10.12871/CLICIT2014155

"Il Piave mormorava…": Recognizing Locations and other Named Entities in Italian Texts on the Great War

Lucia Passaro
CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica,
University of Pisa (Italy)
lucia.passaro@for.unipi.it

Alessandro Lenci
CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica,
University of Pisa (Italy)
alessandro.lenci@ling.unipi.it

Abstract

English. Increasing amounts of sources about World War I (WWI) are nowadays available in digital form. In this paper, we illustrate the automatic creation of a NE-annotated domain corpus used to adapt an existing NER to Italian WWI texts. We discuss the annotation of the training and test corpus and provide results of the system evaluation.

Italiano. *Negli ultimi anni, si sono resi disponibili in formato digitale un numero sempre maggiore di materiali riguardanti la Prima Guerra Mondiale. In questo lavoro illustriamo la creazione automatica di un corpus di addestramento per adattare un NER esistente a testi italiani sulla Prima Guerra Mondiale e presentiamo i risultati della valutazione del nostro sistema addestrato sul nuovo corpus.*

1 Introduction

Increasing amounts of sources about World War I (WWI) are nowadays available in digital form. The centenary of the Great War is also going to foster this trend, with new historical sources being digitized. This wealth of digital documents offers us an unprecedented possibility to achieve a multidimensional and multiperspectival insight on war events, understanding how soldiers and citizens of different countries and social conditions experienced and described the events in which they were involved together, albeit on opposite fronts and with different roles. Grasping this unique opportunity however calls for advanced methods for the automatic semantic analysis of digital historical sources. The application of NLP methods and tools to historical texts is indeed attracting growing interest and raises interesting and highly challenging research issues (Piotrowsky 2012).

The research presented in this paper is part of a larger project dealing with the digitization and computational analysis of Italian War Bulletins of the First World War (for details see Boschetti et al. 2014). In particular, we focus here on the domain and language adaptation of a Named Entity Recognizer (NER) for Italian. As a byproduct of this project, we illustrate the automatic creation of a NE-annotated domain corpus used to adapt the NER to the WWI texts.

War bulletins (WBs) were issued by the Italian Comando Supremo "Supreme Headquarters" during WWI and WWII as the official daily report about the military operations of the Italian armed forces. They are plenty of Named Entities, mostly geographical locations, often referring to small places unmarked in normal geographic maps or with their name changed during the last century because of geopolitical events, hence hardly attested in any gazetteer.

To accomplish the Named Entity Recognition task, several approaches have been proposed such as Rule Based Systems (Grover et al., 2008; Mikheev et al., 1999a; Mikheev et al., 1999b), Machine Learning based (Alex et al., 2006; Finkel et al., 2005; Hachey et al., 2005; Nissim et al., 2004, including HMM, Maximum Entropy, Decision Tree, Support Vector Machines and Conditional Random Field) and hybrid approaches (Srihari et al., 2001). We used a Machine Learning approach to recognize NEs.

Rule-based systems usually give good results, but require long development time by expert linguists. 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.

The problems of the NER in WWI bulletins are larger than those encountered in modern texts. The language used in such texts is early

20th century Italian, which is quite different from contemporary Italian in many respects and belongs to the technical and military jargon. These texts are therefore difficult to analyze using available Italian resources for NER, typically based on contemporary, standard Italian. Grover et al. (2008) describe the main problems encountered by NER systems on historical texts. They evaluated a rule-based NER system for person and place names on two sets of British Parliamentary records from the 17th and 19th centuries. One of the most important issues they had to deal was the gap between archaic and contemporary language.

This paper is structured as follows: In section 2, we present the CoLingLab NER and in section 3 we describe its adaptation to WWI texts.

2 The CoLingLab NER

The CoLingLab NER is a NER for Italian developed with the Stanford CoreNLP NER (Finkel et al., 2005). The Stanford NER, also known as CRFClassifier, is a Java implementation of Named Entity Recognizer (NER) available for download under the GNU General Public License.

The classification algorithm used in Stanford NER is a Conditional Random Field (CRF) as in Lafferty et al. (2001). This model represents the state of the art in sequence modeling, allowing both a discriminative training, and a calculation of a flow of probability for the entire sequence.

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

Entity	P	R	F1	TP	FP	FN
B-GPE	0.828	0.765	0.795	870	181	267
B-LOC	0.767	0.295	0.426	46	14	110
B-ORG	0.726	0.65	0.684	834	315	455
B-PER	0.881	0.82	0.85	1892	255	413
I-GPE	0.73	0.583	0.649	84	31	60
I-LOC	0.833	0.366	0.508	30	6	52
I-ORG	0.556	0.442	0.493	192	153	242
I-PER	0.835	0.862	0.848	891	176	143
MicroAVG	0.811	0.735	0.771	4839	1131	1742

Table 1– CoLingLab NER trained and tested on I-CAB.

Table 1 reports the performance of the CoLingLab NER and Table 2 compares it with other state-of-the-art NER systems for Italian in EVALITA 2011.[1]

Participant	FB1	Precision	Recall
FBK_Alam_ro1	63.56	65.55	**61.69**
UniPi_SimiDeiRossi_ro1	58.19	65.90	52.09
UniPi_SimiDeiRossi_ro2	52.15	54.83	49.72
CoLingLab	**65.66**	**76.96**	59.76
BASELINE	44.93	38.84	53.28

Table 2 – Comparison between the CoLingLab NER and the 3 top models in Evalita (2011)

3 Adapting the NER to WWI bulletins

We use as test corpus (WB1) the Italian WBs of WWI. These texts come from the digitization of bulletins published in "I Bollettini della Guerra 1915-1918", preface by Benito Mussolini, Milano, Alpes, 1923 (pages VIII + 596).

To speed up the creation of the gold standard annotated corpus, these texts were first tagged semi-automatically with the existing NER, and then manually checked by an annotator to fix the incorrect tags and to add missing annotations.

The tagset consists of five entity classes, with begin-internal notation: Locations (LOC; e.g., *Monte Bianco*), Persons (PER; e.g., *Brandino Brandini*), Military Organizations (MIL; e.g., *Brigata Sassari,* Sassari Brigade), Ships (SHP; e.g., *Czepel*), Airplanes (PLN; e.g., *Aviatik*). The final test corpus consists of 1361 bulletins, covering the period from May 24, 1915 up to November 11, 1918 (1282 days).

In particular, the corpus is composed of 189,783 tokens, with the following NE distribution: 24 PER (13 B-PER – 11 I-PER); 19,171 LOC (12,542 B-LOC – 6,629 I-LOC); 38 SHP (33 B-SHP – 5 I-SHP); 1,249 MIL (615 B-MIL – 634 I-MIL); 54 PLN (52 B-PLN – 2 I-PLN). The corpus was automatically POS tagged with the Part-Of-Speech tagger described in Dell'Orletta (2009) and dependency-parsed with the DeSR parser (Attardi et al., 2009) using Support Vector Machines as learning algorithm.

In the following sections, we describe 3 experiments. First, we annotated the WBs using the existing CoLingLab NER trained on I-CAB. Then, we combined the I-CAB resource with WB2, a new domain NE-annotated corpus creat-

[1]The presented results have been produced according to the 'open' modality, therefore, with the possibility of using any type of supplementary data.

ed *ad hoc* from a digitized version of WBs of WWII. In the last experiment, we annotated WB1 with the NER trained on WB2 only. Tables 4, 5 and 6 report the following values to evaluate the performance of the various models: precision (P), recall (R), F1-Measure (F1), true positives (TP), false positives (FP), false negatives (FN), and microaveraged total scores (MicroAVG).

3.1 Features

In all the three experiments we used the same feature sets: morphological and orthographical features, information about the word shape, the part-of-speech tag, named entity tag, and contextual features.

In particular, we trained the models with the following types of features:

Word features. We used two different features: the first one considered next and previous words. For example, in the expression *"capitano [David Frazer]"* (captain David Frazer), the presence of the word "captain" helps to determine that the following words belong to the class PERSON. The second one considered a window of 6 words (3 preceding and 3 following the target word). It is useful to deal with cases like "*capitano di corvetta [Bandino Bandini]* PER" (Lieutenant Commander [Bandino Bandini] PER).

Orthographic features. We considered "word shape" features such as spelling, capital letters, presence of non–alphabetical characters etc.

Linguistic features. We used 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 et al., 2014). For example, the expression *"capitano di corvetta"* (Lieutenant Commander) is recognized by the system as a single item.

It is worth stressing that no information from gazetteers was used in the experiments reported below. It is clear that the system could be extended using lists of names of people, military groups, places, planes, and ships taken by several sources.

3.2 Experiment 1

In this experiment we tagged WWI texts using the CoLingLab NER (see Section 2) trained on a modified version of I-CAB in which we merged Locations with Geopolitical Entities, and we mapped I-CAB's Organizations into Military Organizations. Table 3 shows the mapping between I-CAB NEs and WBs NEs, which produced the following distribution of NEs: 10,487 PER (6,955, B-PER – 3,532 I-PER); 5,636 LOC (4,474 B-LOC – 1,162 I-LOC); 8,304 MIL (4,947 B-MIL – 3,357 I-MIL).

I-CAB		WWII-Bulletins	
B-LOC	LOC	LOC	B-LOC
I-LOC			
B-GPE	GPE		I-LOC
I-GPE			
B-ORG	ORG	MIL	B-MIL
I-ORG			I-MIL
B-PER	PER	PER	B-PER
I-PER			I-PER
-	-	SHP	B-SHP
-			I-SHP
-	-	PLN	B-PLN
-			I-PLN

Table 3– Mapping I-CAB and WB2 classes

Table 4 shows the results obtained using this mapped version of I-CAB as training corpus and the bulletins of WWI as test:

Entity	P	R	F1	TP	FP	FN
B-LOC	0.879	0.425	0.573	5327	732	7210
B-MIL	0.056	0.111	0.075	68	1142	541
B-PER	0.005	0.692	0.01	9	1747	4
I-LOC	0.827	0.433	0.568	2116	442	2771
I-MIL	0.077	0.093	0.084	33	395	323
I-PER	0.006	0.37	0.0118	3	498	5
Micro AVG	0.604	0.408	0.487	7556	4956	10950

Table 4–Annotation results using mapped I-CAB

In this experiment, the CoLingLab NER did not achieve good results on WB1. In fact, we can notice a significant decrease in the system ability to identify all kinds of NEs. This is due to the huge difference between the training and the test corpus, both in the language (modern Italian and generalist in I-CAB, archaic and military in WB1) and in the distribution of NEs, which in WB1 are strongly biased towards Locations.

3.3 Experiment 2

Given the unsatisfactory results obtained by annotating WB1 with a NER trained on a corpus from modern standard Italian, we have retrained the classifier using texts more similar to the test corpus.

Since the process of building annotated corpora is very expensive, we created a new automatically annotated training corpus (WB2) in a very fast way. We started from an html version of World War II Bulletins freely available[2] on the

[2]http://www.alieuomini.it/pagine/dettaglio/bollettini_di_guerra

Web, which includes an index containing different classes of NEs attested in the bulletins. The WW II bulletins were automatically downloaded and cleaned of html tags. The NE index was projected on WB2 to create a new training corpus, which was linguistically annotated with the same tools used for WB1. WB2 consists of 1,201 bulletins covering the time span from June 12th 1940 to September 8th 1943 (typically a bulletin per day), for a total of 211,778 tokens. WB2 is annotated with the same five classes as WB1, i.e. PER, LOC, MIL, PLN, and SHP. The class LOC includes both geo-political entities (e.g. *Italia*) and real locations (e.g. *Monte Bianco*), because such distinction was not marked in the original resource we used for the automatic construction of WB2.

We made a first experiment on this dataset using 10-fold cross-validation, obtaining a F1-Measure ~95%. This good result encouraged us to use WB2 as a gold standard to annotate WB1.

The model in the second experiment has been trained on the combination of I-CAB and WB2. Therefore, we mapped I-CAB's classes to WB2 classes as described in Table 3. The results obtained in this experiment are shown in Table 5. The combined corpora allowed us to increase the performances by 19%. It is worth noticing the significant improvement on Locations. This means that the new corpus provides the NER with much more evidence to identify this class. However, this improvement did not affect the recognition of PER and MIL. In these cases, in fact, we can observe a great number of false positives surely due to fact that I-CAB is very biased towards this class. Moreover, some semantic classes are not recognized because of the dearth of examples in the training data.

Entity	P	R	F1	TP	FP	FN
B-LOC	0.886	0.649	0.75	8141	1044	4396
B-MIL	0.174	0.186	0.18	113	537	496
B-PER	0.016	0.846	0.031	11	695	2
I-LOC	0.846	0.579	0.688	2831	517	2056
I-MIL	0.226	0.216	0.221	77	264	279
I-PER	0.02	0.625	0.038	5	250	3
Totals	0.772	0.604	0.678	11178	3307	7328

Table 5 – Annotation results using I-CAB + WB2

3.4 Experiment 3

In the last experiment, we trained our NER only on the WB2 corpus. This has the advantage of containing texts temporally and thematically closer to WB1, and a more balanced proportion of entity types. Results are presented in Table 6. For the sake of comparison with the previous experiments, we only provide a report for Locations, Persons and Military Organizations, leaving aside the identification of the SHP and PLN classes.

Entity	P	R	F1	TP	FP	FN
B-LOC	0.816	0.82	0.818	10279	2312	2258
B-MIL	0.474	0.074	0.128	45	50	564
B-PER	0.151	0.615	0.242	8	45	5
I-LOC	0.783	0.687	0.732	3359	929	1528
I-MIL	0.34	0.0899	0.144	32	57	324
I-PER	0.098	0.625	0.169	5	46	3
Totals	0,8	0.746	0.772	13728	3439	4682

Table 6 – Annotation results using WB2

The global scores obtained in this third experiment are higher than those in the second one, with a much lower amount of FPs per Persons and Military Organizations.

3.5 Discussion

Analyzing the results of the three experiments, the adapted NER performs better for Location names. This may reflect the sparsity of the data in the other classes.

It should be noticed that in the experiments 1 and 2, the number of false positives for persons and military organizations is very high. This seems to be a direct consequence of the different distribution of the observations in I-CAB compared to WBs.

Unsurprisingly, the best performing model is the one that has been entirely domain-tuned.

We are confident that new lexicons and gazetteers could help us to improve the identification of Locations and other Named Entities.

4 Conclusion

Location names play an important role in historical texts, especially in those - like WBs - describing the unfolding of military operations.

In this paper, we presented the results of adapting an Italian NER to Italian texts about WWI through the automatic creation of a new NE-annotated corpus of WBs. The adapted NER shows a significantly increased ability to identify Locations.

In the near feature, we aim at processing other types of texts about the Great War (e.g., letters, diaries and newspapers) as part of a more general project of information extraction and text mining of war memories.

References

Attardi, G., Dell'Orletta, F., Simi, M., Turian, J. (2009). Accurate Dependency Parsing with a Stacked Multilayer Perceptron. In Proceedings of EVALITA 2009, Reggio Emilia, Italy.

Bartalesi Lenzi, V., Speranza, M., and Sprugnoli, R. (2011). EVALITA 2011: Description and Results of the Named Entity Recognition on Transcribed Broadcast News Task. In Working Notes of EVALITA 2011, 24-25th January 2012, Rome, Italy.

Boschetti F., Cimino A., Dell'Orletta F:,Lebani G., Passaro L., Picchi P., Venturi G., Montemagni S., LenciA. (2014). Computational Analysis of Historical Documents: An Application to Italian War Bulletins in World War I and II. In Proceedings of the LREC 2014 Workshop on "Language resources and technologies for processing and linking historical documents and archives- Deploying Linked Open Data in Cultural Heritage", Reykjawik, Iceland.

Dell'Orletta, F. (2009). Ensemble system for Part-of-Speech tagging. In Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence, 12th December 2009, Reggio Emilia, Italy.

Finkel J.R., Grenager T. and Manning C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005).

Grover C., Givon S., Tobin R. and Ball J. (2008). Named Entity Recognition for Digitised Historical Texts. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco.

Hachey B., Alex B., and Becker M. (2005). Investigating the effects of selective sampling on the annotation task. In Proceedings of the 9th Conference on Computational Natural Language Learning.

Lafferty J., McCallum A., and Pereira F.(2001). Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th ICML. Morgan Kaufmann, San Francisco, CA.

Magnini B., Pianta E., Speranza M., Bartalesi Lenzi V. and Sprugnoli V. (2011).ITALIAN CONTENT ANNOTATION BANK (I-CAB): Named Entities.

Nissim M., Matheson C. and Reid J. (2004). Recognising geographical entities in Scottish historical documents. In Proceedings of the Workshop on Geographic Information Retrieval, SIGIR ACM 2004.

Mikheev A., Grover C. and Moens M. (1999a). XML tools and architecture for named entity recognition. Journal of Markup Languages: Theory and Practice, 1(3).

Mikheev A., Grover C. and Moens M. (1999b). Named entity recognition without gazetteers. In Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL'99).

Passaro, L., Lebani, G.E. and Lenci A. (2014). Extracting terms with EXTra. submitted.

Piotrowsky M. (2012).Natural Language Processing for Historical Texts, Morgan & Claypool.

Srihari R., Niu C., and Li W., (2001). A hybrid approach for named entity and sub-type tagging. In Proceedings of the 6th Applied Natural Language Processing Conference, pages 247–254, Seattle.