Cooperatively evaluating Portuguese morphology

Diana Santos
Luís Costa
Paulo Rocha
www.linguateca.pt

Morfolimpíadas: cooperatively evaluating morphological analysers for Portuguese

- Evaluation contest paradigm
  - Importance for science and for community building
  - Shared task, consensual result, objective measures, knowledgeable organization
- Why morphology
  - Mildly inflected language (70 verb forms)
  - Simple and well defined (?) problem, no infinite set of members
  - Traditionally the first module in a set of NLP tools
  - The task for which there was greater interest
- Goals
  - Exemplify the paradigm with a relatively short schedule
  - Assess the state of the art in morphology (also looking at tokenization)
  - Measure the problem

What is Linguateca

- Improve Portuguese processing
  - Dissemination
  - Resource creation
  - Evaluation
- A virtual organization with four nodes
  - Oslo, Braga, Lisbon, Oporto, ...
  - Collaboration partners in more locations: Odense, Lisbon, São Carlos, Porto Alegre, ...
- A follow-up of the *Computational Processing of Portuguese* project, created in 1998, by the then Ministry of Science and Technology

Assumptions of Linguateca

- First things first
  - Find out what are the problems and bottlenecks of Portuguese processing
- International entities or bodies cannot solve our problems
  - In any case not better than us
  - Resource building is time consuming, and "market driven"
- Language (and not region, or nation) should be the unit for natural language processing
  - So Brazil and Portugal should cooperate closely
- Public resources are a must for scientific progress
  - There are enough barriers already
- One needs organizers to do evaluation

What did we learn?

- Organizing an evaluation contest

5-10 times more time consuming than

- Resource building, maintaining and giving support

5-10 times more time consuming than

- Portal organizing and maintaining

Challenge(s) with *Morfolimpíadas*

- Produce informative and intuitively satisfying measures
- Satisfy participants while at the same time showing problems and remaining work
- Produce quantitative and qualitative data that can be used beyond the actual contest
- Make it interesting enough to have further contests in the future, with more participants (e.g. from industry) and maybe several tracks
- Reuse the experience gained in the organization of other evaluation contests

Why morphology

- There is enough work for Portuguese here
  - nouns x 2 x 4
  - adjectives x 4 x 4 x 2
  - verbs ~ 70 forms
  - enclitics and contractions
- Morphology in our definition included
  - lemmatization
  - PoS classification
- There were enough systems to be interesting to compare
- It can be discussed with a layman!
- There are obvious direct uses of a morphological analyser for Portuguese (spellchecking, IR)

Is morphology well defined?

- NO.
- What is the realm of morphology?
- What is in the scope of a morphological analyser?

Two-step process

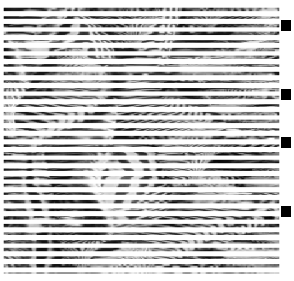
- Setup: organizers and participants take their places...
- Trial
  - Hopefully just like the contest but results are not made public
  - Some options weighed
  - The machinery is put in place
  - Eventual problems discussed with the participants
- Real contest
  - The results are interesting
  - The participants and the organization join to reflect together about what happened and eventually to plan a new way of proceed
- The paper was about the trial, but the presentation is mainly about the real contest

1.as Morfolimpíadas: overview

- Seven participating systems, out of 16-20 out there
  - 3 Portugal 2 Brazil 2 Int
  - 5 “real” morphological analysers, 1 spellchecker and 1 stemmer
- Organization: Linguateca Oslo (+Oporto+contractors)
- Setup:
  - Registration, providing some data
  - Ran their system over 80,000 running text words, in three different formats
  - Processing:

uts.SYSTEM.def.preze.ze.hi.gr.un.le

Zebras: transform into an internal format

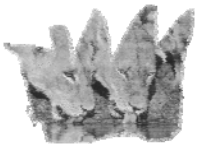


- Every system (with a wildly different output format) is turned into “zebraic” format
- Every zebra output is apparently similar but intriguingly different
- Zebras may still require hienas to deal with complex issues (clitics and contractions)
- Zebra programming requires a full understanding of the high and low level details of the systems (underlying linguistic conception, tokenization [illegible])

Further processing

- Grammatical analyses are turned into one analysis named GRAM
- Some sets of always ambiguous interpretations in the verbal paradigm are turned into one
  - first and third person singular of some tenses
  - personal and impersonal infinitive
  - third person plural of Perfeito and Mais que perfeito
- Numbers are dealt with in a simple form
- Punctuation marks and proper names are handled to yield a hopefully more similar output
- Tokenization problems are dealt with to some extent

Leoas: tearing files to pieces

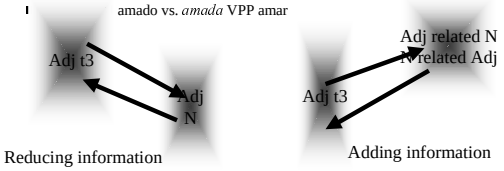


- Distribution by text
- Distribution by variant
- Distribution by genre
- Distribution by medium

- Rationale:
  Is system performance correlated with type of text? Variant?

Domadores: still more is required

- Partially agreeing pieces of information (systems more informative than others)
- ADJ of kind t3 : Noun and Adj
- VPP vs ADJ: VPP and ADJ: only VPP



System's signature

- No. of tokens
- No. of analyses
- Distribution of analyses per form
- Distribution of PoS ambiguity
- Distribution of lemma ambiguity
- No. of verbs
  - No. of tokens which can be analysed as verbs
  - No. of verb analyses
- No. of guessed analyses
- No. of derived analyses
- ...

System comparison

- Qualitative: different kinds of information
- Using the raw output
  - ranking systems in terms of tokenization, verbishness, etc. indexed per text genre, variant, etc.
- Using a golden list: a set of manually agreed upon "right answers" to input forms
  - (Extremely) time consuming task
  - Large room for disagreement
  - Several decision sources (dictionaries, Web, own intuition)
  - Large gray zones (foreign terms, colloquial language, specialized words, PoS classification vs. the flexibility of natural language, common faults, tokenization)
- Using sets of cleverly chosen forms from the automatic output conflation

Tokenization and morphological analysis

- Conceptually, two different things.
- In practice, every MA embodies a tokenizer (or relies implicitly on some tokenization options).
- So, the approach of furnishing our own tokenizer did not help at all. In fact, it was just one more tokenizer to help show that real text tokenization is far from trivial.
  - Percentage of sentences that had "-,(, numbers, . and slash):
- With six/eight different systems, we had in average 80% agreement
- Different systems had different capabilities: NER, compound words, etc.

In our system we do it in a postprocessing phase!

Clitics processing

- In principle, just a fixed set of rules...
- In practice, a hard problem
  - two systems handled clitics as is
  - the others had to be "helped" -- in fact, the organization had to do quite a heavy programming job in order to get a proper output, and with mixed results

- Some blunders

atenta-se ADJ+CL
ei-lo Vt + CL
há-de PREP+PREP
encontrá-mo-nos V + CL + CL +CL
leia-se 1S+3P
devem-se devir IMP2 +3S

Normalization

- Another hard problem for the organization is the fact that most systems do not output their input, but only the result of some amount of normalization
  - NORMALIZAÇÃO -> Normalização
  - Si -> si
  - IDICT -> Idict
  - disto -> de isto
  - alhos-franceses -> alhos franceses
- Or they use an implicit lemma with different capitalization
  - cunha -> Cunha
  - se -> SE
- Capitalization is also the hidden source of other problems, namely the heuristics the systems employ widely differ when the capitalized word is not in their internal lexicon. It can be PROP, X and/or the classifications of the non-capitalized word

Lemmatization

- This is where most *a priori* disagreement could be found
  - of adverbs in *mente*
  - of acronyms and abbreviations
  - of derived words
  - of grammatical words and clitics
  - of numerals
  - of guessed words (adjectives and nouns)

PoS attribution

- Fuzziness between adjective and noun
- Fuzziness between adjective and adverb
- The ponderous problem of past participles
- Grammatical classifications
- Is there an INTERJ PoS?
- How proper are Proper nouns?
- Acronyms are always PROPs ? *V.S.F.F., IA*
- Numbers and dates are NUM?

Interesting problems

- Past participle vs. adjective
  - sweeping generalization: all cases are considered BOTH V and ADJ
  - but neither of these three ways is foolproof
  - inalterada ADJ (no verb *inalterar*)
  - atempada ADJ (no verb *atempar*)
  - ferida ADJ VPP (lemmata: *ferido ferir*)
  - impressa ADJ VPP (lemmata: *impresso imprimir*)
  - intervindo VPP (adjective: *interveniente*)
  - resultado VPP (adjective: *resultante*)

Rare and non-standard

- What to do with them?
- Depending on what a system does
  - it is better to have wide coverage and accept non-standard language
  - it is better to be focussed on standard language so that errors can be found
  - it is better to accept every possible meaning albeit rare
  - it is better to concentrate on modern present day language and on the meanings that are actually in use
- In the golden list we tried to annotate rare and non-standard cases, so that comparisons can be made including or excluding them

Conclusion?

- This is just the beginning
- **See you at the final session of *Morfolimpíadas*, tomorrow!**
- AVALON'2003 - avaliação conjunta para o português