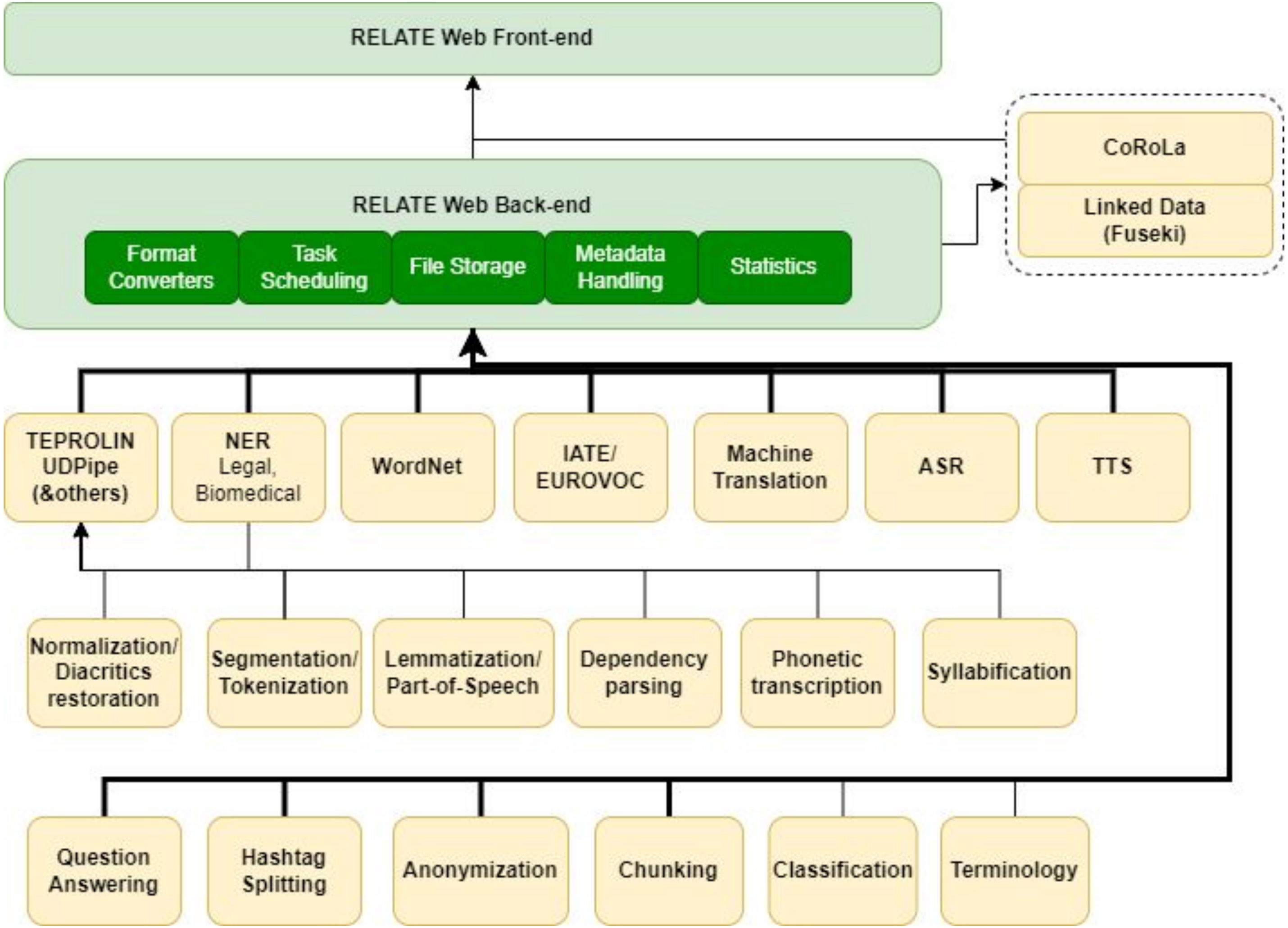
https://relate.racai.ro
https://github.com/racai-ai/RELATE

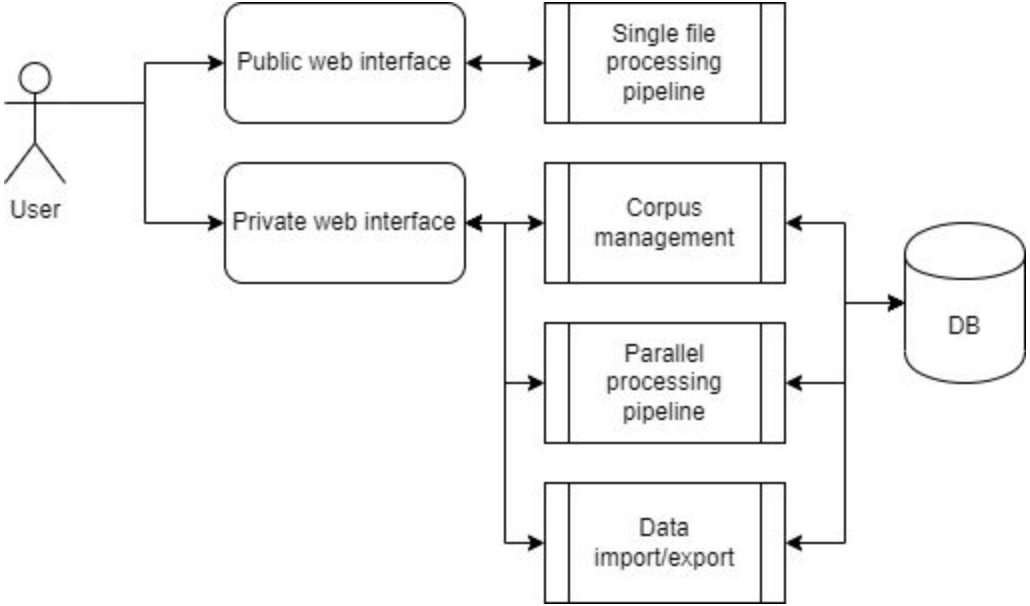
RELATE - overview papers

Vasile Păiș, Radu Ion, and Dan Tufiș. “**A Processing Platform Relating Data and Tools for Romanian Language**”. English. In:Proceedings of the 1st International Workshop on Language Technology Platforms. Marseille, France: European Language Resources Association, 2020, pp. 81–88. URL:https://www.aclweb.org/anthology/2020.iwltp-1.13

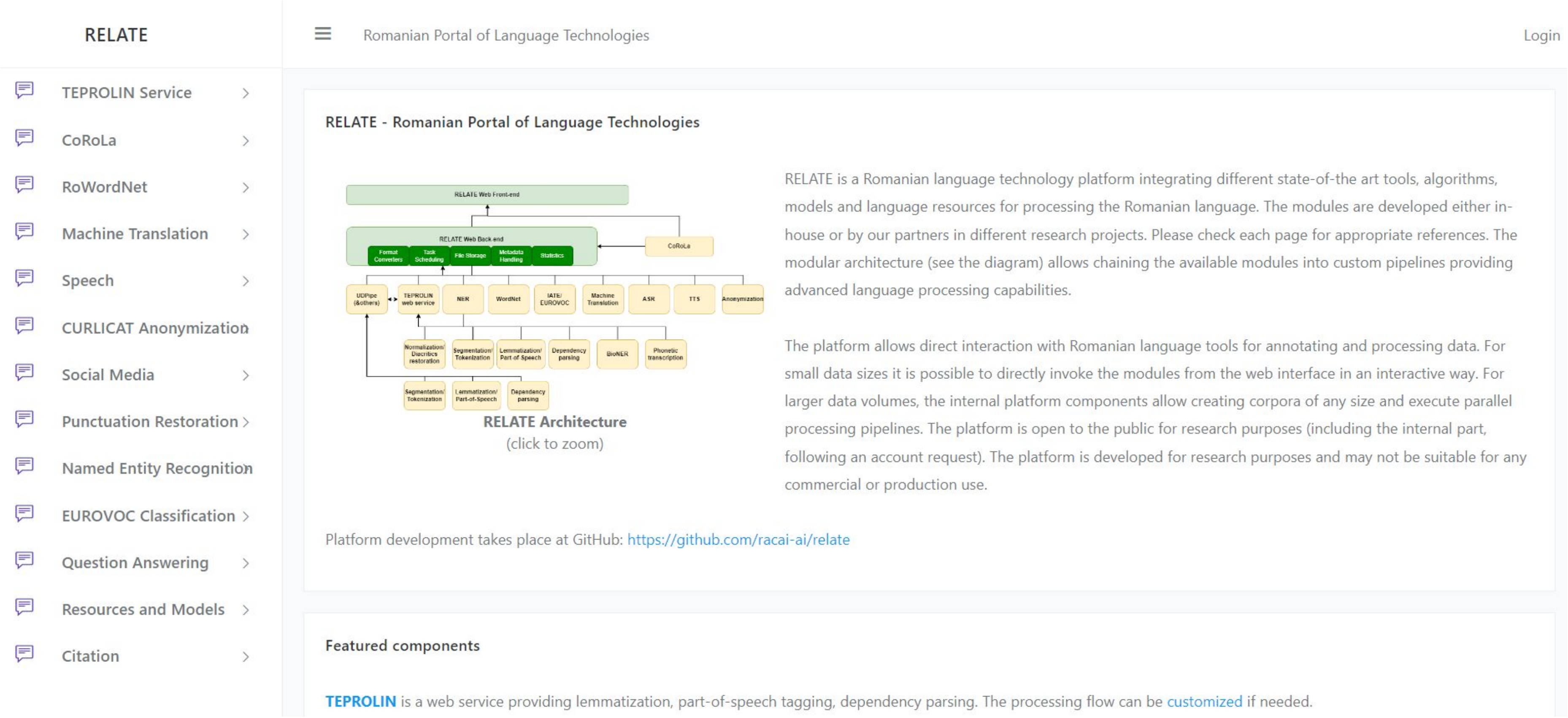
Vasile Păiș. “**Multiple annotation pipelines inside the RELATE platform**”. In:The 15th International Conference on Linguistic Resources and Tools for Natural Language Processing. 2020, pp. 65–75. URL:https://profs.info.uaic.ro/~consilr/wp-content/uploads/2021/03/volum-ConsILR-v-4-final-revizuit.pdf#page=73 .

Vasile Păiș, Dan Tufiș, and Radu Ion. “**Integration of Romanian NLP tools into the RELATE platform**”. In:International Conference on Linguistic Resources and Tools for Natural Language Processing. 2019, pp. 181–192. URL: https://profs.info.uaic.ro/~consilr/2019/wp-content/uploads/2020/01/ConsILR2019_final_BTT-60-ex-B5.pdf#page=189 .


RELATE Web Front-end
CoRoLa
Linked Data (Fuseki)
RELATE Web Back-end
Format Converters
Task Scheduling
File Storage
Metadata Handling
Statistics
TEPROLIN UDPipe (&others)
NER Legal, Biomedical
WordNet
IATE/ EUROVOC
Machine Translation
ASR
TTS
Normalization/ Diacritics restoration
Segmentation/ Tokenization
Lemmatization/ Part-of-Speech
Dependency parsing
Phonetic transcription
Syllabification
Question Answering
Hashtag Splitting
Anonymization
Chunking
Classification
Terminology


User
Public web interface
Private web interface
Single file processing pipeline
Corpus management
Parallel processing pipeline
Data import/export
DB

RELATE - public view



RELATE - Romanian Portal of Language Technologies

RELATE Architecture

(click to zoom)

RELATE is a Romanian language technology platform integrating different state-of-the art tools, algorithms, models and language resources for processing the Romanian language. The modules are developed either in-house or by our partners in different research projects. Please check each page for appropriate references. The modular architecture (see the diagram) allows chaining the available modules into custom pipelines providing advanced language processing capabilities.

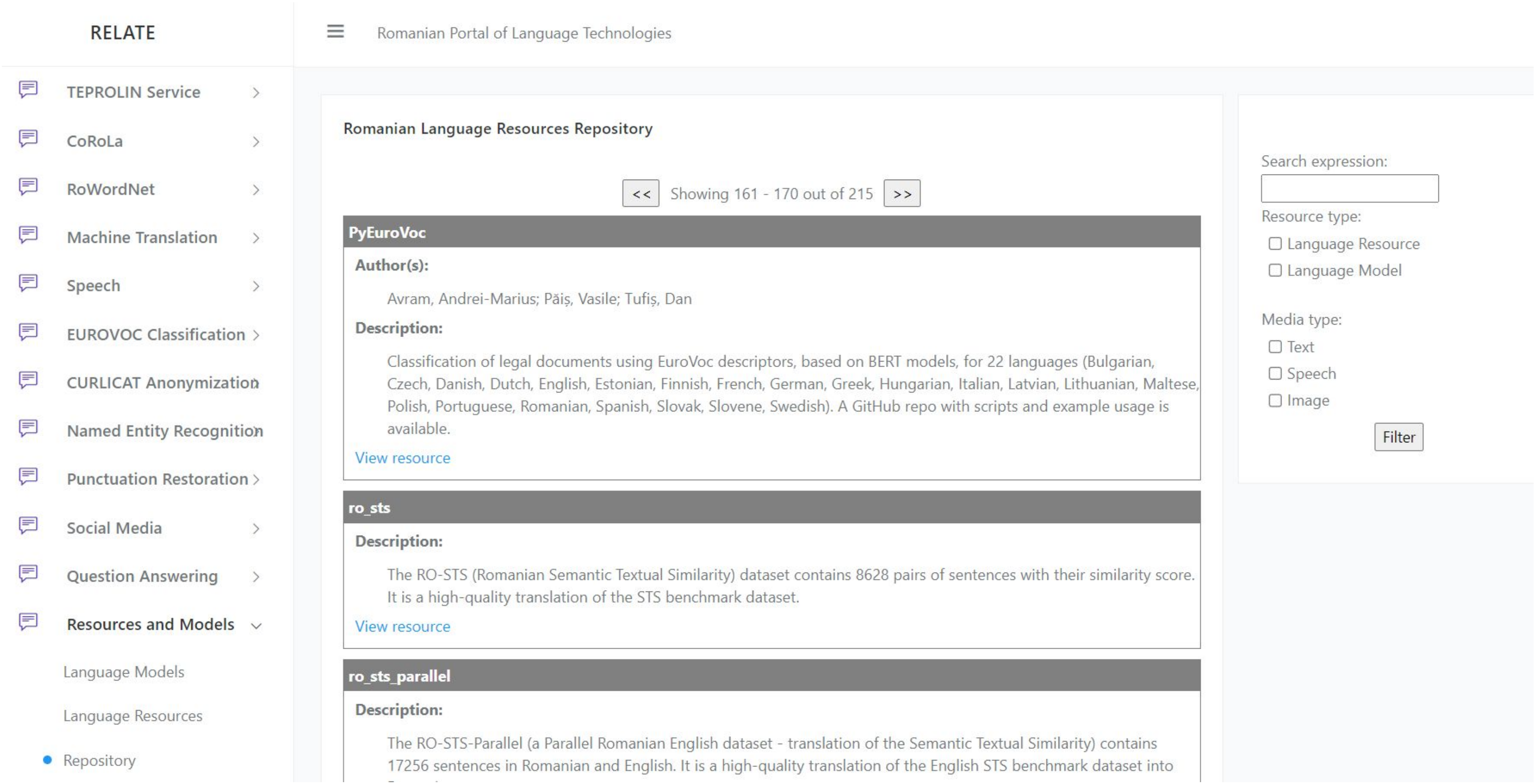
The platform allows direct interaction with Romanian language tools for annotating and processing data. For small data sizes it is possible to directly invoke the modules from the web interface in an interactive way. For larger data volumes, the internal platform components allow creating corpora of any size and execute parallel processing pipelines. The platform is open to the public for research purposes (including the internal part, following an account request). The platform is developed for research purposes and may not be suitable for any commercial or production use.

Platform development takes place at GitHub: https://github.com/racai-ai/relate

Featured components

TEPROLIN is a web service providing lemmatization, part-of-speech tagging, dependency parsing. The processing flow can be customized if needed.

Language resources and pre-trained models



Romanian Language Resources Repository

Showing 161 - 170 out of 215

PyEuroVoc

Author(s):

Avram, Andrei-Marius; Păiș, Vasile; Tufiș, Dan

Description:

Classification of legal documents using EuroVoc descriptors, based on BERT models, for 22 languages (Bulgarian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Spanish, Slovak, Slovene, Swedish). A GitHub repo with scripts and example usage is available.

View resource

ro_sts

Description:

The RO-STS (Romanian Semantic Textual Similarity) dataset contains 8628 pairs of sentences with their similarity score. It is a high-quality translation of the STS benchmark dataset.

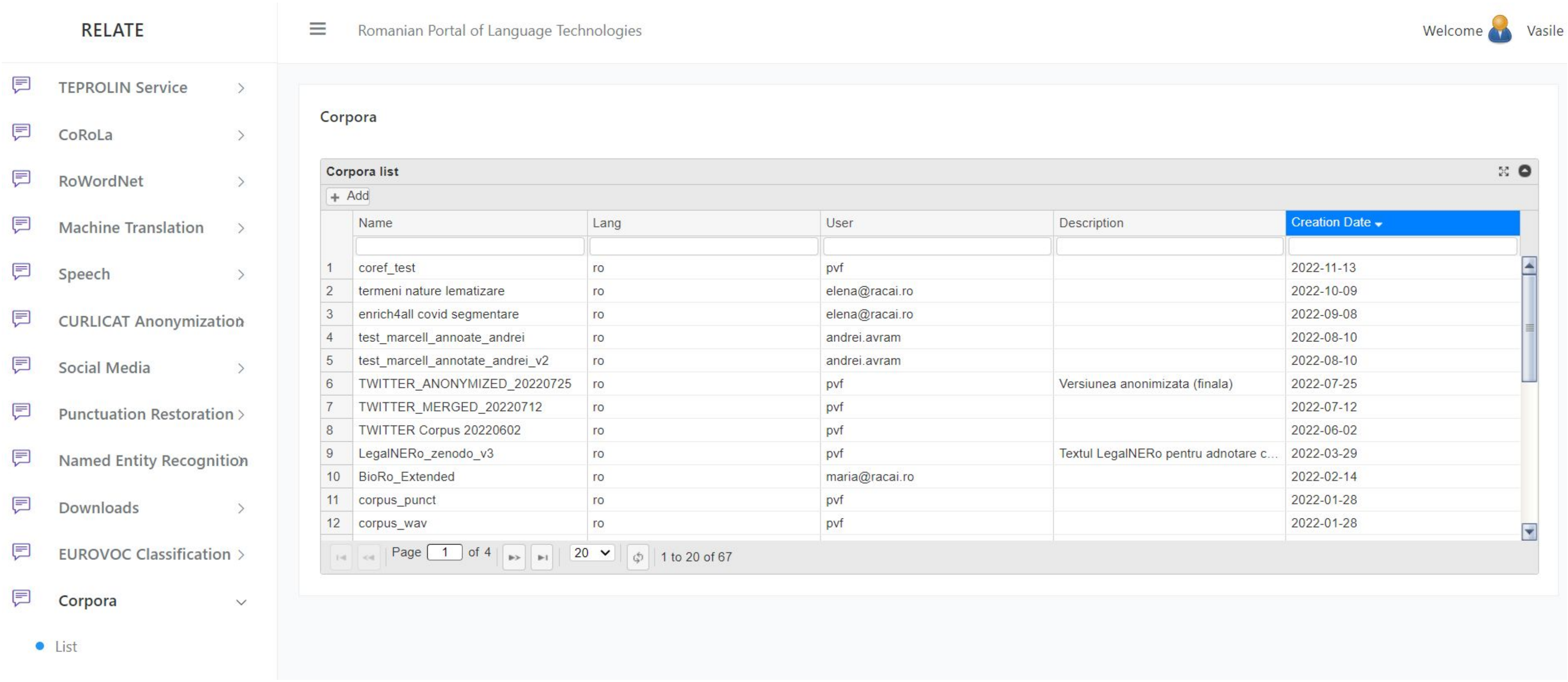
View resource

ro_sts_parallel

Description:

The RO-STS-Parallel (a Parallel Romanian English dataset - translation of the Semantic Textual Similarity) contains 17256 sentences in Romanian and English. It is a high-quality translation of the English STS benchmark dataset into

RELATE - authenticated view



Corpora

	Name	Lang	User	Description	Creation Date
1	coref_test	ro	pvf		2022-11-13
2	termeni nature lematizare	ro	elena@racai.ro		2022-10-09
3	enrich4all covid segmentare	ro	elena@racai.ro		2022-09-08
4	test_marcell_annoate_andrei	ro	andrei.avram		2022-08-10
5	test_marcell_annotate_andrei_v2	ro	andrei.avram		2022-08-10
6	TWITTER_ANONYMIZED_20220725	ro	pvf	Versiunea anonimizata (finala)	2022-07-25
7	TWITTER_MERGED_20220712	ro	pvf		2022-07-12
8	TWITTER Corpus 20220602	ro	pvf		2022-06-02
9	LegalNERo_zenodo_v3	ro	pvf	Textul LegalNERo pentru adnotare c…	2022-03-29
10	BioRo_Extended	ro	maria@racai.ro		2022-02-14
11	corpus_punct	ro	pvf		2022-01-28
12	corpus_wav	ro	pvf		2022-01-28

Page 1 of 4 20 1 to 20 of 67

RELATE - Large corpora

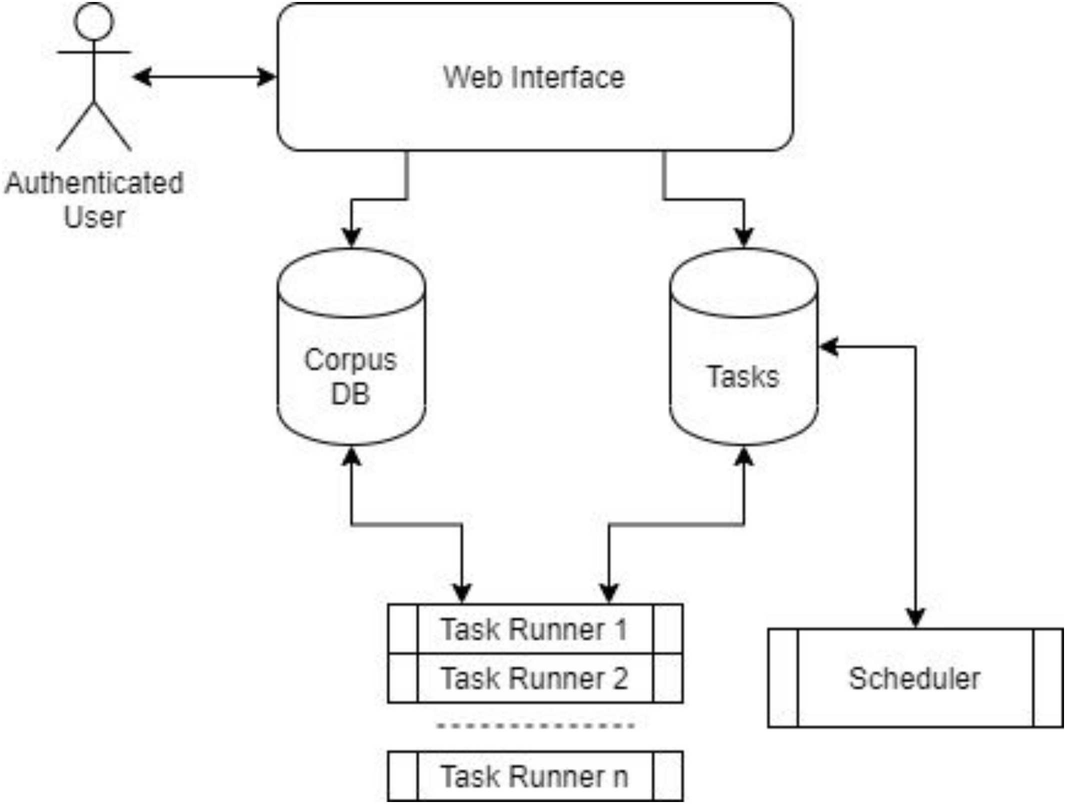
Efficiently stores and handles large corpora:

- text only
- text + audio
- text + PDF
  - + metadata
  - in the future => other types of bimodal/multimodal data

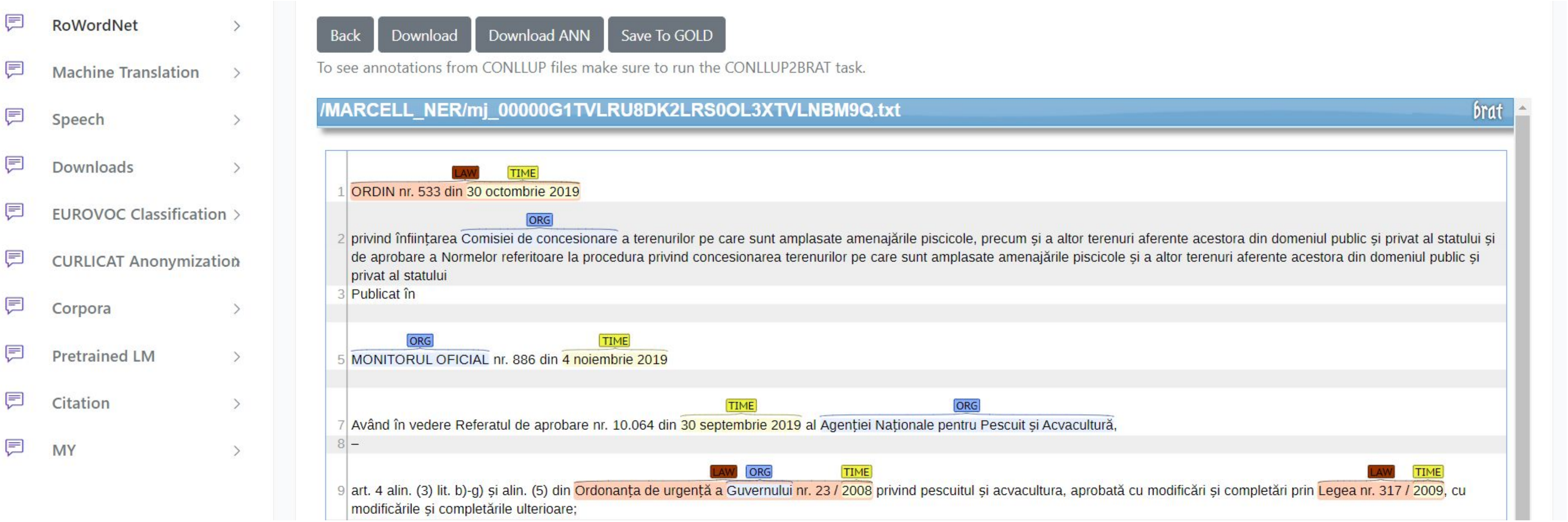
Parallel processing

- task scheduling
- services from multiple nodes

RELATE Task-based processing



Gold corpora creation - NER - LegalNERo



To see annotations from CONLLUP files make sure to run the CONLLUP2BRAT task.

/MARCELL_NER/mj_00000G1TVLRU8DK2LRS0OL3XTVLNBM9Q.txt

ORDIN nr. 533 din 30 octombrie 2019

privind înființarea Comisiei de concesionare a terenurilor pe care sunt amplasate amenajările piscicole, precum și a altor terenuri aferente acestora din domeniul public și privat al statului și de aprobare a Normelor referitoare la procedura privind concesionarea terenurilor pe care sunt amplasate amenajările piscicole și a altor terenuri aferente acestora din domeniul public și privat al statului

Publicat în

MONITORUL OFICIAL nr. 886 din 4 noiembrie 2019

Având în vedere Referatul de aprobare nr. 10.064 din 30 septembrie 2019 al Agenției Naționale pentru Pescuit și Acvacultură,

–

art. 4 alin. (3) lit. b)-g) și alin. (5) din Ordonanța de urgență a Guvernului nr. 23 / 2008 privind pescuitul și acvacultura, aprobată cu modificări și completări prin Legea nr. 317 / 2009, cu modificările și completările ulterioare;

- Păiș, Vasile and Mitrofan, Maria and Gasan, Carol Luca and Ianov, Alexandru and Ghiță, Corvin and Coneschi, Vlad Silviu and Onuț, Andrei (2021). *Romanian Named Entity Recognition in the Legal domain (LegalNERo)*. Zenodo, https://doi.org/10.5281/zenodo.4772094
- Păiș, Vasile and Mitrofan, Maria. (2021) *Towards a named entity recognition system in the Romanian legal domain using a linked open data corpus*. In Workshop on Deep Learning and Neural Approaches for Linguistic Data, pp. 16--17.
- Păiș, Vasile and Mitrofan, Maria and Gasan, Carol Luca and Coneschi, Vlad and Ianov, Alexandru (2021). *Named Entity Recognition in the Romanian Legal Domain*. In Proceedings of the Natural Legal Language Processing Workshop, pp. 9--18. https://aclanthology.org/2021.nllp-1.2

LegalNERo - Context

Multilingual Resources for CEF.AT in the legal domain (MARCELL)

- https://marcell-project.eu/
- large comparable corpus of legal documents for 7 languages
  - Bulgarian, Croatian, Hungarian, Polish, **Romanian**, Slovak, Slovenian
- existing NER system for Romanian
  - offered low precision (64.1%)
  - not aware of any specific entities present in legal documents

LegalNERo corpus

- Sub-corpus of the MARCELL-RO corpus
- 5 annotators supervised by 2 senior researchers, Cohen Kappa 0.89
- 5 entity classes: Person, Organization, Location, Time, Legal Reference
- Legal Reference
  - laws, government decisions, orders, …
  - similar to Jörg Landthaler, Bernhard Waltl, and Florian Matthes. *Unveiling references in legal texts-implicit versus explicit network structures*. In IRIS: Internationales Rechtsinformatik Symposium, volume 8, pages 71–8, 2016
  - similar to the coarse-grained class of Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. *Fine-grained named entity recognition in legal documents*. In Semantic Systems. The Power of AI and Knowledge Graphs,pages 272–287, Springer International Publish-ing, 2019.

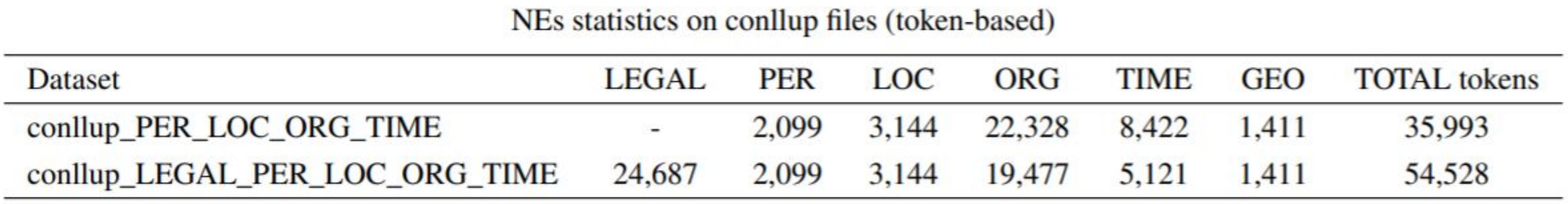
LegalNERo corpus

- Legal Reference
  - *“Legea nr. 13/2008”, “Ordinul nr. 625 din 25 aprilie 2019”, “Referatul de aprobare al Direcției relații cu presa nr. S8 4.536/4.04.2019”*
  - may contain Organizations and Time expressions
    - these were annotated, if present
  - resulted two ways to exploit the corpus
    - PER,LOC,ORG,TIME
    - PER,LOC,ORG,TIME,LEGAL_REF

LegalNERo corpus

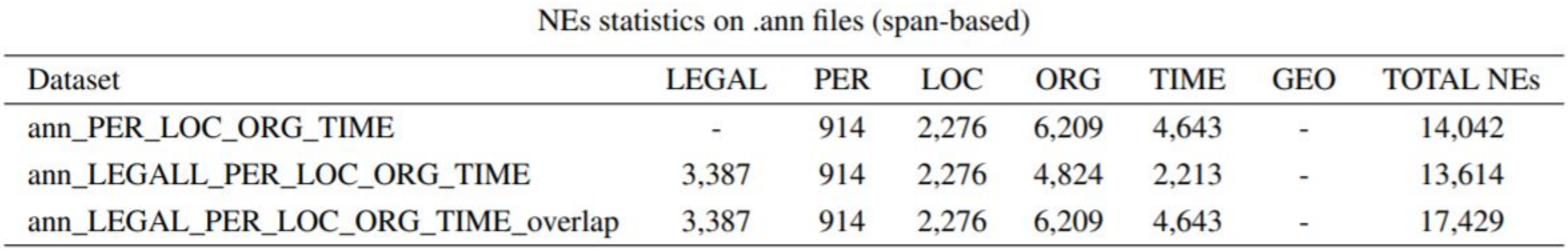
- Depending on the NER system being developed
  - span-based annotations
    - this is the format produced by annotation
  - token-based annotations
    - UDPipe was used for tokenization
    - span-based mapped to token-based
      - resulting in additional views over the corpus
- Locations were linked to GeoNames where possible

LegalNERo corpus



NEs statistics on conllup files (token-based)

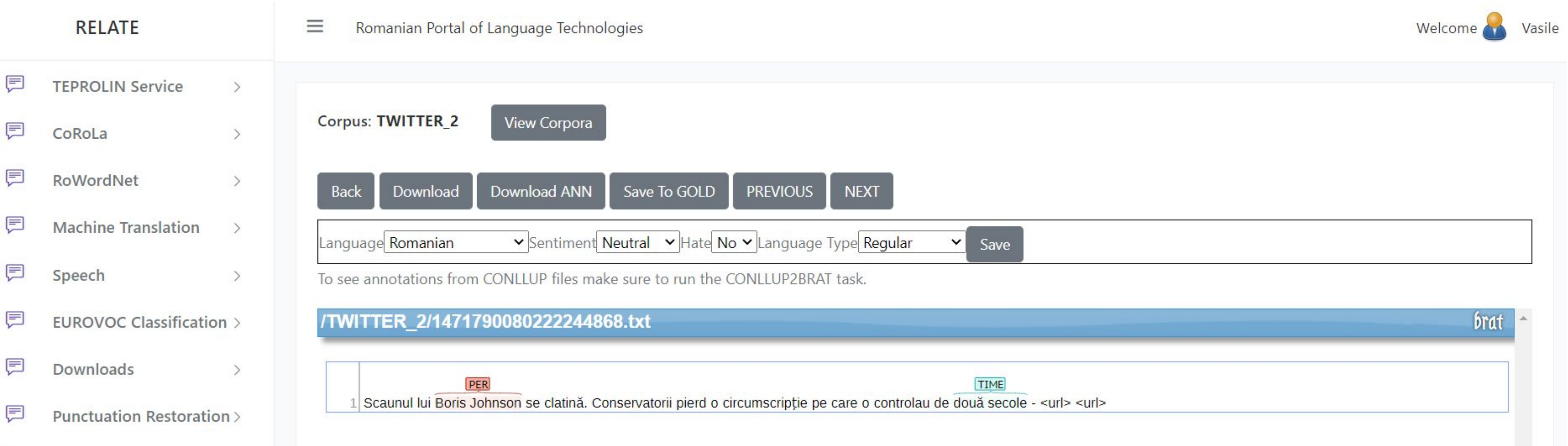
Dataset	LEGAL	PER	LOC	ORG	TIME	GEO	TOTAL tokens
conllup_PER_LOC_ORG_TIME	-	2,099	3,144	22,328	8,422	1,411	35,993
conllup_LEGAL_PER_LOC_ORG_TIME	24,687	2,099	3,144	19,477	5,121	1,411	54,528



NEs statistics on .ann files (span-based)

Dataset	LEGAL	PER	LOC	ORG	TIME	GEO	TOTAL NEs
ann_PER_LOC_ORG_TIME	-	914	2,276	6,209	4,643	-	14,042
ann_LEGALL_PER_LOC_ORG_TIME	3,387	914	2,276	4,824	2,213	-	13,614
ann_LEGAL_PER_LOC_ORG_TIME_overlap	3,387	914	2,276	6,209	4,643	-	17,429

Gold corpora creation - NER - MicroBloggingNERo



Corpus: TWITTER_2

Language Romanian Sentiment Neutral Hate No Language Type Regular

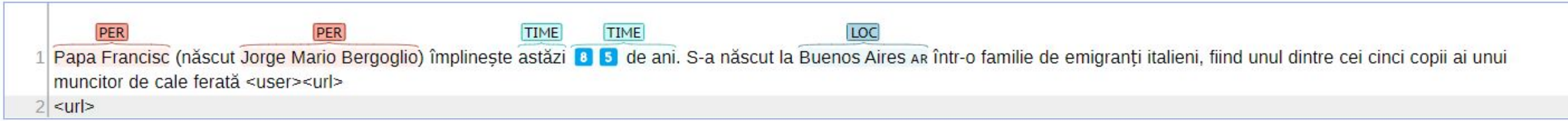
To see annotations from CONLLUP files make sure to run the CONLLUP2BRAT task.

/TWITTER_2/1471790080222244868.txt

1 Scaunul lui Boris Johnson se clatină. Conservatorii pierd o circumscripție pe care o controlau de două secole - <url> <url>

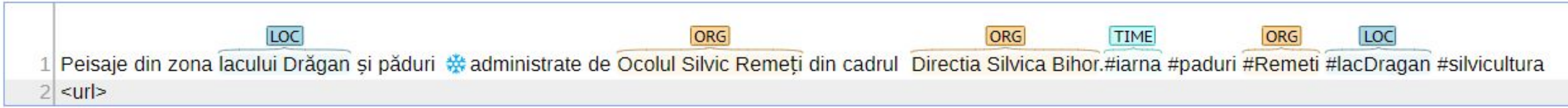
- Păiș, Vasile and Mitrofan, Maria and Barbu-Mititelu, Verginica and Irimia, Elena and Gasan, Carol Luca and Micu, Roxana and Marin, Laura and Dicusar, Maria and Florea, Bianca and Badila, Ana (2022). *Romanian micro-blogging named entity recognition (MicroBloggingNERo)*. Zenodo, https://doi.org/10.5281/zenodo.6905235
- Păiș, Vasile and Barbu Mititelu, Verginica and Irimia, Elena and Mitrofan, Maria and Gasan, Carol Luca and Micu, Roxana (2022). *Romanian micro-blogging named entity recognition including health-related entities*. In Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, pp. 190--196, https://aclanthology.org/2022.smm4h-1.49

Examples from MicroBloggingNERo



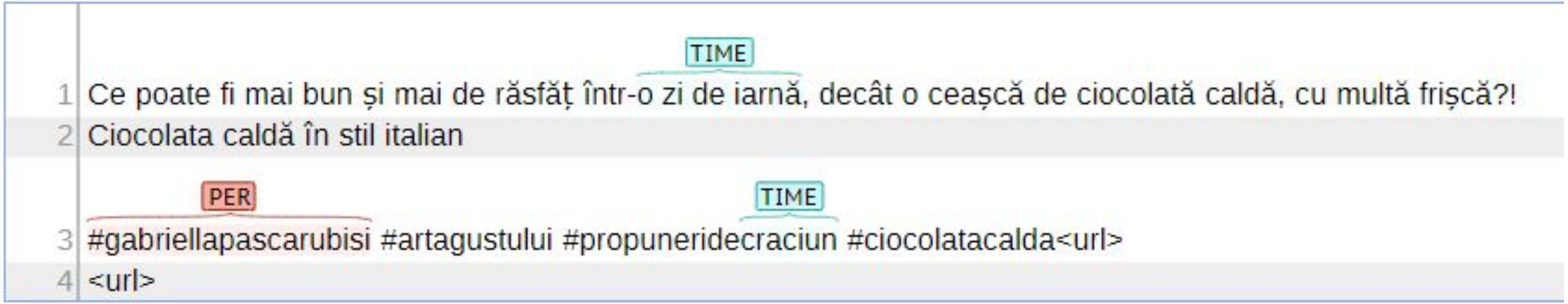
Papa Francisc (născut Jorge Mario Bergoglio) împlinește astăzi 8 5 de ani. S-a născut la Buenos Aires AR într-o familie de emigranți italieni, fiind unul dintre cei cinci copii ai unui muncitor de cale ferată <user><url>

<url>



Peisaje din zona lacului Drăgan și păduri ❄ administrate de Ocolul Silvic Remeți din cadrul Directia Silvica Bihor.#iarna #paduri #Remeti #lacDragan #silvicultura

<url>



Ce poate fi mai bun și mai de răsfăț într-o zi de iarnă, decât o ceașcă de ciocolată caldă, cu multă frișcă?!

Ciocolata caldă în stil italian

#gabriellapascarubisi #artagustului #propuneridecraciun #ciocolatacalda<url>

<url>

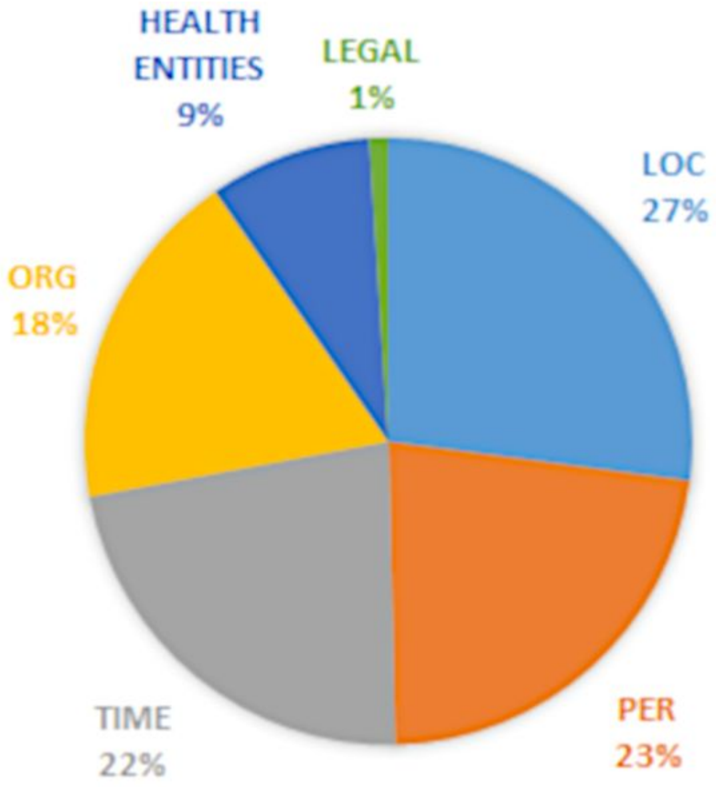
Păiș, Vasile and Mitrofan, Maria and Barbu Mititelu, Verginica and Irimia, Elena and Micu, Roxana and Gasan, Carol Luca. Challenges in Creating a Representative Corpus of Romanian Micro-Blogging Text. In Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-10). European Language Resources Association, Marseille, France, pp. 1--7, June 2022 https://aclanthology.org/2022.cmlc-1.1/

MicroBloggingNERo dataset

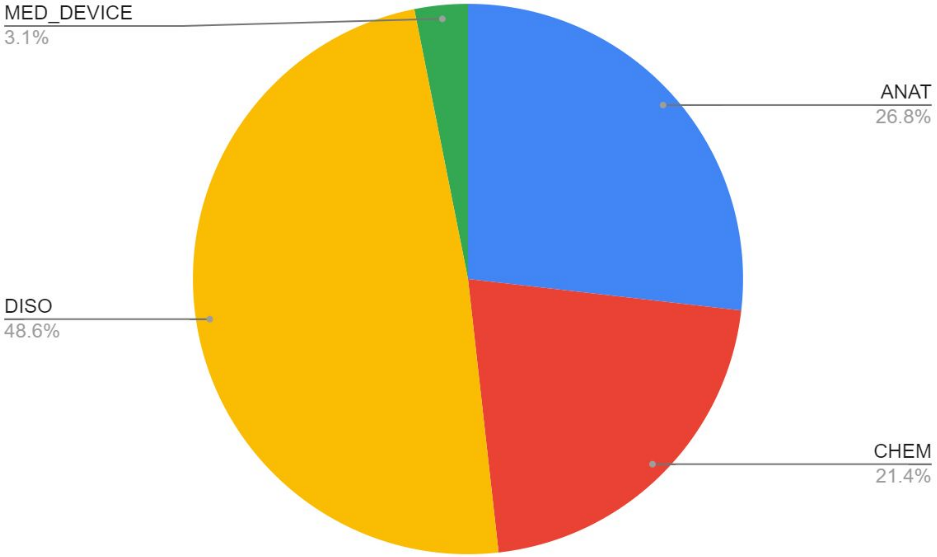
- Gathered via crawling
  - based on search API, using specific Romanian language queries
  - did not rely on platform language detection => wanted to gather code-mixed data as well
- Anonymized
  - removed platform-specific IDs: message id, user, url
  - names and specific locations replaced with pseudonims
- Annotation scheme adapted from:
  - SiMoNERo
    - medical texts (scientific books, journal articles and blog posts)
    - https://github.com/UniversalDependencies/UD_Romanian-SiMoNERo
  - LegalNERo
    - legal-domain texts https://doi.org/10.5281/zenodo.4772094
- https://doi.org/10.5281/zenodo.6905234

MicroBloggingNERo dataset

- 7,800 messages; 11,099 annotations
- 9 Entity classes
  - General:
    - Organization (ORG)
    - Person (PER)
    - Location (LOC)
    - Time expressions (TIME)
  - Legal references (LEGAL)
  - Health-related:
    - Anatomical parts (ANAT)
    - Chemical and drugs (CHEM)
    - Disorders (DISO)
    - Medical devices (MED_DEVICE)



MicroBloggingNERo dataset



ANAT	257
CHEM	205
DISO	466
MED_DEVICE	30

- 7 annotators
- Cohen Kappa 0.8
  - According to (Cohen, 1960), a value between 0.61–0.80 indicates substantial agreement between the annotators
- Health entities seemed more difficult to annotate

Data Augmentation

- Larger datasets (usually) produce better models
- Not always possible to manually annotate very large datasets

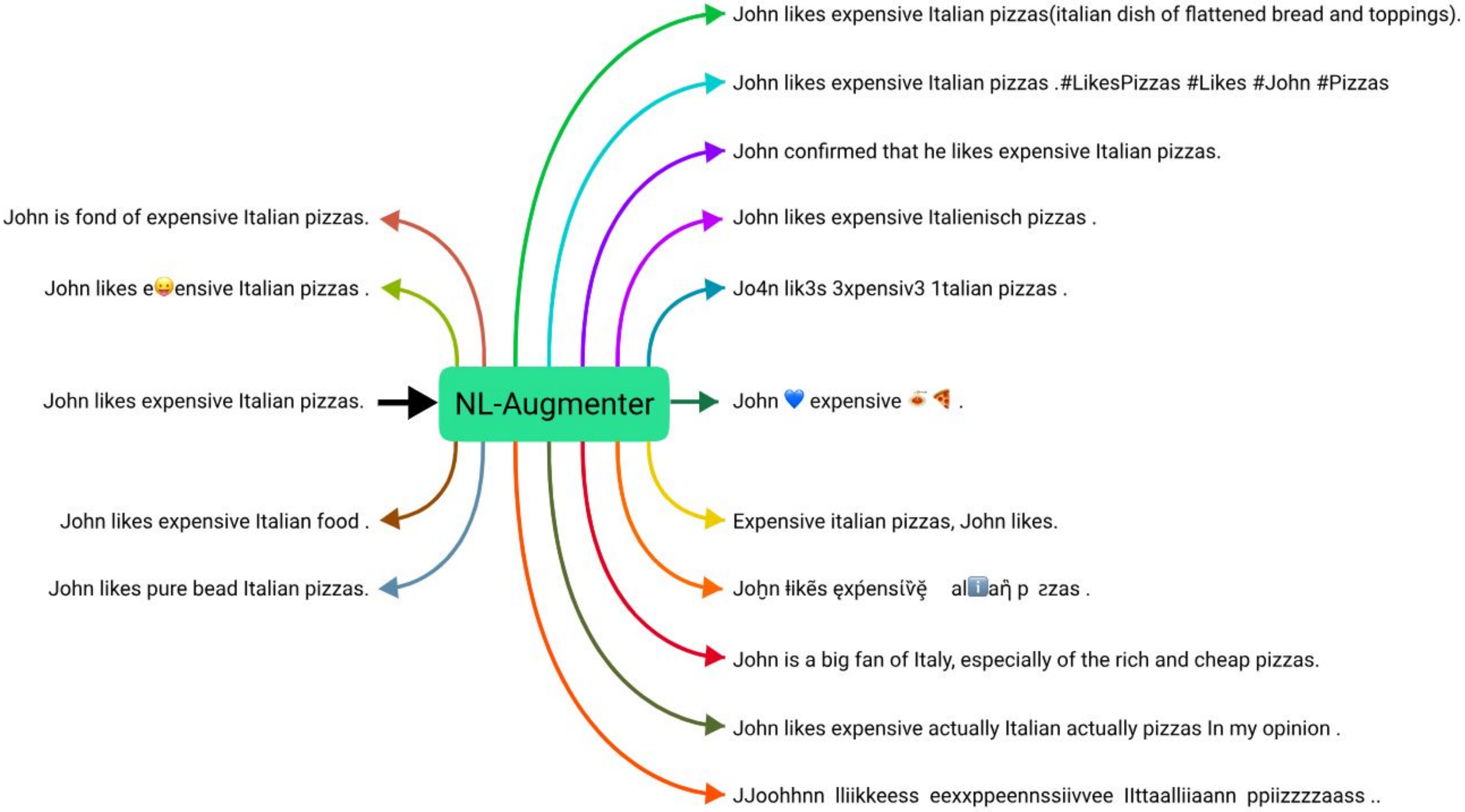
- Data Augmentation
  - Generate new sentences based on content already present in the dataset
  - Simple perturbations
    - Simulate typing mistakes, include social media specifics (hashtags, emojis), lowercase/uppercase, etc.
  - Generative AI approaches
    - Paraphrasing, sentence completion, machine translation, back-translation, etc.

NL-Augmenter framework https://github.com/GEM-benchmark/NL-Augmenter

Dhole, K., Gangal, V., Gehrmann, S., Gupta, A., Li, Z., Mahamood, S., Mahadiran, A., Mille, S., Shrivastava, A., Tan, S., Wu, T., Sohl-Dickstein, J., Choi, D.J., Hovy, E., Dušek, O., Ruder, S., Anand, S., Aneja, N., Banjade, R., Barthe, L., Behnke, H., Berlot-Attwell, I., Boyle, C., Brun, C., Cabezudo, S.A.M., Cahyawijaya, S., Chapuis, E., Che, W., Choudhary, M., Clauss, C., Colombo, P., Cornell, F., Dagan, G., Das, M., Dixit, T., Dopierre, T., Dray, P., Dubey, S., Ekeinhor, T., Giovanni, D.M., Goyal, T., Gupta, R., Hamla, L., Han, S., Harel-Canada, F., Honoré, A., Jindal, I., Joniak, K.P., Kleyko, D., Kovatchev, V., Krishna, K., Kumar, A., Langer, S., Lee, R.S., Levinson, J.C., Liang, H., Liang, K., Liu, Z., Lukyanenko, A., Marivate, V., Melo, d.G., Meoni, S., Meyer, M., Mir, A., Moosavi, S.N., Meunnighoff, N., Mun, H.S.T., Murray, K., Namysl, M., Obedkova, M., Oli, P., Pasricha, N., Pfister, J., Plant, R., Prabhu, V., *Păiș, V.*, Qin, L., Raji, S., Rajpoot, K.P., Raunak, V., Rinberg, R., Roberts, N., Rodriguez, D.J., Roux, C., S., H.P.V., Sai, B.A., Schmidt, M.R., Scialom, T., Sefara, T., Shamsi, N.S., Shen, X., Shi, Y., Shi, H., Shvets, A., Siegel, N., Sileo, D., Simon, J., Singh, C., Sitelew, R., Soni, P., Sorensen, T., Soto, W., Srivastava, A., Srivatsa, A.K., Sun, T., T, V.M., Tabassum, A., Tan, A.F., Teehan, R., Tiwari, M., Tolkiehn, M., Wang, A., Wang, Z., Wang, J.Z., Wang, G., Wei, F., Wilie, B., Winata, I.G., Wu, X., Wydmanski, W., Xie, T., Yaseen, U., Yee, A.M., Zhang, J. and Zhang, Y.. **NL-Augmenter: A framework for task-sensitive natural language augmentation**. In *Northern European Journal of Language Technology*. Northern European Association of Language Technology, vol. 9, no. 1, pp. 1--41, 2023.
https://nejlt.ep.liu.se/article/view/4725

NL-Augmenter author affiliations

ACKO, Agara, Amelia R&D, New York, Applied Research Laboratories, The University of Texas at Austin, Bloomberg,Brigham Young University, Carnegie Mellon University, Center for Data and Computing in Natural Sciences, Universität Hamburg, Charles River Analytics, Charles University, Prague, Columbia University, Council for Scientific and Industrial Research, **DeepMind**, Department of Computer Science, University of Pretoria, Drexel University, Eberhard Karls University of Tübingen, Edinburgh Napier University, Emory University, Fablab by Inetum in Paris, Fraunhofer IAIS, Georgia Tech, **Google Brain**, **Google Research**, Harbin Institute of Technology, Hasso Plattner Institute / University of Potsdam, Hong Kong University of Science and Technology, **IBM Research**, IIIT Delhi, IIT Delhi, IIT Madras, Illinois Mathematics and Science Academy, **Imperial College, London**, Independent, Indian Institute of Science, Bangalore, Institut Teknologi Bandung, Institute of Data Science, National University of Singapore, International Institute of Information Technology, Hyderabad, Jagiellonian University, Poland, Jean Monnet University, Johns Hopkins’, KTH Royal Institute of Technology, KU Leuven, MTS AI, France, **Microsoft**, Redmond, WA, Mphasis NEXT Labs, National University of Ireland Galway, National University of Science and Technology, Pakistan, National University of Singapore, Naver Labs Europe, Peking University, Politecnico di Milano and University of Bologna, Polytechnic Institute of Paris, Pompeu Fabra University, Pontifical Catholic University of Minas Gerais, Brazil, **Princeton University**, Rakuten India, *Research Institute for Artificial Intelligence Mihai Drăgănescu, Romanian Academy*, Rutgers University, Siemens AG, **Stanford University**, SyNaLP, LORIA, TU Darmstadt, Technical University of Braunschweig, The Alan Turing Institute, The University of Texas at Austin; (University of Barcelona, University of Birmingham), **The University of Tokyo**, Toyota Technological Institute at Chicago, UC Berkeley, UC Santa Barbara / **Google**, UCLA, UMass Amherst, UT Austin, UnifyID, Universiti Brunei Darussalam, University of California, Berkeley and Research Institutes of Sweden, University of Illinois, Urbana Champaign, University of Memphis, University of Pittsburgh, University of São Paulo, University of Toronto, University of Washington, University of Wisconsin–Madison, University of Würzburg, University of Edinburgh, VMware, Vade, Westlake Institute for Advanced Study, Whirlpool Corporation, reciTAL, trivago N.V., Salesforce Research Asia, University of Michigan



John likes expensive Italian pizzas.

NL-Augmenter

John is fond of expensive Italian pizzas.

John likes e😛ensive Italian pizzas .

John likes expensive Italian food .

John likes pure bead Italian pizzas.

John likes expensive Italian pizzas(italian dish of flattened bread and toppings).

John likes expensive Italian pizzas .#LikesPizzas #Likes #John #Pizzas

John confirmed that he likes expensive Italian pizzas.

John likes expensive Italienisch pizzas .

Jo4n lik3s 3xpensiv3 1talian pizzas .

John 💙 expensive 🍝 🍕 .

Expensive italian pizzas, John likes.

Joẖn łikẽs ęxṕensíṽĕ   alℹ️aῇ p ƨzas .

John is a big fan of Italy, especially of the rich and cheap pizzas.

John likes expensive actually Italian actually pizzas In my opinion .

JJoohhnn lliikkeess eexxppeennssiivvee IIttaalliiaann ppiizzzzaass ..

NL-Augmenter

- 117 transformations
  - Generate new sentence(s) given a seed sentence
  - Perturbations / transformations / generators
  - Language-independent / multilingual / monolingual (usually English)
  - May be chained

- 23 filters
  - May remove sentences based on different criteria
  - May be chained with filters/transformations

Example transformations

- Abbreviations
  - homework -> hwk, You -> yu, miles per hour -> mph
- Add hash tags
  - I love domino’s pizza.-> #LovePizza #Love #I #Pizza
- Adjectives antonyms switch
  - Based on WordNet
  - Amanda’s mother was very beautiful -> ugly.
- Back-translation
  - English -> German -> English
  - Andrew finally returned -> eventually gave Chris the French book.
- BackTranslation for NER

Example transformations

- Butter fingers
  - *Sentences -> Senhences* with gapping, such as Paul likes *coffee -> coffwe* and Mary tea, lack an overt predicate to *indicate -> indicatx* the *relation -> relauion* between two or more *arguments -> argumentd*.
- Change Person Named Entities
  - Based on lexicon of person names
- City Names Transformation
  - replaces instances of populous and well-known cities in Spanish and English sentences with instances of less populous and less well-known cities
- Concatenate two random sentences
  - Generates large sentences from small ones

Example transformations

- Concept2Sentence
  - It works by extracting keyword concepts fromthe original sentence, passing them into a BART transformer trained on CommonGen to generate a new, related sentence which reflects the extracted concepts.
  - Original Sentence: "*a disappointment for those who love alternate versions of the bard, particularly ones that involve deep fryers and hamburgers.*"
    Original Label: 0 # (sst2 dataset 0=negative, 1=positive)

    Extracted Concepts: ['disappointment', 'for']
    New Sentence: "*A man is waiting for someone to give him a look of disappointment.*"
    New Label: 0

Example transformations

- Contextual meaning perturbation
  - This transformation changes the meaning of the sentence while avoiding grammar, spelling and logical mistakes.
    “**During his vacation, John visited Brașov, a very nice city in Romania.**” ->
    “During his vacation, John visited Brașov, a very nice town in Serbia.”
  - "**Brașov is a city situated in the central part of Romania.**" ->
    “Craiova is a city situated in the western part of Slovenia.”
- Diverse paraphrase
  - “While on holiday, John visited Brasov, a very beautiful city in Romania.”
  - “Braşov is a city in central Romania.” / “Braşov is a city located in the centre of Romania.”

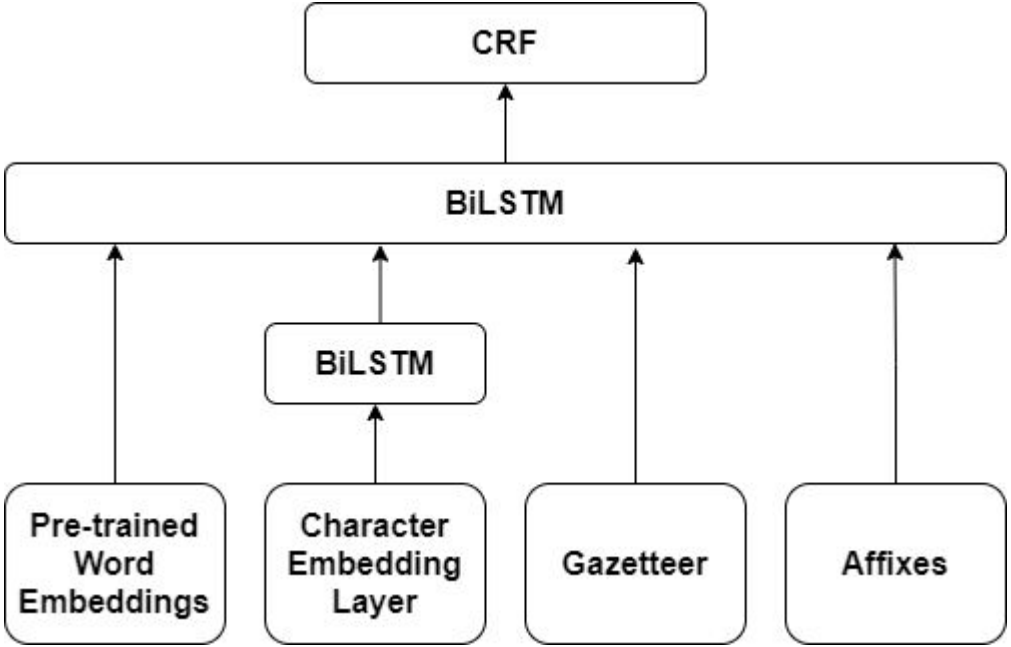
Example transformations

- Emojify
  - Apple is looking at buying U.K. startup for $132 billion.
  - 🍎 is 👀 at 🛍️ 🇬🇧 startup for $1️⃣3️⃣2️⃣ billion.
- GeoNames
  - augments the input sentence with information based on location entities (cities and countries) available in the GeoNames database
  - “Mangalia is situated on the shores of the Black Sea”
  - *Mangalia, a city in Romania, is situated on the shores of the Black Sea*
  - *Mangalia, a city with a population of 39,619, is situated on the shores of the Black Sea*
  - *Mangalia, a city in Romania, with a population of 39,619 inhabitants, is situated on the shores of the Black Sea*
  - *Mangalia, a city in Europe, is situated on the shores of the Black Sea*
  - *Mangalia, a city in Europe, with a population of 39,619 inhabitants, is situated on the shores of the Black Sea*

Model Training

- LegalNERo
  - Păiș, Vasile and Mitrofan, Maria and Gasan, Carol Luca and Coneschi, Vlad and Ianov, Alexandru. Named Entity Recognition in the Romanian Legal Domain. In Proceedings of the Natural Legal Language Processing Workshop 2021. Association for Computational Linguistics, Punta Cana, Dominican Republic, pp. 9--18, nov 2021 https://aclanthology.org/2021.nllp-1.2
- MicroBloggingNERo
  - Păiș, Vasile and Barbu Mititelu, Verginica and Irimia, Elena and Mitrofan, Maria and Gasan, Carol Luca and Micu, Roxana. Romanian micro-blogging named entity recognition including health-related entities. In Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task. Association for Computational Linguistics, Gyeongju, Republic of Korea, pp. 190--196, oct 2022 https://aclanthology.org/2022.smm4h-1.49

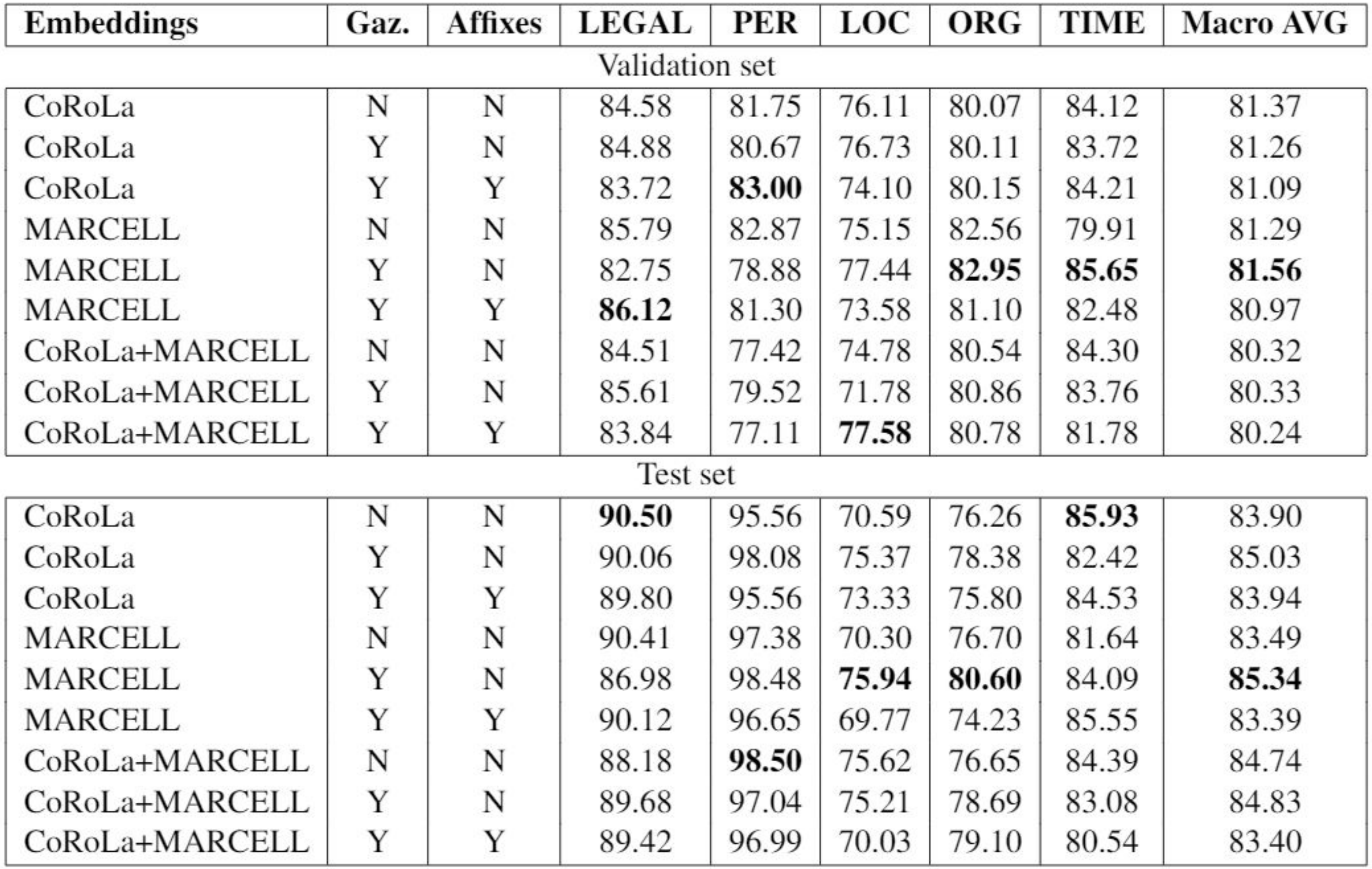
LegalNERo NER system



- ➢ Pre-trained word embeddings from:
  - ○ CoRoLa
  - ○ Marcell
  - ○ Combinations

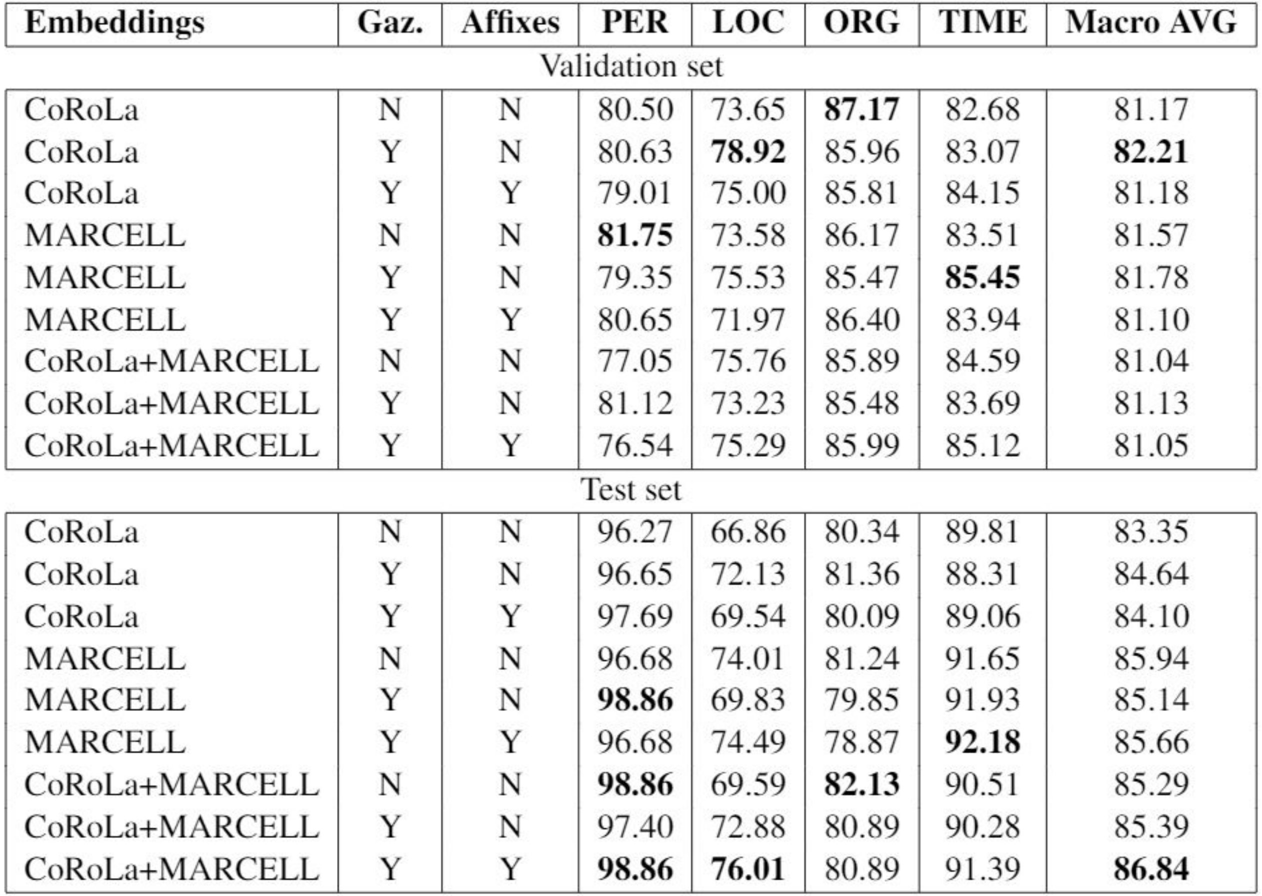
- ➢ Implemented using a modified NeuroNER package
  - ○ inspired from: “PharmacoNER Tagger: a deep learning-based tool for automatically finding chemicals and drugs in Spanish medical texts”

Results



Embeddings	Gaz.	Affixes	LEGAL	PER	LOC	ORG	TIME	Macro AVG
Validation set								
CoRoLa	N	N	84.58	81.75	76.11	80.07	84.12	81.37
CoRoLa	Y	N	84.88	80.67	76.73	80.11	83.72	81.26
CoRoLa	Y	Y	83.72	**83.00**	74.10	80.15	84.21	81.09
MARCELL	N	N	85.79	82.87	75.15	82.56	79.91	81.29
MARCELL	Y	N	82.75	78.88	77.44	**82.95**	**85.65**	**81.56**
MARCELL	Y	Y	**86.12**	81.30	73.58	81.10	82.48	80.97
CoRoLa+MARCELL	N	N	84.51	77.42	74.78	80.54	84.30	80.32
CoRoLa+MARCELL	Y	N	85.61	79.52	71.78	80.86	83.76	80.33
CoRoLa+MARCELL	Y	Y	83.84	77.11	**77.58**	80.78	81.78	80.24
Test set								
CoRoLa	N	N	**90.50**	95.56	70.59	76.26	**85.93**	83.90
CoRoLa	Y	N	90.06	98.08	75.37	78.38	82.42	85.03
CoRoLa	Y	Y	89.80	95.56	73.33	75.80	84.53	83.94
MARCELL	N	N	90.41	97.38	70.30	76.70	81.64	83.49
MARCELL	Y	N	86.98	98.48	**75.94**	**80.60**	84.09	**85.34**
MARCELL	Y	Y	90.12	96.65	69.77	74.23	85.55	83.39
CoRoLa+MARCELL	N	N	88.18	**98.50**	75.62	76.65	84.39	84.74
CoRoLa+MARCELL	Y	N	89.68	97.04	75.21	78.69	83.08	84.83
CoRoLa+MARCELL	Y	Y	89.42	96.99	70.03	79.10	80.54	83.40

Results



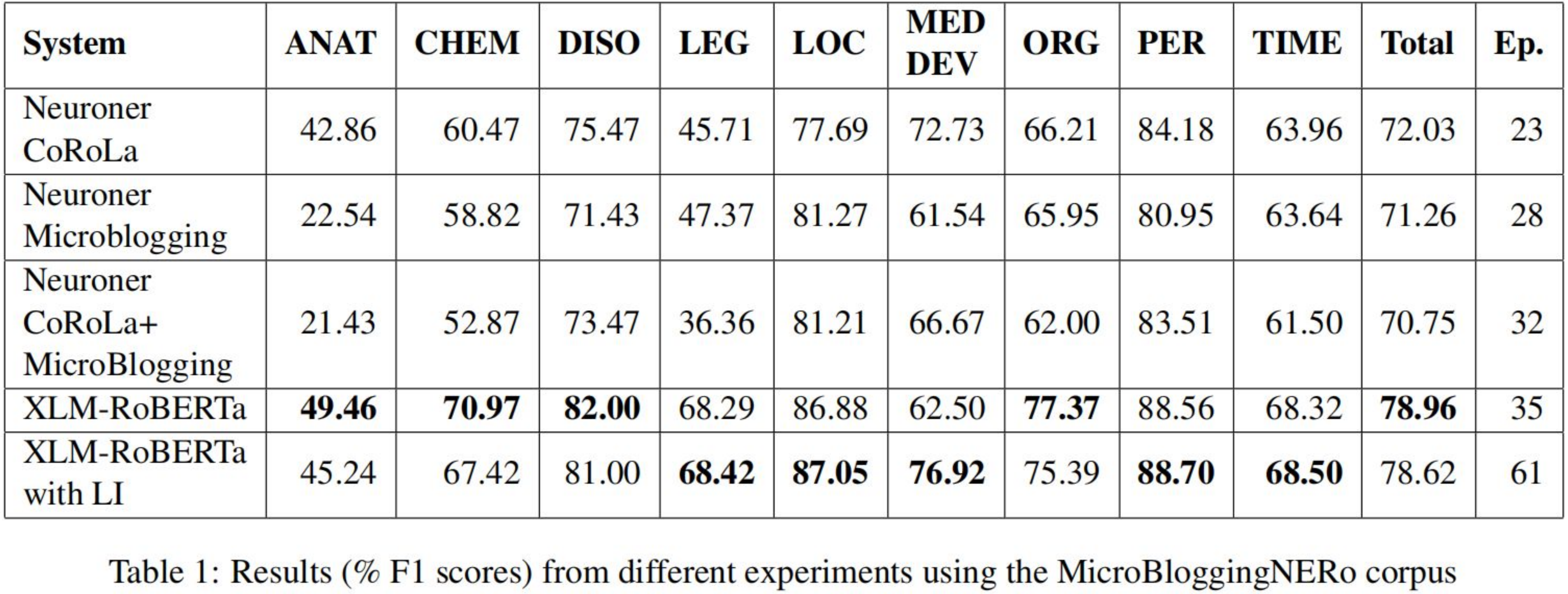
Embeddings	Gaz.	Affixes	PER	LOC	ORG	TIME	Macro AVG
			Validation set				
CoRoLa	N	N	80.50	73.65	**87.17**	82.68	81.17
CoRoLa	Y	N	80.63	**78.92**	85.96	83.07	**82.21**
CoRoLa	Y	Y	79.01	75.00	85.81	84.15	81.18
MARCELL	N	N	**81.75**	73.58	86.17	83.51	81.57
MARCELL	Y	N	79.35	75.53	85.47	**85.45**	81.78
MARCELL	Y	Y	80.65	71.97	86.40	83.94	81.10
CoRoLa+MARCELL	N	N	77.05	75.76	85.89	84.59	81.04
CoRoLa+MARCELL	Y	N	81.12	73.23	85.48	83.69	81.13
CoRoLa+MARCELL	Y	Y	76.54	75.29	85.99	85.12	81.05
			Test set				
CoRoLa	N	N	96.27	66.86	80.34	89.81	83.35
CoRoLa	Y	N	96.65	72.13	81.36	88.31	84.64
CoRoLa	Y	Y	97.69	69.54	80.09	89.06	84.10
MARCELL	N	N	96.68	74.01	81.24	91.65	85.94
MARCELL	Y	N	**98.86**	69.83	79.85	91.93	85.14
MARCELL	Y	Y	96.68	74.49	78.87	**92.18**	85.66
CoRoLa+MARCELL	N	N	**98.86**	69.59	**82.13**	90.51	85.29
CoRoLa+MARCELL	Y	N	97.40	72.88	80.89	90.28	85.39
CoRoLa+MARCELL	Y	Y	**98.86**	**76.01**	80.89	91.39	**86.84**

MicroBloggingNERo Word Embeddings

- a larger micro-blogging corpus
  - 853k messages
- static embeddings trained with FastText
  - https://relate.racai.ro/resources/microblogging/
  - dimension = 300
  - minCount = 20

- Existing word embeddings
  - CoRoLa (Representative Corpus of Contemporary Romanian Language) embeddings
  - contains a social component gathered from blogs
  - https://academiaromana.ro/sectii2002/proceedings/doc2018-2/Art12Pais.pdf
  - XLM-ROBERTA

Experiments



System	ANAT	CHEM	DISO	LEG	LOC	MED DEV	ORG	PER	TIME	Total	Ep.
Neuroner CoRoLa	42.86	60.47	75.47	45.71	77.69	72.73	66.21	84.18	63.96	72.03	23
Neuroner Microblogging	22.54	58.82	71.43	47.37	81.27	61.54	65.95	80.95	63.64	71.26	28
Neuroner CoRoLa+ MicroBlogging	21.43	52.87	73.47	36.36	81.21	66.67	62.00	83.51	61.50	70.75	32
XLM-RoBERTa	**49.46**	**70.97**	**82.00**	68.29	86.88	62.50	**77.37**	88.56	68.32	**78.96**	35
XLM-RoBERTa with LI	45.24	67.42	81.00	**68.42**	**87.05**	**76.92**	75.39	**88.70**	**68.50**	78.62	61

Table 1: Results (% F1 scores) from different experiments using the MicroBloggingNERo corpus

XLM-RoBERTa with Lateral Inhibition: V. Păiș. 2022. RACAI at SemEval-2022 task 11: Complex named entity recognition using a lateral inhibition mechanism. https://aclanthology.org/2022.semeval-1.215/ ; Mitrofan & Păiș. 2022. Improving Romanian BioNER Using a Biologically Inspired System https://aclanthology.org/2022.bionlp-1.30

System for the Anonymisation of Romanian Jurisprudence (SAROJ)

- The anonymisation tool developed under this grant will lead to the full automation of the anonymisation of judicial decisions in Romania.
    - Improved protection of personal data in judicial decisions, more efficient publication of decisions leading to greater transparency of the Romanian judiciary and improved access to justice

- Beneficiary of the tool: Superior Council of Magistracy (CSM)

- The anonymisation tool will be integrated in the ReJust portal https://www.rejust.ro/

SAROJ https://www.racai.ro/p/saroj/

Persons: name and alias

Organizations: private companies, public organizations only if part in the judicial process

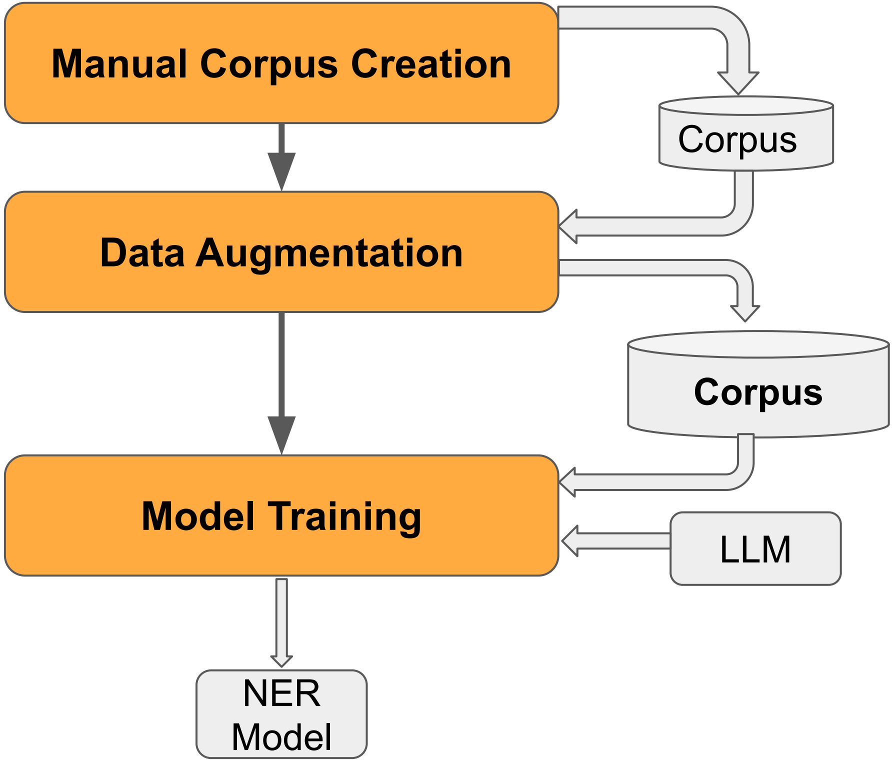
Locations: primarily addresses, but not locations of events (accidents, etc.)

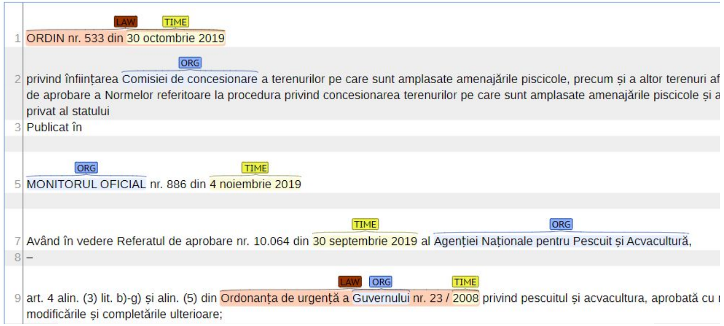
References: cases, decisions, European Case Law Identifiers, but not decisions by CCR, CJUE, CEDO

Date: birthdates

CNP, **phone**, **car registration**, **identity documents**, **email**, **CIF**, **cadastral number**

Recap





ORDIN nr. 533 din 30 octombrie 2019

privind înființarea Comisiei de concesionare a terenurilor pe care sunt amplasate amenajările piscicole, precum și a altor terenuri af
de aprobare a Normelor referitoare la procedura privind concesionarea terenurilor pe care sunt amplasate amenajările piscicole și a
privat al statului

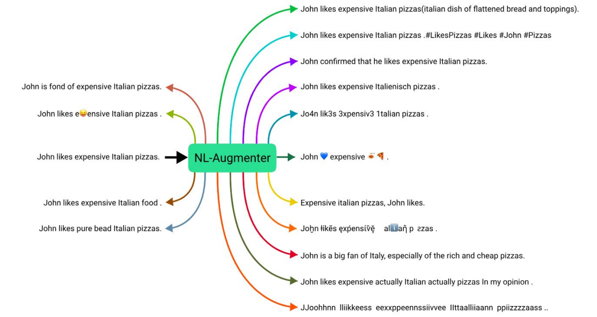
Publicat în

MONITORUL OFICIAL nr. 886 din 4 noiembrie 2019

Având în vedere Referatul de aprobare nr. 10.064 din 30 septembrie 2019 al Agenției Naționale pentru Pescuit și Acvacultură,

–

art. 4 alin. (3) lit. b)-g) și alin. (5) din Ordonanța de urgență a Guvernului nr. 23 / 2008 privind pescuitul și acvacultura, aprobată cu
modificările și completările ulterioare;



John likes expensive Italian pizzas.

John likes expensive Italian pizzas(italian dish of flattened bread and toppings).

John likes expensive Italian pizzas .#LikesPizzas #Likes #John #Pizzas

John confirmed that he likes expensive Italian pizzas.

John is fond of expensive Italian pizzas.

John likes expensive Italienisch pizzas .

John likes e😛ensive Italian pizzas .

Jo4n lik3s 3xpensiv3 1talian pizzas .

John 💙 expensive 🍝🍕 .

John likes expensive Italian food .

Expensive italian pizzas, John likes.

John likes pure bead Italian pizzas.

Joḫn łikẽs ȩxṗensív̌ȩ alℹ️ań p ɀzas .

John is a big fan of Italy, especially of the rich and cheap pizzas.

John likes expensive actually Italian actually pizzas In my opinion .

JJoohhnn lliikkeess eexxppeennssiivvee IIttaalliiaann ppiizzzzaass ..

BERT, DistilBERT, XLM-RoBERTa, etc.

Q
A