Squibs and Discussions

Unsupervised Named Entity Recognition Using Syntactic and Semantic Contextual Evidence

Alessandro Cucchiarelli*
Università di Ancona

Paola Velardi†
Università di Roma 'La Sapienza'

Proper nouns form an open class, making the incompleteness of manually or automatically learned classification rules an obvious problem. The purpose of this paper is twofold: first, to suggest the use of a complementary "backup" method to increase the robustness of any hand-crafted or machine-learning-based NE tagger; and second, to explore the effectiveness of using more fine-grained evidence—namely, syntactic and semantic contextual knowledge—in classifying NEs.

1. Proper Noun Classification

In this paper we present a corpus-driven statistical technique that uses a learning corpus to acquire contextual classification cues, and then uses the results of this phase to classify unrecognized proper nouns (PN) in an unlabeled corpus. Training examples of proper nouns are obtained using any available named entity (NE) recognizer (in our experiments we used a rule-based recognizer and a machine-learning-based recognizer). The contextual model of PN categories is learned without supervision.

The approach described in this paper is complementary to current methods for NE recognition: our objective is to improve, without additional manual effort, the robustness of any available NE system through the use of more "fine-grained" contextual knowledge, best exploited at a relatively late stage of analysis. The method is particularly useful when an available NE system must be rapidly adapted to another language or to another domain, provided the shift is not dramatic.

Furthermore, our study provides experimental evidence relating to two issues still under debate: i) the effectiveness, in practical NLP applications, of using syntactic relations (most systems use plain collocations and morphological features), and ii) context expansion based on thesauri. While we do not provide a definitive argument in favor of syntactic contexts and semantic expansion for word sense disambiguation tasks in general, we do show that they can be successfully used for unknown proper noun classification. Proper nouns have particular characteristics, such as low or zero ambiguity, which makes it easier to characterize their contexts.

2. Description of the U_PN Classification Method

In this section we briefly summarize the corpus-based tagging technique for the classification of unknown proper nouns (for more details, see Cucchiarelli, Luzi, and Velardi [1998]).

* Istituto di Informatica, Via Brecce Bianche I-60131 Ancona, Italy. E-mail: alex@inform.unian.it
† Dipartimento di Scienze dell'Informazione, Via Salaria 113, I-00198 Roma, Italy. E-mail: velardi@dsi.uniroma1.it

2.1 Learning Contextual Sense Indicators

Our method proceeds as follows: first, by means of any available NE recognition technique (which we will call an **early NE classifier**), at least some examples of PNs in each category are detected. Second, through an unsupervised corpus-based technique, typical PN syntactic and semantic contexts are learned. Syntactic and semantic cues can then be used to extend the coverage of the early NE classifier, increasing its robustness to the limitations of the gazetteers (PN dictionaries) and domain shifts.

In phase one, a learning corpus in the application domain is morphologically processed. The gazetteer lookup and the early NE classifier are then used to detect PNs. At the end of this phase, "some" PNs are recognized and classified, depending upon the size of the gazetteer and the actual performance (in the domain) of the NE classifier.

In phase two, the objective is to learn a contextual model of each PN category, augmented with syntactic and semantic features. Since the algorithm is unsupervised, statistical techniques are applied to smooth the weight of acquired examples as a function of semantic and syntactic ambiguity.[1]

Syntactic processing is applied over the corpus. A shallow parser (see details in Basili, Pazienza, and Velardi [1994]) extracts from the learning corpus **elementary syntactic relations** such as Subject-Object, Noun-Preposition-Noun, etc.[2] An **elementary syntactic link** (esl) is represented as:

$$esl(w_i, mod(type_i, w_k))$$

where w_j is the headword, w_k is the modifier, and $type_i$ is the type of syntactic relation (e.g. Prepositional Phrase, Subject-Verb, Verb-Direct-Object, etc.). For example, esl(*close* mod(G_N_V_Act *Xerox*)) reads: *Xerox* is the modifier of the head *close* in a Subject-Verb (G_N_V_Act) syntactic relation.

In our study, the **context** of a word w in a sentence S is represented by the esls including w as one of its arguments (w_j or w_k). The esls that include semantically classified PNs as one of their arguments are grouped in a database, called **PN_esl**. This database provides contextual evidence for assigning a category to unknown PNs.

2.2 Tagging Unknown PNs

A corpus-driven algorithm is used to classify unknown proper nouns recognized as such, but not semantically classified by the early NE recognizer.[3]

- Let **U_PN** be an unknown proper noun, i.e., a single word or a complex nominal. Let $C_{pn} = (C_{pn1}, C_{pn2}, \ldots, C_{pnN})$ be the set of semantic categories for proper nouns (e.g. Person, Organization, Product, etc.). Finally, let **ESL** be the set of esls (often more than one in a text) that include **U_PN** as one of their arguments.
- For each esl_i in **ESL** let:

$$esl_i(w_j, mod(type_i, w_k)) = esl_i(x, U\_PN)$$

1 We say the algorithm is unsupervised because neither the NE items detected by the early recognizer nor the extracted syntactic contexts are inspected for correctness.

2 Shallow, or partial parsers are a well-established technique for corpus parsing. Several partial parsers are readily available—for example, the freely downloadable LINK parser.

3 A standard POS tagger augmented with simple heuristics is used to detect possible instances of PNs. Errors are originated only by ambiguous sentence beginners, as "Owens Illinois" or "Boots Plc" causing partial recognition.

where $x = w_j$ or $x = w_k$ and U_PN=w_k or w_j (the unknown PN can be either the head or the modifier), $type_i$ is the syntactic type of esl (e.g. N-of-N, N_N, V-for-N, etc.), and furthermore let:

$$pl(esl_i(x, U\_PN))$$

be the **plausibility** of a detected esl. Plausibility is a measure of the statistical evidence of a detected syntactic relation (Basili, Marziali, and Pazienza 1994; Grishman and Sterling 1994) that depends upon local (i.e., sentence-level) syntactic ambiguity and global corpus evidence. The plausibility accounts for the **uncertainty** arising from syntactic ambiguity.

- Finally, let:
  - $\mathbf{ESL_A}$ be a set of esls in PN_esl (the previously learned contextual model) defined as follows: for each $esl_i(x, U\_PN)$ in ESL, put in ESL_A the set of $esl_j(x, PN_j)$ with $type_j = type_i$, x in the same position as esl_i, and PN_j a known proper noun, in the same position as U_PN in esl_i.
  - $\mathbf{ESL_B}$ be a set of esls in PN_esl defined as follows: for each $esl_i(x, U\_PN)$ in ESL put in ESL_B the set of $esl_j(w, PN_j)$ with $type_j = type_i$, w in the same position as x in esl_i, $Sim(w, x) > \delta$, and PN_j a known proper noun, in the same position as U_PN in esl_i. $Sim(w, x)$ is a similarity measure between x and w. In our experiments, $Sim(w, x) > \varepsilon$ iff w and x have a common hyperonym H in WordNet. The generality of H (i.e., the number of levels from x to H) is made parametric, to analyze the effect of generalization.
- For each semantic category C_{pnj} compute $evidence(C_{pnj})$ as:

$$evidence(C_{pnj}) = \alpha \frac{\sum_{esl_i \in ESL_A, C(PN_j)=C_{pnj}} weight_{ij}(x) D(x, C(PN_j))}{\sum_{esl_i \in ESL_A} weight_{ij}(x) D(x, C(PN_j))} + \beta \frac{\sum_{esl_i \in ESL_B, C(PN_j)=C_{pnj}} weight_{ij}(x) D(x, C(PN_j))}{\sum_{esl_i \in ESL_B} weight_{ij}(x) D(x, C(PN_j))}$$

where:

- $weight_{ij}(x) = weight_{ij}(esl_i(x, PN_j)) = pl(esl_i(x, PN_j)) \cdot (1 - \frac{amb(x)-1}{2k-1})$
- $weight_{ij}(w) = weight_{ij}(esl_i(w, PN_j)) = pl(esl_i(w, PN_j)) \cdot (1 - \frac{amb(w)-1}{k-1})$
- $pl(esl_i(x, PN_j))$ is the plausibility and $amb(x)$ is the ambiguity of x in esl_i
- k is a constant factor used to incrementally reduce the influence of ambiguous words. The smoothing is tuned to be higher in ESL_B
- α and β are parametric, and can be used to study the evidence provided by ESL_A and ESL_B

— $D(x, C(PN_j))$ is a discrimination factor used to determine the saliency (Yarowsky 1992) of a context $esl_i(x, \_)$ for a category $C(PN_j)$, i.e., how good a context is at discriminating between $C(PN_j)$and the other categories.[4]

The selected category for U_PN is

$$C = argmax(evidence(C_{pnk}))$$

When grouping all the evidence of a U_PN in a text, the underlying hypothesis is that, in a given linguistic domain (finance, medicine, etc.), a PN has a *unique* sense. This is a reasonable restriction for Proper Nouns, supported by empirical evidence, though we would be more skeptical about the applicability of the one-sense-per-discourse paradigm (Gale, Church, and Yarowsky 1992) to generic words. We believe that it is precisely this restriction that makes the use of syntactic and semantic contexts effective for PNs.

Notice that the formula of the evidence has several smoothing factors that work together to reduce the influence of unreliable or uninformative contexts. The formula also has parameters (k, α, β), estimated by running systematic experiments. Standard statistical techniques have been used to balance experimental conditions and the sources of variance.

3. Using WordNet for Context Generalization

One of the stated objectives of this paper is to investigate the effect of context generalization (the addend ESL_B in the formula of the evidence) on our sense tagging task.

The use of on-line thesauri for context generalization has already been investigated with limited success (Hearst and Schuetze 1993; Brill and Resnik 1994; Resnik 1997; Agirre and Rigau 1996). Though the idea of using thesauri for context expansion is quite common, there are no clear indications that this is actually useful in terms of performance. However, studying the effect of context expansion for a PN tagging task in particular is relevant because:

- PNs may be hypothesized to have a unique sense in a text, and even in a domain corpus. Therefore, we can reliably consider as potential sense indicators *all* the contexts in which a PN appears. The only source of ambiguity is then the word w_i co-occurring in a syntactic context with a PN, $esl_i(w_i, U\_PN)$, but since in ESL_B we group several contexts, hopefully spurious hyperonyms of w_i will gain lower evidence. For example, consider the context *"division of American_Brands_Inc"*. *Division* is a highly ambiguous word, but, when generalizing it, the majority of its senses appearing in the same type of syntactic relation with a Proper Noun (e.g. *branch of Drexel_ Burnham_Lambert_Group_Inc, part of Nationale_ Nederlanden_Group*) are indeed pertinent senses.

4 For example, a Subject_Verb phrase with the verb *make* (e.g., *Ace made a contract*) is found with almost equal probability with Person and Organization names. We used a simple conditional probability model for $D(x, C(PN_j))$, but we believe that more refined measures could improve performance.

- PN categories (e.g., Person, Location, Product) exhibit a more stable and less ambiguous contextual behavior than other more vague categories, such as *psychological_feature*.[5]
- We can study the degree of generalization at which an optimum performance is achieved.

4. Experimental Discussion

The purpose of experimental evaluation is twofold:

- To test the improvement in robustness of a state-of-the-art NE recognizer.
- To study the effectiveness of syntactic contexts and of a "cautious" context generalization on the performance of the U_PN tagger, analyzed in isolation. The effect of generalization is studied by gradually relaxing the notion of similarity in the formula of evidence and by tuning, through the factors α and β, the contribution of generalized contexts to the formula of evidence.

In our experiment, we used the Italian Sole24Ore half-million-word corpus on financial news, the one-million-word *Wall Street Journal* corpus, and WordNet, as standard on-line available resources, as well as a series of computational tools made available for our research:

- the VIE system (Humphreys et al. 1996) for initial detection of Proper Nouns from the learning corpus; for the same purpose we also used a machine learning method based on decision lists, described in Paliouras, Karkaletsis, and Spyropolous (1998).
- the SSA shallow syntactic analyzer (Basili, Pazienza, and Velardi 1994) for surface corpus parsing.[6]
- the tool described in Cucchiarelli and Velardi (1998) for corpus-driven WordNet pruning.[7]

4.1 Experiment 1: Improving Robustness of NE Recognizers

The objective of Experiment 1 is to verify the improvement in robustness of existing NE recognizers, through the use of our tagger. In Figure 1, three testing experiments are shown. The table measures the local performance of the NE tagging task achieved by the early NE recognizer, by our untrained tagger, and finally, the joint performance of the two methods.

In the first test, we used the Italian Sole24Ore corpus. Due to the unavailability of WordNet in Italian, we used a dictionary of strict synonyms for context expansion. In this test, we "loosely" adapted the English VIE system (as used in MUC-6) to Italian.

5 In Velardi and Cucchiarelli (2000) we formally studied the relation between category type and learnability of contextual cues for WSD.

6 We also used the GATE partial parser. We were not as successful with this parser because it is not designed for high-performance VP_PP and NP_PP detection, but prepositional contexts are often the most informative indicators.

7 This method produces a 20–30% reduction of the initial WordNet ambiguity, depending on the specific corpus.

	A	B	C	D	E	F	G	H	I	J	K	L
Test 1	239	3 55	67.32%	339	70.50%	60	83	72.29%	75	80.00%	84.23%	88.20%
Test 2	650	7 93	81.90%	759	85.63%	67	83	80.72%	80	83.75%	90.42%	94.47%
Test 3	3,040	4,168	72.94%	3,233	94.03%	585	935	62.57%	810	72.22%	86.97%	89.66%

Legend

A: PNs correctly tagged by the early NE recognizer

B: Total PNs in the Test Corpus

C: Local Recall of the early NE recognizer (A/B)

D: Total PNs detected by the early NE recognizer (D = A + A1 (errors) + G(unknown)

E: Local Precision of the early NE recognizer (A/D)

F: UPNs correctly tagged by the UPN tagger in the Test Corpus

G: Total UPNs not detected by the early NE recognizer

H: Local recall of UPN tagger (Phase2) (F/G)

I: Total UPNs for which a decision was possible by the UPN tagger

J: Local precision of the UPN tagger

K: Joint Recall of the two methods (A + F)/ B

L: Joint Precision of the two methods (A+F)/D

Figure 1
Outline of results on the Sole24Ore corpus.

We used the English gazetteer as it was and we applied simple "language porting" to the NE grammar (e.g., replacing English words and prepositions with corresponding Italian words, and little more).[8] This justifies the low performance of the rule-based classifier. Note that our context-based tagger produces a considerable improvement in performance (around 18%), therefore the global performance (column K and L) turns out to be comparable with state-of-the-art systems, without a significant readaptation effort.

In the second test, we used again VIE, on the English *Wall Street Journal* corpus. We used a version of VIE that was designed to detect NE in a management succession domain (we are testing the effect of a domain shift here). Local performance was somewhat lower than in MUC-6. Again, we measured a 9% improvement using our tagger, and very high global performance.

The third test was the most demanding. Here, we used only half of the named entity gazetteer used in previous experiments. The purpose of this test was also to verify the effect on performance of a poorly populated gazetteer. In this test, rather than using LASIE, we used a machine learning method described in Paliouras, Karkaletsis and Spyropolous (1998). This method uses as a training set the available half of the gazetteer to learn a context-based decision list for NE classification.

As shown in Test 3, column B, the initial number of PNs in the test corpus is now considerably higher. The decision-list classifier is tuned to classify with high precision and lower recall. Therefore, only the "hardest" cases are submitted to our untrained classifier. In fact, local performance of our classifier is around 10% lower than for previous tests, but nevertheless, global performance (in terms of joint precision and recall) shows an improvement. Finally, we observe that the performance figures reported in Figure 1 say nothing about the various sources of errors. Errors and misses occur both during the off-line learning phase (as we said, NE instances and syntactic contexts

8 Most location and company names known worldwide (e.g., NewYork, IBM) are in fact mentioned in economic journals regardless of the language.

are not inspected for correctness, therefore the contextual knowledge base is error prone) and prior to the U_PN tagging phase: a compound PN may be incompletely recognized during POS tagging, causing the generation of an uninformative syntactic context (e.g., "Owens Illinois" at the beginning of a sentence is recognized as "owens Illinois", causing a spurious N_N(owen,Illinois) context to be generated).

Because all these "external" sources of noise are not filtered out, we may then reliably conclude that our tagger is effective at improving the robustness of proper noun classification, though clearly the amount of improvement depends upon the baseline performances of the early method used for PN classification.

Although the classification evidence provided by syntactic contexts is somewhat noise prone, it proves to be useful as a "backup," when other "simpler" contextual evidence does not allow a reliable decision.

4.2 Effectiveness of Syntactic and Semantic Cues for Semantic Classification

In a second experiment, we used the experimental set up of Test 2 (WSJ+VIE described above) to evaluate the effectiveness of context expansion on system performance. We applied a pruning method on WordNet (Cucchiarelli and Velardi 1998) to reduce initial ambiguity of contexts. This pruning method allowed an average of 27% reduction in the initial ambiguity of the total number of the 13,428 *common* nouns in the Wall Street Journal corpus. The objective of this experiment was to allow a more detailed evaluation of our method, with respect to several parameters.

We built four test sets with the same distribution of PN categories and frequency distribution as in the application corpus. We selected four frequency ranges (1, 2, 3–9, $\geq$ 10) and in each range we selected 100 PNs, reflecting the frequency distribution in the corpus of the three main PN semantic categories—Person, Organization, and Location. We then built another test set, called TSAll, with 400 PNs again reflecting the frequency and category distribution of the corpus. The 400 PNs were then removed from the set of 37,018 esls extracted by our parser and from the gazetteer (whenever included).

In this experiment, we wanted to measure the performance of the U_PN tagger over the 400 words in the test set, in terms of F-measure, according to several varying factors:

- the category type;
- the amount of initial contextual evidence (i.e., the frequency range, reflected by the different test sets);
- the factors α and β, i.e., the influence of local and generalized contexts;
- the level of generalization L.

Figures 2 summarizes the results of the experiment. Figure 2(a) shows the increase in performance as a function of the values of α and β and the generalization level. N means no generalization, only the evidence provided by ESL_A is computed; 0 means that ESL_B collects the evidence provided by contexts in which w is a **strict synonym** of x according to WordNet; 1, 2, and 3 refer to incremental levels of generalization in the (pruned) WordNet hierarchy. The figure shows that context generalization produces up to 7% improvement in performance. Best results are obtained with L = 2 and $\alpha = 0.7$, $\beta = 0.3$. Further generalization may cause a drop in performance. High ambiguity is the cause of this behavior, despite WordNet pruning (without WordNet pruning, we observed a performance inversion at level 1; this experiment is not reported due to

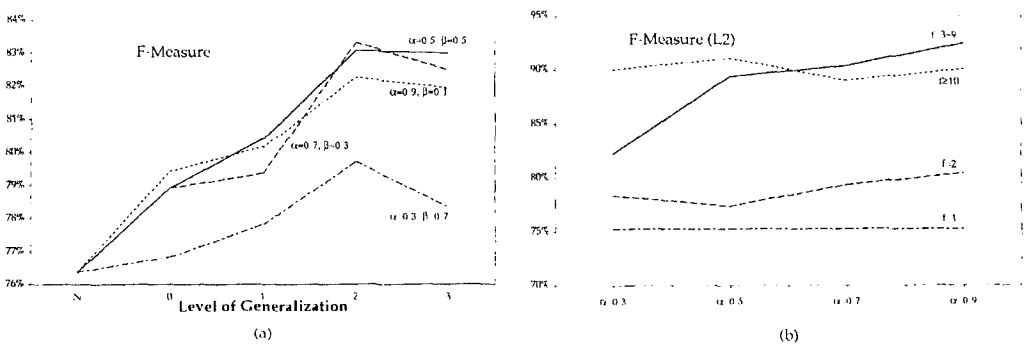

Figure 2
Evaluation of the effectiveness of context expansion.

limitations of space). Figure 2(b) illustrates the influence of initial contextual evidence. Recognition of singleton PNs remains almost constant as the contribution of generalized and nongeneralized contexts varies. Looking more in detail, we observe that recall increases with $\beta = (1-\alpha)$, but precision decreases. Generalization on the basis of a unique context does not allow any filtering of spurious senses, while when grouping several contexts, spurious senses gain lower evidence (as anticipated in Section 3).

Finally, we designed an experiment to evaluate the influence of the test set composition on the U_PN tagger performances. We performed an analysis of variance (ANOVA test [Hoel 1971]) on the results obtained by processing nine different test sets of 400 PNs each, selected randomly. In all our experiments the details of which we omit, for lack of space), we found that the U-PN tagging method performances were independent of the variations of the test set.

References

Agirre, Eneko and German Rigau. 1996. Word Sense Disambiguation using Conceptual Density. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING '96)*, Copenhagen, Denmark.

Basili, Roberto, Alessandro Marziali, and Maria Teresa Pazienza. 1994. Modelling syntax uncertainty in lexical acquisition from texts. *Journal of Quantitative Linguistics*, 1(1).

Basili, Roberto, Maria Teresa Pazienza, and Paola Velardi. 1994. A (not-so) shallow parser for collocational analysis. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING '94)*, Kyoto, Japan.

Brill, Erik and Philip Resnik. 1994. A transformation-based approach to prepositional phrase attachment disambiguation. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING '94)*, Kyoto, Japan.

Cucchiarelli, Alessandro, Danilo Luzi, and Paola Velardi. 1998. Automatic semantic tagging of unknown proper names. In *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montreal, Canada.

Cucchiarelli, Alessandro and Paola Velardi. 1998. Finding a domain-appropriate sense inventory for semantically tagging a corpus. *International Journal on Natural Language Engineering*, December.

Gale, William, Kenneth Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the DARPA Speech and Natural Language Workshop*. Harriman, NY.

Grishman, Ralph and John Sterling. 1994. Generalizing automatically generated selectional patterns. *Proceedings of the 15th International Conference on Computational Linguistics (COLING '94)*, Kyoto, Japan.

Hearst, Marti and Hinrich Schuetze. 1993. Customizing a lexicon to better suite a computational task. In *Proceedings of ACL–SIGLEX Workshop on Lexical Acquisition from Text*. Columbus, OH.

Hoel, Paul Gerhard. 1971. *Introduction to*

Mathematical Statistics. John Wiley & Sons Inc., New York.

Humphreys, Kevin, Robert Gaizauskas, Hamish Cunningam, and Sheila Azzan. 1996. *Technical Specifications, 1996/10/1815*. ILASH, University of Sheffield, UK.

Paliouras, George, Vangelis Karkaletsis, and Constantine Spyropolous. 1998. Results from the named entity recognition task. In Deliverable 3.2.1 of the European project ECRAN LE 2110. Available at: http://www2.echo.lu/langeng/en/le1/ecran/ecran.html.

Resnik, Philip. 1997. Selectional reference and sense disambiguation. In *Proceedings of the ACL Workshop Tagging Text with Lexical Semantics: Why, What, and How?* Washington, DC.

Velardi, Paola and Alessandro Cucchiarelli. 2000. A theoretical analysis of contextual-based learning algorithms for word sense disambiguation. In *Proceedings of ECAI 2000*, Berlin, Germany. (To appear.)

Yarowsky, David. 1992 Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING '92)*, Nantes, France.