„Al. I. Cuza" University, Iasi, Romania

Faculty of Computer Science

# Grammar-based Automatic Extraction of Definitions. Applications for Romanian

**Adrian Iftene, Diana Trandabăţ, Ionuţ Pistol**
**{adiftene, dtrandabat, ipistol}@info.uaic.ro**

September, 26, 2007, Borovets, Bulgaria

# Overview

- Introduction
- Romanian grammar
  - Categorization of Definitions
  - Distribution of the definitions into types
  - Rules
  - Evaluation
- Applications
  - Question Answering
  - Textual Entailment

# Introduction

- Under the framework of the project LT4eL was created an environment for collecting and (semi)automatic exploiting language resources (Monachesi et al, 2006)
- 9 languages involved (bul, cze, dut, eng, ger, mal, pol, por and rom)
- Manually annotation of keywords, definitions of various terms and semantic concepts
- A grammar was created for the automatic identification of definitions in texts

# Categorization of Definitions

- **"is_def"** – "*HTML* ***este*** *tot un protocol folosit de World Wide Web."* (HTML **is** also a protocol used by World Wide Web).
- **"verb_def"** – "*Poşta electronică* ***reprezintă*** *transmisia mesajelor prin intermediul unor reţele electronice*." (Electronic mail represents sending messages through electronic networks).
- **"punct_def"** – "*Bit* **–** *prescurtarea pentru binary digit*" (Bit **–** shortcut for binary digit)

# Categorization of Definitions (cont…)

- **“layout_def”**

| | |
|---|---|
| **Ro:** | |
| ***Organizarea datelor*** | ***Cel mai simplu mod de organizare este cel secvențial.*** |
| **En:** | |
| **Data organizing** | **The simplest method is the sequential one.** |

- **“pron_def”** – *“…definirii conceptului de baze de date. **Acesta** descrie metode de ….”* (…defining the database concept. **It** describes methods of ….)
- **“other_def”** – “*triunghi echilateral, **adică** cu toate laturile egale*” (equilateral triangle **i.e.** having all sides equal).

# Distribution of the definitions

| Type | Manual | % | Automatic | % |
|---|---|---|---|---|
| is_def | 70 | 33.8 | 204 | 32.8 |
| **verb_def** | 116 | **56.0** | 272 | **43.8** |
| **punct_def** | 15 | **7.2** | 124 | **20.0** |
| layout_def | 2 | 1.0 | 21 | 3.4 |
| pron_def | 4 | 2.0 | 0 | 0.0 |
| **Total** | **207** | | **621** | |

# Rules

- Simple grammar rules
- Composed grammar rules
- “is_def” grammar rule:

```
<rule name="may_be_term">
  <seq>
    <query match="tok[@base='fi' and
                  substring(@ctag,1,5)='vmip3']"/>

    <first>
        <ref name="UndefNominal" />
        <ref name="DefNominal" />
    </first>
  </seq>
</rule>
```

# Evaluation

- Lxtransduce (Tobin 2005) is used to match the grammar in files

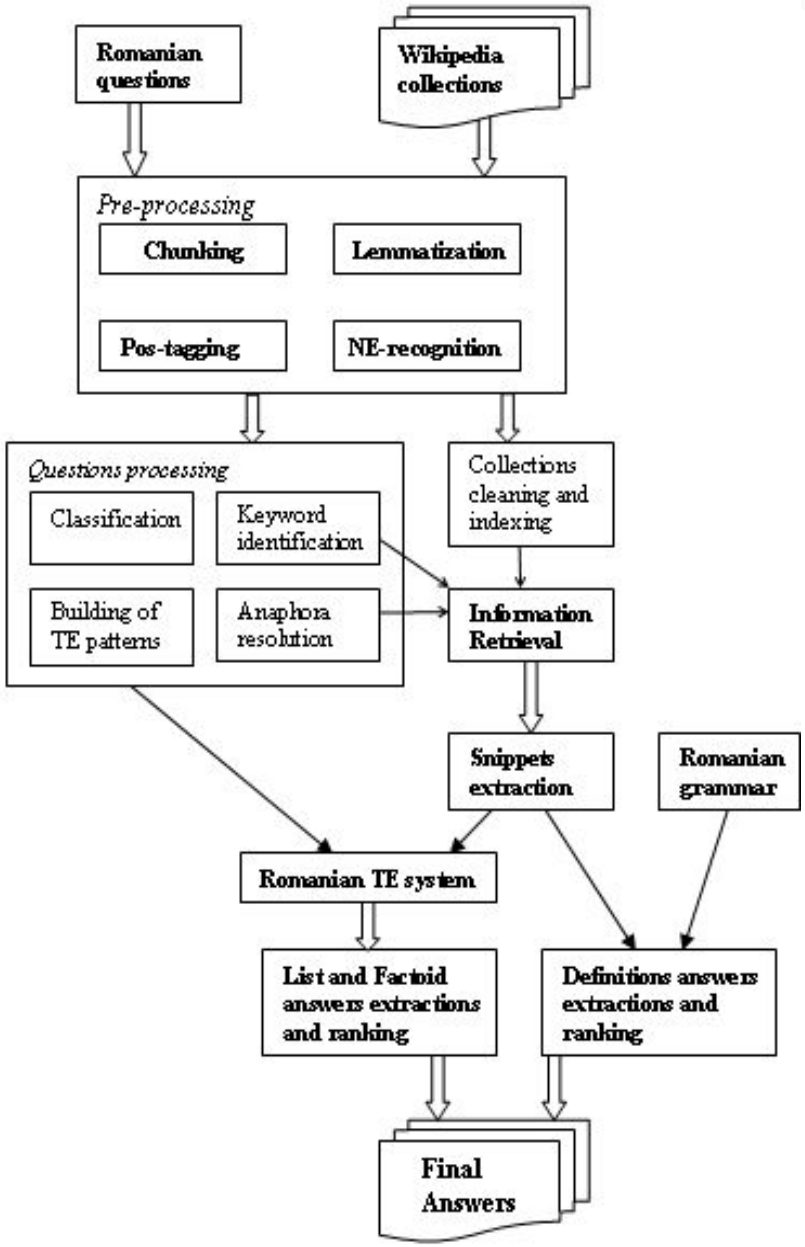
| Definition Type | Result |
|---|---|
| is_def | **Sentence-level matching:**<br>P: 0.5366, **R: 1.0**, F2: 0.7765<br>**Token-level matching:**<br>P: 0.0648, R: 0.3328, F2: 0.14 |
| **verb_def** | **Sentence-level matching**<br>P: 0.7561, **R: 1.0**, F2: 0.9029<br>**Token-level matching**<br>P: 0.0471, R: 0.1422, F2: 0.085 |
| punct_def | **Sentence-level matching**<br>P: 0.1463, **R: 1.0**, F2: 0.3396<br>**Token-level matching**<br>P: 0.0025, R: 0.1163, F2: 0.0072 |
| layout_def | **Sentence-level matching**<br>P: 0.0488, **R: 1.0**, F2: 0.1333<br>**Token-level matching**<br>P: 0.0007, R: 0.1020, F2: 0.0022 |

# Question Answering

- Accordingly to the answer type, we have the following type of questions (Harabagiu, Moldovan 2007):
  - ***Factoid*** – "*Who discovered the oxygen?*" or "*When did Hawaii become a state?*" or "*What football team won the World Coup in 1992?*".
  - ***List*** – "*What countries export oil?*" or "*What are the regions preferred by the Americans for holidays?*".
  - ***Definition*** – "*What is quasar?*" or "*What is a question-answering system?*"

# QA: System architecture

- Pre-processing
- Questions processing
- Collection indexing
- Information retrieval
- Snippet extraction
- Answer extraction



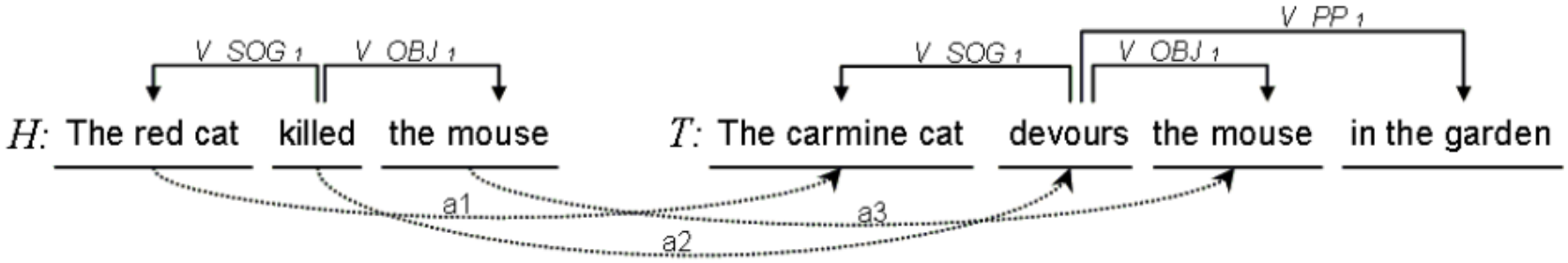
# QA – Example

- Question: Cine este Zeus?

  D: 0026: (Cine, zeus, PERSON)
- Snippet:
  0026#10014#1.0#Zeus#Zeus\zeus\ASN\P este\fi\V3\ cel\cel\TSR\ mai\mai\R\ puternic\puternic\ASN\ dintre\dintre\S\ olimpieni\olimpieni\NPN\ ,\,\COMMA\ socotit\socoti\VP\ drept\drept\S\ stăpânul\stăpân\NSRY\ suprem\suprem\ASN\ al\al\TS\ oamenilor\om\NPOY\ şi\şi\CR\ al\al\TS\ zeilor\zeu\NPOY\ .\.\PERIOD\
- Our pattern for "is_def" (0026.*\zeus\.*\P .*\fi\V3\ (.*)) match the snippet

# Textual Entailment

- TE is defined (Dagan et al., 2006) as a directional relation between two text fragments, termed *T* (text) - the entailing text, and *H* (hypothesis) - the entailed text.
- It is then said that *T* entails *H* if, typically, a human reading *T* would infer that *H* is most likely true.
- Example:
  - T: The carmine cat devours the mouse in the garden.
  - H: The red cat killed the mouse.



# TE - Background Knowledge

- Snippet extraction for NEs from H without corresponding value in T. For each such context:
  a) "core" identification
  b) left NEs extraction
  c) right NEs extraction
  d) calculate LNEs X RNEs

# BK - Example

<pair id="748" entailment="YES">

<T>Argentina President Carlos Menem has ordered an 'immediate' investigation into war crimes allegedly committed by British troops during the 1982 Falklands War.</T>

<H>Argentine demanded an investigation of alleged war crimes during the Falklands War.</H>

</pair>

# BK - Example

"Argentine": Extracted Snippets from Wikipedia:

```
 ar |calling_code = 54 |footnotes = Argentina also has a
territorial dispute
 Argentina', , NaciÃ³n Argentina (Argentine Nation) for many
legal purposes), is
 in the world. Argentina occupies a continental surface area of
 Argentina national football team
```

| |
|---|
| Argentine [is] Argentina |
| Netherlands [is] Holland |
| 2 [is] two |
| Los Angeles [in] California |
| Chinese [in] China |

```
Netherlands [is] Dutch
Netherlands [is] Nederlandse
Netherlands [is] Antillen
Netherlands [in] Europe
Netherlands [is] Holland
Antilles [in] Netherlands
```

# Conclusions

- We presented the Romanian grammar used in the European LT4eL project
- The definitions were devised in 6 types
- Applications: QA and TE
- Apply the grammar to a new corpus

# Acknowledgments


- Special thanks goes to the other members of the Romanian team in the LT4eL project, Dan Cristea and Corina Forăscu
- We also acknowledge the help provided by Claudia Borg

# THANK YOU!