SIEMÊS - a named-entity recognizer for Portuguese relying on Similarity Rules

Luís Sarmento
FEUP & Linguateca

Outline

- Introduction to NER
- Motivation for using "Similarity Rules"
- Inside SIEMÊS:
  - Architecture: 3 Stages
  - Classification Options and "Similarity Rules"
  - Disambiguation in Context
- Past and Current Evaluation
- Future Work
- Conclusions

What is NER?

- Goal of NER is to identify and classify entities that traditionally correspond to proper names and numerical and temporal expressions:
  - People, Places, Orgs, TimEx, NuMex
- But also:
  - Abstract Things, Events, Products, etc..
- Important in pre-processing stages of:
  - Information Extraction
  - Machine Translation, …

How to Build a NER System?

- *Usually* NER systems are built employing:
  - a set of rules regarding NE morphology and context
  - one or more gazetteers
- For developing the rule set, developers may:
  - manually encode the rules
  - apply ML techniques *if a tagged corpus is available*
- To obtain gazetteers, developers may:
  - Rely on existing (official) resources
  - Compile their own gazetteers from various sources

Dealing with a large number of rules

- Rule sets tend to grow exceedingly, especially if:
  - the NER task involves more than the "traditional" entities. Ex: "9th Symphony", "Alzheimer's disease"
  - the text to be analyzed is not just "well-behaved" newspaper text. Ex: web-pages, blogs, etc.
- Developers usually end up with a large rule set
  - difficult to maintain
  - difficult to debug
  - difficult to expand

A possible alternative approach

- Use higher-level rules for classifying NEs instead of a set of very specific (and potentially over-fitting) rules.
- Relying on "similarity" between internal features of "similar" NE's
- Classification becomes a task of assigning the same tag to NE's that are "similar" to other known NE's:
  - This leads to a smaller set of rules
- This strategy helped us building SIEMÊS, one of the top scoring systems in HAREM 2005

Ambiguity in NER task

- But: NOT that easy because NER task involves many ambiguous cases. At least:
  - **Type I Ambiguity:** the same name is used by two different type of entities (ex: "Lima")
  - **Type II Ambiguity**: the same entity may be mentioned with different *semantic roles* which need to be differentiated (ex: "Teatro de S. João")
  - **Type III Ambiguity:** the same name is used to refer to different, yet closely related entities (ex: "Picasso", "Kispo")

Hypotheses underlying SIEMÊS

- Disambiguation ***seems*** to involve deciding between just between a few (i.e. less than four) "reasonable" Classification Options
- Classification Options can be generated using similarity rules, even if we totally disregard the context
- Disambiguation among previously generated Classification Options can then be performed by looking at simple contextual clues

* Hypothesis *

SIEMÊS: NER in 3 Stages

- Stage 1: Identification of Candidate NE's (and Classification of NuMex)
- Stage 2: Generation of Classification Hypothesis based on similarity rules and a wide-scope gazetteer
- Stage 3: Classification in Context / Disambiguation

Stage 1: Identification of Candidate NE's

- Identification of NE "seeds": valid ones are uppercased words and numerical tokens
- "Seed growing": words and tokens surrounding "seeds" are absorbed according to simple regular expression grammars and a list of possible "linkage" structures. Ex: *de*, *para a*, etc.
- Numerical expressions are also classified in this stage: ambiguity is less problematic for NumEx
- Note: if no classification is found for that seed, splitting may occur later in Stage 3

Stage 2: Generation of Classification Hypotheses

- The goal is to formulate reasonable hypotheses ("educated guesses") for classifying NE candidates:
  - *José da Silva* "possibly" refers to a Person
  - *Fundação José da Silva* "possibly" refers to an Org.
- Using similarity rules:
  - based on "internal evidence" of the candidate
  - disregarding the context (which may be absent or may be even more difficult to analyze…)
- Ex: *Satini* ***GTI*** -> may be a car (i.e. Product)

But similarity to what?

- Similarity is calculated over the content of gazetteers that serve as "knowledge bases"
- Moves classification effort **from** hard-coded rules over found NE's **to** similarity rules over the gazetteer
- The gazetteer will need to be representative
  - store many different examples to cover many different possibilities

REPENTINO

- The gazetteer behind SIEMÊS:
  - wide-scope: 11 top categories -> 102 subcategories
  - compiled mostly by extracting names from corpora and from content-specific web sites

Top-Category	# Sub-Categories / #examples	Top-Category	# Sub-Categories / #examples
Abstractions	13 / 5,832	Paperworks	9 / 4,439
Art/Med/Com	9 / 15,358	Products/Brands	15 / 9,262
Events	8 / 25,424	Beings	6 / 287,707
Places	16 / 50,810	Substances	4 / 1,472
Nature	5 / 869	Others	6 / 1,809
Organizations	11 / 47,143	Total	102 / ~ 450k

How similarity is used in SIEMÊS

- SIEMÊS tries to discover possible sub-categories for a given NE candidate using REPENTINO.
- Interesting sub-categories are those which include examples "similar" to the candidate
- 5 high-level similarity rules: first-to-be-matched policy
- SIEMÊS may find more than one sub-category for a given NE:
  - that information is used in Stage 3 (Disambiguation)

The five kinds of similarity rules of SIEMÊS

Rules, from the most restrictive to the least:

1. Exact match: a given candidate is exactly matched
   - Several possible matches may occur. Ex: *América*
2. Same N words in the beginning
   - Tries to match the longest possible substring
3. Same N words in the ending
   - Tries to match the longest possible substring
4. Number of common N-Grams
   - this rule generates several partial comparisons for each candidate.
5. Frequent word(s) in certain subclasses.
   - this rule tries to match the candidate with items in REPENTINO that share *any* word in common with it.

Rules 2 and 3

- Intended to deal with highly regular cases that are very frequent in Portuguese.
  - Rule 2 is especially suited to cover Organizations and Events with long names
    - "Universidade Federal do...", "Associação Regional de…".
  - Rule 3 deals with cases that have "standard" endings such as brands or company names
    - "… Ltd.", "... GTI", "... Corp."

Rules 4 and 5

- Rule 4 explores certain regularities that are often found in titles (books, movies, computer games)
  - Ex: "O Regresso do Herói" / "A Fúria do Herói"
- Rule 5 tries do deal those cases that do not have enough regularity except for a word that is highly discriminative and which may occur anywhere in the candidate
  - Ex: *Intercooler*, *Pentium*.

Stage 3 - Classification in Context / Disambiguation (1)

- Stage 2 provides several possible classification options
- Stage 3 performs disambiguation in context using:
  - Rules that disambiguate between two or more classification hypotheses
  - rules that deal with cases that may be ambiguous even if only one classification option is found
- In Stage 3 SIEMÊS also tries to classify candidates for which no relevant classification evidence has been found in Stage 2: "Last Attempt Rules"

Stage 3 - Disambiguation rules

- usually considering only one or two words preceding the candidates (prepositions or other function words)
  - some important information has already been collected during Stage 2
  - Example (disambiguate Place-Person ambiguity):
    - if the Top 2 classification Options for a given candidate (found in Stage 2) are Place and Person
    - and in this particular context the candidate is preceded by *no*, *na* or *em*, then the candidate is tagged as Place;
    - if it is preceded by *o* or *a*, then tag it as Person;
    - otherwise, tag the candidate as the highest scoring classification hypothesis, as given by Stage 2

Stage 3 - Disambiguation rules (2)

- Equally simple disambiguation rules for
  - Company-Brand/Product
  - Place vs. Other
  - …
- However:
  - For many of those cases we were not able to resolve ambiguity using these simple rules
  - The "Last Attempt Rules" were also too noisy
  - A lot of work is still needed to improve Stage 3

Evaluation of SIEMÊS (1)

- Participation in HAREM (2005):

Category	Rank	Precision (%)	Recall (%)	F-Measure
PERSON	4	65.29	52.20	0.5801
ORG	2	57.63	41.17	0.4803
TIME	4	55.81	61.35	0.5845
PLACE	**1**	**64.09**	**69.83**	**0.6683**
WORKS	**1**	**29.75**	**11.96**	**0.1706**
EVENT	**1**	**47.26**	**43.05**	**0.4506**
ABSTRACT	2	41.80	28.60	0.3396
THING	2	30.00	13.33	0.1846
VALUE	**8**	**53.32**	**37.42**	**0.4398**

- Overall 2nd

Evaluation of SIEMÊS (2)

- Participation in Mini-HAREM (2006) with a completely re-engineered version: SIEMÊS2
- Chance to perform component analysis:
  - Test relative importance of individual rules (1,2,3)
  - Test two implementations for Rule 5
  - Test an additional "high-precision" classification layer, based on very explicit contexts
- Results will be known soon…

Future work

- Extend "Similarity Rules" to Stage 3.
  - Use information of entities or objects found in "Similar Contexts" to disambiguate NE.
- Example "A aldeia foi inundada pelo Lima" "The village was flooded by Lima"
  - Searching BACO for {foi inundada pelo X}, X =
    - "rio/river","lago" / "lake", "mar" / "sea", etc…
    - "Rio Jaboatão", "Rio Beberibe", etc.
  - We may be able disambiguate several cases by looking at information occurring in very "similar" contexts

Conclusions

- By implementing five kinds of high-level similarity rules and using a wide-scope gazetteer we were able to develop one of the top-scoring NER systems in HAREM
- Similarity rules:
  - exploit certain regularities that exist in names
  - help generating a set of "reasonable" classification options
- Classification options are disambiguated through simple contextual rules that focus on frequent ambiguous cases
- Further improvement may be achieved by extending Similarity Rules to context analysis procedures and disambiguation