Zapisnik 13. sestanka DS3

15. junij 2021, 9:00 @ Zoom

Prisotni:

Slavko Žitnik
Matej Ulčar
Matic Kavaš
Milan Ojsteršek
Mladen Borovič
Aleš Žagar
Marko Robnik- Šikonja
Polona Gantar
Matej Martinc
Bojan Klemenc
Simon Krek
Iztok Kosem
Cyprian Laskowski
Jaka Čibej

Opravičeni:

Aljaž Košmerlj

Dnevni red:

1. Pregled izvedbe zadolžitev
2. Poročanje stanja po aktivnostih, ki trenutno potekajo
3. Razno

Sestanek:

- **Ad1:**
  - (**Milan Ojsteršek**) - pripravi dokument in orodja za predprocesiranje različnih tipov vhodov (PDF, ...) v formate, primerne za nadaljnje obdelave.
    - Dostopno na: https://git.lhrs.feri.um.si/ivan.kovacic/textapi
    - TODO: še dodati stvari, ki so implementirane, nato se predstavi ...
  - (**Milan Ojsteršek**) Priprava končnega seznama virov in določitev obsega dodatnih odgovorov na vprašanja do sestanka RSDO DS3 v juliju.
  - (**Ojsteršek, Simon, Aljaž, Slavko, Marko RŠ**):
    - Sestanek bo **24. 6. ob 9:00**. Slavko bo še poslal mail z vabilom in opisom predlaganih aktivnosti.
  - (**VSI**) Pripraviti predloge programskih vmesnikov za razvita orodja

    - https://docs.google.com/document/d/1KinWqrF9vMyPaanQT5PV48qcBw44-ddcqJdLC-bSUSg/edit#

- **Ad2:**
    - **A3.1**
        - Posodobljena infrastruktura Centra za jezikovne vire in tehnologije na področju “digitalne slovarske baze” in orodij za označevanje (**M0-M34, Krek**) in Digitalna slovarska baza s odprtimi in povezanimi podatki iz odkupljenih jezikovnih virov (**M0-M34, Krek**)
            - odkupi in odstopi - aktivnosti potekajo v DS7, v CJVT se bodo pojavili viri, ko bodo na voljo (verjetno pred poletjem):
                - --> reševanje administrativnih težav (prvi krog)
            - naslednji krog jeseni
        - Sloleks (**-M34, Krek**)
            - Na Sloleksu dela predvsem Jaka Čibej, skupaj z Nejcem Robido, drugi je zadolžen za govorjeni del (naglasi, transkripcija).
            - Avtomatsko pridobivanje oblik za podano besedo (Matic Kavaš)
                - TODO: obširnejši uvodni sestanek?

    - **A3.2**
        - Orodje za prepoznavanje imenskih entitet (**M3-M31, Žitnik**)
            - Prvi izdelek narejen.
        - Orodje za ekstrakcijo povezav (**M3-M31, Žitnik**)
            - Identificiranih 5 pristopov za "medjezikovni prenos" označevanja relacij
                - TODO: preskus nekaj orodij (prenos vložitev, prevajanje in poravnava)
            - TODO: izdelava validacijske množice na podlagi identificiranih pravil
            - --> delo se bo nadaljevalo v avgustu (Timotej Knez, Miha Štravs)
        - Orodje za odkrivanje koreferenčnosti (**M3-M31, Žitnik**)
            - Poslan posodobljen članek v revijo Repozitorij bo kmalu na voljo (Klemen, Žitnik)
            - TODO: identifikacija omenitev + priprava docker vsebnika (Martin Čebular)
        - Orodje za avtomatsko ekstrakcijo relacij za gradnjo semantične mreže in Semantična mreža oz. avtomatsko zgrajena baza znanja (**M3-M31, Ojsteršek**).
            - Ojsteršek, Simon, Aljaž, Slavko, Marko RŠ:
                - Sestanek 24. 6. ob 9:00 - https://docs.google.com/document/d/19PJYUYnm0ahPnga4APS0dwITAJL7SiDyot47b3mBV_Q/edit#heading=h.vobvjicctcl

        - @IJS&UMB: uskladitev glede nalog in priprava časovnice (orodje za posodabljanje baze najkasneje konec 2021), poročanje na rednih mesečnih sestankih, razmislek o integraciji (označevanje in posodabljanje baze v živo)
        - @UMB: priprava procesov in zaslonskih mask. Potrebno je tudi razmisliti glede preverjanja označevanja (tehnična specifikacija) in pregledati način priprave izvorne baze.
    - Predstavljena zasnova orodja

- **A3.3**
    - Vnaprej izračunane kontekstne vložitve tipa BERT (**M0-M13, Robnik-Šikonja**) in Orodje za izračun kontekstnih vložitev (**M0-M13, Robnik-Šikonja**)
        - KONČANO (SloBERTa 1.0 in SloBERTa 2.0)
    - Orodje za razdvoumljanje (**M0-M31, Robnik-Šikonja**)
        - / (čaka se na podatke - cca. 2000 stavkov)
        - znotraj te množice bodo označene tudi imenske entitete (hkrati z nalogo razdvoumljanja)
        - @Krek: pregled projekta in rezultatov (do sept. 2021) - Franček
    - Semantični premiki, diahrone analize, testna množica (**M9-M31, Pollak**)
        - Seznam za evalvacijo na Redmine (Polona Gantar)?
            - -> seznam pridobljen, potrebno zgraditi validacijsko/testno množico
        - Preskušen SloBERTa 2.0
            - TODO: kako bi se odkrilo besede, ki spreminjajo pomen skozi čas?
        - V jeseni več aktivnosti ...

- **A3.4**
    - Orodje za avtomatsko ekstraktivno in abstraktivno povzemanje krajših in daljših besedil (**M10-M31, Robnik-Šikonja**)
        - Gradnja pristopa za daljša besedila in delo na segmentaciji. Izdelani osnovni modeli.
    - Orodje za avtomatsko odgovarjanje na vprašanja (**M10-M31, Ojsteršek**)
        - Začetek dela: Junija 2021
        - Prva naloga:
            - @Ojsteršek: Priprava končnega seznama virov in določitev obsega dodatnih odgovorov na vprašanja
                - -> do naslednjega sestanka (Sandi Majniger) se pošlje obseg in vire za pridobitev QA

        - CoPA, BoolQ, MultiRC, ReCORD - naloge SuperGLUE kot izhodišče
        - Definirati sistem:
            - vhod: besedilo + vprašanje
            - izhod: odgovor

    - **A3.5**
        - Orodje za delo z evalvacijskimi nalogami (**M0-M13, Robnik-Šikonja**)
            - KONČANO (Orodje ;;; Navodila NER, [REL], COREF)
            - Aktivnosti/izboljšave še potekajo (naloga sicer traja do konca projekta).
                - SuperGLUE se še dalje čisti + testiranje modelov za posamezne naloge - dolgoročna aktivnost (tudi izzven RSDO)

    - **Druge aktivnosti:**
        - Programski vmesniki za orodja (Q3-4 2021)
            - https://docs.google.com/document/d/1KinWqrF9vMyPaanQT5PV48qcBw44-ddcqJdLC-bSUSg/edit#
                - **@Slavko**: uskladiti predloge, pripraviti zaslonske maske in arhitekturo za zajemanje rezultatov (slednje poenoteno za vse sklope).

- **Ad3:**
    - Naslednji sestanek bo **14. september 2021 ob 9:00.**
    - Primer formata za objavo virov: Multiword Expressions lexicon extracted from the Gigafida 2.1 corpus, http://hdl.handle.net/11356/1421

Zadolžitve:

- (**Milan Ojsteršek**) - pripravi dokument in orodja za predprocesiranje različnih tipov vhodov (PDF, ...) v formate, primerne za nadaljnje obdelave.
    - Dostopno na: https://git.lhrs.feri.um.si/ivan.kovacic/textapi
    - TODO: še dodati stvari, ki so implementirane, nato se predstavi ...
- (**Milan Ojsteršek**) Priprava končnega seznama virov in določitev obsega dodatnih odgovorov na vprašanja do sestanka RSDO DS3 v juliju.
- **(Slavko Žitnik)**: uskladiti predloge, pripraviti zaslonske maske in arhitekturo za zajemanje rezultatov (slednje poenoteno za vse sklope).
- (**Simon Krek**) Pregled projekta Franček glede možnosti pridobitve dodatnih virov (naj bi bili objavljeni do sept. 2021) in uporaba v okviru RSDO.
    - -> pogledamo septembra
-
-