Language Independent NER using a Unified Model of Internal and Contextual Evidence

Silviu Cucerzan and David Yarowsky
Department of Computer Science and
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD 21218, USA
{silviu,yarowsky}@cs.jhu.edu

Abstract

This paper investigates the use of a language independent model for named entity recognition based on iterative learning in a co-training fashion, using word-internal and contextual information as independent evidence sources. Its bootstrapping process begins with only seed entities and seed contexts extracted from the provided annotated corpus. F-measure exceeds 77 in Spanish and 72 in Dutch.

1. Introduction

Our aim has been to build a maximally language-independent system for named-entity recognition using minimal supervision or knowledge of the source language. The core model utilized, extended and evaluated here is based on Cucerzan and Yarowsky (1999). It assumes that only an entity exemplar list is provided as a bootstrapping seed set.

For the particular task of CoNLL-2002, the seed entities are extracted from the provided annotated corpus. As a consequence, the seed examples may be ambiguous and the system must therefore handle seeds with probability distribution over entity classes rather than unambiguous seeds. Another consequence is that this approach of extracting only the entity seeds from the annotated text does not use the full potential of the training data, ignoring contextual information. For example, *Bosnia* appears labeled 9 times as LOC and 5 times as ORG and the only information that would be used is that the word *Bosnia* denotes a location 64% of the time, and an organization 36% of the time, but not in which contexts is labeled one way or the other. In order to correct this problem, an improved system also uses context seeds if available (for this particular task, they are extracted from the annotated corpus). Because the representations of entity candidates and contexts are identical, this modification imposes only minor changes in algorithm and code.

Because the core model has been presented in detail in Cucerzan and Yarowsky (1999), this paper focuses primarily on the modifications of the algorithm and its adaptation to the current task. The major modifications besides the seed handling include a different method of smoothing the distributions along the paths in the tries, a new 'soft' discourse segmentation method, and use of a different labeling methodology, as required by the current task i.e. no overlapping entities are allowed (for example, the correct labeling of *colegio San Juan Bosco de Mérida* is considered to be ORG(*colegio San Juan Bosco*) *de* LOC(*Mérida*) rather than ORG(*colegio* PER(*San Juan Bosco*) *de* LOC(*Mérida*))).

2. Entity-Internal Information

Two types of entity-internal evidence are used in a unified framework. The first consists of the prefixes and suffixes of candidate entities. For example, in Spanish, names ending in *-ez* (e.g. *Alvarez* and *Gutiérrez*) are often surnames; names ending in *-ia* are often locations (e.g. *Austria*, *Australia*, and *Italia*). Likewise, common beginnings and endings of multiword entities (e.g. *Asociación de la Prensa de Madrid* and *Asociación para el Desarrollo Rural Jerez-Sierra Suroeste*, which are both organizations) are good indicators for entity type.

3. Contextual Information

An entity's left and right context provides an essentially independent evidence source for model bootstrapping. This information is also important for entities that do not have a previously seen word structure, are of foreign origin, or polysemous. Rather than using word bigrams or trigrams, the system handles the context in the same way it handles the entities, allowing for variable-length contexts. The advantages of this unified approach are presented in the next paragraph.

4. A Unified Structure for both Internal and Contextual Information

Character-based tries provide an effective, efficient and flexible data structure for storing both contextual and morphological patterns and statistics.

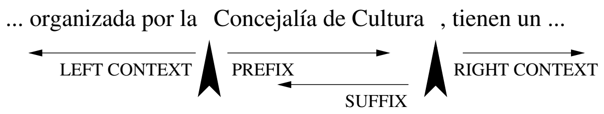

Figure 1: An example of entity candidate and context and the way the information is introduced in the four tries (arrows indicate the direction letters are considered)

They are very compact representations and support a natural hierarchical smoothing procedure for distributional class statistics. In our implementation, each terminal or branching node contains a probability distribution which encodes the conditional probability of entity classes given the sistring corresponding to the path from the root to that node. Each such distribution also has two standard classes, named "questionable" (unassigned probability mass in terms of entity classes, to be motivated below) and "non-entity" (common words).

Two tries (denoted PT and ST) are used for internal representation of the entity candidates in prefix, respectively suffix form, respectively. Other two tries are used for left (LCT) and right (RCT) context. Right contexts are introduced in RCT by considering their component letters from left to right, left contexts are introduced in LCT using the reversed order of letters, from right to left (Figure 1). In this way, the system handles variable length contexts and it attempts to match in each instance the longest known context (as longer contexts are more reliable than short contexts, and also the longer context statistics incorporate the shorter context statistics through smoothing along the paths in the tries).

The tries are linked together into two bipartite structures, PT with LCT, and ST with RCT, by attaching to each node a list of links to the entity candidates or contexts with, respectively in which the sistring corresponding to that node has been seen in the text (Figure 2).

5. Unassigned Probability Mass

When faced with a highly skewed observed class distribution for which there is little confidence due to small sample size, a typical response is to back-off or smooth to the more general class distribution. Unfortunately, this representation makes problematic the distinction between a back-off conditional distribution and one based on a large sample (and hence estimated with confidence). We address this problem by explicitly representing the uncertainty as a class, called "questionable". Probability mass continues to be distributed among the primary entity classes proportional to the observed distribution in the data, but with a total sum that reflects the confidence in the distribution and is equal to $1 - P_{node}(quest)$.

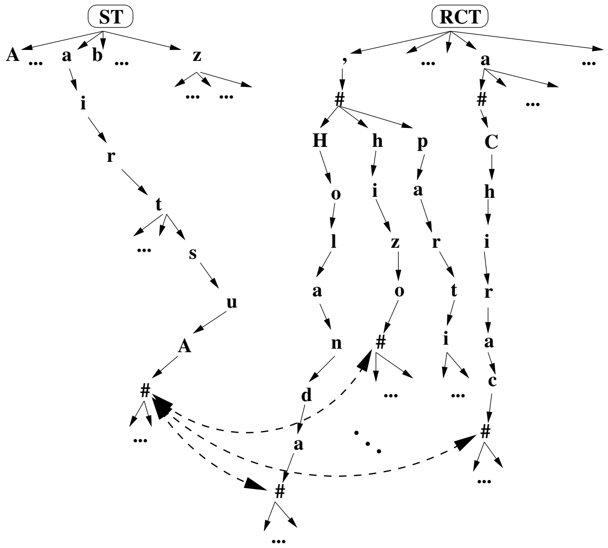

Figure 2: An example of links between the Suffix Trie and the Right Context Trie for the entity candidate *Austria* and some of its right contexts as observed in the corpus (< , *Holanda* >, < , *hizo* >, < *a Chirac* >)

Incremental learning essentially becomes the process of gradually shifting probability mass from questionable to one of the primary classes.

6. Smoothing

The probability of an entity candidate or context as being or indicating a certain type of entity is computed along the path from the root to the node in the trie structure described above. In this way, effective smoothing can be realized for rare entities or contexts. A smoothing formula taking advantage of the distributional representation of uncertainty is presented below.

For a sistring $l_1 l_2 ... l_n$ (i.e. the path in the trie is $root - l_1 - l_2 - ... - l_n$) the general smoothing model for the conditional class probabilities is given by the recursive formula:

$$\widehat{P}(c_j|l_1 l_2 ... l_i) = \frac{1}{Z}\Big(P_{node}(c_j|l_1 l_2 ... l_i) + \beta P_{node}(quest|l_1 l_2 ... l_i)^{\alpha} \widehat{P}(c_j|l_1 l_2 ... l_{i-1})\Big) \quad (1)$$

where Z is a normalization factor and $\beta \in [0,1]$, $\alpha \geq 1$ are model parameters.

7. One Sense per Discourse

Clearly, in many cases, the context for only one instance of an entity and the word-internal information is not enough to make a classification decision. But, as noted by Katz (1996), a newly introduced entity will be repeated, "if not for breaking the monotonous effect of pronoun use, then for

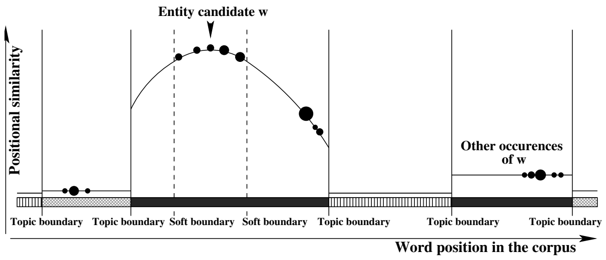

Figure 3: Using contextual clues from all instances of an entity candidate in the corpus. Each instance is depicted as a disc with the diameter representing the confidence of the classification of that instance using word-internal and local contextual information.

emphasis and clarity". We use this property in conjunction with the *one sense per discourse* tendency noted by Gale et al. (1992). The later paradigm is not directly usable when analyzing a large corpus in which there are no document boundaries, like the one provided for Spanish. Therefore, a segmentation process needs to be employed, so that all the instances of a name in a segment have a high probability of belonging to the same class. Our approach is to consider a 'soft' segmentation, which is word-dependent and does not compute topic/document boundaries but regions for which the contextual information for all instances of a word can be used jointly when making a decision. This is viewed as an alternative to the classical topic segmentation approach and can be used in conjunction with a language-independent segmentation system (Figure 3) like the one presented by Richmond et al. (1997).

After estimating the class probability distributions for all instances of entity candidates in the corpus, a re-estimation step is employed. The probability of an entity class c_j given an entity candidate w at position pos_i is re-computed using the formula:

$$\widehat{P'}(c_j|w,pos_i) = \frac{1}{Z}\sum_{k=1}^{n}\widehat{P}(c_j|w,pos_k)\cdot sim(pos_i,pos_k)\cdot conf(w,pos_k) \qquad (2)$$

where $pos_1,...,pos_n$ are the positions of all instances of w in the corpus, sim is the positional similarity, encoding the physical distance and topic (if topic or document boundary information exists), *conf* is the classification confidence of each instance (inverse proportional to the the $\widehat{P}(quest|w,pos_k)$, Z is a normalization factor.

8. Entity Identification / Multiple-Word Entities

There are two major alternatives for handling multiple-word entities. A first approach is to tokenize the text and classify each individual word as being or not part of an entity, process followed by an entity assemblance algorithm. A second alternative

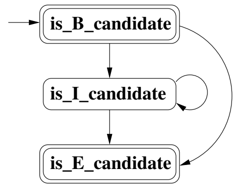

Figure 4: The structure of an entity candidate represented as an automaton with two final states

is to consider a chunking algorithm that identifies entity candidates and classify each of the chunks as Person, Location, Organization, Miscellaneous, or Non-entity. We use this second alternative, but in a 'soft' form; i.e. each word can be included in multiple competing chunks (entity candidates). This approach is suitable for all languages including Chinese, where no word separators are used (the entity candidates are determined by specifying starting and ending character positions). Another advantage of this method is that single and multiple-word entities can be handled in the same way.

The boundaries of entity candidates are determined by a few simple rules incorporated into three discriminators: *is_B_candidate* tests if a word can represent the beginning of an entity, *is_I_candidate* tests if a word can be the end of an entity, and *is_E_candidate* tests if a word can be an internal part of an entity. These discriminators use simple heuristics based on capitalization, position in sentence, length of the word, usage of the word in the set of seed entities, and co-occurrence with uncapitalized instances of the same word. A string is considered an entity candidate if it has the structure shown in Figure 4.

An extension of the system also makes use of Part-of-Speech (POS) tags. We used the provided POS annotation in Dutch (Daelemans et al., 1996) and a minimally supervised tagger (Yarowsky and Cucerzan, 2002) for Spanish to restrict the space of words accepted by the discriminators (e.g. *is_B_candidate* rejects prepositions, conjunctions, pronouns, adverbs, and those determiners that are the first word in the sentence).

9. Algorithm Structure

The core algorithm can be divided into eight stages, which are summarized in Figure 5. The bootstrapping stage (5) uses the initial or current entity assignments to estimate the class conditional distributions for both entities and contexts along their trie paths, and then re-estimates the distributions of the contexts/entity-candidates to which they are linked, recursively, until all accessible nodes are reached, as presented in Cucerzan and Yarowsky (1999).

1	Extract the entity (and context) seed sets from the annotated data
2	Read the text to be annotated and extract all entity-candidates
3	Extract the sets LC and RC of all contexts of entity candidates
4	Build the tries using all individual words and entity candidates, and all instances of the elements LC and RC from the text
5	Apply the bootstrapping procedure using the seed data
6	Classify each entity-candidate in isolation
7	Re-classify each entity-candidate by using formula (2)
8	Resolve conflicts between competing entity candidates

Figure 5: Algorithm structure

10. Results

We compare the results of two variants of the described model on the development and test sets provided (Table 1). The first one uses only exemplar entity and context seeds extracted from the training corpus. The second also employs POS information to rule out unlikely entity candidates.

The system was built and tested initially utilizing only the provided Spanish data. The parameters were estimated using an 80/20 split of the training data (*esp.train* and *ned.train*). The dev-test data (*testa*) were not used during the parameter estimation phase. The programs were run once on the final test data (files *testb*). We allocated only one person-day to adapt the system for Dutch and tune the parameters to this language in order to show functional language independence. We opted not to make a detailed study of parameter variation on test data to avoid any potential for tuning to this resource and preserve its value for future system development.

The following table further details the types of errors made by the algorithm (full system on Spanish dev-set). E_1 represents the number of over-generated and under-generated entities in the precision and recall rows (respectively). E_2 represents the number of entities with correctly identified boundaries, but wrong classifications.

$E_1 + E_2$	PER	LOC	ORG	MISC
Precision	43 + 153	73+118	87+341	76+170
Recall	43+123	34+310	112+279	22+70

Because our system takes seed lists rather than annotated text as input, additional entity lists can be used by the system. By employing such lists of countries, major cities, frequent person names and major companies (extracted from the web), significant improvements can be obtained (preliminary tests show as much as 2.5 F-measure improvement on a 80/20 split of the training data in Dutch).

11. Conclusion

This paper has presented and evaluated an extended bootstrapping model based on Cucerzan and Yarowsky (1999) that uses a unified framework of both entity internal and contextual evidence. Starting only with entity and context seeds extracted from training data and the addition of part-of-speech information, system performance exceeds 77 and 72 F-measure for Spanish and Dutch respectively.

Spanish Dev-set	without POS information			with POS information		
	Precision	Recall	F-meas.	Precision	Recall	F-meas.
LOC	69.11	80.41	74.33	69.77	80.61	74.80
MISC	66.89	44.49	53.44	68.38	44.72	54.08
ORG	73.26	71.71	72.47	76.49	74.82	75.65
PER	85.39	83.22	84.29	86.07	83.96	85.00
Overall	75.08	74.13	74.60	78.82	75.62	**76.22**

Spanish Test	without POS information			with POS information		
	Precision	Recall	F-meas.	Precision	Recall	F-meas.
LOC	78.62	73.62	76.04	79.66	73.34	76.37
MISC	63.73	38.24	47.79	64.22	38.53	48.16
ORG	74.86	78.50	76.64	76.79	81.07	78.87
PER	80.63	87.21	83.79	82.57	88.30	85.34
Overall	76.62	74.96	75.78	78.19	76.14	**77.15**

Dutch Dev-set	without POS information			with POS information		
	Precision	Recall	F-meas.	Precision	Recall	F-meas.
LOC	73.30	70.38	71.81	76.87	73.32	75.05
MISC	64.08	57.64	60.69	68.16	63.14	65.55
ORG	67.34	53.76	59.79	70.63	55.96	62.45
PER	63.17	79.94	70.57	64.99	80.51	71.92
Overall	66.10	65.01	65.55	69.14	67.84	**68.49**

Dutch Test	without POS information			with POS information		
	Precision	Recall	F-meas.	Precision	Recall	F-meas.
LOC	73.65	77.56	75.55	77.72	80.54	79.11
MISC	70.10	57.29	63.05	74.67	62.34	67.95
ORG	69.78	62.14	65.74	72.12	64.88	68.31
PER	67.62	79.26	72.98	69.39	80.71	74.62
Overall	69.95	68.49	69.21	73.03	71.62	**72.31**

Table 1: Results on the development sets (files *esp.testa* and *ned.testa*) and on the test sets (files *esp.testb* and *ned.testb*)

12. Acknowledgements

This work was supported by NSF grant IIS-9985033 and ONR/MURI contract N00014-01-1-0685.

References

S. Cucerzan and D. Yarowsky. 1999. Language independent named entity recognition combining morphological and contextual evidence. In *Proceedings of the Joint SIGDAT Conference on EMNLP and VLC 1999*, pages 90–99.

W. Daelemans, J. Zavrel, and S. Berck. 1996. Mbt: A memory-based part of speech tagger-generator. In *Proceedings of the 4th Workshop on Very Large Corpora*, pages 14–27.

W. Gale, K. Church, and D. Yarowsky. 1992. One sense per discourse. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, pages 233–237.

S. M. Katz. 1996. Distribution of context words and phrases in text and language modeling. *Natural Language Engineering*, 2(1):15–59.

K. Richmond, A. Smith, and E. Amitay. 1997. Detecting subject boundaries within text: a language independent statistical approach. In *Proceedings of EMNLP 1997*, pages 47–54.

D. Yarowsky and S. Cucerzan. 2002. Bootstrapping a multilingual part-of-speech tagger in 1 person-day. In *Proceedings of CoNLL 2002*.