Spanish Nested Named Entity Recognition using a Syntax-dependent Tree Traversal-based Strategy

Yunior Ramírez-Cruz and Aurora Pons-Porrata

Center for Pattern Recognition and Data Mining
Universidad de Oriente
Santiago de Cuba, Cuba, 90500
{yunior, aurora}@cerpamid.co.cu

Abstract. In this paper, we address the problem of nested Named Entity Recognition (NER) for Spanish. Phrase syntactic structure is exploited to generate a tree representation for the set of phrases that are candidate to be named entities. The classification of all candidate phrases is treated as a single problem, for which a globally optimal solution is approximated using a strategy based on the postorder traversal of that representation. Experimental results, obtained in the framework of SemEval 2007 Task 9 NER subtask, demonstrate the validity of our approach.

1 Introduction

Named Entity Recognition (NER) is a basic step for a number of tasks such as Information Extraction, Question Answering and Automatic Summarization. The classic NER task was introduced in the Message Understanding Conferences and consists in detecting and classifying elemental information units contained in text documents such as proper names (persons, organizations, locations, etc.), quantities and temporal expressions. The classic problem definition considers no nesting or overlapping between different Named Entities (NE's).

Extensive work has been conducted on NER. The greatest effort has been focused on English documents covering the biomedical and newswire domains. While the annotation of benchmark corpora for English NER in the biomedical domain considers nested NE's, this is not common for the newswire domain. Because of this, research efforts to deal with the problem of recognizing nested structures have been largely confined to the former domain, although nested NE's are likely to occur in any knowledge domain.

Until recently, Spanish corpora have often lacked the annotation of nested NE's. Consequently, a small number of works have addressed the problem for that language. However, being able to recognize all NE's is crucial for other tasks depending on it, such as coreference resolution and scenario template matching, since nested structures implicitly contain relations that may help improve their performance.

In this paper, we focus on the nested NER problem for the Spanish language. Unlike most common approaches, which treat the problem either as a postprocessing stage of the classic NER problem or as a combination of several instances of it, we address the recognition of all NE's included in a nested structure as a single problem. For each sentence, phrases that are candidate to be NE's are detected in the deep constituency tree. A second tree containing the representations of all candidate phrases through a set of boolean features is generated. The structural and functional relations imposed to the candidate phrases by the syntax are encoded in that tree. Finally, a globally optimal classification for all candidate phrases is approximated using a strategy based on the postorder traversal of the representation tree.

For evaluation purposes, we use the SemEval 2007 Task 9 Spanish dataset for the NER subtask. We evaluate the impact of several elements in our model and establish a comparison between the results obtained by our method and those reported for that subtask.

The rest of the paper is structured as follows. In Section 2 we review previous work on nested NER. Section 3 is devoted to describing our approach, whereas Section 4 contains the description of the experiments that we carried out. Finally, we expose our conclusions in Section 5.

2 Related Work

As we mentioned before, a considerable part of the research effort on nested NER has been focused on English documents in the biomedical domain, mainly due to the availability of corpora such as GENIA [1], which contains MEDLINE abstracts annotated with nested NE's. Aiming to reuse well known, successful techniques, common approaches to nested NER treat the problem either as a separate postprocessing stage of the classic NER problem or as the combination of several instances of the classic problem.

The first type of approaches consists in extending classic NE recognizers through a mechanism for merging together several NE's or detecting new NE's that either are embedded in the original ones or contain them. For instance, Zhang et al. [2] address English biomedical NER by using Hidden Markov Models (HMM's) to recognize the innermost NE's. As a postprocessing stage, they apply a set of rules extracted from the training data to detect other NE's that contain the initial ones.

The second type of approaches consists in handling different nesting levels or NE types as separate problems by combining several instances of a classic recognizer. Zhang et al. [2] also propose a method consisting in the application of several passes of an HMM-based recognizer, adding a new nesting level in each one. The process starts by recognizing the innermost phrases. After a pass is completed, the input sequence is modified in such a way that recognized entities are treated as special tokens, and then a new pass is performed to obtain a new nesting level. The process ends when no new NE's are found after a pass is completed. As the authors point out, errors in each pass affect the recognition

process in further passes. Alex et al. [3] discuss three modeling techniques using the BIO encoding in English biomedical NER. Two of these techniques fall into this type of approach: *layering*, where each nesting level is modeled as a separate BIO problem; and *cascading*, where entity types are divided into groups and a separate model is trained for each group. Layering is affected by the same situation mentioned before: errors in a level negatively impact the recognition process for further levels; whereas cascading is unable to recognize NE's that are embedded in other NE's of the same type.

The third modeling technique discussed by Alex et al., *joined label tagging*, does not belong to any of these two types of approaches. It consists in creating a new tagging scheme, where BIO tags for all nesting levels are concatenated to process all levels in a single pass. As they point out, this technique is prone to be affected by data sparseness. A similarity between this idea and our approach is that the recognition process for all nesting levels is performed in a single pass. However, Alex et al. address the problem by tagging a word sequence using a modification of the BIO tagset, whereas we classify entire phrases instead of independent words. Besides, they follow the order of words in the sentence, whereas we apply a syntax-dependent tree traversal-based strategy.

For Spanish, due in part to the lack of annotated corpora containing nested NE's, a small number of works have addressed the problem. The MICE system [4] relies on the distinction between strong NE's, which contain proper nouns, and weak NE's, which are constructions containing trigger words and optionally other embedded NE's. AdaBoost is used to recognize and classify strong NE's whereas a handcrafted context independent grammar is used to recognize weak NE's. A complete quantitative evaluation of this method is not presented in [4] due to the early stage of the corpus development.

Recently, within the framework of SemEval 2007 Task 9 [5], Spanish and Catalan nested NER was addressed as a subtask. The UPC system [6] was the only one that submitted results for NER. Based on the distinction between strong and weak NE's that characterizes the competition corpus, UPC performs NER in two stages, both using AdaBoost. The first stage deals with the classification of strong NE's, whereas the second deals with the detection and classification of weak NE's. UPC, as well as MICE, are examples of the first type of approaches to nested NER mentioned earlier.

The two types of approaches to nested NER that we have discussed here tackle the complexity of the problem by splitting it into subproblems that are solved separately. However, in doing this, it is possible that useful interactions between these subproblems are lost. We consider that approaches intending to find near-to-globally optimal solutions may exploit these interactions in order to obtain better results.

3 Our Approach

In the proposed method, the recognition process is carried out in a sentence by sentence basis. For each sentence, candidate phrases, i.e., those that may be

NE's, are detected in the deep constituency tree. Candidate phrase detection is performed based on the syntactic labels of phrases, following the criterion that any definite noun phrase may be considered as a candidate. The set of candidate phrases is represented using a tree, where each node contains the representation of one phrase and nesting is encoded by the parent-child relation between nodes. A set of boolean features is used to describe phrases, thus the representation of a phrase is a boolean vector containing the results of evaluating these features on a vicinity of the phrase in the deep constituency tree. Using a strategy based on the postorder traversal of the representation tree, the classification of the set of candidate phrases is carried out in such a way that the set of classes given to all candidate phrases in a nested structure approximates the global optimum. Next we describe this process in detail.

3.1 Obtaining the representation

Let $S = w_1 w_2 ... w_n$ be a sentence. Consider two phrases $P = w_i w_{i+1} ... w_{i+k}$ ($i \geq 1$, $i + k \leq n$) and $P^{'} = w_j w_{j+1} ... w_{j+k'}$ ($j > 1$, $j + k^{'} < n$) contained in S. If $j > i$ and $j + k^{'} < i + k$, we say that $P^{'}$ is embedded in P. If no phrase $P^{''}$ embedded in P exists such that $P^{'}$ is embedded in $P^{''}$, we say that $P^{'}$ is immediately embedded in P.

For each sentence, once the deep constituency tree has been obtained and candidate phrases are detected in it, the representation tree is constructed in such a way that each candidate phrase is represented by a node. If the node N represents phrase P, its children are those nodes $N_1, N_2, ..., N_e$ representing phrases $P_1^{'}, P_2^{'}, ..., P_e^{'}$ that are immediately embedded in P. Unless the whole sentence is a candidate phrase, an artificial root node is added such that all nodes representing candidate phrases that are not immediately embedded in any other are its children.

For example, consider the sentence *Agüero visitó Trinidad y Tobago* (*Agüero visited Trinidad and Tobago*). The named entities contained in that sentence are *Agüero* (person name), *Trinidad y Tobago* (location name), *Trinidad* and *Tobago* (location names embedded in *Trinidad y Tobago*). Figure 1 shows the representation tree for that sentence. For clarity, in the figure every node (except the artificial root node) is labeled as *Repr: P*, where P is the candidate phrase represented by it. Notice that what the nodes in the tree actually contain are boolean vectors representing the phrases.

Let $\wp$ be the set of all possible candidate phrases. A phrase P will be represented in terms of a set of q boolean features $f_i : \wp \rightarrow \{0, 1\}$ ($1 \leq i \leq q$). Thus, the representation of phrase P is a boolean vector $r = (f_1(P), f_2(P), ..., f_q(P))$.

We consider two types of features. The first one includes features that are evaluated locally, i.e., on the phrase per se. We use POS tags, trigger word dictionaries and gazetteers of location and organization names.

The second type are syntax-dependent features. A set of these features checks whether the syntactic function of the candidate phrase or, alternatively, that of a prepositional phrase in which it is immediately embedded, is one of the following:

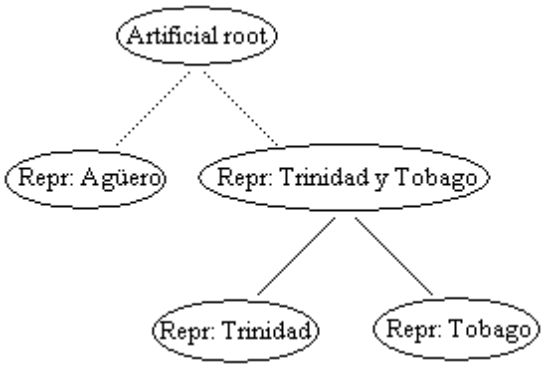

Fig. 1. Representation tree obtained for the sentence *Agüero visitó Trinidad y Tobago*.

subject, direct object, indirect object, adjunct or agent complement. Additionally, a set of verb lemma dictionaries is collected from the training corpus and a feature is defined for each dictionary. Each one of these features checks whether the lemma of the verb governing the clause where the candidate phrase is embedded occurs in its associated dictionary. In order to collect the set of dictionaries, a mapping is constructed between entity types and syntactic functions. Thus, dictionaries are constructed for "person as subject", "organization as object", and so on. For each entity type T and each syntactic function F, a dictionary is constructed which contains the lemmas of verbs occurring in at least one clause in the training corpus where a NE of type T is found being its syntactic function, or that of a prepositional phrase where it is immediately embedded, F. Auxiliary verbs are disregarded.

For example, when obtaining the representation for the phrase *Agüero* in the sentence considered in Figure 1, suppose a feature f_i is associated with a dictionary containing the verb *visitar* (lemma of *visitó*). This feature will yield the value 1 when evaluated on the phrase because it belongs to a clause governed by *visitó*. In constructing the dictionaries, if this sentence were a part of the training data, the dictionary corresponding to the mapping "person as subject" would contain *visitar*, because *visitó* is the verb governing the clause and the syntactic function of *Agüero* (a person name) is subject.

The combination of syntactic function information and verb lemma dictionaries is expected to help take into account the behavior of candidate phrases. Intuitively, verb lemma dictionaries are expected to provide information about the action in which the phrases are involved and the syntactic function is expected to provide a rough approximation of the role that each phrase is playing.

While locally evaluated features are expected to help determine the literal type of a NE, syntax-dependent features should allow to discriminate between literal and metonymic readings of NE's.

3.2 Traversing the representation tree

In selecting an appropriate order to traverse the representation tree, we consider the following ideas. First, in determining the type of a NE, it is convenient that the types of its embedded phrases are considered. Second, since errors in irrevocably classifying a phrase cause a negative impact in the classification of the phrases in which it is embedded, the best classification is the one that satisfies some global optimality measure.

In order to allow the syntactic structure of the sentence to guide the classification process, we propose a strategy that consists in generating a sequence of candidate phrase representations following a postorder traversal of the representation tree. The sequence thus obtained is such that the node representing a phrase is located after those representing its embedded phrases and after those representing preceding phrases. Figure 2 illustrates the situation for the previously analyzed sentence.

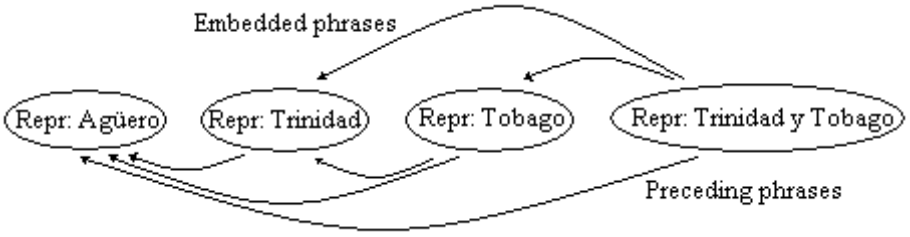

Fig. 2. Dependences expected to be captured by the postorder traversal strategy.

3.3 Classification

We treat the classification of the set of candidate phrases as a sequence classification problem by using a Markovian approach, according to which only one independence assumption needs to be inserted. A Viterbi search is performed on that sequence to capture the desired dependences of a candidate phrase to its embedded phrases, as well as those to previoulsy occurring phrases, which are commonly used in Markovian approaches. We combine all candidate phrases in a single sequence and approximate a globally optimal solution without making the classification of any level irrevocable. This aims to avoid the problems of iteratively classifying superimposed nesting levels pointed out in Section 2.

For the classification process, the class set C contains one class for each NE type, plus an extra class *NONE* to handle candidate phrases that are not NE's.

Given a sequence of boolean vectors $r_1, r_2, ..., r_m$, which represent candidate phrases obtained following the postorder traversal of the representation tree, we use a Conditional Markov Model (CMM) to obtain the class sequence

$c_1, c_2, ..., c_m$ that maximizes the probability $P(c_1, c_2, ..., c_m | r_1, r_2, ..., r_m)$. We follow the independence assumption adopted by Punyakanok and Roth [7]:

$$P(c_t | r_t, ..., r_1, c_{t-1}, ..., c_1) = P(c_t | r_t, c_{t-1}) \quad (1)$$

That is, at step t, the probability of the phrase represented by r_t to be classified as c_t depends on the representation itself and on the class given to the phrase represented by the previous vector in the sequence.

In their work, Punyakanok and Roth split $P(c_t | r_t, c_{t-1})$ into $|C|$ functions $P_{c_{t-1}}(c_t | r_t)$. Here we do not follow this approach to prevent class unbalance from causing data sparseness in estimating $P_{c_{t-1}}(c_t | r_t)$ for some values of c_{t-1}. Instead, we codify c_{t-1} in terms of $|C|$ additional features $f'_{q+j} : C \rightarrow \{0, 1\}$ $(1 \leq j \leq |C|)$ such that

$$f'_{q+j}(c_{t-1}) = \begin{cases} 1 \text{ if } c_{t-1} = c_j \\ 0 \text{ otherwise} \end{cases} \quad (2)$$

and obtain a new vector $r'_t = (f_1(P_{r_t}), ..., f_q(P_{r_t}), f'_{q+1}(c_{t-1}), ..., f'_{q+|C|}(c_{t-1}))$, where P_{r_t} is the phrase whose representation is r_t. The new vector is an extension of r_t, which allows us to express $P(c_t | r_t, c_{t-1}) = P(c_t | r'_t)$.

We use the variant of the Viterbi algorithm described by Punyakanok and Roth to approximate the globally optimal solution.

In this paper, we consider two variants of our postorder traversal-based classification strategy. The first one consists in constructing a single sequence containing all the nodes in the representation tree, except the root node when it was artificially added. In what follows, we will refer to this variant as sentence-level postorder traversal-based strategy. The second variant limits interactions to those between phrases that belong to the same nested structure. In this variant, for the most common case when the root node is artificial, a separate sequence is obtained for each subtree corresponding to a child node of the root. Each sequence thus obtained is processed independently during both training and classification. Even though the whole sentence is not processed at once by using this variant, contextual information may still be captured if syntax-dependent features are used in describing phrases. We will refer to this variant as nested structure-level postorder traversal-based strategy. Notice that if the root is not an artificial node, both variants are equivalent.

Probability estimation is carried out using Maximum Entropy (ME). For describing phrases, ME relies on a set of functions that depend not only on the result of evaluating a boolean feature on the phrase but also on the class for which the probability is estimated. We integrate our feature set into the ME framework following the idea described by McCallum et al. [8]. Thus, for each component $r'_{t,i}$ $(1 \leq i \leq q + |C|)$ of the extended representation vector r'_t and each class $c_j \in C$, a function

$$g_{i,j}(r'_t, c_t) = \begin{cases} 1 \text{ if } r'_{t,i} = 1 \text{ and } c_t = c_j \\ 0 \text{ otherwise} \end{cases} \quad (3)$$

is constructed. We train the estimators using Generalized Iterative Scaling [9].

4 Experiments

The purpose of our experiments is three-fold. Firstly, we evaluate the usefulness of syntax-dependent features for describing candidate phrases. Secondly, we compare the behavior of the postorder traversal-based classification strategy, both the sentence-level and the nested structure-level variants, against two bottom-up strategies using the same probability estimator. Finally, we compare our results to those reported for the NER subtask at SemEval 2007 Task 9 [5]. The Spanish dataset for that subtask is used throughout our experiments.

4.1 Experimental setting

The Spanish dataset used for SemEval 2007 Task 9 is a subset of the CESS-ECE corpus [10]. This subset contains 101,136 words in 3,611 sentences. The corpus is annotated with POS tags, lemmas, syntactic constituents, syntactic functions, named entities, verb argument structure, thematic roles, semantic classes of verbs and WordNet synsets for the 150 most frequent nouns. The training/test corpus size ratio is 9:1. The test corpus is furtherly split into two subsets, in-domain and out-of-domain test corpora. The in-domain test corpus contains documents covering the same domain as those of the training corpus, whereas the out-of-domain test corpus contains documents from a different domain in order to assess the systems adaptability.

Six types of named entities are annotated, namely **person**, **organization**, **location**, **number**, **date** and **other**. Table 1 shows the distribution of the different entity types throughout the corpus. In the table, *test.in* stands for in-domain test corpus whereas *test.out* stands for out-of-domain test corpus.

Table 1. Distribution of NE types in the Spanish corpus for SemEval 2007 Task 9.

NE type	training	test.in	test.out
person	1,953	116	72
organization	1,346	234	6
location	944	173	67
number	758	95	5
date	500	58	6
other	769	95	26

4.2 Results and discussion

Evaluation was carried out in terms of precision, recall and F_1 measures in the same way that it was done in SemEval 2007 Task 9. Table 2 shows the results obtained in our experiments.

The first section of the table contains the results for several variations of the tree traversal-based strategy. **PO_sent** represents the sentence-level postorder

Table 2. Experimental results over the Spanish dataset of SemEval 2007 Task 9 corpus.

	In-domain			Out-of-domain			Global		
	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
PO_struct_LS	74.48	70.04	72.19	68.79	53.30	60.06	73.56	66.84	70.04
PO_sent_LS	73.76	69.26	71.44	67.61	52.75	59.26	72.75	66.11	69.27
BU_glob_LS	70.01	68.74	69.37	63.76	52.20	57.40	68.98	65.58	67.24
BU_loc_LS	67.60	69.00	68.26	61.49	54.40	57.73	66.56	66.21	66.39
PO_sent_L	64.94	63.68	66.67	64.86	52.75	58.18	69.06	61.59	65.11
PO_struct_L	69.89	63.81	66.71	64.19	52.20	57.58	68.90	61.59	65.04
BU_loc_L	65.75	68.48	67.09	56.04	56.04	56.04	63.96	66.11	65.02
BU_glob_L	68.65	64.20	66.35	61.49	54.40	57.73	67.35	62.33	64.74
UPC	72.53	68.48	70.45	62.03	53.85	57.65	70.65	65.69	68.08
baseline	-	-	-	-	-	-	71.88	12.07	20.66

traversal-based strategy, whereas **PO_struct** represents the nested structure-level postorder traversal-based strategy.

BU_glob stands for a bottom-up strategy where two indepedence assumptions are considered. First, the class of a node is assumed to depend only on the set of classes of its children. All class combinations are considered. Since the average number of children of non-leaf nodes in the representation trees obtained from the test corpus is 1.34 and the maximum is six, running times for this method still remain reasonable in spite of its asymptotically exponential nature. To prevent data sparseness in considering a set of classes, the second assumption consists in considering the classes of all children of a node mutually independent. Under these considerations, a bottom-up dynamic programming method is used over each nested structure to approximate a globally optimal set of classes for all nodes.

On the other hand, **BU_loc** stands for a bottom-up strategy where no dynamic programming is used. In classifying a node, the locally most probable class according to the estimator is taken. The class given to a node is considered when classifying its parent but the classification process is iterative and irrevocable. This may be seen as an instance of the layering modeling technique [3] discussed in Section 2.

The suffix **L** means that only locally evaluated features (POS tags, trigger word dictionaries and gazetteers) were used to describe the candidate phrases. The suffix **LS** means that both locally evaluated and syntax-dependent features (syntactic functions and verb lemma dictionaries) were used.

The second section of the table contains results reported for the NER subtask at SemEval 2007 Task 9. UPC stands for the only system presented to the competition [6]. Additionally, the baseline results for the competition are presented. The baseline consisted on collecting a gazetteer of NE's from the training corpus and recognizing those segments in the test corpus that had the longest match with some of the NE's in this gazetteer.

Columns 2 to 4 show precision, recall and F_1, in that order, for the in-domain test corpus. Similarly, columns 5 to 7 show the results for the out-of-domain test

corpus and columns 8 to 10 for the entire test corpus. In each section, results are sorted by global F_1. Only results on the entire corpus are available for the baseline.

Several observations can be made after analyzing these results. First, we can corroborate the possitive impact of using syntax-dependent features to describe candidate phrases for the classification process. Global F_1 values using both locally evaluated and syntax-dependent features are always greater (between 1.37% and 5%) than those where only locally evaluated features were used.

A second important observation is that the postorder traversal-based classification strategy outperformed the bottom-up strategy in all cases sharing the same feature set. We consider that the main reason for this phenomenon is that the latter includes more independence assumptions than the former. Notice that when syntax-dependent features are used, the differences between global F_1 values are greater (between 2.03% and 3.65%) than those for the case when only locally evaluated features are used (between 0.02% and 0.31%). This suggests that syntax-dependent features enhance the postorder traversal-based classification strategy.

Regarding the postorder traversal-based strategy, we can also observe that when syntax-dependent features are used, the nested structure-level variant works better than the sentence-level variant whereas the opposite situation occurs when only locally evaluated features are used. Although we consider that these results cannot be taken to be conclusive in that aspect as yet because the differences are not very considerable, there are some ideas that might explain the situation. When only locally evaluated features are used, constructing a single sequence for the whole tree may help the classification process by providing more contextual information. On the other hand, syntax-dependent features contain contextual information per se and, in this case, it is possible that some amount of noise is added if we also consider dependences between phrases that may not be directly related by the clause-governing verb.

As can be largely expected, results over the in-domain corpus are better than those over the out-of-domain corpus in all cases.

Finally, when syntax-dependent features were used, the two variants of the postorder traversal-based classification strategy outperformed the results reported for the UPC system that participated in the SemEval competition. Regarding the competition baseline, the conservative nature of the method results in very low recall, which causes the F_1 value to be very low. According to F_1 values, this baseline is largely outperformed by the proposed method as well as by the UPC system. However, it should be noticed that UPC obtains a precision value below that of the baseline, whereas our syntax-dependent postorder traversal-based method outperforms the baseline in both precision and recall.

5 Conclusions

In this work, Spanish nested NER has been addressed. Unlike most approaches to nested NER, our method treats the classification of all phrases in a nested

structure as a single problem in order to obtain a near-to-globally optimal solution. We propose a tree representation for the set of candidate phrases in which syntactic information, both structural and functional, is encoded. In our opinion, the main contribution of this work is the proposal of a global classification strategy based on the postorder traversal of this representation tree.

Experimental results on the Spanish dataset for the SemEval 2007 Task 9 NER subtask show the validity of our postorder traversal-based classification strategy and the usefulness of syntactic information in describing phrases for the classification process.

In order to improve these results, an attractive direction for future work is the use of semantic information, such as verb senses, verb semantic classes, noun senses for trigger words, semantic roles, etc., in order to turn syntax-dependent features into semantic features. For example, if verb sense information is combined with verb lemma dictionaries in an appropriate way, results are likely to improve. Similarly, it may be helpful to use semantic roles instead of (or in combination with) syntactic functions.

References

1. Ohta, T., Tateisi, Y., Kim, L., Mima, H., Tsujii, J.: The GENIA corpus: An annotated research abstract corpus in molecular biology domain. In: 2nd International Conference on Human Language Technology Research, pp. 82–86. San Diego, CA, USA (2002).
2. Zhang, J., Shen, D., Zhou, G., Su, J., Tan, C.: Enhancing HMM-based biomedical named entity recognition by studying special phenomena. Journal of Biomedical Informatics 37, 411–422 (2004).
3. Alex, B., Haddow, B., Grover, C.: Recognising Nested Named Entities in Biomedical Text. In: BioNLP Workshop on Biological, translational, and clinical language processing, pp. 65–72. Prague, Czech Republic (2007).
4. Arévalo, M., Civit, M., Martí, M.A.: MICE: A module for Named Entities Recognition and Classification. International Journal of Corpus Linguistics 9:1, 53–68 (2004).
5. Màrquez, L., Martí, M.A., Taulé, M.: SemEval-2007 Task 09: Multilevel Semantic Annotation of Catalan and Spanish. In: 4th International Workshop on Semantic Evaluations, pp. 42–47. Prague, Czech Republic (2007).
6. Màrquez, L., Padró, L., Surdeanu, M., Villarejo, L.: UPC: Experiments with Joint Learning within SemEval Task 9. In: 4th International Workshop on Semantic Evaluations, pp. 426–429. Prague, Czech Republic (2007).
7. Punyakanok, Vasin and Dan Roth. The use of Classifiers in Sequential Inference. Advances in Neural Information Processing Systems 13, 995–1001 (2001).
8. McCallum, A., Freitag D., Pereira, F.: Maximum Entropy Markov Models for Information Extraction and Segmentation. In: 17th International Conference on Machine Learning, pp. 591–598. Stanford, CA, USA (2000).
9. Darroch, J., Ratcliff, D.: Generalized iterative scaling for log-linear models. The Annals of Mathematical Statistics 43, 1470–1480 (1972).
10. Martí, M.A., Taulé M., Màrquez L., Bertran, M.: CESS-ECE: A Multilingual and Multilevel Annotated Corpus. http://www.lsi.upc.edu/∼mbertran/cess-ece/publications (2007).