REDIT: A Tool and Dataset for Extraction of Personal Data in Documents of the Public Administration Domain

Teresa Paccosi, Alessio Palmero Aprosio
Fondazione Bruno Kessler
Via Sommarive 18, Trento (Italy)
[tpaccosi|aprosio]@fbk.eu

Abstract

English. New regulations on transparency and the recent policy for privacy force the public administration (PA) to make their documents available, but also to limit the diffusion of personal data. The present work displays a first approach to the extraction of sensitive data from PA documents in terms of named entities and semantic relations among them, speeding up the process of extraction of these personal data in order to easily select those which need to be hidden. We also present the process of collection and annotation of the dataset.

Italiano. *Le nuove regolamentazioni sulla trasparenza e la recente legislazione sulla privacy hanno spinto la pubblica amministrazione a rendere i loro documenti pubblicamente consultabili limitando però la diffusione di dati personali. Presentiamo qui un primo approccio all'estrazione di questi dati da documenti amministrativi in termini di named entities e relazioni semantiche tra di esse, in modo da facilitare la selezione dei dati che devono rimanere privati. Presentiamo inoltre il processo di collezione e annotazione del dataset.*

1 Introduction

In recent years, public administrations (PA) in the Italian government have been forced to publish a huge amount of documents, to make them available to citizens, organisations, and authorities. This is the result of the recent legislation about the transparency. For instance, municipalities have to share their documents in a virtual place called *Albo Pretorio*. In most cases, the online publication of these acts is a necessary condition for their purposes to become effective.[1]

On the other side, the General Data Protection Regulation (GDPR), approved in 2016 by the European Union, enhances individuals' control and rights over their personal data, limiting its diffusion over any medium (especially including online platforms such as websites and social networks).

In this context, it is important for the public servants within the PA to amend some documents by hiding the data that cannot be publicly published. Nowadays, most of this work is done manually, hiding the sensitive information document by document. This procedure is clearly time-consuming, non-scalable, and error-prone.

Natural Language Processing (NLP) techniques can be seen as a watershed between a manual management of the PA documents and a new generation of instruments that will finally speed up the process, leaving manual effort as the sole final check just before the publication of the data.

This is not the first time this problem is tackled using NLP, but past works are mainly focused on English and limited to the entity extraction task (Guo et al., 2021).

Our approach to the extraction of personal data from documents focuses on a combination of three NLP instruments:

- **Named-entity Recognition (NER).** This task consists in seeking texts in natural language to locate and classify named entities (NE) mentioned in them. This search is usually limited to a few needed categories: the most common are persons, locations, and organisations. Several approaches have been

[1] In the Italian legislation, this is called "referto di pubblicazione". See also: http://qualitapa.gov.it/

used in literature, between completely rule-based (Appelt et al., 1993; Budi and Bressan, 2003) and machine learning-based (Chiu and Nichols, 2016; Strubell et al., 2017; Devlin et al., 2019), including some hybrid approaches, for example using gazettes of known entities belonging to a particular category (Finkel et al., 2005). In this paper, we use the last approach, mixing a Conditional Random Fields (CRF) algorithm (Lafferty et al., 2001) with the addition of a list of entities, extracted from various knowledge bases, that describe persons, companies, and locations. We describe this process in detail in Section 5.

- **Structured-entity Identification.** A parallel rule-based task is used to extract entities that can easily be recognised without the need of training data. Among them: dates and times, numbers, email addresses, Italian "codice fiscale", that are based on textual patterns; roles and document types, that are based on prepacked lists.

- **Relation extraction (RE).** It is the task of extracting semantic relationships from text. Extracted relationships usually occur between two or more entities of a certain type (for example persons, locations, etc., see previous points), and fall into a number of semantic categories (such as birth location, role in a company, etc.). Relation extraction is widely used also in specific domains such as medicine (Giuliano et al., 2007) and finance (Vela and Declerck, 2009). Successful experiments made use of Conditional Random Fields (Surdeanu et al., 2011), Dependency-based Neural Networks (Liu et al., 2015), and transformers like BERT (Baldini Soares et al., 2019).

In this paper we present REDIT (Relation and Entities Dataset for Italian with Tint), a complete framework that aims to solve the personal data identification in textual documents. The software is mainly based on Tint (Palmero Aprosio and Moretti, 2018), an NLP pipeline specifically designed for Italian and based on Stanford CoreNLP (Manning et al., 2014). REDIT includes part of the annotated dataset (Section 4), the compiled model, and the supporting Java code. It is available for free on Github (see Section 6).

The content is structured as follows. Section 2 presents in detail how we collected the documents that are annotated and how we used fictitious data to make the resource available for download. In Section 3, we describe the process used to annotate the data. Section 4 illustrates the dataset, giving some statistics on the entities and relations included in it. In Section 5 we give some results on the performance of the resulting entity extraction and relation extraction system. The downloadable package (that contains the dataset, the model and the Java code) is finally described in Section 6.

2 Data Collection

The corpus is composed of documents taken from different institutions of the public administration. The documents with which we have worked are different types of forms, varying from license for parking to adoption forms, school enrollments, marriage licenses and so on.

Starting from this set, we create two datasets. One is composed of documents compiled with real data and one with documents compiled by us with fictitious data, using lists of all the Italian streets and surnames in order to guarantee the diversification of the data in the compiled forms, and to not exclusively rely on the annotators' fantasy. The fictitious compilation aims to avoid using sensitive data in terms of privacy issues, leading to the possibility of publicly releasing the dataset. The documents which contain real data are indeed not included in the public dataset. For instance, a sentence such as *Il sottoscritto Gianluca Freschi, nato a Pesaro il 12/12/1990 e residente in Pesaro, Via Virgilio n.76* presents data whose association was invented by the annotator. It could be possible that a person called *Gianluca Freschi* exists in real world but it is almost impossible that he would fit with the rest of the data since they all derive by annotator's fantasy. However, as we can see from the example, while the data are fictitious the structure of the document is identical to that of real ones.

3 The Annotation

Each document in the set is annotated both with entities and relations between them.

For the annotation of entities we adopt the guidelines already used for KIND (Paccosi and Palmero Aprosio, 2021), a corpus containing NE on documents taken from Wikinews. The named entities included in KIND belong to the standard

NE classes and are of three types: **LOC**, **PER**, and **ORG**. As already noticed by (Passaro et al., 2017), these categories are quite unsatisfactory to deal with the information contained in the PA documents, since the model is not designed at capturing information such as laws or protocols. In REDIT we then distinguish different types of ORGs, differentiating public offices and municipality and companies: the former is annotated as **ENTE**, while the latter as usual (**ORG**). Finally, we add a label to mark laws and protocols, **LEX**, so that in the present work there are five types of annotated entities: **LOC**, **PER**, **ORG**, **LEX**, and **ENTE**. The original guidelines used in KIND have therefore been slightly modified to meet our needs (see Section 5.1 for more details).

In addition to the NE annotation, we are interested in annotating the relations among them. In particular, we need to develop a system of relations which links the person with its personal data or with its role in terms of responsibility of the company/public administration or in terms of relative/family relationships.

Since for the annotation task a relation must connect two entities, some additional entity types are annotated only when involved in a relation (see below). The list of additional entities includes **ROLE** for personal and organisation roles (for example, words such as "responsabile", "titolare", "genitore", and so on, representing the role of a person in a company, in the PA domain, or in a family), **DOCTYPE** for document types (such as "passaporto", "patente"), **EMAIL** for e-mail addresses, **DATE** for dates, **NUMBER** for generic numbers (such as VAT), **CF** for the Italian "codice fiscale" sequence of chars.

Regarding relations, `address` is used for instance to link a LOC entity representing an address to the person or company to which the address belongs, while `birthDate`, `birthLoc` link respectively the date and location of birth.

Table 1 shows the complete list of the relations included in the dataset.

The annotation is performed by a domain expert using INCEpTION (Klie et al., 2018), a web-based text-annotation environment which allows users to: (i) select a group of tokens and assign a label to it (entities); (ii) connect two entities among them and assign a label to the link (relations).

This is an example of NER annotation:

Al [Comune di Alessandria]$_{ENTE}$.
[Casale Monferrato]$_{LOC}$, 20 settembre 2021.
Il sottoscritto [Davide Aiello]$_{PER}$, nato a [Milano]$_{LOC}$ il [31/07/1985]$_{DATE}$, [titolare]$_{ROLE}$ della ditta [Aiello Ceramiche S.r.l.]$_{ORG}$, ai sensi dell' [art. 76 del D.P.R. n. 445/2000]$_{LEX}$, dichiara di voler partecipare all'evento "Il mercante in Fiera".

These are the corresponding relations:

- `birthLoc` (Davide Aiello, Milano)
- `birthDate` (Davide Aiello, 31/07/1985)
- `companyRole` (Davide Aiello, titolare)
- `personInOrg` (Davide Aiello, Aiello Ceramiche S.r.l.)

In the example, "31/07/1985" is tagged as DATE, since it is involved in the *birthDate* relation. On the contrary, since no relations include "20 settembre 2021", it's not mandatory, for the annotator, to mark it as DATE.

The system uses two different approaches to identify entities. Entities such as DATE or ROLE are annotated only when involved in a relation because they are labels identified through a rule-based approach which can be easily recognised without the need of training data. For what concerns instead PER, LOC, ORG, ENTE and LEX the identification occurs using a machine-learning technique and they need to be always annotated.

4 The Dataset

As we have seen in Section 2, the complete dataset consists of two parts: the first one presents the documents fictitiously compiled and it is publicly released; the latter, on the contrary, comprehends instances compiled with real data and is not released. Nevertheless, we consider also the unreleased dataset in training the model, so that the amount of annotated relations in the final dataset is 7,821, while that of annotated entities is 21,307. The released one presents 1,439 annotated entities and 1,476 annotated relations.

Looking at the data in Table 1, it is possible to notice that the amount of annotations referring to some relations (marked with *) are considerably fewer than others. Despite the small amount, we have already annotated them in the view of future works on these relations but we do not consider them in the experiments.

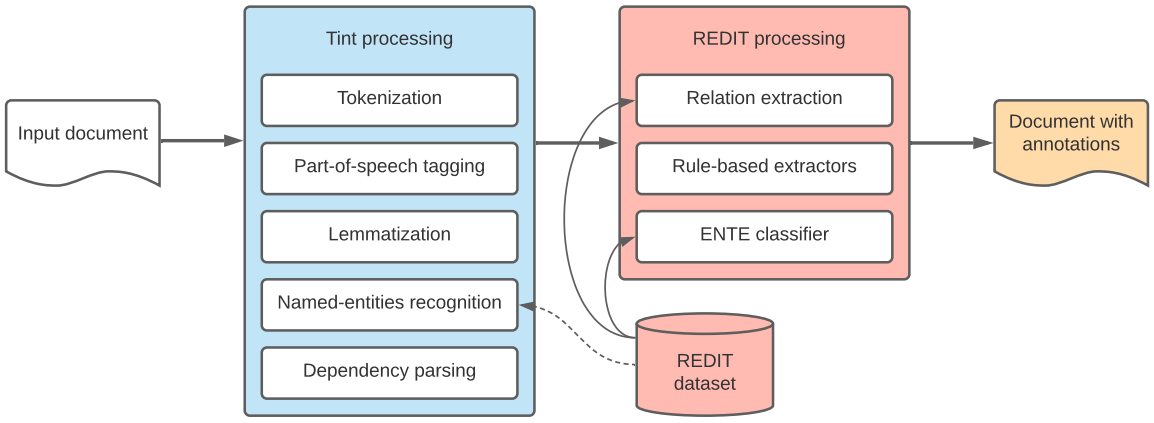

Figure 1: A chart depicting the REDIS architecture and its interaction with Tint.

Relation name	Released ds	Complete ds
address	451	2,115
birthDate	160	639
birthLoc	221	678
codiceFiscale	84	902
companyRole	59	99
deathDate (*)	5	6
deathLoc (*)	6	6
docExpDate (*)	2	2
docID (*)	13	13
docIssueDate (*)	8	8
docIssueLoc (*)	9	9
docType (*)	13	13
email	76	213
name (*)	28	28
personalRole	122	2,192
personInOrg	54	109
relative (*)	28	98
telephone	100	325
vat	37	366
Total	1,476	7,821

Table 1: Amount of annotated relations in the dataset.

Relation name	Released ds	Complete ds
ENTE	192	2829
LEX	214	8314
LOC	743	3788
ORG	62	2179
PER	228	4197
Total	1,439	21,307

Table 2: Amount of annotated entities in the dataset.

5 The Pipeline

To work properly, REDIT relies on a complex pipeline that includes various steps, very different in structure and management (see Figure 1). Most of the steps are performed using well-known tools and algorithms (sometimes not reaching state-of-the-art accuracy), so that the whole program does not need particular hardware (such as the GPUs needed in environments using deep learning and transformers) and is easy to run on almost every common software environment.

1. First, the input text is parsed with Tint (Palmero Aprosio and Moretti, 2018) using these annotators: tokenizer, sentence splitter, truecaser, part-of-speech tagger, lemmatizer, dependency parser.

2. Named-entities are extracted using the CRF implementation included in Stanford NER (Finkel et al., 2005) and the model trained on the annotated dataset (see Subsection 5.1).

3. A second run on named-entities, with the rule-based Stanford TokensRegex software (Chang and Manning, 2014), is performed (see Subsection 5.2)

4. ORG and LOC entities are passed into a Support Vector Machines classifier (Cortes and Vapnik, 1995) to extract ENTE entities (see Subsection 5.3).

5. Finally, the Stanford Relation Extractor (Surdeanu et al., 2011) is used to find relations between entities in the text (see Subsection 5.4).

Source	Tag	Labels
Wikipedia	LOC	377,611
Wikipedia	PER	608,547
Wikipedia	ORG	84,887
OpenStreetMap	LOC	389,649

Table 3: Items added to the NER training taken from gazettes.

5.1 The CRF Named-Entities Tagger

Since the sole REDIT dataset is not sufficient to train a robust NER tagger, we use it in combination with KIND (see Section 3). Guidelines for the two datasets are, of necessity, slightly different, therefore we need to use some precautions in merging them.

Sometimes, the entities annotated as ORG in KIND (such as "Unione Europea") should have been annotated as ENTE in REDIT. We then decided, in the training phase, to merge all ENTE entities into ORG. We then trained a classifier dedicated to the ENTE tag (see Subsection 5.3), trained on REDIT dataset only, that performs the sole disambiguation between ORG and ENTE.

To enhance the classification, Stanford NER also accepts gazettes of names labelled with the corresponding tag. We collect a list of persons, organizations and locations from the Italian Wikipedia using some classes in DBpedia (Auer et al., 2007): `Person`, `Organisation`, and `Place`, respectively. In addition to this, we collect the list of streets from OpenStreetMap (OpenStreetMap contributors, 2017), limiting the extraction to Italian names. Table 3 shows statistics about the gazettes.

The evaluation is performed by randomly splitting the dataset into train/dev/test using 80/10/10 ratio. During training phase, we tried some sets of features choosing among the ones available in Stanford NER. We obtained the best results (considering also a good balance between training/testing time and performances) with word shapes, n-grams with length 6, previous, current, and next token/lemma/class. Table 4 displays the results of the NER module.

5.2 The Rule-Based Named-Entities Tagger

As said in Section 3, there is the need for more entity types, because in the training phase we need to have both arguments of a relation annotates as an entity (of any type). For this reason, we use a rule-based approach to annotate DATE, ROLE, DOCTYPE, EMAIL, NUMBER, and CF.

Relation	P	R	F-score
LEX	0.762	0.760	0.761
LOC	0.830	0.811	0.820
ORG	0.832	0.821	0.826
PER	0.868	0.894	0.881
Total (micro)	**0.805**	**0.799**	**0.802**
Total (macro)	**0.823**	**0.821**	**0.822**

Table 4: Evaluation of the entity tagger.

- Tint TIMEX annotator is used to tag DATE entities.
- ROLE and DOCTYPE entities are extracted given a list of roles taken from the annotated training set.
- Numbers, e-mail addresses and Italian codice fiscale are tagged using regular expressions.

5.3 The SVM Classifier for ENTE Entities

After the previous steps, the entities that should be marked with ENTE now falls into the ORG or LOC entity sets. We then use a simple SVM classifiers (using shallow features, such as words, bigrams, previous and following content words, etc.) that, given an entity tagged as LOC or ORG, return whether it should be annotated as ENTE. The training set used by the classifier consists in entities taken from REDIT and annotated as ORG, LOC, and ENTE. The first two categories represent the zero class, while entities tagged with ENTE represent the other class. It is therefore a binary classifier. In a 10-fold cross-validation environment, results shows a F-score equals to 0.978 (precision 0.981, recall 0.974).

5.4 Relation Extractor Module

The last module in REDIT is Stanford Relation Extractor (Surdeanu et al., 2011), used to train and extract relations in the text.

Similarly to the NER training, we test approaches with different sets of features, obtaining the best results with unigrams/bigrams, adjacent words, argument words, argument class, dependency path between the arguments, entities and concatenation of POS tags between arguments.

Table 5 shows the results on the relation extractor (the evaluation is performed using gold-labeled entities).

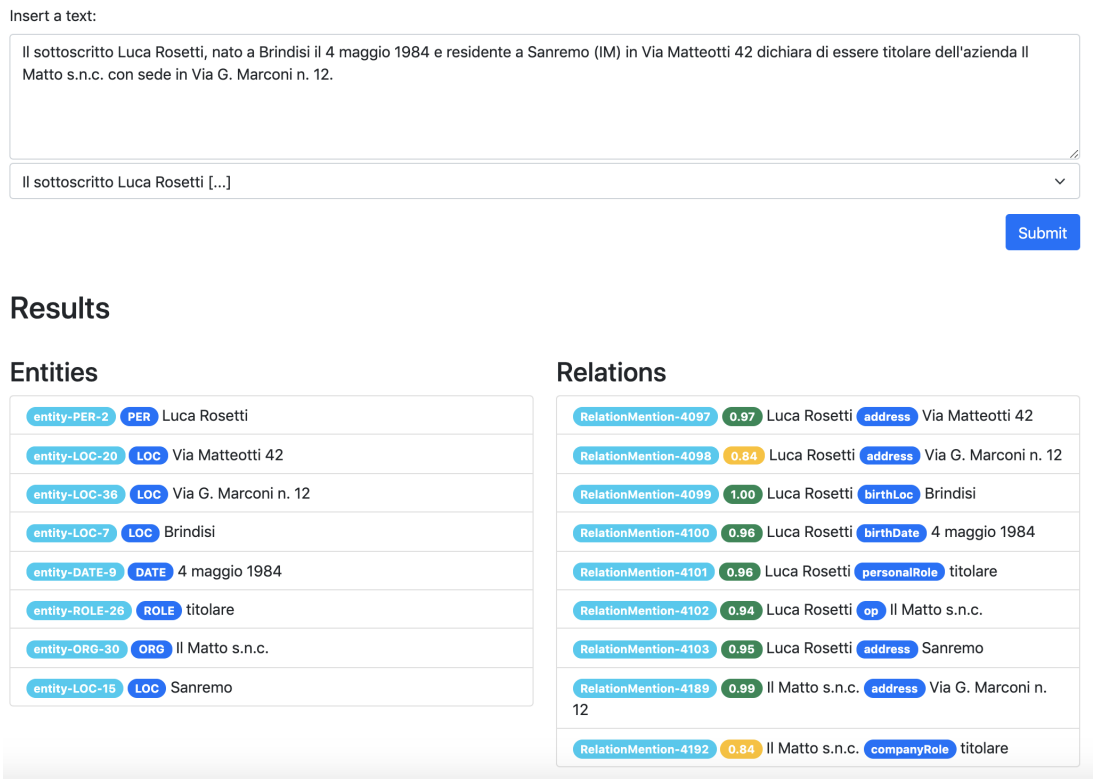

Figure 2: A screenshot of the demo interface.

Relation	P	R	F-score
address	0.929	0.908	0.918
birthDate	0.907	0.907	0.907
birthLoc	0.902	0.874	0.888
codiceFiscale	0.854	0.752	0.800
companyRole	0.902	0.676	0.773
email	0.865	0.421	0.566
personalRole	0.892	0.892	0.892
personInOrg	0.909	0.674	0.774
tel	0.935	0.580	0.716
vat	0.964	0.870	0.915
Total (micro)	**0.914**	**0.841**	**0.876**
Total (macro)	**0.906**	**0.755**	**0.824**

Table 5: Evaluation of the relation extractor.

6 The Release

All parts of REDIT (except part of the annotated dataset, see Section 4) are released for free under the CC BY 4.0 license,[2] and can be downloaded on Github.[3] These include the annotations, in WebAnno format (Yimam et al., 2013), the gazettes, both the NER and the RE models (created using the whole corpus), and the source code, written in Java, used to parse the files and run the classifiers.

A working demo of the tool is available online (See Figure 2).[4] Its web interface is written with VueJS/Boostrap and it is available for download in the Github project page.

7 Conclusion and Future Work

In this paper we present a completely automatic approach to extract personal data (view as entities) and relations between them from documents of the public administration written in Italian texts. The pipeline relies on a mix of rule-based and machine learning-base modules. The latter are trained using a manually annotated dataset, which is in part available for download. All the source code, instead, is released and available for download.

In the future, we plan to enhance the coverage of our system by adding more examples on relations that are less represented (see Table 1).

Acknowledgments

The research leading to this paper was partially supported by Wemapp Srl, Potenza, Italy.[5]

[2] https://bit.ly/cc-by-40-intl
[3] https://github.com/dhfbk/redit
[4] https://bit.ly/relation-extraction
[5] https://wemapp.eu/

References

Douglas Appelt, Jerry Hobbs, John Bear, David Israel, and Mabry Tyson. 1993. FASTUS: A Finite-state Processor for Information Extraction from Real-world Text. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1172–1178, 01.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *The Semantic Web*, pages 722–735, Berlin, Heidelberg. Springer Berlin Heidelberg.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy, July. Association for Computational Linguistics.

I. Budi and S. Bressan. 2003. Association rules mining for name entity recognition. In *Proceedings of the Fourth International Conference on Web Information Systems Engineering, 2003. WISE 2003.*, pages 325–328.

Angel X. Chang and Christopher D. Manning. 2014. TokensRegex: Defining cascaded regular expressions over tokens. Technical Report CSTR 2014-02, Department of Computer Science, Stanford University.

Jason P. C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Claudio Giuliano, Alberto Lavelli, Daniele Pighin, and Lorenza Romano. 2007. FBK-IRST: Kernel methods for semantic relation extraction. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 141–144, Prague, Czech Republic, June. Association for Computational Linguistics.

Yongyan Guo, Jiayong Liu, Wenwu Tang, and Cheng Huang. 2021. Exsense: Extract sensitive information from unstructured data. *Comput. Secur.*, 102:102156.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics, Juni.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Yang Liu, Furu Wei, Sujian Li, Heng Ji, Ming Zhou, and Houfeng Wang. 2015. A dependency-based neural network for relation classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 285–290, Beijing, China, July. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

OpenStreetMap contributors. 2017. Planet dump retrieved from https://planet.osm.org . `https://www.openstreetmap.org`.

Teresa Paccosi and Alessio Palmero Aprosio. 2021. KIND: an Italian Multi-Domain Dataset for Named Entity Recognition. In *arXiv preprint*.

Alessio Palmero Aprosio and Giovanni Moretti. 2018. Tint 2.0: an all-inclusive suite for nlp in italian. In *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it*, volume 10, page 12.

Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. 2017. Fast and accurate entity recognition with iterated dilated convolutions.

Mihai Surdeanu, David McClosky, Mason Smith, Andrey Gusev, and Christopher Manning. 2011. Customizing an information extraction system to a new domain. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*, pages 2–10, Portland, Oregon, USA, June. Association for Computational Linguistics.

Mihaela Vela and Thierry Declerck. 2009. Concept and relation extraction in the finance domain. In H. Bunt, V. Petukhova, and S. Wubben, editors, *Proceedings of the Eighth International Conference on Computational Semantics (IWCS-8). International Conference on Computational Semantics (IWCS-8), January 7-9, Tilburg, Netherlands*, pages 346–351. Tilburg University, 1.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. WebAnno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6, Sofia, Bulgaria, August. Association for Computational Linguistics.