

---

# Extraction of Linguistic Information for Temporal QA: Ambiguity Resolution of Dense Verbal Inflection



**António Branco**

University of Lisbon

Department of Informatics

NLX – Natural Language and Speech

Group

joint work with

Pedro Martins and Filipe Nunes



- 
- Question answering
  - QA and temporal processing
  - Verbal featurization in context
    - Tagging
    - Problem space
    - Cascaded MLE
    - Naive bag-of-words similarity



# Question answering

---

- QA: to obtain
  - Not documents from keywords
  - But answers from questions
  
- QuexTing
  - Portuguese
  - Questions
    - Written
    - Factoid
    - Open domain
  - Answers
    - Extraction-based
    - Sentence + short exact answer
    - Text documents
    - Web



# Quem é Fernando Pessoa?

---

-----  
RESPOSTA #1 (valor 5)  
-----

Resposta exacta: o primeiro português

Frase completa: Fernando Pessoa é o primeiro português a figurar na Plêiade ( Collection Bibliothèque de la Pléiade ), prestigiada coleção francesa de grandes nomes da literatura.

Data: Mon Nov 12 19:19:03 WET 2007

Documento: [pt.wikipedia.org/wiki/Fernando\\_Pessoa](http://pt.wikipedia.org/wiki/Fernando_Pessoa)

Motor: Google (rank 1)

Busca e parse de resultado: 496 ms.

Download e parse de documento: 356 ms.

Extraccao de resposta: 16 ms.

Tempo total: 1019 ms.

-----  
RESPOSTA #2 (valor 5)  
-----

Resposta exacta: o poeta

Frase completa: Fernando Pessoa é o poeta do anti-sentimentalismo, da evocação da infância como símbolo da felicidade perdida e do fingimento enquanto alienação de si próprio e processo criativo.

Documento: [www.angelfire.com/ar/andret/fernando.html](http://www.angelfire.com/ar/andret/fernando.html)

Motor: Yahoo (rank 10)

Busca e parse de resultado: 764 ms.

Download e parse de documento: 969 ms.

Extraccao de resposta: 36 ms.

Tempo total: 1920 ms.

...

# QA Procedures

---

- Question processing
  - Keywords for document retrieval
  - Expected answer type
  
- Passage retrieval
  - Web services: Third-party search engines
  
- Answer extraction
  - Sentence extraction
  - Answer extraction and ranking

# Development directions

---

## □ Breadth

- Extraction-based answers:
  - Factoid questions: Quem...?, Quando...?, Que X...?,...
  - Definition questions: O que é um/o X?
- Operational system
  - Trade-off accuracy vs. usability
  - 2-4 sec. to display answer

## □ Depth

- Generated answers
  - Temporal questions



# Pattern-based Temporal QA

---

## □ Pushing pattern-based extraction

- **Q:** Quando nasceu Fernando Pessoa?
  - Fernando Pessoa nasceu em 1888...
  - Fernando Pessoa, nascido em 1888,...
  - Fernando Pessoa (1888-1935)...
  - ...
- **A:** 1988
  
- **Q:** Quando viveu Juan Carlos em Lisboa, durante o exílio?
  - Durante o exílio, Juan Carlos viveu em Lisboa entre 1945 e 1948.
- **A:** Entre 1945 e 1948



# Semantic-based Temporal QA

---

## □ Beyond pattern-based...

- **Q:** Juan Carlos morou em Lisboa em 1946?
  - Juan Carlos morou em Lisboa entre 1945 e 1948.
- **A:** Sim
  
- **Q:** Juan Carlos deixou Lisboa em 1952?
  - Juan Carlos deixou Lisboa entre 1948 e 1954.
- **A:** ? \_

## □ ... and clearly beyond

- **Q:** Juan Carlos morou em Lisboa em 1946?
  - Juan Carlos chegou a Lisboa em 1945... Em 1948 Juan Carlos deixava Lisboa para...
  - Juan Carlos chegou a Lisboa em 1945, onde permaneceu os 3 anos seguintes.

# QA meets Temporal processing

---

## □ Tasks and tools

- Detect document date
- Detect and normalize temporal expressions
- Determine relations btw events and temporal entities
- Semantic representation and temporal inferencing
- Standards: TIMEX2; Competition: TERN2004

## □ Relations btw events and temporal entities

- Duration:
  - O parlamento preparou o referendo ao aborto em 2 meses.
- Temporal localization:
  - O parlamento esteve parado em 2 meses.



# Tense and aspect

---

## □ Relations btw events

- |    |   |                  |
|----|---|------------------|
| 1. | (a) Ele entrou na sala. (b) Ela saiu.           | $a < b$          |
| 2. | (a) Ele caiu. (b) Ela empurrou-o.               | $b < a$          |
| 3. | (a) Ele entrou na sala. (b) Ela tinha saído.    | $b < a$          |
| 4. | (a) Ele entrou na sala. (b) Ela estava sentada. | $b$ overlaps $a$ |

## □ Linguistic information needed

- |   |                   |                             |                   |
|---|-------------------|-----------------------------|-------------------|
| ■ | 1. Sentence order | $a < b$                     |                   |
| ■ | 2. Causality      | cause(empurrar, cair)       | ontology          |
| ■ | 3. Tense          | pretérito mqp composto      | verbal featurizer |
| ■ | 4. Aspect         | pretérito imperfeito verbal | featurizer        |
| ■ | ...               |                             |                   |

## □ Tools

- Deep linguistic processing grammar
- Temporal inference



# Verbal featurization as tagging

□ <http://lxsuite.di.fc.ul.pt>

```
<p>  
<s>  
Esta/DEM#fs  
frase/FRASE/CN#fs  
serve/SERVIR/V#pi-3s  
para/PREP testar/TESTAR/V#INF-nInf  
o/DA#ms  
funcionamento/FUNIONAMENTO/CN#ms  
de_/PREP  
a/DA#fs suite/SUITE/CN#fs  
.*//PNT  
</s>  
<s> Esta/DEM#fs  
outra/OUTRO/ADJ#fs  
frase/FRASE/CN#fs  
faz/FAZER/V#pi-3s  
o/LDEM1  
mesmo/LDEM2  
./PNT  
</s>  
</p>
```



# Tagging with best algorithms

## □ HMM-based featurization

- Manually annotated corpus
- 425 000 tokens
- 50 000 verb forms
- 10-fold cross-validation
- "Word": word + pos tag
- Tag: verbal infl feature bundle
- Tagset: 100
  
- Thorsten Brants' TnT
  
- Precision 1: **94.8%**
  - input with manual, correct POS
- Precision 2: **91.7%**
  - input with automatic, noisy POS

## □ HMM-based POS tagging

- Manually annotated corpus
- 260 000 tokens
- 10-fold cross-validation
- "Word": raw token
- Tag: POS categories
- Tagset: 60
  
- Thorsten Brants' TnT
  
- Precision: **96.9%**

## □ Featurtization

- **> 5 points below**
- just a matter of tagset size?



# More tagging techniques

## □ ME-based POS Tagging

- Ratnaparkhi's MXPOST
- Featurization: 89.2%
- POS: 97.1%
- **Featurization: >> 8 points below now!**
  - Most likely not just a matter of tagset size

## □ Tagging

<i>Accuracy</i>	TnT	MXPOST
POS #Tagset=60 200KToken	96.9%	97.1%
Featurization #Tagset=100 400M / 50KToken	94.8% (91.7%)	89.2%



# Is featurization harder than POS tagging?

## □ Baselines

- From the admissible tags/feature bundles for a given token, assign the most frequent one

<i>F-Score</i>	Baseline	TnT	MXPOST
POS	<b>90.4%</b>	96.9%	97.1%
Featurization	<b>88.5% (*)</b>	94.8% (91.7%)	89.2%

- (\*) featurization and lemmatization actually

## □ What to learn from these scores?



# Insight into the problem space

---

## □ Ambiguity classes

### ▪ Lemma-only (a WSD task)

consumo → consumir»IndPres1S; consumir»IndPres1S  
bate → bater/vencer/...»IndPres3S; bater/remexer/...»IndPres3S

### ▪ Termination-only

Mood: dê → dar»ConjPres3S; dar»Imp2cS  
Polarity: dêmos → dar»ImpAfirm1P; dar»ImpNeg1P  
Tense: deram → dar»IndPretPerf3P; dar»IndPretMPerf3P  
Person: dava → dar»IndPretImp1S; dar»IndPretImp3S  
Number: parti → partir»IndPretPerf1S; partir»Imp2P  
Gender: assente → assentar»PartP3SMasc; assentar»PartP3SFem

### ▪ Termination-and-lemma

virei → vir»IndFut1S; virar»IndPretPerf1P

## □ Which task?

- Featurization + lemmatization
- Tagging all words all tags: explosion of the tagset

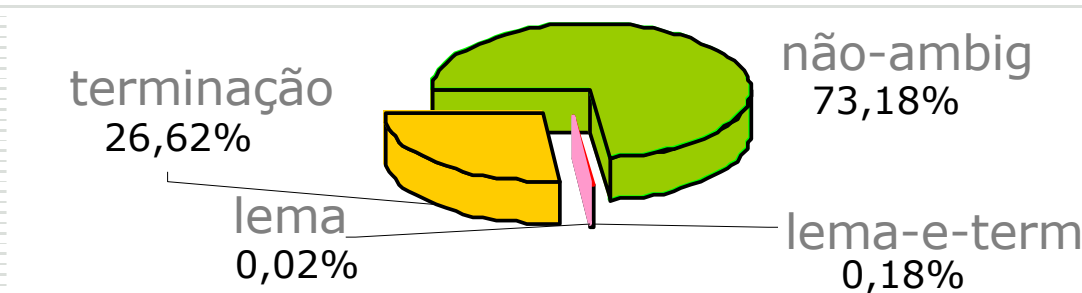


# Lexical ambiguity

## □ Verbal lexicon

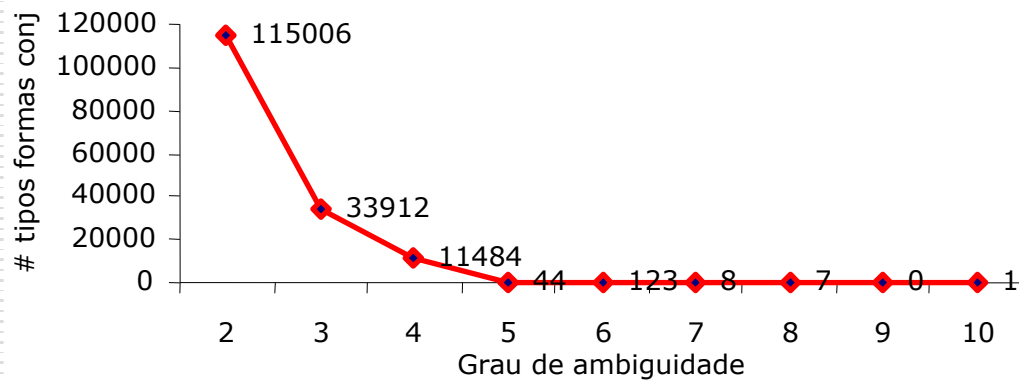
- Lemmas/infinitives
  - 11 350 types
- Verb forms/conjugated forms
  - 816 130 tokens ; 598 651 types
  - 1.36 lexical ambiguity ratio
  - ignoring compound tenses, and 2nd Pers Courtesy (= 3 Pers)

## □ Lexical ambiguity per classes



# Degrees of lexical ambiguity

## □ Verbal lexicon



# Verbal lexicon put to use

- Annotated corpus
  - ca. 250 000 MTokens
  - ca. 35 000 verb forms (13.5%)

- Verbal lexicon put to use

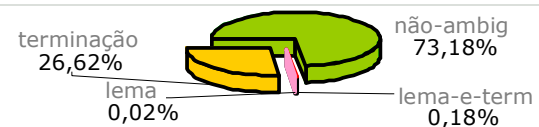
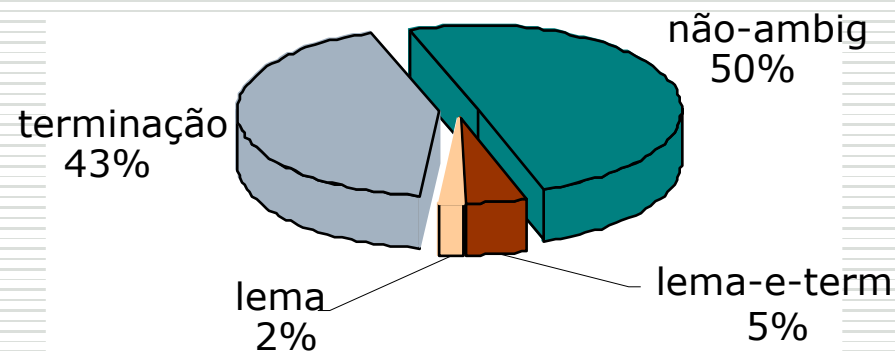
Types	Lemmas	Feature bundles	Verb forms with feature bundle	Ambiguous verb forms
# in lexicon	11 350	109	816 830	160 587
# in corpus	1 951	82	8 635	4 142
Usage rate	<b>17.19%</b>	<b>75.23%</b>	<b>1.06%</b>	<b>2.58%</b>



# Ambiguity in context

## □ Ambiguity per class

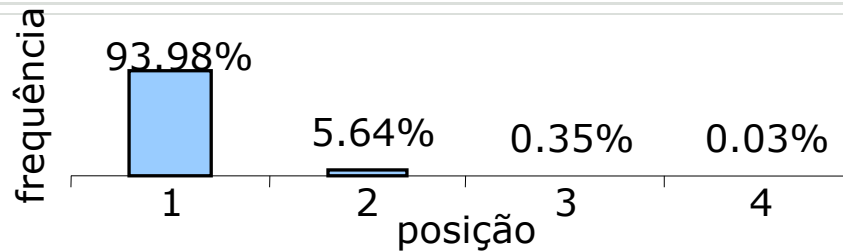
- Slightly more cases to decide (ca 50%) than in POS tagging (ca 40%)



# More on ambiguity in context

---

- Degrees of ambiguity



# Trying to making sense of first results

<i>F-Score</i>	Baseline	TnT	MXPOST
POS	<b>90.4%</b>	96.9%	97.1%
Featurization	<b>88.5%</b>	94.8%	89.2%

- MXPOST: t-2 t-1, t-1, w-2, w-1, w+1, w+2, prefix, suffix
- TnT: trigrams with backoff, suffix
- Drop from POS to Feat by MXPOST
  - Too much for local context ...
  - ... non-local, topical context more important to verbal featurization
- Slight drop by TnT, in contrast to MXPOST
  - From the local context features, suffix is very important given its more frequent feature bundle is very frequent

# Two directions to explore

---

- Suffix-driven MLE
  - Suffix is very important
  - Which smoothing?
  
- Featurization as WSD
  - Topical context more important
  - Which features?



# Cascaded MLE

---

## □ Algorithm

- Most frequent feature bundle and lemma pair
- Else
  - Obtain candidate solutions with lemmatizer
  - Discard cand. with unknown lemmas if there are known ones
  - Select remaining cand. with most frequent feature bundle
  - Select remaining cand. with most frequent lemma
  - Select remaining cand. with lemma ending "-ar"

## □ Evaluation

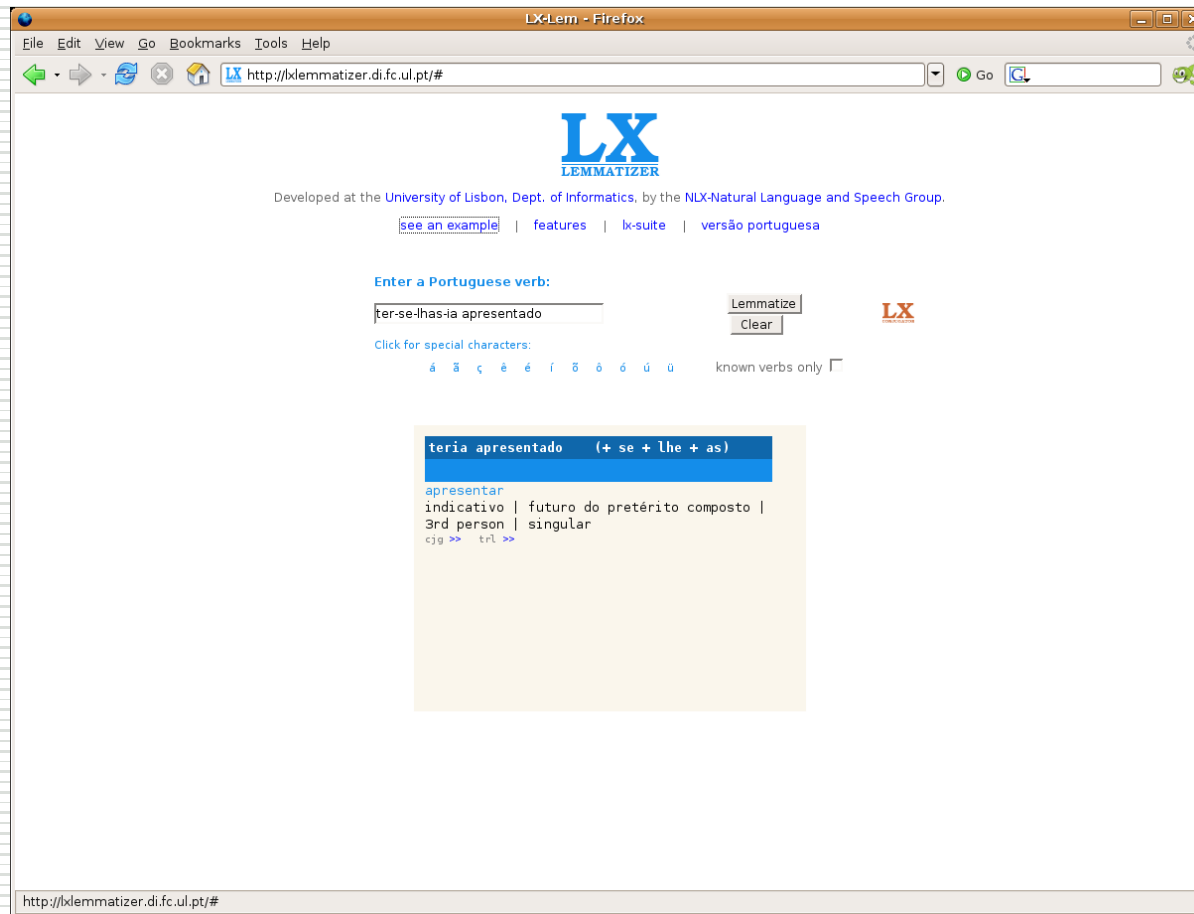
- 96.0% Precision, 95.9% Recall
- **96.0% F-score**
  
- Naive classifier, yet very good results!
  - > 94.8% TnT, second best
- Featurization and lemmatization: lemmatization errors also included

## □ Out of context lemmatizer needed



# LX-Lemmatizer

<http://lxlem.di.fc.ul.pt>



The screenshot shows a web browser window titled "LX-Lem - Firefox" with the address bar containing "http://lxlemmatizer.di.fc.ul.pt/#". The page features the "LX LEMMATIZER" logo and text indicating it was developed at the University of Lisbon. It includes a form for entering a Portuguese verb, with "ter-se-lhas-ia apresentado" entered. A "Lemmatize" button is visible, along with a "Clear" button and a "LX" logo. Below the input field, there are links for "see an example", "features", "lx-suite", and "versão portuguesa". A "Click for special characters" section lists characters: á, ã, ç, ê, é, í, ï, ó, ô, ú, ü, and a "known verbs only" checkbox. The output area shows the lemmatized form "teria apresentado (+ se + lhe + as)" in a blue box, followed by the base form "apresentar" and its grammatical details: "indicativo | futuro do pretérito composto | 3rd person | singular". At the bottom of the output area, there are small navigation arrows: "< jg >>" and "tr >>".



# Naive bag-of-words similarity

---

## □ Algorithm

- Compare sentence with each stored sentences
  - Score 1 for each matching word
- Keep the most similar
  - Prefer similar sized sentence in case of a tie
- Assign feature bundle and lemma pair occurring there

## □ Evaluation

- 92.2% Precision, 80.3% Recall
- 85.8% F-score
- < 88.5% baseline: naiveté didn't pay off here



# So far

<i>F-Score</i>	Naive Bag-of-words Similarity	Baseline	MXPOST	TnT	Cascaded MLE
POS		90.4%	97.1%	96.9%	
Featurization	85.8% (*)	88.5% (*)	89.2%	94.8% (91.7%)	96.0% (*)

- (\*) Featurization + lemmatization



# Related work

---

- (Chrupala, 2006)
  - Hard to compare, though:
    - lemmatization only
    - nominal and verbal
    - infinitives skipped
  - Best scores:
    - 94.64% Catalan
    - 92.48% Spanish
    - ...
    - 91.21% Portuguese



# Convenient

---

- (Out of context) lemmatizer+featurizer
  - Narrow space of classes for each decision
  - Neologisms/out-of-lexicon items
  - Featurization in tandem with lemmatization
  
- Inter-annotator agreement
  - Likely < 100%
  - annotation of infinitive 3PerSing vs non-inflected
  - ...



# On going work

---

- Cascaded MLE: push further
  - Error analysis:
    - 46% 3Pers assigned: should be 1Pers
    - 31% non-inflected infinitive: should be inflected 3Per
  
- Take featurization as WSD seriously
  - Adapt best performing classification algorithms: NB, SVM,...
  - Feature selection
    - Strictly local
    - Non local: which, besides bag-of words?
      - Negation (Imperatives): Faz isso / Não faça isso
      - “talvez” (Conjuntivo): Talvez chova amanhã / \*Talvez chove amanhã
      - ...



---

# Thank you



**António Branco**

University of Lisbon

**Department of Informatics**

**NLX – Natural Language and Speech**

**Group**

joint work with

**Pedro Martins and Filipe Nunes**

