Effects of Semantic Analysis on Named-Entity Recognition with Conditional Random Fields – Extended Abstract –

Sonia Bergamaschi[1], Andrea Cappelli[2*], Antonio Circiello[1*], Marco Varone[2]

[1]Dipartimento di Ingegneria "Enzo Ferrari"
Università di Modena e Reggio Emilia - Italy
sonia.bergamaschi@unimore.it
a.circiello@outlook.com
[2]Expert System S.p.A.
Modena - Italy
mvarone@expertsystem.com
[*]affiliation when the work was performed

Abstract. We propose a novel Named Entity Recognition (NER) system based on a machine learning technique and a semantic network. The NER system is able to exploit the advantages of semantic information, coming from Expert System proprietary technology, Cogito. NER is a task of Natural Language Processing (NLP) which consists in detecting, from an unformatted text source and classify, Named Entities (NE), i.e. real-world entities that can be denoted with a rigid designator. To address this problem, the chosen approach is a combination of machine learning and deep semantic processing. The machine learning method used is Conditional Random Fields (CRF).
CRF is particularly suitable for the task because it analyzes an input sequence considering the whole sequence, instead of one item at a time. CRF has been trained not only with classical information, available after a simple computation or anyway with little effort, but with semantic information too. Semantic information is obtained with Sensigrafo and Semantic Disambiguator, which are the proprietary semantic network and semantic engine of Expert System, respectively. The results are encouraging, as we can experimentally prove the improvements in the NER task obtained by exploiting semantics.

1 Introduction

In this work we tackle the Named Entity Recognition (NER) task by combining machine learning with a heuristic disambiguation approach based on deep semantic analysis. NER is a well-known Natural Language Processing (NLP) task consisting in detecting and classifying Named Entities (NE) from free, unstructured text. Named Entities are typically defined as real-world concepts that can be referred to via rigid designators. We address this problem by employing a powerful machine learning algorithm, Conditional Random Fields (CRF), and providing it with data enriched by deep semantic processing. NER can be mapped into a labeling problem. CRF are very well suited to this kind of problems since they can analyze sequences as a whole, choosing labels for the sequence items with globally optimal choices instead of one item at a time. We employ Cogito, Expert System proprietary linguistic analysis technology, to enrich our machine learning pipeline with semantic information. In this work, CRF are fed with both standard linguistic features and semantic information obtained from Sensigrafo and Semantic Disambiguator, the proprietary semantic network and semantic engine of Expert System, respectively.

In this paper, results are presented about the performance improvements obtained by using semantic data together with standard features, as compared with the case in which only traditional features are employed.

Moreover, in the extended version of this paper submitted to the WIMS 2017 conference we performed more experiments varying the size of the corpus used for the analysis and showed the fundamental role semantics plays in many cases.

The rest of the paper is organized as follows.

Section 2 briefly describes our supervised Named Entity Recognition approach that is trained on both standard features and semantic information obtained from text analysis performed with the well-known Cogito linguistic analysis engine, developed by Expert Systems, an international Text Analytics and Cognitive Computing Company. Section 3 is devoted to the Experimental Results obtained on a reduced version of the larger Reuters Corpus, a collection of Reuters news articles. The documents in the corpus are related to various categories, from politics to sports. The training set is composed by one thousand documents while the test set is composed by four hundred documents. Finally, Section 4 outlines conclusions and future work.

2 The Method

The approach we propose is based on a CRF algorithm [11], that is trained on both standard features and semantic information obtained from text analysis performed with the Cogito linguistic analysis engine [3][12].

NER is a field that has been extensively explored in the last years. But the use of Expert System's technology, that is not simple semantic technology, but a complex system aimed, designed and optimized to create an optimal disambiguation process, could lead to very interesting outcomes. Our goal is to combine this resource to the tunability and domain adaptability of a machine learning algorithm such as CRF.

The final goal is to devise a new supervised Named Entity Recognition method, paying attention to the role of semantics in condition of scarse available training data. CRFs are a state-of-the-art class of machine learning algorithms to solve sequence labeling problems. They are part of the more general category of graphical models and are widely used in the domain of NLP, particularly as regards Part-Of-Speech tagging and Named Entity Recognition. Labels are obtained for an input sequence by evaluating label probabilities for a token given

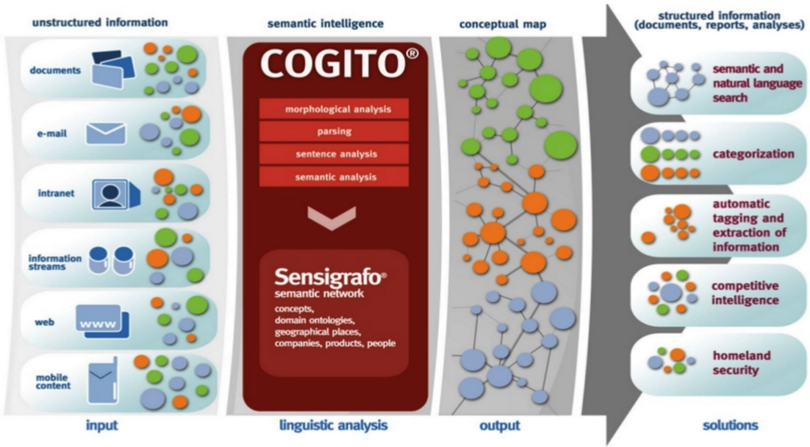


Fig. 1: Architecture of Expert System's semantic technology.

the surrounding tokens, their properties and earlier labels in the sequence. The most likely sequence is finally chosen based on an overall optimization over all possible label sequences.

The semantic analysis of the Cogito component named Semantic Disambiguator allows to associate words in the analysed text to *syncons*, a concept similar to WordNet synsets [8], which are related to each other via semantic links (hyperonymy, meronymy and others) in a proprietary semantic network called Sensigrafo. The linguistic engine is used also for basic linguistic tasks like tokenisation and POS-tagging, and for subject-verb-object relations detection. It also performs text categorisation. All of the information described above is combined in order to get a data matrix of linguistic information for each word in the text [5], so that features can be generated from it to train the CRF and finally detect entities.

Specifically, semantic information was employed to take advantage of the rich hyperonymy/hyponymy relations encoded in the semantic network. For that purpose, some columns of the data matrix were built as follows: for a given word in the text, its meaning ID was retrieved thanks to disambiguation [7][4][10]. Using it, the whole hyperonymy chain for that meaning was obtained, from the concept itself to its more abstract semantic ancestor (i.e. the last of its hypernyms of hypernyms, etc.). Then, moving top-down from that ancestor, up to four levels of ancestors were selected. The choice was limited to some specific nodes of the semantic network that are internally marked as *category nodes*, i.e. well representing a specific class of meanings (e.g.: verbs of communication or invertebrates). Each of such retrieved ancestor meanings (max 4) was used as one separate column in the data matrix. In other words, the four farthest hypernyms of the current meaning were retrieved (if present), that also are marked as "category nodes".

The rationale behind this procedure was to permit the clustering of the meaning of the words in the text at different levels of fine-graining, subsequently leav-

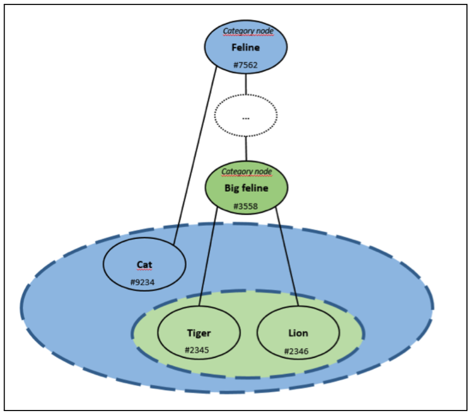


Fig. 2: Example of category nodes. Each node that is under this kind of nodes represents a concepts that semantically belongs to the category expressed by the category node.

ing the CRF the task of deciding which levels to pay more attention to. In this way, e.g., both the word lion and the word tiger are associated with the semantic father *feline* (as well as *vertebrate* and *animal*), and the CRF is enabled to determine the importance of such common property of the two different word, if the training subsequently highlights such importance. The procedure was used for all parts of speech for which a meaning was recognised.

A variation of the meaning clustering algorithm described above was employed for subject-verb and verb-object relations: words that were recognised as subjects or objects of a verb in the text were enriched with up to four more data columns, populated with the semantic ancestors of the verb they were subjects or objects of (the same logic described above applies). This allowed annotating words in the text with classes of verbs they are typically subject or object of. E.g. John plays basketball → John is annotated as a subject of the verb "to play", basketball is annotated as an object of the verb "to play".

Finally, semantic analysis provided categorisation: each document was given a category label (e.g. sports, news, medicine, science, etc.) based on the linguistic engine internal taxonomy. Such tag constituted one more data column for each word found in that document. Such feature was included in order to help recognise the different role of same words in different global context, with the category providing a context discriminator.

Standard data columns used besides semantic ones include: the form with which each word appears in the text, the lemma of the word (its normalised form), the part of speech, a list of regex-based columns (beginsWithUppercase, allUppercase, containsNumbers, allNumbers, etc.), character-type patterns, both extended and reduced (LeBron → extended: AaAaaa, reduced: AaAa; James → extended: Aaaaa, reduced: Aa).

The data matrix constructed with all of these data columns for each word was then used to generate CRF features: for each word, features were generated starting from data for that same word and for surrounding words (typically in a range of -2 to 2 position shifts, -5 to 5 for some cases). For the training phase, true labels in the IOB2 standard [9] were also included in such features. This is done by adding the letters B or I ahead of a label of a word in order not to loose information about multi-word named entities. The standard states that:

- The first word of a named entity is annotated with a B-label;
- Following words of the same entity, if they exist, are annotated with a I-label;
- A word that does not belong to any entity is annotated with O.

" O / A O / U.S. B-LOC /F-14 B-MISC/ military O/ plane O/ while O/ landing O/ at O/ Ben B-LOC/ Gurion I-LOC/ airport O/ blew O/ a O/ wheel O/ and O/ a O/ fire O/ broke O/ out O/, O/ " O/ said O/ spokesman O/ Yehiel B-PER/ Amitai I-PER.

Features were of the label unigram type, in the sense that the correct label of the preceding word was not included in the feature itself, except for the feature composed of the current label and the preceding label alone. As an example, O/O, O/B-LOC and B-LOC/I-LOC could actually occur while O/I-LOC could not, and this feature allowed to account for this. The CRF engine chosen for the experiments was the Wapiti[1] implementation [6]. Elastic-net regularisation was employed [1]. Elastic-net regularisation is a combination of the two regularisations L1 and L2, whose operating parameters are respectively $\rho 1$ and $\rho 2$. Parameters $\rho 1$ and $\rho 2$ were chosen via 10-fold cross-validation. For each fold, 7/10 of the training data were used for training, 2/10 for validation and convergence checks during training, and 1/10 for metric evaluation for the current fold (accuracy in our case, taking care of the macro F-1, the average F-1 score computed across all label types). Predictions were performed using Wapiti's posterior decoding option (some experiments were conducted also with Viterbi decoding, no significant variation was seen).

After parameters selection, quality metrics were assessed over a held out test set, prepared for all corpora used for the experiments.

3 Experimental Results

3.1 Corpus used for the experiments

For these experiments, we used the corpus prepared for the CoNLL 2003 workshop [2]. This corpus is a reduced version of the larger Reuters Corpus, a collection of Reuters news articles. The documents in the corpus are related to various categories, from politics to sports. The training set is composed by one thousand documents while the test set is composed by four hundred documents.

The documents of the CoNLL 2003 corpus are manually annotated (we did not do the annotation) with a label set comprising the following labels:

- PER, tag that represents human beings;

[1] https://wapiti.limsi.fr/

- ORG, tag that indicates companies, industries and other organizations;
- LOC, tag for geographic places;
- MISC, tag that represents other named entities not included in the previous categories;
- O, label that indicates a word not belonging to a named entity.

The training files have been formatted in order to respect the IOB2 format for representing the words belonging to a named entity.

We performed the following experiment on this corpus: A comparison between the performance of CRFs trained with non-semantic features and the performance of CRFs trained adding semantic features to the features set.

The same comparison between the two types of models, this time repeated with models trained on various different sizes of the training corpus (the original corpus has been artificially reduced) and its results are reported in the extended version of this paper, submitted to the 7th ACM International Conference On Web Intelligence, Mining and Semantics.

3.2 Comparison between models trained with and without semantics

The purpose of this experiment is to compare the results obtained with models trained with semantics features and the ones obtained training CRF without the use of semantic technology.

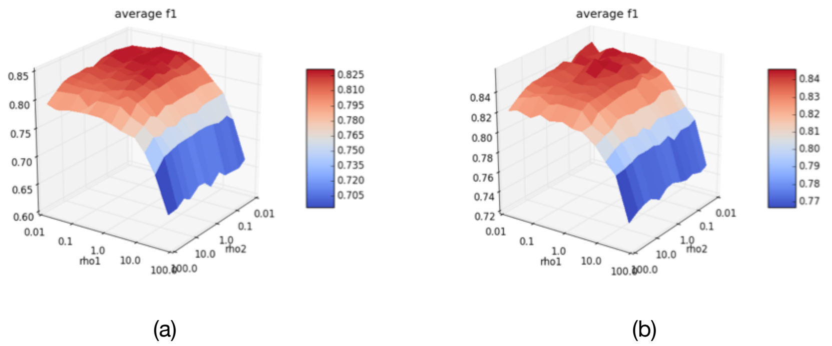


Fig. 3: (a) Case without semantics. (b) Case with semantics. The trends in the figures are similar but the values are different. Specifically, the semantic case leads to better performance, showing increasing values, with a difference of 0.015 - 0.020 with respect to the non-semantic case.

As previously explained, $\rho 1$ and $\rho 2$ parameters allow to configure the contributions of L1 and L2 regularizations. We trained different models using two ranges of these parameters. Doing this, we aimed at identifying the acceptable values for $\rho1$ and $\rho2$ that could lead to better performance of the models. The ranges used for the two parameters are: $\rho1 = [\ 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0, 20.0, 50.0\ ]$, $\rho2 = [\ 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0, 20.0, 50.0\ ]$.

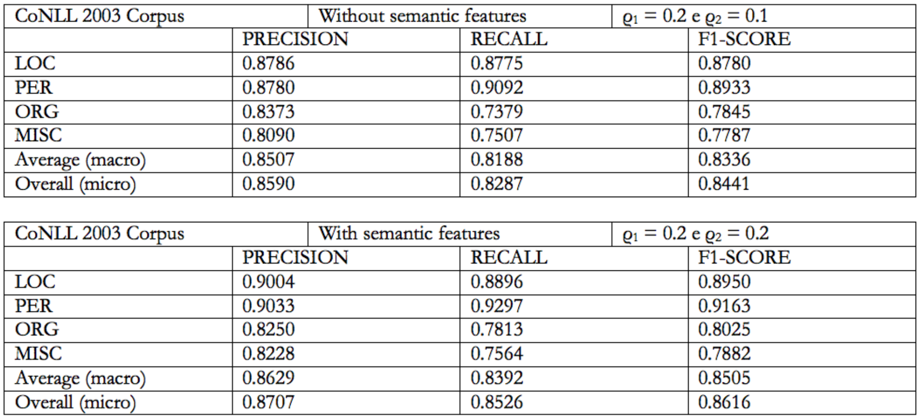
This results into 144 different models for each case (semantics and not).

The curves in Figure 3a and Figure 3b show the performance in the two cases. On the x and y axis there are the $\rho1$ and $\rho2$ values, while on the z axis there is the macro F-1 measure (calculated on the F-1 measure of each label). Each point in the graph represents a model, trained with the respective values of $\rho1$ and $\rho2$.

The better performances are concentrated near the origin of the axis. With ρ values greater than 0.5 the performances get worse. This is clearer with the parameter $\rho1$, whose bigger values lead to the worst performance.

Figure 4 reports numeric values for each tag in two cases: the best pair of $\rho1$ and $\rho2$ for the semantic case and for the non-semantic case. The best pair of ρ is the one which leads to the best result in terms of macro F-1 score.

Both from the plots and from the tables, we can see how the semantic features lead to better performance of the models. The semantic case shows increasing percentages, with a difference of 1.5 - 2 percentage points respect to the non-semantic case.



CoNLL 2003 Corpus	Without semantic features		$\varrho_1 = 0.2$ e $\varrho_2 = 0.1$
	PRECISION	RECALL	F1-SCORE
LOC	0.8786	0.8775	0.8780
PER	0.8780	0.9092	0.8933
ORG	0.8373	0.7379	0.7845
MISC	0.8090	0.7507	0.7787
Average (macro)	0.8507	0.8188	0.8336
Overall (micro)	0.8590	0.8287	0.8441

CoNLL 2003 Corpus	With semantic features		$\varrho_1 = 0.2$ e $\varrho_2 = 0.2$
	PRECISION	RECALL	F1-SCORE
LOC	0.9004	0.8896	0.8950
PER	0.9033	0.9297	0.9163
ORG	0.8250	0.7813	0.8025
MISC	0.8228	0.7564	0.7882
Average (macro)	0.8629	0.8392	0.8505
Overall (micro)	0.8707	0.8526	0.8616

Fig. 4: Numerical results for the comparison between the two cases (with and without semantics). Here we can better see how semantics leads to higher performance, in particular from the last column of the tables, which shows higher values in the semantic case.

With this experiment, we have identified the cases in which the category nodes help to better classify the words. For example, the adjectives of nationality (such as japanese or korean) are labeled as MISC. A simple CRF sees those adjectives as simple words, so if a different adjective of nationality has to be classified (such as chinese), the system will fail to identify the entity. Our NER system, instead, can recognize that those are not simple words, but they belong to a particular category node (the concept adjective of nationality), and that all the concepts belonging to that category node are classified as MISC; thus if a new adjective of nationality is presented in the test phase, the system correctly labels it as MISC. In the case of other MISC entity types (such as tournaments or public events), where fewer or less clear-cut examples are available, this generalization property seems to be less effective. However, generally speaking, this behavior improves performances and is one of the clearest advantages of our system.

4 Conclusions

We proposed a novel Named Entity Recognition (NER) system based on the combination of a machine learning algorithm with the semantic proprietary technology, Cogito, coming from Expert System S.r.l.

NER is the task of Natural Language Processing (NLP) which consists in detecting, from an unformatted text source, and classify Named Entities (NE), real-world entities that can be denoted with a rigid designator. To address this problem, the chosen approach was a combination of machine learning and deep semantic processing. The machine learning method used is Conditional Random Fields (CRF), particularly suitable for the task because it analyzes an input sequence considering it as a whole, instead of one item at a time. CRF has been trained both with classical information, available after a simple computation or anyway with little effort, and with semantic information too. Semantic information comes from Sensigrafo and Semantic Disambiguator, which are the proprietary semantic network and semantic engine of Expert System, respectively. We experimentally evaluated the NER system trained with and without semantic information and compared the results obtained. The results were promising, as we were able to experimentally prove the improvements in the NER task obtained by exploiting semantics. The extended version of this paper submitted to the WIST 2017 conference confirmed and improved these results, showing that when the size of the corpus used for the training is limited, the role of semantics plays is fundamental to improve the NER task.

Our approach can be used to tackle several problems within the category of sequencing problems, as for example, sentiment analysis, text segmentation, direct-speech extraction, etc. We think that, in all these research areas, semantics can help, e.g. by recognizing different words as synonyms or denoting similar concepts, as well as distinguishing different meanings of the same word. Combining words‘ semantics with machine learning tuning could permit to capture different nuances of words based on context, which might be difficult to model with hand-written rules. Other more general applications domains that can benefit froma our hybrid approach are keyword search on the deep web [4] and Entity Resolution [10].

References

1. http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2005.00503.x/full.

2. http://www.cnts.ua.ac.be/conll2003/ner/.
3. http://www.expertsystem.com/it/.
4. S. Bergamaschi, F. Guerra, M. Interlandi, R. T. Lado, and Y. Velegrakis. QUEST: A keyword search system for relational data based on semantic and machine learning techniques. *PVLDB*, 6(12):1222–1225, 2013.
5. J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289, 2001.
6. T. Lavergne, O. Cappé, and F. Yvon. Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513. Association for Computational Linguistics, July 2010.
7. M. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation, SIGDOC 1986, Toronto, Ontario, Canada, 1986*, pages 24–26, 1986.
8. G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244, 1990.
9. D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.
10. G. Simonini, S. Bergamaschi, and H. V. Jagadish. BLAST: a loosely schema-aware meta-blocking approach for entity resolution. *PVLDB*, 9(12):1173–1184, 2016.
11. C. A. Sutton and A. McCallum. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373, 2012.
12. M. Varone. Method and system for automatically extracting relations between concepts included in text, Mar. 1 2011. US Patent 7,899,666.