Tehnici de Ingineria Limbajului Natural

Curs 2
Proiecte

Curs: Dan Cristea

Laboratoare: Diana Trandabăț, Mihaela Onofrei, Daniela Gîfu, Ionuț Pistol

PROIECT 1

Name entity recogniser (NER)

- NER: un modul capabil să clasifice mențiunile de entități cu nume
- Utilizare:
  - input: text
  - output: entitățile marcate XML
- Exemple

```
<ENTITY ID="…" TYPE="PERSON">George Ionescu</ENTITY>
<ENTITY ID="…" TYPE="ORGANISATION">
    <ENTITY ID="…" TYPE="PERSON">Alexandru Ioan Cuza</ENTITY>
    din <ENTITY ID="…" TYPE="PLACE" SUBTYPE="TOWN">Iași</ENTITY>
 </ENTITY>
```

PROIECT 1

Name entity recogniser (NER)

- Tehnologii și resurse
  - Expresii regulate: Graphical Grammar Studio (GGS)
  - Gazeteers (liste de nume publice)
    - Geonames (Pentru România, Geonames include peste 25.000 de entităţi unice, cu peste 45.000 de denumiri alternative)
- Referințe:
  - Download GGS: https://sourceforge.net/projects/ggs/
  - Gazetteer: https://drive.google.com/open?id=1oTBmSX_93cnVIwqxJcAoqQU7W4qDBFEV
  - convenții XML de adnotare a entităților cu nume: https://drive.google.com/drive/folders/1S4Mfj-4hfCXavAEkY8VQGvBjNxlcdOTF
  - D. Cristea, D. Gîfu, I. Pistol, D. Sfirnaciuc, M. Niculiță (2016). A Mixed Approach in Recognising Geographical Entities in Texts. in D.Trandabăț and D.Gîfu (eds.): Proceedings of the Workshop on Social Media and the Web of Linked Data, RUMOUR-2015, A satellite event of EUROLAN-2015, Sibiu, Romania, July 2015, Springer International Publishing, https://profs.info.uaic.ro/~dcristea/papers/RUMOUR-Cristea%20et%20al.pdf

PROIECT 2

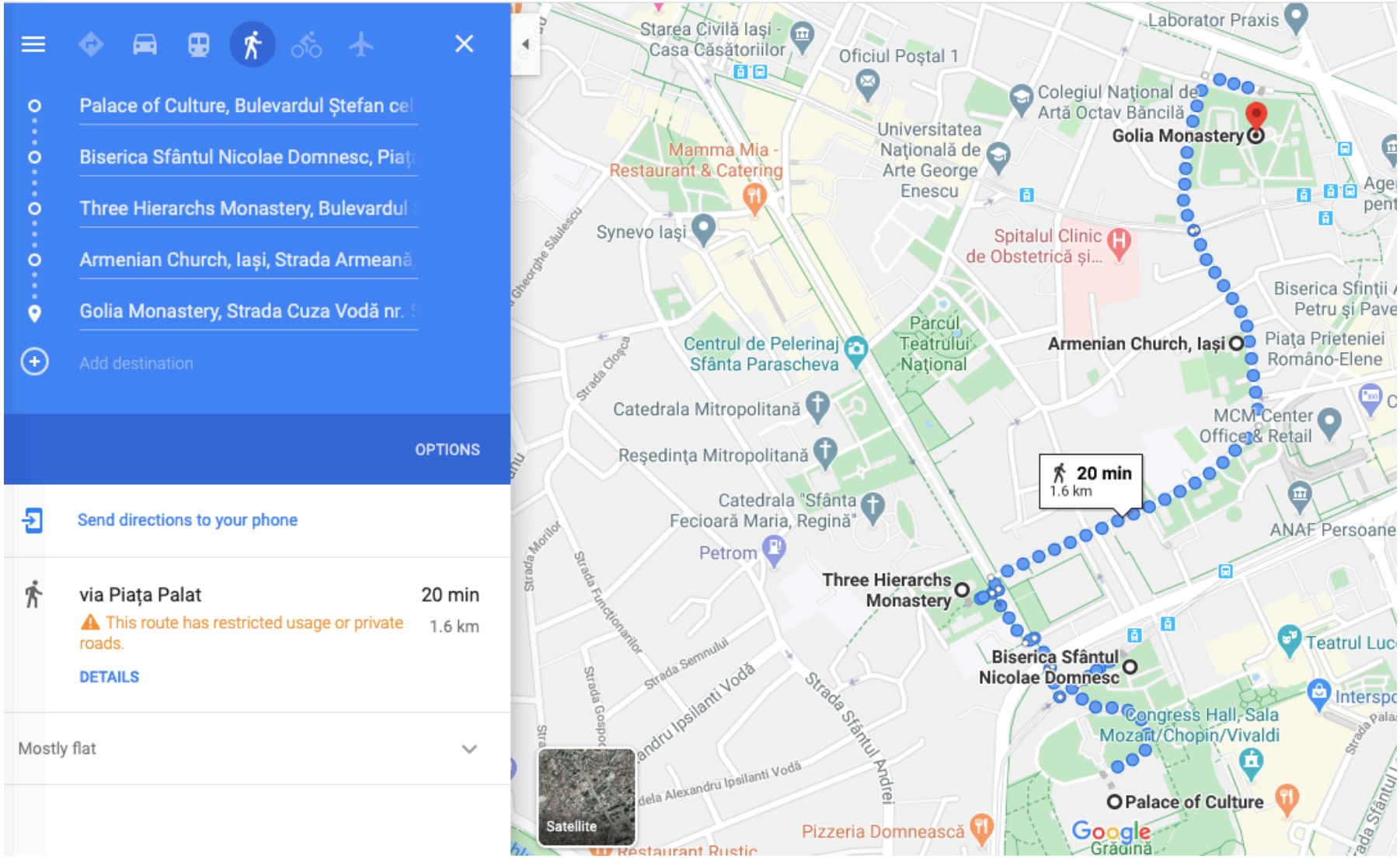
Route tracer following text descriptions

- Un modul capabil să deseneze trasee plecând de la descrierile lor în text
- Utilizare:
  - input: text adnotat XML cu nume de entități geografice ori de instituții
  - output: o hartă cu trasee (Google Maps)
- Exemplu

"Biserica din cărămidă de la sfârșitul secolului XV-lea, de lângă Palatul Culturii, este Biserica Sf. Nicolae… O plimbare de 5 minute spre nord, pe Bulevardul Ștefan cel Mare, te duce la Biserica Trei Ierarhi (str. Ștefan cel Mare și Sfânt nr. 28)... Biserica Armenească de la începutul secolului XIX-lea se află pe Strada Armenească, o plimbare de 8 minute la nord-est de Piața Palatului, pe Strada Costache Negri… Mergi puțin mai departe spre nord, până pe Strada Cuza Vodă nr. 51, unde se înalță Mănăstirea Golia."

PROIECT 2

Route tracer following text descriptions



PROIECT 2

Route tracer following text descriptions

- Tehnologia
  - Extrageți descrieri din cărți, Wikipedia, ghiduri de călătorie etc.
  - Notați în XML pe ele, în convențiile de la Proiectul NER, entități de interes turistic
  - Folosiți Google Maps sau o altă aplicație web capabilă să deseneze trasee
  - Un bonus dacă descrierile rezultate sunt în termeni de relații spațiale
- Referințe
  - Pentru convenții XML de adnotare a relațiilor spațiale: raport MappingBooks 2015, la: https://drive.google.com/drive/folders/1S4Mfj-4hfCXavAEkY8VQGvBjNxlcdOTF

PROIECT 3

Recognising time in texts

- Un modul capabil să adnoteze XML, în maniera TimeML, expresiile temporale găsite într-un text
- Utilizare:
  - input: text, eventual adnotat TOK, POS
  - output: expresii temporale marcate TIMEX3
- 4 tipuri de expresii temporale TIMEX3:

(a) specificate complet (DATE): 11 iunie, 1989, vara anului 2002;

(b) nespecificate, relative la momentul curent (TIME): luni, luna viitoare, anul trecut, acum două zile;

(c) durate (DURATION): 3 luni, doi ani, o săptămână;

(d) cu repetare (REPEAT): în fiecare miercuri, anual.

PROIECT 3

Recognising time in texts

- Exemple

```
<TIMEX3 TID="t…" TYPE="DATE" VALUE="25D, 02M, 2020Y">25.02.2020</TIMEX3>
<TIMEX3 TID="t…" TYPE="TIME" VALUE="+1D">mâine</TIMEX3>
<TIMEX3 TID="t…" TYPE="TIME" VALUE="-2D">acum două zile</TIMEX3>
<TIMEX3 TID="t…" TYPE="DATE" VALUE="19C">secolul al XIX-lea</TIMEX3>
<TIMEX3 TID="t…" TYPE="TIME" VALUE="90-99Y, -1C">deceniul '90 al secolului trecut</TIMEX3>
<TIMEX3 TID="t…" TYPE="TIME" VALUE="+1Y">la anul</TIMEX3>
<TIMEX3 TID="t…" TYPE="TIME" VALUE="20D">în data de 20 a lunii</TIMEX3>
<TIMEX3 TID="t…" TYPE="REPEAT" VALUE="1W, M">în primul weekend al fiecărei luni</TIMEX3>
<TIMEX3 TID="t…" TYPE="DURATION" VALUE="1M">o lună</TIMEX3>
```

PROIECT 3

Recognising time in texts

- Referințe
  - James Pustejovsky *et al*. (2003). TimeML: Robust Specification of Event and Temporal Expressions in Text, AAAI Technical Report SS-03-07, https://www.aaai.org/Papers/Symposia/Spring/2003/SS-03-07/SS03-07-005.pdf
  - Sutime Python: https://github.com/FraBle/python-sutime

PROIECT 4

Temporal aligner

- Un modul capabil să determine relații temporale între mențiuni de evenimente și de expresii temporale.
- Utilizare:
  - input: text adnotat TOK, POS, TIMEX3
  - output: marcaje EVENT, SIGNAL și TLINK

PROIECT 4

Temporal aligner

- Exemplu:

```
La
<TIMEX3 TID="t01" TYPE="DURATION" VALUE="1Y">un an</TIMEX3>
<SIGNAL SID="s01" TYPE="AFTER">după</SIGNAL>
ce s-a
<EVENT ID="e01">înființat</EVENT>
Institutul, acesta
<EVENT ID="e02">avea</EVENT>
deja 20 de cercetători.
<TLINK LID="L01" eventID="e01" relatedToEvent="e02" relType="AFTER"
VALUE="1Y">
```

PROIECT 4

Temporal aligner

- Referințe
  - Verhagen & Pustejovsky: Temporal Processing with the TARSQI Toolkit, Coling 2008, https://www.aclweb.org/anthology/C08-3012.pdf
  - Marc Verhagen, Robert Knippen, Inderjeet Mani, James Pustejovsky: Annotation of Temporal Relations with Tango.
  - https://github.com/tarsqi/ttk

PROIECT 5

Automatic writing of the history of a place

- Un modul capabil să sintetizeze o istorie, derularea unor evenimente legate între ele, un timeline, corelând mențiuni de evenimente, de persoane implicate în acele evenimente și de momente ori intervale de timp în care au ele loc, găsite în mai multe documente.
- Utilizare:
  - input: o colecție de texte adnotate la TOK, POS, EVENT pe verbe, TIMEX3, TLINK
  - output: EVENT cu roluri și apoi text

PROIECT 5

Automatic writing of the development of a story

- Exemplu: evoluția epidemiei de Coronavirus (SARS-CoV-2)
  - Luați primele 10 articole din Google găsite cu secvența de căutare “epidemia de coronavirus în Italia”
    - Deșteptarea, 24 februarie: “*Sunt peste 200 de persoane infectate cu coronavirus în Italia, iar patru au murit. Italia este în acest moment cel mai mare focar de coronavirus din Europa. Autoritățile italiene au decis ca 11 orașe să intre în carantină, multe școli au fost închise, iar carnavalul de la Veneția s-a încheiat mai devreme.*”
    - Radio Europa Liberă Moldova, 25 februarie: “*În Italia au fost înregistrate 219 infectări, iar cinci oameni au murit…*”
  - Modificați manual exprimările pentru a le aduce la o formă mai simplă: “*200 de persoane au fost infectate cu coronavirus în Italia*”

PROIECT 5

Automatic writing of the development of a story

- *"200 de persoane au fost infectate cu coronavirus în Italia"*
  - Extrageți:
    - Grupuri nominale: <NP nID="n01">200 de persoane</NP>

  - extrageți evenimentele:

```
<TIMEX3 tID="t01" TYPE="DATE" VALUE="2020.02.24"/>
<EVENT eID="e01" TIME="t01" REC="n01" VERB>
```

PROIECT 6

I listen my speaking agent reading fragments as I walk by

- Aveți o colecție de texte care abundă în entități geografice, marcate XML explicit, textele fiind însoțite de metadate care descriu: autorul și titlul cărții, anul de apariție și editura. Aplicația va semnala proximitatea telefonului față de locațiile menționate în texte și vă va citi acele fragmente care includ mențiunilor respective. În felul acesta, o plimbare printr-un mare oraș se poate transforma într-o călătorie literară.

PROIECT 6

I listen my speaking agent reading fragments as I walk by

- Referințe:
  - rapoartele proiectului ReTeRom (v. laborator)
  - lucrări MappingBooks (a se consulta Cristea et al., în https://profs.info.uaic.ro/~dcristea/publications.html)

PROIECT 7

Cultural routes creator