RAPORTARE ŞTIINŢIFICĂ

Proiect complex ReTeRom. Proiect component CoBiLiRo

Activitatea 3.3 - ***Realizarea de statistici privind corpusul bimodal voce/text***

Faza de predare: noiembrie 2020

Autori: Pistol Ionuţ, Scutelnicu Andrei, Serban Boghiu, Felix Cristian Pericică

1. Rezumatul etapei

În această etapă a proiectului complex ReTeRom consorţiul şi-a propus să consolideze şi apoi să exploateze rezultatele acumulate în primii doi ani, cu obiectivele: existenţa unui Portal pregătit a primi şi prelucra resurse bilingve româneşti, dezvoltarea în continuare a unei colecţii de resurse care să corespundă formatului agreat de consorţiu, perfecţionarea lanţurilor de prelucrări lingvistice şi sonore, atât asupra componentelor textuale cât şi vocale ale resurselor bimodale, care să permită alinieri între componentele vocale şi textuale, recunoaşterea cu minimum de erori a vocii, generarea expresivă a vocii şi antamarea de aplicaţii bazate pe aceste tehnologii.

A treia etapă (2020) a proiectului CoBiLiRO prevede completarea inventarului de resurse disponibile pe portal cât şi valorificarea lor atât în cadrul platformei (statistici şi instrumente integrate) precum şi în afara platformei, propunând o serie de proiecte ce utilizează tehnologiile dezvoltate în cadrul proiectelor partenere. Similar celorlalte etape, este prevăzută şi o activitate de diseminare, atât la evenimente ştiinţifice cât şi în mass-media. De interes special pentru platforma dezvoltată este respectarea drepturilor de autor şi a anonimizării solicitate pentru contribuitorii de resurse pe platformă.

2. Rezumatul activităţii

Activitatea 3.3 are ca obiectiv dezvoltarea platformei CoBiLiRo prin oferirea unor statistici privind resursele disponibile, atât la nivel individual cât şi la nivelul întregului corpus. Statisticile prezintă, la cererea utilizatorului, parametri calculaţi din exploatarea informaţiilor conţinute în fişierele text şi audio, în alinierile dintre cele două, cât şi în metadatele aferente.

3. Descrierea ştiinţifică şi tehnică

3.1 Scurtă descriere a tehnologiilor relevante

Componenta de procesare lingvistică necesară obţinerii de valori statistice, este realizată sub forma unui script Python, încadrat într-un web API, care poate fi accesat folosind principiile arhitecturii Restful, apelat doar la nivel intern (pe server); aici s-a folosit tehnologia **Flask Restful**.

Pentru realizarea prelucrării şi procesării la nivel de limbaj natural a textelor a fost utilizată librăria open-source **Spacy**, mai exact modelul **ro_core_news_sm**; acest model atribuie vectori de tokeni specifici contextului şi etichete ale părţilor de vorbire având o acurateţe de 95.62%. Ca alternativă poate fi folosită componenta de analiză a textului produsă de TEPROLIN; motivul pentru care nu a fost inclusă în versiunea iniţială este că nu toate fişierele de pe platformă au fost procesate de acea componentă. Pentru finalul proiectului avem în plan tranziţia spre tehnologia TEPROLIN.

3.2 Statistici la nivelul întregii colecţii de resurse

La nivelul întregii colecţii de resurse se pot aplica mai multe tipuri de filtre, ele întorcând valori numerice rezultate din numărarea resurselor pentru care filtrele întorc valori *true*. Utilizatorul poate afla numele fişierelor pentru care filtrele dau valoarea *true*:

- Pe tipul de segmentare a fişierelor audio:
  - **“File”** - un fişier audio corespunde unei zone din text. Alinierea se face prin marcajul *file* ce indică fişierul audio corespunzător unei zone din text.
  - **“Start-Stop”** - un singur fişier audio corespunde întregului text. Alinierea se face prin marcaje *start* şi *stop* ce indică momentul de început şi de sfârşit din înregistrarea conţinută în fişierul audio ce corespunde unei zone din text.
  - **“File-Start-Stop” -** mai multe fişiere audio corespund întregului text. Alinierea se face prin marcajul *file* ce indică fişierul audio corespunzător şi prin marcajele *start* şi *stop* ce indică momentul de început şi de sfârşit din înregistrarea conţinută în fişierul audio ce corespunde unei zone din text.
- Pe tipul de fişier audio disponibil:
  - **“wav”** (Waveform Audio File Format) - calitate foarte bună dar dimensiune mare.
  - **“mp3”** (MPEG Audio Layer 3) - calitate medie, dimensiune scăzută.
- Pe nivelul de adnotare:
  - **“Propoziţie”** - alinierea dintre audio şi text se face la nivel de propoziţie.
  - **“Cuvânt”** - alinierea dintre audio şi text se face la nivel de cuvânt.
  - **“Fonem”** - alinierea dintre audio şi text se face la nivel de fonem.
  - **“Prozodic”** - alinierea dintre audio şi text marchează pe text adnotările de prozodie din semnalul sonor.

Pe lângă filtrele de mai sus platforma include şi statistici ce indică numărul de resurse disponibile pentru fiecare tip menţionat mai sus. De asemenea este calculată şi durata totală a fişierelor audio de fiecare tip, folosind *ffprobe*[1] - un multimedia stream analyzer.

Fig.1: Statistici generale

3.3 Statistici la nivelul fişierelor text

Înainte de a afişa statisticile pe interfaţă, fişierele text asociate resurselor sunt prelucrate şi procesate pentru a obţine rând pe rând următoarele componente:

- **Densitatea lexicală** a textului transcris asociat resursei - după procesul de tokenizare, se aplică etichetarea părţilor de vorbire având grijă ca tokenii să fie clasificaţi în două mari grupe (cuvinte lexicale şi cuvinte funcţionale); rezultatul final este dat de factorii din grupa tokenilor lexicali, prin formula: *nr. de cuvinte lexicale / nr. total de cuvinte * 100*;
- **Raportul type/token** - după ce se calculează numărul total de cuvinte, se grupează cele care apar o singură dată, în forma flexionară; rezultatul final este dat de formula: *nr. cuvintelor care apar o dată / nr. total de cuvinte * 100*;
- **Numărul de tokeni**;
- **Numărul de lemme** - se aplică procesul de lematizare şi o lemă se ia în considerare doar o singură dată;
- **Numărul formelor flexionare distincte** - se aplică o procedură similară ca şi în cazul numărului de leme;
- **Media formelor flexionare distincte** - este dată de formula *nr. total de forme flexionare distincte / nr. total de tokeni*;
- **Media de lemme din text** - componentă ce returnează rezultatul calculului: *nr. de lemme distincte / nr. total de tokeni* ; se iau în considerare toate lemmele din text, numărate o singură dată;

[1] https://ffmpeg.org/ffprobe.html

- **Numărul de lemme care apar de mai multe ori** - după ce se aplică procesul de lematizare, se grupează lemmele în două categorii: cele care apar o singură dată (*hapax legomenon*) şi cele care apar de mai multe ori; este luată în considerare ultima grupă;
- **Raportul de lemme care apar de mai multe ori** - redat prin formula *nr. de lemme care apar de mai multe ori / nr. total de lemme luate în considerare o singură dată * 100*. În abordarea prelucrărilor lingvistice ulterioare se va avea în vedere şi inventarierea lemmelor unicate, urmând ca pe interfaţă să fie afişată o nouă componentă - *hapax legomenon*.

Toate fişierele sunt curăţate în prealabil de semnele de punctuaţie şi eventualele spaţii şi/sau simboluri nejustificate.

3.4 Statistici la nivelul fişierelor audio

Din cauza numărului mare de fişiere disponibile în platformă şi a dimensiunii ridicate a acestora, taskul care calculează durata fişierelor a trebuit paralelizat (folosind librăria *TPL*[2]) şi totodată aplicată o *tehnică de caching*, deoarece calculul total al duratelor este un proces foarte costisitor. Prin urmare, am decis utilizarea unui *Background Job Worker* (folosind librăria *HangFire*[3]), care recalculează zilnic durata totală a fişierelor încărcate pe platformă.

Pentru a putea accesa metadatele interne ale înregistrărilor şi a afla durata acestora (mp3, mpeg, mp4, wav), a fost folosit analizatorul de stream-uri *ffprobe*. Cu ajutorul lui, am reuşit să extragem durata fiecărui fişiere sunet.

Una dintre provocările întâlnite a fost configurarea lui pe serverul CentOS şi integrarea lui în mediul de dezvoltare .NET Core. A fost necesară modelarea şi implementarea unor optimizări pentru a face faţă problemelor generate de memoria aflată la dispoziţie şi de numărul maxim de fişiere ce puteau fi deschise simultan.

3.5 Statistici la nivelul alinierilor

Pentru alinierile la nivel de cuvânt, cele mai relevante în majoritatea aplicaţiilor care valorifică alinieri (antrenarea *text-to-speech* şi *speech-to-text*, aliniatoare automate), am pregătit un set de statistici ce pot fi generate automat. Scopul principal al acestora este oferirea unei priviri de ansamblu asupra calităţii alinierilor disponibile. Trebuie remarcat faptul ca această calitate depinde nu numai de procesul de aliniere (fie automată sau manuală) cât şi de calitatea versiunilor paralele text şi voce. Pentru o discuţie mai detaliată despre posibilelor probleme, a se vedea raportul A3.4 *Proiectare de aplicaţii de exploatare a corpusului bimodal şi a tehnologiilor de prelucrări textuale şi voce, create în proiectele P2, P3, P4*.

---

[2] https://docs.microsoft.com/en-us/dotnet/standard/parallel-programming/task-parallel-library-tpl

[3] https://www.hangfire.io

File	AutoText	AlignedText	LongestUnmatch	StartsAt	SingleUnmatched
Alma-Mater-Iassiensi-24-Facultatea-de-Drept	5379	3761	44	30.51	193
Alma-Mater-Iassiensis-1-Aniversarile-Universitatii	7078	4400	121	2944.26	248
Alma-Mater-Iassiensis-11-Prof-Univ-Dr-Andrei-Marga	7072	4947	65	2688.87	266
Alma-Mater-Iassiensis-12-Facultatea-de-Fizica	5232	3312	52	3282.42	166
Alma-Mater-Iassiensis-13-Universitatea-din-Iasi-prezent-si-perspective	7006	5220	61	2109.15	281
Alma-Mater-Iassiensis-14-Universitatea-din-Iasi-Dimensiunea-culturala	6301	4558	41	2090.97	265

Fig.2 : Statistici pe alinieri la nivel de cuvânt

În scopul obţinerii acestor statistici a fost utilizată alinierea produsă de TADARAV ce aliniază textul extras şi aliniat automat de tehnologia TADARAV cu transcrierea manuală disponibilă pentru resursa respectivă [1].

În figura 2 semnificaţia coloanelor este:

- *AutoText* - dimensiunea (în cuvinte) a textului extras automat din fişierul audio;
- *AlignedText* - dimensiunea (în cuvinte) a textului aliniat cu transcrierea originală;
- *LongestUnmatch* - cea mai lungă secvenţă de cuvinte din transcrierea originală pentru care nu există nici o aliniere;
- *StartsAt* - indexul cuvântului de la care începe secvenţa de mai sus;
- *SingleUnmatched* - numărul de cuvinte din transcrierea originală care nu au fost aliniate, deşi cuvântul imediat precendent şi cel imediat ulterior au fost aliniate.

O serie de observaţii pot fi făcute pe aceste statistici, de exemplu se observă că procentul cuvintelor aliniate corect (pe cele 6 fişiere din figura 2) este de aproximativ 70%. Se observă că în fişierul al doilea există o zonă lungă de cuvinte nealiniate (121 cuvinte) ce poate indica prezenţa în transcrierea manuală a unui fragment care nu se regăseşte şi în înregistrarea audio. O altă observaţie ce ar putea fi făcută pe aceste statistici ar fi că în fişierul 5 există un număr relativ mare de cuvinte izolate nealiniate (281) ceea ce ar putea indica o calitate scăzută a semnalului audio sau un accent/prozodie nestandard a vorbitorului.

Aceste statistici sunt implementate şi în curs de integrare cu platforma online, versiunea finală a platformei, ce va fi predată în aprilie 2021, va include şi aceste statistici.

3.6 Concluzii

Oferirea unor statistici relevante pentru o colecţie de resurse contribuie atât la mărirea încrederii unor potenţiali noi contributori cât şi la potenţialul resurselor din colecţie de a-şi găsi utilitatea în diverse contexte. Statisticile deja incluse pe platformă precum şi cele care vor fi incluse până în aprilie 2021 promit să asigure acest avantaj pentru colecţia de resurse bimodale, principalul rezultat al proiectului CoBiLiRO.

Până la terminarea proiectului avem în vedere ca procesele de determinare a lungimii fişierelor să fie lansate doar asupra fişierelor nou introduse sau updatate. Serverul e capabil să urmărească fişierele asupra cărora s-a umblat şi să calculeze/recalculeze lungimile doar pentru acestea. Odată aflată o lungime, ea se va înscrie în metadate, ceea ce înseamnă că nu va trebui recalculată la fiecare cerere de lungimi din partea utilizatorului.

Toate obiectivele incluse în plan la această activitate au fost realizate.

Bibliografie

[1] C. Burileanu, D. Burileanu, H. Cucu (2019). Raportul aferent activităţii A2.11 *Proiectarea şi implementarea unei soluţii de bază de filtrare şi aliniere a transcrierilor aproximative cu semnalul de vorbire*, proiectul RETEROM