Multi-Level NER for Portuguese in a CG framework

Eckhard Bick

Institute of Language and Communication,
Southern Denmark University
lineb@hum.au.dk, http://visl.sdu.dk

Abstract. This paper describes and evaluates a linguistically based NER system for Portuguese, based on lexico-semantical information, pattern matching and morphosyntactic, context driven Constraint Grammar rules. Preliminary F-scores for cross-domain news texts, when distinguishing six different name types, were 91.85 (raw) and 93.6 (subtyping of ready-chunked proper nouns).

1 Introduction

Named entity recognition (NER) in running text is a complex task with a number of obvious applications - semantic corpus annotation, summarisation, text indexing, to name but a few. This work focuses on Portuguese, and strives to distinguish between 6 main name type categories by linguistic rather than statistical means.

1.1 Previous work

In recent years, a number of different approaches have been carried out and evaluated by the NLP community. Thus, at the MUC-7 confeence (1998), competing systems were tested on broadcast news. The best performing system (LTG, Mikheev et. al. 1998) used both sgml-manipulating hand-crafted symbolic rules, a Hidden Markov Modeling (HMM) POS-tagger, name lists, partial probabilistic matching and semantic suffix classes, achieving an overall F-measure[1] of 93.39, with recall/precision rates of 95/97, 91/95 and 95/93 for person, organisation and location, respectively. HHM-results in isolation ("learned-statistical") can be regarded as a kind of baseline against which more sophisticated systems should be measured. A well performing example is the Nymbel system (Bikel et. al. 1997), which achieved F-scores of 93 and 90 for English and Spanish news text, respectively. Also, results for English were shown to be fairly stable down to a training corpus size of 100.000, indicating the cost/performance efficiency of the approach. Another automatic learning method, based on maximum entropy training (MENE), is described by Borthwick et. al. (1998). This system showed a clear increase in performance with growing training corpora, with in-domain F-scores of 89.17 for 100.000 tokens and 92.20 for 350.000 tokens. The authors also stress the potential of hybrid systems, with a maximum F-

[1] Defined as 2 x Recall x Precision / (Recall + Precision)

score of 97.12 when feeding information from other MUC-7-systems into MENE. One possible weakness of trained systems is indicated by the fact that in MUC's cross-domain formal test, F-scores dropped to 84.22 and 92 for pure and hybrid MENE, respectively. Another interesting base line is human annotators' F-score, which at MUC-7 was reported as 96.95 and 97.60 (Marsh & Perzanowski, 1998).

1.2 Methodological and data framework

In this paper, I shall present a mostly linguistic approach to NER, combining different lexical, pattern and rule based strategies in a multi-level Constraint Grammar framework (CG). This approach has previously been succesfully carried out for Danish (Bick, 2002) within the Scandinavien NER project *Nomen Nescio*. For Portuguese, the system builds on a pre-existing general Constraint Grammar parser (*PALAVRAS*, Bick 2000) with a high degree of robusticity and a comparatively low percentage of errors (less than 1% for word class). Tag types are word based and cover part of speech, inflexion and syntactic function, as well as dependency markers.

The language data used in this article are drawn from the *CETEM Público* news text corpus (Rocha & Santos, 2000), which has been grammatically annotated in a joint venture between the VISL project at Southern Denmark University and the Linguateca-AC/DC project at SINTEF, Norway (Santos & Bick, 2000).

1.3 Discussion of name categories

For this project, proper nouns are defined as upper case non-inflecting words or word chains with upper case in the first and last parts. Simplex names in lower case (e.g. pharmaceuticals) are treated as nouns, as are nouns with upper case initial in mid-sentence, though the latter may be marked as <prop> with a secondary tag for later filtering by corpus users. In agreement with general *Nomen Nescio* strategy, 6 core categories were used (human, place, organisation, event, title, and brand/object):

Human personal names <hum>. This category covers (chains of) personal names (Christian, middle and surnames), possibly interspaced with structural lower case particles *(de, da, von, ten, ...)* or prefixes *(McNamara).* Human name chains can be complements of title nouns, as in *Dr. Mota* and profession nouns (*o jornalista Nelson Aguiar*). Since the latter allow interfering modifiers (*o jornalista português Nelson Aguiar*), only the former were fused into the name chain. Lumped into the <hum> category are other **animate names <A>**, both scientific species names (*Mus musculus*), pet names, gods and mythical beasts.

Place names <top>. As a prototype, this category has a clear syntactic and semantic context. However, human settlements (countries, towns, villages etc.) may also function as +HUM subjects of cognitive verbs, blurring the distinctional line between <hum> and <top>. For semantic reasons, and to allow for CG rules using +HUM and -HUM syntactic contexts, I have therefore introduced the **civitas <civ>** category for these names: *Dinamarca, EUA, Coreia do Norte*.

Buildings which metaphorically can *offer*, *invite* or *earn*, receive a separate subcategory, **institution <inst>**, a kind of hybrid between <org> and <top>.

Organisations <org>. This is another +HUM category, covering companies, sports clubs, movements and the subcategory **<party>**, often as abbreviations *(ONU, FIFA)*. A special case is the **<media>** category (newspapers, radio channels and tv stations), which is systematically ambiguous (<org>/<tit>).

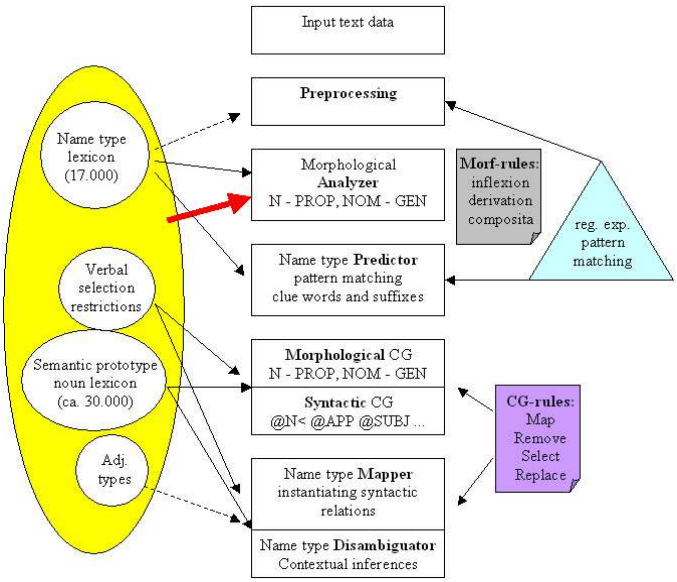
Events and occasions. This category is used for both natural and organized events, and includes the experimental weather subcategory **<wea>** (*El Nino*, hurricane names). There is a certain metaphoric transfer from sites to events *(desde Maastricht)*, and some event names contain ordinal or year markers *(Expo 98)*.

. **Semantic products, book, film and music titles <tit>.** Semantically and formally, a distinction can be made between literary "running text" titles on the one hand *("O Grito Silencioso")*, and "classifiers" on the other. This latter category is rarely quote marked, does not exceed np-structure, and can usually be recognized by a classifying key nominal element: *a Lei Áurea*. A related, though more generic - "nominal" - (sub)category is that of **<genre>** *(Anatomia, Islã, Judo, Funk)*. **<disease>** and **<ling>** (languages) are experimental subcategories of <tit>.

Object and brand names <brand>. This is the default object and waste bin category, containing besides brand names *(Coca Cola, Linux*), also the subclass of **vehicles <V>,** covering both brand and individual names *(Columbia, Santa Maria)* and names ambiguous with or derived from company names *(Peugeot, Konica E240)*.

2 System architecture and strategies

The system treats NER as a distributed task, matching the progressive level architecture of the parser itself, applying different techniques at different levels.



2.1 Preprocessing

Besides more "ordinary" preprocessing tasks like sentence separation, this first module creates '='-linked name chains based on pattern matching *(Edvard=Munch).* Upper case is the obivous trigger (with some sentence-initial ambiguity), but certain lower case particles *(da, di, von, ibn)* are allowed in the chain, as are numericals in certain patterns *(version numbers, car names).* A particular problem is the recognition of in-name punctuation *(initials, Sta., H.M.S., jr., & Co., d'Ávila, web-urls, e-mails).* Though the preprocessor does check its chunking against full name entries in the lexicon, proper nouns are a very productive class, and heuristic patterns may lead to overchunking *(diretor de marketing para a Europa da Logitech).* Here, a second lexicon-lookup checks for recognizable parts of a name chain candidate, and re-splits it.

Palmer & Day (1997) compared the coverage of inter-corpus name vocabulary transfer in 6 languages and found the second highest transfer rate for NEs in Portuguese (61.3%), after Chinese (73.2%) and way above English (21.2%), suggesting the importance of a lexicon module and gazeteer lists in Portuguese NER.

2.2 The name type predictor

Some frequent names receive a semantic type tag already from the lexicon based morphological analyzer module (otherwise handling lemmatizing, inflexion and derivation). However, most proper nouns have to be typed later. The name type predictor is a semi-heuristic module, which has its own lexicon (ca. 16.000 entries), enabling it to match *parts* of names, for instance recognizing person names by looking up Christian names first parts. Similarly, *Shell=Portuguesa* is typed as <org>, because *Shell* is recognized as a company. Hyphen-bound pairs of recognized place names will be typed as event/path candidates *(o corredor mediterráneo Valencia-Barcelona),* an ambiguity that will be tackled later by contextual CG rules. In other cases, the type predictor tries to instantiate morphological and pattern based type clues for the different categories, using both reg. expressions and lists of key-words:

<tit> e.g. quotes, in-name function words (articles, pronouns etc.), "semantic things" *("Voo das Borboletas", II Lei da Programação Militar, o Pacote Delors)*

<media> e.g. *Diário=, Revista=, =Times, Voz=, Channel= ...*

<occ> e.g. *Conferência=, Expo=, Guerra=, Rali=, =\'[0-9][0-9] ...*

<V> e.g. *Boeing/Mercedes/Toyota=, =Combi, =Sedan, HMS=, USS ...*

<brand> e.g. *Macintosh/Sanyo=[0-9],*quality markers:*=Extra, =de=Luxe*

<hum> e.g. suffixes:*-sen, -sson, -sky, -owa,* infixes: *ibn, van, ter, y, zu, di,* abbreviated and part-of-name titles: *Sra., Madame, Mlle., sr., Mc=, Al=, =Khan*

<civ> e.g. *República=, Puerto=,* suffixes: *-town, -ville, -polis* (a number of these will receive both <civ> and <hum> tags for later disambiguation

<top> e.g. *Cabo=, Praça=, Rua=, =Island, =de=Leste, =do=Sul*

<org> e.g. in-word capitals: *[a-z][A-Z] (MediaSoft),* "suffixes": *Cia., & Co ...,* type indicators: *=Holding, =Clube, FC=,* morphological indicators: *-com, -tech, -soft*

<inst> e.g. *Universidade=, Tribunal=, =Hilton, Aeroporto= ...*

Of course, there may be interferences and contradictions between the patterns, so matchings are ordered, conditioned and iterated, and they do allow some ambiguity. Finally, the type predictor uses non-alphabetic characters, paired upper case, function words etc. to assign <non-hum> tags, preventing over-usage of this most common category in the cg-based part of the system.

2.3 The CG modules

CG adds, selects or removes tags from words, i.e. performs word based annotation by mapping and resolving ambiguities. Rules use sentence-wide context and have access to word class, inflexion, verbal and nominal valency potential as well as - in the Portuguese system - semantic prototype information for nouns and some verbal selection patterns. The "ordinary" (preexisting), morphological and syntactic CG levels consist of about 7000 rules. Though only a small part of these tackles proper nouns, it is much safer to contextually disambiguate, say, sentence initial imperatives from heuristic proper nouns, than at the pattern matching stages. Of course, proper nouns can also for their part form valuable context for the disambiguation of other classes, and besides functioning as subjects and objects like other np's, they can fill certain more specific syntactic slots:

@N< (valency governed nominal dependents): *o presidente americano* ***George Bush***
@APP (identifying appositions): *uma moradora do palácio,* ***Júlia Duarte****, ...*
@N<PRED (predicating appositions)
*Os marchadores Susana Feitor (****CN Rio Maior)*** *e José Magalhães* ***(Alfenense)***

Syntactic relations like the above, once established, can later be used by the name CG proper to derive semantic name types from lexical semantic information residing in the corresponding noun heads. The name grammar consists of both a mapping CG and a disambiguation CG. The former is capable of adding to or even overriding semantic class types from the preceding levels, while the latter removes or selects semantic type tags where the lexion, heuristic type predictor or mapping grammar created ambiguities. The name grammar is still small (ca. 400 rules) and not optimised for Portuguese, since many rules were ported and adapted from corresponding rules for Danish.

Cross-nominal prototype transfer. One major technique used by the grammar is cascading nominal inference, exploiting preexisting safe semantic noun contexts and coordination, and creating more safe context in the process. The following simplified example rules map <top> (PLACE) onto names , if they are direct (i) or prepositional (ii) dependents of place-nouns, or if they are predicated as places (iii).

MAP (<top>) TARGET (PROP @N<) (-1 N-TOP) ;
MAP (<top>) TARGET (PROP @P<) (-1("de" PRP @N<) (-2 N-TOP) ;
*SELECT (<top>) (0 @SUBJ>) (*1 @<SC BARRIER @SUBJ LINK 0 N-TOP) ;*

More detailed rules can match such information between main and relative clauses.

Coordination based type inference. Drawing on and matching syntactic tags from the syntactic CG-module, the name type-mapper first establishes a secondary tag for "close/safe coordinators" (&KC-CLOSE), with one rule for each matched syntactic function, then uses it for disambiguatione:

REMOVE %non-h (0 %hum-all) (-1 &KC-CLOSE BARRIER @NON->N LINK *-1C %hum OR N-HUM BARRIER @NON-N<);*

*SELECT (<top>) (1 &KC-CLOSE) (*2C <top> BARRIER @NON->N) ;*

PP-contexts. Postnominal prepositions link their argument names to the semantic information residing in their noun heads, and to a certain degree prepositions may themselves provide semantic clues. Thus, the following rule derives event-hood from temporal prepositions, if they are not objects (@PIV).

MAP (%occ) TARGET (PROP @P<) (-1 PRP-TEMP BARRIER @NON-N< LINK NOT 0 @<PIV);*

The preposition "de" covers what in an equivalent germanic grammar would be **genitive mapping,** exploiting, for instance, that companies have directors *(o director da Logitech),* or that non-HUM names can't de-attach to thought products *(plano).*

Selection restrictions. A number of rules exploit verbal selection restrictions. Thus, the lexicon marks +HUM subject selections, and some other selection sets are defined in the grammar itself. The following rules work on negative set targets, not human (i) and not organisation (ii), discarding several tag candidates at a time:

*REMOVE %non-hum (0 @SUBJ> LINK 0 %hum-all) (*1 @MV BARRIER ser/estar/ficar LINK 0 V-HUM);* [@MV = main verb, @SUBJ = subject]

REMOVE %non-org (0 @<ACC LINK 0 %org/inst) (-1 @MV LINK 0 V-ADMIN);* [@<ACC = accusative/direct object]

4 Evaluation

Performance. Though the NER module of PALAVRAS is an unfinished project, a pilot evaluation study was performed on a 45.000 word chunk from the CETEM Público corpus, containing 2672 name chains (4764 tokens). In the light of the above mentioned difference between MUC F-scores for same domain and cross-domain, it has to be stressed that the Público sample was domain-mixed ("politics", "culture", "social issues", "economy", "opinion" and sports). The table below indicates the relative weight of two different types of errors, one preprocessor and PoS-based (2.2 % chunking and proper name errors as such), the other name typing error themselves (6 %), which together make for error rate of 8.2%. More than half of all name tokens had no lexicon entries, and as might be expected, recall was lower for these names. (89.1% as opposed to 94.9%). However, the 5% typing error rate even for lexicon[2] based names indicate problems with ambiguity resolution, lexicon errors and a certain price

[2] One reason is that the name lexicon was in large parts compiled automatically from a variety of text sources, and manual checking so far has been incomplete at best.

for the fact that contextual rules are allowed to override the lexicon. Interestingly, few errors occur between subcategories of the same major category.

Table 1

Público (jan. 2003) 45099 words	all PROP (5.9% of words) 4764 tokens		heuristic PROP (i.e. no lexicon entry) 52.7 %	
	cases	percent	cases	percent
correct found (6 classes) of these non-heur alone	2453 1198	91.8 % 94.9 %	1255	89.1%
wrong major class (6 classes)	168	6.3 %	118	8.4 %
wrong subclass (same major)	14	0.5 %	14	1.0 %
false positive PROP reading (incl. “overchunking”)	10+23=33	1.2 %	5+14=19	1.3 %
false negative (missing) PROP (incl. “underchunking”)	13+14=27	1.0 %	0+12=12	0.9 %
all evaluated proper nouns[3]	**2672**		2672	
of these: not in lexicon	1409		**1409**	

Recall and precision. Since types were almost completely disambiguated, and since false positive and false negative chunking errors were of similar frequency, recall and precision were similar, 91.8% and 91.9%, respectively, resulting in an F-score of 91.85. For subtyping alone (of correctly chunked and prerecognized proper nouns), an F-Score of 93.6 was measured. The table below shows distribution and performance for the 6 super-categories used.

Table 2

Name type	distribution	Recall	Precision	F-Score
Person <hum> <A> <B>	35.5 %	97.9 %	87.7 %	92.5
Organisation <org> <party> <media>	22.6 %	93.3 %	95.4 %	94.3
Place <top> <civ> <inst>	34.5 %	93.3 %	96.9 %	95.1
Event <occ> <wea>	2.2 %	82.1 %	96.5 %	88.7
Abstracta <tit> <genre> <ling> ...	5.4 %	84.3 %	84.3 %	84.3
Objects <brand> <V> <common> ...	1.1%	53.8 %	60.9 %	57.1

It is interesting that major categories outperformed minor ones, suggesting a systematic/heuristic rule bias towards the former. Especially <hum> usurps other NEs from other categories, giving it high recall, but low precision. Place names <top>, on the other hand, are strong on precision, possibly because they are easier to lexicalize than person names, and share a more specific syntactic context.

Moreover, the table shows that the person and place NE categories make up roughly a third of all names each, while organisation names account for a fifth. Interestingly, corresponding counts for Danish news corpora vary in two major aspects, yielding 50% person names and only 16% organisations (Bick 2003).

[3] Ignoring 17 cases of garbled corpus input with upper case.

5 Conclusion

This paper has presented a linguistic approach to NER, showing how lexical, pattern and rule based name typing tools can be integrated into a multi-level CG system for Portuguese. Though still immature, the name recognizer module has demonstrated promising results on mixed-domain news texts, with an overall F-Score of 91.85 with a 6-way category distinction. An F-Score of 93.6 for name type recognition of correctly chunked proper nouns, and a 2% chunking error rate suggest that improvements in lexicon enhanced preprocessing might improve overall performance. Future work will also focus on tuning CG name rules to Portuguese language data, and proofreading the name lexicon. Since recall and precision varied significantly across categories, rules should concentrate on precision for person names, and recall for the minor categories, in particular.

References

1. Bick, Eckhard: The Parsing System 'Palavras' - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press, Århus (2000)
2. Bick, Eckhard: "Named Entity Recognition for Danish". I: *Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004*. Forthcoming (2003).
3. Bikel, Daniel M. & Miller, Scott & Schwartz, Richard & Weischedel, Ralph: Nymble: a High-Performance Learning Name-finder. In: Proc. of the Conf. on Applied Natural Language Processing 1997
4. Borthwick, Andrew & Sterling, John & Agichtein, Eugene & Grishman, Ralph: NYU: Description of the MENE Named Entity System as Used in MUC-7. In: Proc. of the 7th Message Understanding Conf. (MUC7), April 29th - May 1st, Fairfax (1998)
5. Iason, Demiros et. al.: Named Entity Recognition in Greek Texts. In: Proceedings of the 2nd Int. Conference on Language Resources & Evaluation (LREC), 2000
6. Marsh, E. & Perzanowski, D.: MUC-7 evaluation of I.E. Technology: Overview of Results. In: Proc. of the 7th Message Understanding Conf. (MUC7), April 29th - May 1st, Fairfax (1998)
7. Mikheev, Andrei & Grover, Claire & Moens, Marc: Description of the LTG System used for MUC-7. In: Proceedings of the 7th Message Understanding Conference (MUC7), April 29th - May 1st, Fairfax (1998)
8. Palmer, David D. & Day, David S.: A Statistical Profile of the Named Entity Task. In: Proceedings of the Fifth Conference on Applied Natural Language Processing March 31st - April 3rd 1997
9. Rocha, Paulo A. & Santos, Diana: CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In: Maria das Graças Volpe Nunes (ed.): Actas do V. PROPOR, Nov. 19th-22nd, Atibaia (2000), pp. 131-140
10. Santos, Diana & Bick, Eckhard: Providing Internet access to Portuguese corpora: the AC/DC project. In Gavriladou et al. (eds.): *Proc. 2nd International Conf. on Language Resources and Evaluation, LREC2000* (Athens, 2000), pp. 205-10.
11. Stevenson, Mark & Gaizauskas, Robert: Using Corpus-derived Name Lists for Named Entity Recognition. In: Proc. of the Sixth Conf. on Applied Natural Language Processing, Seattle, 2000