INFORMed PA: A NER for the Italian Public Administration Domain

Lucia C. Passaro[⋆], **Alessandro Lenci**[⋆], **Anna Gabbolini**[⋆⋆]
[⋆] Dipartimento di Filologia, Letteratura e Linguistica, University of Pisa (Italy)
[⋆⋆] ETI^3 | Evolution, Technology & Innovation
lucia.passaro@for.unipi.it
alessandro.lenci@unipi.it
anna.gabbolini@eti3.it

Abstract

English. In this paper, we illustrate the creation of a NER for the Public Administration (PA) domain. We discuss the creation of an annotated corpus with documents from the Italian *Albo Pretorio Nazionale* and provide results of the system evaluation.

Italiano. *In questo lavoro mostriamo la creazione di un NER per il dominio della Pubblica Amministrazione (PA). Presentiamo la creazione del corpus formato da documenti dell'Albo Pretorio Nazionale e mostriamo i risultati della valutazione del sistema.*

1 Introduction

In the Public Administration (PA) domain, the rapid adoption of the new legislation about the governance transparency has been forcing Italian municipalities to produce their acts in a digital form and to make them available for both citizens and authorities. However, the acts delivered by PAs are typically in a free-text electronic format, which is not convenient for searching, decision-support, and data analysis. Therefore, the development of NLP tools to extract high-quality structured information, including Named Entities (NEs) such as Persons and Organizations, represents a key factor to enable the access to the wealth of information produced by PAs, and a crucial step in turning the keyword of "transparency" into reality. The potentialities of NLP tools can be exploited to mine the large document repositories produced by PA daily, with the aim of identifying trends in their activity, suggesting possible synergies to increase their efficiency, and raising "red flags" about suspicious behaviors, especially for their relationships with private companies.

In this paper, we focus on Named Entity Recognition (NER) for PA. Several approaches have been proposed in literature including Rule-based, Machine Learning-based and Hybrid methods.

Hand-made Rule-based NERs focus on extracting names using lots of human-made rules. In general, these systems consist of a set of patterns based on grammatical (e.g., part of speech), syntactic (e.g., word precedence) and orthographic features (e.g., capitalization) in combination with dictionaries (Budi and Bressan, 2003; Appelt et al., 1993; Grishman, 1995). These approaches usually give good results, but require long development time by expert linguists. On the one hand, these systems have better results for restricted domains, being capable of detecting very complex entities, but, on the other one, they lack portability and robustness and do not necessarily adapt well to new domains and languages.

Machine learning techniques, on the contrary, use a collection of annotated documents for training the classifiers. Therefore the development time moves from the definition of rules to the preparation of annotated corpora (Bikel et al., 1997; Borthwick et al., 1998; McCallum and Li, 2003). The systems identify and classify nouns using machine learning algorithms such as Maximum Entropy (Berger et al., 1996), Support Vector Machines (Cortes and Vapnik, 1995) and Conditional Random Field (Lafferty et al., 2001). More recently, also deep learning architectures have been proposed for Named Entity Recognition (Chiu and Nichols, 2015; Strubell et al., 2017).

Finally, Hybrid NER systems, combine rule-based and machine learning-based methods, and make new methods using strongest points from each method (Srihari et al., 2000).

Existing general purpose Italian corpora annotated with NEs such as I-CAB (Magnini et al., 2006) are not optimal for training a NER for the domain of PA because of the gap between bu-

reaucratic language and standard Italian, and also because of the lack of important classes such as act and normative references, that are very useful in PA-oriented applications. To tackle these problems, we decided to create a new corpus from scratch starting from: (i) administrative documents belonging to the *Italian Albo Pretorio*; (ii) the CoLingLab NER, a general NER trained on I-CAB, from which we took the initial configuration of features. The corpus of PA documents written in Italian "bureaucratese", has the characteristics described in Brunato (2015):

1. Pseudo-technicisms or collateral technicisms (e.g., *balneazione, fattispecie*);
2. Abstract nouns with *-zione/-mento* suffixes (e.g., *stipulazione, espletamento*), deverbal nouns, usually with zero suffix (e.g., *subentro, scorporo, utilizzo*) and denominal verbs (e.g., *relazionare, disdettare*);
3. Archaic terms (e.g., *allorché, suddetto*) and latinisms (e.g. *una tantum, pro capite*);
4. Forestierisms (e.g., governance, front office);
5. Uncommon and formal terms (e.g., *diniego* for *rifiuto*);
6. Stereotyped phrases (e.g., *entro e non oltre, in riferimento all'oggetto*);
7. Abbreviations and acronyms.

For the creation of a NER for PA, we decided to exploit the existing architecture employed for the project SEMPLICE[1] and in particular we adopted a statistical method based on the Stanford NER (Finkel et al., 2005), a system implemented in Java and available for download under the GNU General Public License. This choice allowed us to easily compare the gain obtained by enriching the training corpus with PA documents and to speed up the development process. Moreover, using a Conditional Random Field (CRF) (Lafferty et al., 2001) as learning algorithm made it possible for us to compare the PA model with other domain-adapted NERs (Passaro and Lenci, 2014).

This paper is structured as follows: In section 2, we present the CoLingLab NER and we show its performance on a sample of PA documents; in section 3 we describe the adaptation of the system to PA texts and its performances (section 4.1). In section 5, we report on the annotation of relations that we performed on a sample of the corpus and finally discuss the results and ongoing work.

2 The CoLingLab NER

The standard Italian CoLingLab NER was trained on the Italian Content Annotation Treebank (I-CAB (Magnini et al., 2006)), a corpus of Italian news, annotated with semantic information at different levels: Temporal Expressions, Named Entities, relations between entities. I-CAB is composed of 525 news documents taken from the local newspaper 'L'Adige' (time span: September-October of 2004). The NEs annotated in the corpus are: Locations (LOC), Geo-Political Entities (GPE), Organizations (ORG) and Persons (PER).

As we said before, this model is unsatisfactory for the domain of Public Administration in two main respects. First, its classes are insufficient to deal with the type of information in the PA documents, that are full of references to other "linked" acts and legislative reference; second, the language used in these documents is a peculiar and highly complex variant of standard Italian (cf. above). In addition, the performance of the model, attested at $\sim$**0.66** of F1-score on a portion of I-CAB decreases dramatically on the PA documents, reaching a F1-score of $\sim$**0.35**. To measure such performances, in the test set we mapped ORG_PA (cf. below) with ORG, and in the training set we mapped GPE with LOC.

3 A NER for PA Documents

The adaptation of the CoLingLab NER to the PA domain included the extension of the standard NE classes (Rau, 1991; Grishman and Sundheim, 1996; Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) to other entity types particularly important in the context of municipalities. In particular, we added the class ACT, to mark other administrative documents (normally, PA texts refer to other documents related to the same procedure), the class LAW for the relevant legislation, and an additional class of organizations, ORG_PA, for municipal departments.

3.1 The PA Corpus

For the creation of the corpus, we used documents taken from the *Albo Pretorio Nazionale* with the

[1]The SEMantic instruments for PubLIc administrators and CitizEns (SEMPLICE; `www.semplicepa.it`) is a 2-year project funded by Regione Toscana in collaboration with IT companies to develop NLP-based tools for knowledge management, information extraction and opinion mining for local public administrations.

aim of capturing the variability of the texts produced by PA. Overall, the corpus includes **460** documents, for a total of **724,623 tokens**, annotated with the following NEs: (i) **ACT**: documents belonging to the Albo Pretorio Nazionale, with their type (optional), number and date: *Determina n. 4 del 12/02/2011*; (ii) **LAW**: legislative references: *art. 183 comma 7 del D.Lgs. n. 267/2000*; (iii) **LOC**: locations and geo-political entities: *Comune di Pisa*; (iv) **ORG_PA**: organizations related to the Public Administration such as municipal Departments: *Sezione Anagrafe*; (v) **ORG**: organizations: *Consip Spa*; (vi) **PER**: physical persons. The corpus has been linguistically annotated by means of a pipeline of general purpose NLP tools and in particular, it has been POS-tagged with the Part-Of-Speech tagger described in Dell'Orletta (2009), dependency parsed with the DeSR parser (Attardi et al., 2009). Finally, complex terms like *forze dell'ordine* (security force) have been identified using the EXTra term extraction tool (Passaro and Lenci, 2016).

3.2 Annotation

NE annotation has been performed by means of an incremental process: first 100 documents have been annotated by 2 annotators (one of them was a domain expert). In a second phase we trained a CRF model on these documents and we used it to automatically annotate new documents. Finally, we identified the most common errors of the classifier and two new annotators manually revised the output. This process has been repeated for each group of 100 documents up to covering the whole corpus that includes 460 distinct documents. The average length of the documents is 1,575.26 tokens and the total number of the tokens is 724,623. Figure 1 shows the distribution of the different NE classes in the corpus.

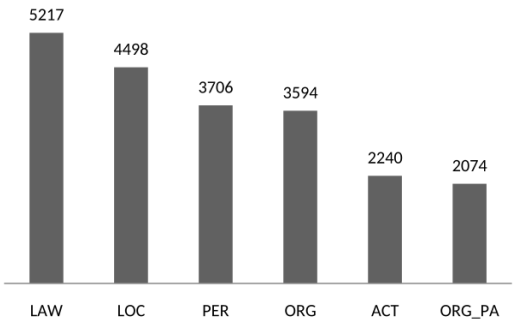

Figure 1: Distribution of the NEs in the corpus.

NEs have been annotated on the CONLL (Nivre et al., 2007) texts using the standard IOB method. In order to deal with acts, we decided to tag them with different "labels" to distinguish their subcomponents: the type (marked with ACT_T), the number (marked with ACT_N), the date (marked with ACT_D), functional tokens (ACT_X) and unparsable tokens (marked wirth ACT_U). For example, the act *Delibera di giunta comunale numero 53 del 23/10/2016* is annotated as follows: *Delibera di giunta comunale (*ACT_T*) numero (*ACT_X*) 53 (*ACT_N*) del (*ACT_X*) 23/10/2016 (*ACT_D*)*, while the act *DD/67/2012* is annotated as ACT_U. This method allows for a simpler normalization of normative references, which is crucial for document retrieval because of the high variability of law mentions in the PA texts.

The inter-annotator agreement between two annotators (attested at $\sim$0.8) has been calculated using the Cohen's K index on a sample of 25 documents of 25 different municipalities, for a total of 26,190 tokens.

4 System Overview

To train the NER, no information from gazetteers was used. The model includes the following groups of features:

SEQUENCES: Next and previous words and a window of 6 words (3 preceding and 3 following the target word) and their classes;

N-GRAMS: Character-level features, i.e., substrings of the word with a maximum length of 6 letters;

ORTHOGRAPHY: "word shape" features such as spelling, capital letters, presence of non–alphabetical characters etc.;

LINGUISTIC FEATURES: The word position in the sentence (numeric attribute), the lemma, and the PoStag (nominal attribute);

TERMS: We employed complex terms as features to train the model. Terms have been extracted with EXTra (Passaro and Lenci, 2016).

4.1 System's Performances

We trained the CRF model based on the CoLingLab NER on the annotated PA corpus, and we tested its performances first with cross-validation and then on a sample of new 25 documents of 25 different municipalities. This choice stems from the fact that very often different municipalities tend to use different templates and different

ways to refer to particular entities. This is particularly common in some NE classes such as ACTS and ORG_PA, that vary a lot across municipalities. For example, some of the analyzed texts contain strings of the form *YYYY/G/NNNNN* to refer to the acts, where the number is actually a string encoding both the date (year: YYYY), a code for the type (G) and the number of the act (NNNNN). Other municipalities instead adopt a less strictly codified pattern to indicate the act such as *Type of act, number N* of DD/MM/YYYY*. Likewise, depending on the writing style (and conventions) of the municipalities, the various departments (i.e., ORG_PA) can include both strings like *Corpo dei Vigili Urbani* and codes like *Tec-01/ICT*. To evaluate the system performance with respect to the variation of the naming conventions adopted by different municipalities, we randomly selected 25 municipalities and one document for each of them balanced for length.

Table 1 reports on the results obtained in cross validation and Table 2 shows the performance on the sample of 25 documents. Figure 2 shows also the confusion matrix for that sample.

In order to investigate the contribution of non-linguistic features, we performed ablation experiments and we tested the results on the sample of 25 documents. The ΔF1-Score for such groups is as follows: SEQUENCES: 3%; N-GRAMS: 1%; ORTHOGRAPHY: 4%. In addition, we performed an additional experiment by training the NER on a combination of I-CAB and the PA documents. In this case, we noticed a ΔF1-Score of 2% by respect to the original model.

	Precision	Recall	F1-Score
ACT	0.7876	0.8914	0.8356
LAW	0.827	0.8423	0.8343
LOC	0.702	0.7398	0.7196
ORG	0.7085	0.689	0.6977
ORG_PA	0.6158	0.7774	0.6855
PER	0.8373	0.8776	0.8567
MacroAVG	**0.7464**	**0.8029**	**0.7716**

Table 1: System results (10-fold cross validation)

5 Towards a Relational Classifier for PA

For a subset of the corpus, we also annotated the semantic relations occurring between two entities in the domain of the PA, using the following scheme:

	Precision	Recall	F1-Score
ACT	0.9747	0.8477	0.9068
LAW	0.9494	0.9615	0.9554
LOC	0.799	0.6913	0.7413
ORG	0.8017	0.7686	0.7848
ORG_PA	0.8706	0.7957	0.8315
PER	0.9142	0.8694	0.8912
MicroAVG	**0.914**	**0.8355**	**0.873**
MacroAVG	**0.8849**	**0.8224**	**0.8518**

Table 2: System results (on a sample of 25 texts)

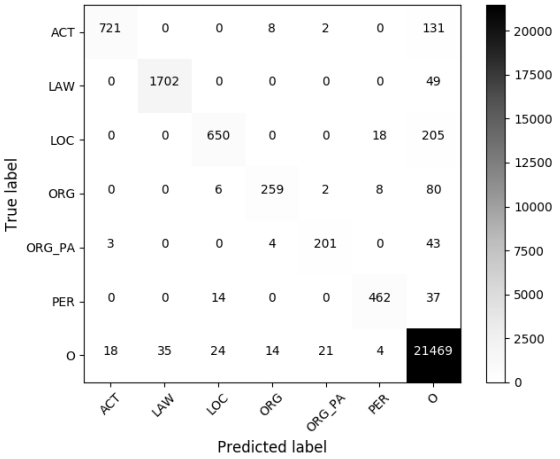

Figure 2: Confusion Matrix (25 texts)

PART OF: the relation of hyponymy, which can occur between: (i) two locations (e.g. a Municipality in Province); (ii) two organizations (e.g. a participated into a holding company); (iii) a person and an organization (e.g. a member of an organization). Implicit attribute for this reation is "work in".

LOCATION: an entity placed into a particular location, occurring between: (i) an organization and a location (e.g. an organization located in a certain region). Possible attributes for this relation are "work in" and "placed in"; (ii) a person and and a location (e.g. a person living in a particular area). Possible attributes are "work in", "born in" and "placed in".

IS RELATED TO: an underspecified relation between any entity pair.

Preliminary experiments have been performed to examine the characteristics of an automatic classifier for extracting relations from administrative acts, and the performance seem to be very promising, despite the size of the training set, which includes in total 100 documents so far. The

extension of the annotated corpus and the training of the relational classifier are currently ongoing.

6 Discussion

The results show that the NER reaches satisfactory results for most of the classes, although legging behind in the recognition of PA Organizations, which, among others, tend to have a higher formal variability, including for example both entities like *Corpo dei Vigili Urbani* and *Tec-01/ICT*. Moreover, in the recognition of Location names in the domain of the PA, the system is expected to detect entities with a non-standard detail level going from the name of the municipalities (e.g. *Comune di Pisa*) to very detailed addresses (e.g., *via S. Maria n. 36, 56126 Pisa (PI) interno 15*). A similar problem occurs in the recognition of very small organizations, whose name contains the name of its founder (i.e., *Mario Rossi snc*). In these cases, especially when *snc* is omitted, the system predicts the class PER instead of the correct class ORG. We are confident that adding lexicons and gazetteers will improve the identification of entities of this kind, but it could be interesting to investigate automatic normalization, disambiguation and entity linking approaches (Hoffart et al., 2011; Han et al., 2011).

7 Conclusions and Ongoing Work

Named entities play an important role in administrative acts, especially in those - like the documents in the Albo Pretorio - describing the main actions taken by Municipalities. This kind of information is very useful to fullfil the obligations related to supervisory monitoring, disclosure, periodic self-assessment, and review of the government decisions.

In this paper, we presented a NER for PA that shows a significant ability to identify the relevant entities, and in particular legislative reference and connected acts. It is important to stress the lexical and syntactic complexity of bureaucratic language represents a big challenge for NLP tools and methods. Such a complexity derives from the technical lexis of other domain-specific languages with which PA deals daily, such as education, environment, ICT technologies, public health and so on. In near feature we plan to explore the possibility of re-engineering our system to take advantage of new algorithms for entity extraction such as neural networks and in particular from character level word embeddings. Moreover, we will focus on the development of classifiers for Relation Extraction and Entity Linking.

Acknowledgments

This research has been supported from the Project SEMantic instruments for PubLIc administrators and CitizEns (SEMPLICE), funded by Regione Toscana, and the Company ETI^3| Evolution, Technology & Innovation. Special acknowledgements go to Roberto Battistelli and Francesco Sandrelli (ETI^3) for support, and to the students Roswita Candusso, Carmela Cinquesanti, Federica Semplici and Ludovica Vasile for manual annotation.

References

Douglas E. Appelt, Jerry R Hobbs, John Bear, David Israel, and Mabry Tyson. 1993. Fastus: A finite-state processor for information extraction from real-world text. In *IJCAI*, volume 93, pages 1172–1178.

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi, and Joseph Turian. 2009. Accurate dependency parsing with a stacked multilayer perceptron. In *EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian 2009*, LNCS, Reggio Emilia (Italy). Springer.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71.

Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: A high-performance learning name-finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 194–201, Washington, DC. Association for Computational Linguistics.

Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. 1998. Nyu: Description of the mene named entity system as used in muc-7. In *In Proceedings of the Seventh Message Understanding Conference (MUC-7*.

Dominique Brunato. 2015. A study on linguistic complexity from a computational linguistics perspective. a corpus-based investigation of italian bureaucratic texts. Ph.D. Thesis, University of Siena.

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

Jason P.C. Chiu and Eric Nichols. 2015. Named entity recognition with bidirectional lstm-cnns. *arXiv preprint arXiv:1511.08308*.

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

Felice Dell'Orletta. 2009. Ensemble system for part-of-speech tagging. In *EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian 2009*, LNCS, Reggio Emilia (Italy). Springer.

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 on Association for Computational Linguistics*, ACL '05, pages 363–370, Ann Arbor, Michigan (USA). Association for Computational Linguistics.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: A brief history. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pages 466–471. Association for Computational Linguistics.

Ralph Grishman. 1995. The nyu system for muc-6 or where's the syntax? In *Proceedings of the 6th Conference on Message Understanding*, pages 167–175, Columbia, Maryland. Association for Computational Linguistics.

Xianpei Han, Le Sun, and Jun Zhao. 2011. Collective entity linking in web text: A graph-based method. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 765–774, Beijing (China).

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh (United Kingdom). Association for Computational Linguistics.

John 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*, pages 282–289, San Francisco, CA (USA). Morgan Kaufmann Publishers Inc.

Bernardo Magnini, Emanuele Pianta, Manuela Speranza, Valentina Bartalesi Lenzi, and Rachele Sprugnoli. 2006. Italian content annotation bank (i-cab): Named entities.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191, Edmonton (Canada). Association for Computational Linguistics.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The conll 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932, Prague (Czech Republic). Association for Computational Linguistics.

Lucia C. Passaro and Alessandro Lenci. 2014. "il piave mormorava...": Recognizing locations and other named entities in italian texts on the great war. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014*, pages 286–290, Pisa (Italy).

Lucia C. Passaro and Alessandro Lenci. 2016. Extracting terms with extra. In *Proceedings of the EUROPHRAS 2015 – Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, pages 188–196, Malaga (Spain).

Lisa F. Rau. 1991. Extracting company names from text. In *Artificial Intelligence Applications, 1991. Proceedings., Seventh IEEE Conference on*, volume 1, pages 29–32. IEEE.

Rohini Srihari, Cheng Niu, and Wei Li. 2000. A hybrid approach for named entity and sub-type tagging. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, pages 247–254, Seattle, Washington (USA). Association for Computational Linguistics.

Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. 2017. Fast and accurate entity recognition with iterated dilated convolutions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2660–2670.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: language-independent named entity recognition. In *Proceedings of the 6th conference on Natural language learning*, volume 31, pages 1–4.