Conditional Random Fields with Semantic Enhancement for Named-Entity Recognition

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
andrea.cappelli86@gmail.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 of tokens considering it as a whole, 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 the addition of semantic information. 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, in particular when the training data size decreases.

CCS CONCEPTS

•**Information systems** → **Information integration; Information integration;** *Information retrieval;*

KEYWORDS

Named Entity Recognition, NLP, CRF, Cogito

ACM Reference format:
Sonia Bergamaschi[1], Andrea Cappelli[2#], Antonio Circiello[1#], Marco Varone[2]. 2016. Conditional Random Fields with Semantic Enhancement for Named-Entity Recognition. In *Proceedings of ACM WIMS, Italy, 2017 (Conference’17),* 7 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In this paper we propose a hybrid system for Named Entity Recognition (NER) combining machine learning techniques with semantic pre-processing and a semantic network. Specifically, the system exploits linguistic analysis and disambiguation performed by Cogito, Expert System proprietary technology. NER is the Natural Language Processing (NLP) task of identifying Named Entities (NE), which can be denoted with a rigid designator, in unstructured text. The chosen machine learning technique is Conditional Random Fields (CRF), since it is particularly suitable to analyzing input sequences (such as text) as a whole, instead of one item at a time [5], [10] [15]. CRFs have been trained on both standard linguistic features and semantic information gathered by processing input text with *Sensigrafo* and *Semantic Disambiguator*, Expert System??s semantic network and engine, respectively.

Employing a supervised sequencing algorithm can allow to tailor entity extraction on specific customer’s needs, even in case the customer is interested in non-standard entity types (e.g. different from People, Places, Organizations). The algorithm requires annotated texts from which to learn what entities are considered interesting, but this can save a significant amount of effort that would be needed to craft case-specific rules or algorithms. In some cases machine learning can generate models which are more fine-tuned than human-generated ones. All of these properties are just baseline advantages that can be improved using extra semantic information.

We show a comparison between systems trained in two cases, namely with and without semantic information. Results are encouraging, improvements due to semantics are shown, particularly when available data size is limited.

The paper pertains to two main conference topics. The first topic is Information Extraction and Knowledge Discovery from Big Data. In fact, this algorithm has a wide applicability in the context of large amount of data that must be processed to extract named entities even when the available training set is small. The second topic is Semantics-Driven Information Retrieval, because we are using

Conference’17, Italy
© 2016 ACM. 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

semantic analysis to improve the quality of our model and this model, despite being an extraction model, can be used to effectively index big amounts of data.

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. Specific information is 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. In particular, in subsection 3.3 the variation of the difference of performance of the models when the training data (i.e. the annotated corpus) size decreases is investigated and the effectiveness of semantics is shown. Finally, Section 4 reviews the related work and Section 5 outlines conclusions and future work.

2 THE METHOD

The approach we propose is based on a CRF algorithm [17], that is trained on both standard features and semantic information obtained from text analysis performed with the Cogito linguistic analysis engine [4][18].

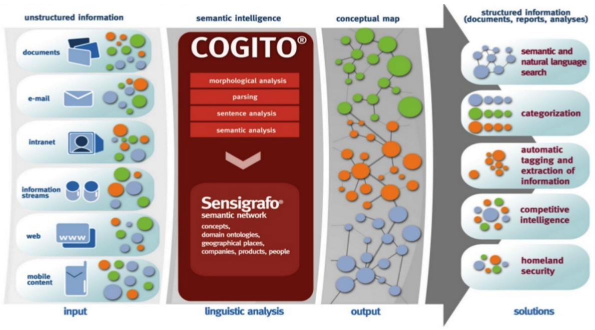


Figure 1: Architecture of Expert System's semantic technology.

The final goal is to devise a new supervised NER 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 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 [13], 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 [10], 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 [12]. 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".

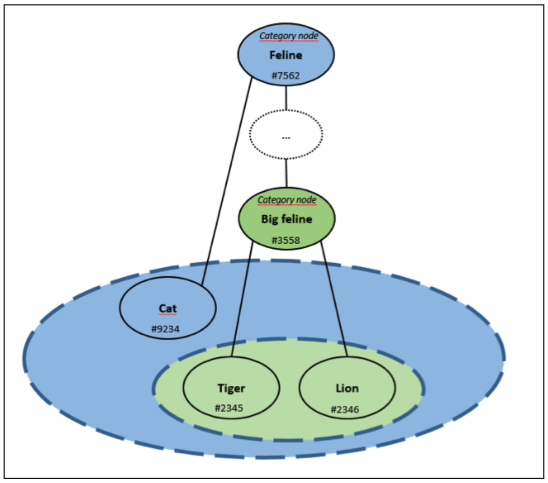


Figure 2: Example of category nodes. Each node that is under this kind of nodes represents a concepts that semantically belongs to the category expressed by the category node.

The rationale behind this procedure was to permit the clustering of the meaning of the words in the text at different levels of finegraining, subsequently leaving 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 words, 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 format [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 [11]. 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 (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 [3]. 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;
- 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.

Then, 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.

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.

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 $\rho 1$ and $\rho 2$ that could lead to better performance of the models. The ranges used for the two parameters are: $\rho 1$ = [0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0, 20.0, 50.0], $\rho 2$ = [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 $\rho 1$ and $\rho 2$ 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 $\rho 1$ and $\rho 2$.

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

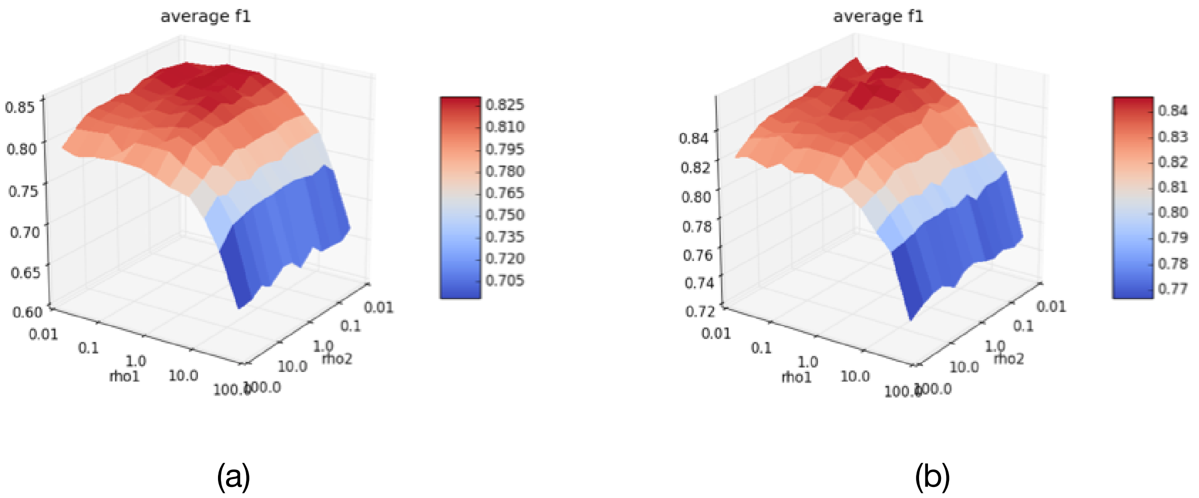


Figure 3: (a) Case without semantics. (b) Case with semantics. The trends in the figures are similar but the values are different. Specifically, the semantic case leads to better performance, showing increasing values, with a difference of 0.015 - 0.020 with respect to the non-semantic case.

The best 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 $\rho 1$, whose bigger values lead to the worst performance.

Figure 4 reports numeric values for each tag in two cases: the best pair of $\rho 1$ and $\rho 2$ 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 with respect to the non-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.

3.3 Comparison between the two cases with varying size of the corpus

The purpose of this experiment is to investigate the variation of the difference between the performance of models trained with and without semantics while the training data decrease. One of the most difficult tasks in machine learning for NLP applications is to find a large enough annotated corpus. This problem could be partly addressed with the use of semantics.

For this second experiment, we prepared multiple corpora, reducing the size of the original one, the CoNLL 2003 corpus. Starting from the original corpus (100%), we prepared 10 corpora, being respectively 10%, 20%, 100% of the size of the original one.

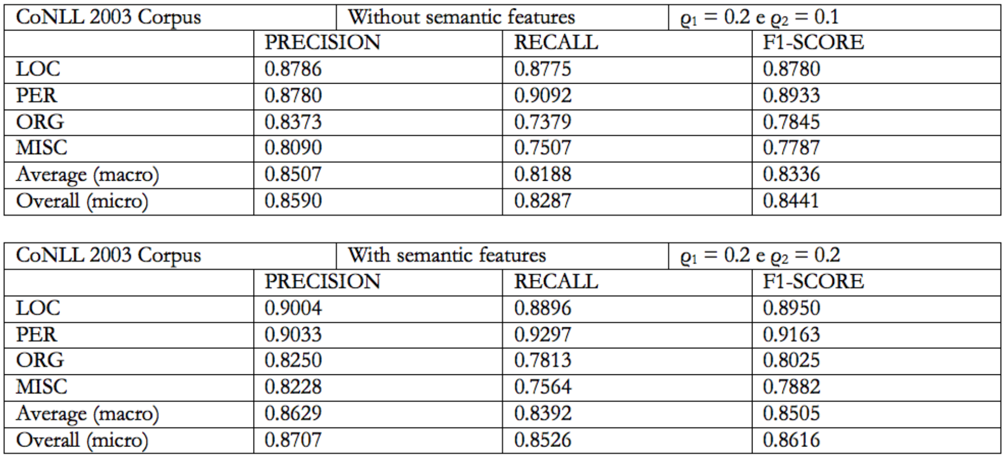
For this second experiment, we used the following ranges for parameters optimizations: $\rho 1$ = [0.01, 0.02, 0.05, 0.1, 0.2, 0.5], $\rho 2$ = [0.01, 0.02, 0.05, 0.1, 0.2, 0.5].

This results in 36 different models for each case (semantics and not) and for each different training corpus.

Figure 5 shows two examples of comparison between the curves that represent the performances of the models trained with and without semantic features. When approaching smaller values for the training set size (from right to left), the two curves exhibit a larger gap.

Figure 6 is a summary table reporting, for each different training corpus, the best average value for F-1 score (the maximum value among the macro F-1 scores of all models for that case) for the semantic case and for the non-semantic case and the difference between these two values (as a percentage).

As we can see from the table, the difference between the performance of the model trained with semantic features and the performance of the one trained without semantic technology increases while the training data size decreases, going from 1.65 for the case with the original corpus size to 6.27 percentage points for the 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

Figure 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.

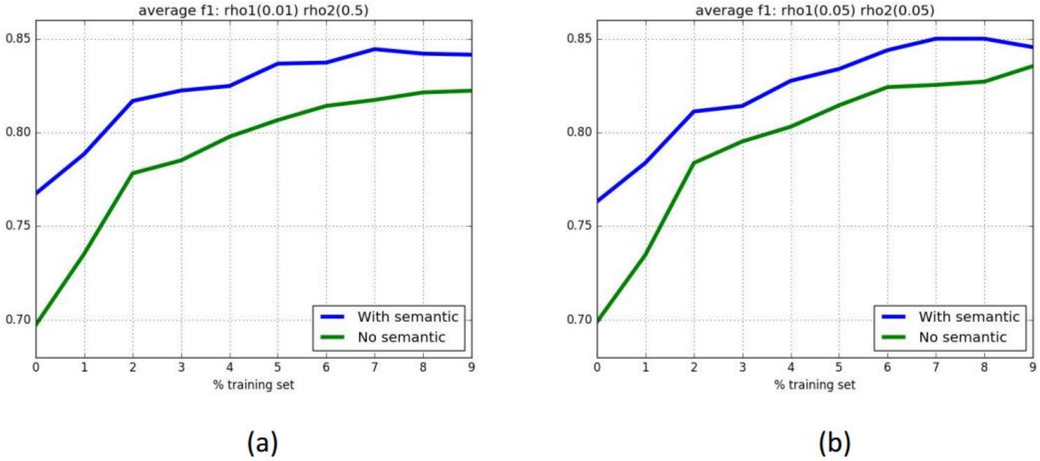


Figure 5: Comparison between the average F-1 scores for a specific pair of ρ while varying the training set size

with the smallest derived training corpus. The trend is exponential. Figure 7 shows this difference for a specific pair of $\rho1$ and $\rho2$.

This shows how the semantic technology plays a fundamental role when the training annotated data are scarce, permitting to address in a more efficient way problems due to lack of data. This is due to the capacity of the semantic analysis to better understand how the words of a text are clustered, linked and classified.

4 RELATED WORKS

Here we report the characteristics of two NER system realized by other organizations/people and documented in literature: the Stanford NER [2] and another system realized for a workshop/competition [14].

Stanford NER is a Java implementation of a Named Entity Recognizer. The software implements linear-chain CRF; when the training is done on annotated data, the code can be used for constructing sequential models for NER or other tasks. Software is realized by Jenny Finkel, Dan Kleid, Christopher Manning, Anna Rafferty and other members of NLP Group of Stanford University.

CoNLL 2003 Corpus			
Corpus size respect to the original corpus	AVERAGE F1-SCORE (without semantics)	AVERAGE F1-SCORE (with semantics)	DIFFERENCE (%)
10 %	0.7069	0.7697	6.27 %
20 %	0.7515	0.7923	4.07 %
30 %	0.7874	0.8183	3.09 %
40 %	0.7975	0.8272	2.96 %
50 %	0.8108	0.8312	2.03 %
60 %	0.8207	0.8387	1.79 %
70 %	0.8254	0.8438	1.84 %
80 %	0.8317	0.8508	1.91 %
90 %	0.8336	0.8529	1.93 %
100 %	0.8387	0.8553	1.65 %

Figure 6: Numerical results for the comparison between the two cases varying training set size. Here, the F1-score averages are performed first on entity labels and then on ρ pairs.

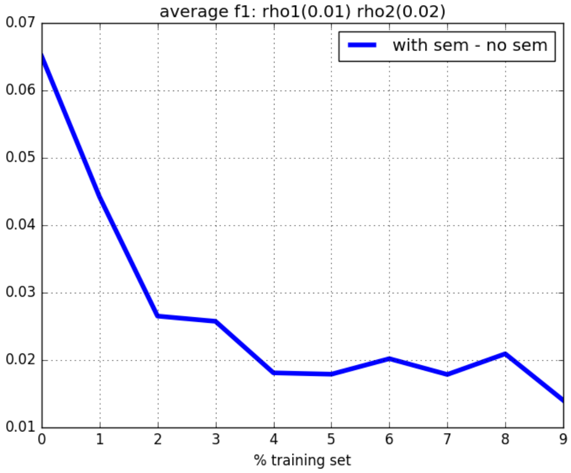


Figure 7: Difference between semantic models and non-semantic models in relation to the training corpus size.

Features provided for Stanford NER are: features related to the words themselves (the word, the next word, the previous word, words contained in a fixed window), orthographic features, prefix and suffix and distributional similarity features. Distributional similarity features represents clusters id; starting from a not annotated corpus, distributions over words contexts are created and, finally, words are clustered based on the similarity of their distibutions.

The structure of the model used by Stanford software is similar to the base model presented in [8]. That paper presents two models suitable for incorporating non local information as CRF features using Gibbs sampling instead of Viterbi algorithm. The model at the base of these two is a local feature model and use Viterbi algorithm. The choice fell on CRFs because it is the state of the art in sequence modeling and permits a bidirectional flow of probabilistic information along the sequence.

The other NER system is applied to biomedical field. The system described in [14] is a framework for recognizing biomedical entities with a CRF. The approach presented brings to F-1 score of circa 0.7. This system has been developed in the context of the competition BioNLP/NLPBA 2004. The model used include a training dictionary, orthographic features based on regular expressions (the token is alphanumeric or not, the token presents roman numerals, etc), features for prefixes and suffixes. If training sequences contain sets of tokens such as ”PML/RAR alpha” or ”beta 2-M” annotated as belonging to category ”protein”, the model could learn that the sets are linked because they contain greek alphabet letters. This kink of knowledge is supplied via lexicons. Lexicons can be inserted manually (like greek letters, chemical elements, known virus and related abbreviations) or extracted from database.

5 CONCLUSIONS AND FUTURE WORK

We proposed a hybrid NER system combining a Conditional Random Fields approach with semantic analysis. The system employs standard linguistic features and enriches them with semantic information provided by the Semantic Disambiguator and Sensigrafo, the core components of the Cogito semantic analysis technology (such information included references to nodes on a semantic network, their hypernyms, text category, etcfb) to improve generalisation capabilities in the entity extraction task. The presented results show the positive effect of adding semantic information to the available features in the NER task and the increasing importance of semantics when the available training dataset is scarce. In that case, support from disambiguation over a semantic network mitigates the problem of incomplete training examples.

The help provided by semantics in conditions of scarce training datasets can prove useful in practical applications: it can be possible to build a NER system working on custom or non-standard entity types, with just a small number of entity examples, tagged by a domain expert. This can be the first stage of an iterative procedure of validation and retraining that can provide good improvements with a relatively small effort.

The algorithm we employed in this work for NER can be used to tackle several other problems that can be mapped into sequencing problems. Examples of these uses can be found in the areas of sentiment analysis, text segmentation, direct-speech extraction, relation extraction, etc... In all of these areas we think that semantics could help, e.g. by recognizing different words as synonyms or similar concepts, as well as distinguishing different meanings of the same word. Combining this with machine learning tuning could allow to capture difficult nuances of terms based on context, which might be difficult to completely model with hand-written rules. We think these might be interesting ideas for future works.

Other application domains can be keyword search on the deep web [7], Entity Resolution [16], and data source topic detection [6].

For example, the QUEST system which exploits HMM for matching keywords and database structures [7]. Through CRFs, the precision of the approach could be improved by considering the order of the keywords.

Finally, BLAST [16], which is the state-of-the-art blocking technique for Entity Reoslution, could be adapted and employed to scale the NER process to huge datasets.

REFERENCES

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

[2] https://nlp.stanford.edu/software/CRF-NER.shtml.

[3] http://www.cnts.ua.ac.be/conll2003/ner/.

[4] http://www.expertsystem.com/it/.

[5] Yasemin Altun, Ioannis Tsochantaridis, and Thomas Hofmann. 2003. Hidden Markov Support Vector Machines. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*. 3–10. http://www.aaai.org/Library/ICML/2003/icml03-004.php

[6] Sonia Bergamaschi, Davide Ferrari, Francesco Guerra, Giovanni Simonini, and Yannis Velegrakis. 2016. Providing insight into data source topics. *Journal on Data Semantics* 5, 4 (2016), 211–228.

[7] Sonia Bergamaschi, Francesco Guerra, Matteo Interlandi, Raquel Trillo Lado, and Yannis Velegrakis. 2013. QUEST: A Keyword Search System for Relational Data based on Semantic and Machine Learning Techniques. *PVLDB* 6, 12 (2013), 1222–1225. http://www.vldb.org/pvldb/vol6/p1222-guerra.pdf

[8] 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)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 363–370. DOI: https://doi.org/10.3115/1219840.1219885

[9] Vijay Krishnan and Vignesh Ganapathy. 2005. Named Entity Recognition. (2005).

[10] 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 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*. 282–289.

[11] Thomas Lavergne, Olivier Cappé, and François Yvon. 2010. Practical Very Large Scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 504–513. http://www.aclweb.org/anthology/P10-1052

[12] Michael Lesk. 1986. 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*. 24–26. DOI: https://doi.org/10.1145/318723.318728

[13] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to WordNet: An on-line lexical database. *International journal of lexicography* 3, 4 (1990), 235–244.

[14] Burr Settles. 2004. Biomedical Named Entity Recognition using Conditional Random Fields and Rich Feature Sets. In *COLING 2004 International Joint workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP) 2004*, Nigel Collier, Patrick Ruch, and Adeline Nazarenko (Eds.). COLING, Geneva, Switzerland, 107–110.

[15] Fei Sha and Lawrence K. Saul. 2007. Comparison of Large Margin Training to Other Discriminative Methods for Phonetic Recognition by Hidden Markov Models. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2007, Honolulu, Hawaii, USA, April 15-20, 2007*. 313–316. DOI: https://doi.org/10.1109/ICASSP.2007.366912

[16] Giovanni Simonini, Sonia Bergamaschi, and H. V. Jagadish. 2016. BLAST: a Loosely Schema-aware Meta-blocking Approach for Entity Resolution. *PVLDB* 9, 12 (2016), 1173–1184. http://www.vldb.org/pvldb/vol9/p1173-simonini.pdf

[17] Charles A. Sutton and Andrew McCallum. 2012. An Introduction to Conditional Random Fields. *Foundations and Trends in Machine Learning* 4, 4 (2012), 267–373. DOI: https://doi.org/10.1561/2200000013

[18] Marco Varone. 2011. Method and system for automatically extracting relations between concepts included in text. (March 1 2011). US Patent 7,899,666.