RAPORTARE ȘTIINȚIFICĂ

Proiect complex ReTeRom. Proiect component CoBiLiRo

Activitate 3.2 - Augmentarea corpusului voce-text: completare de metadate, alinieri prin rularea algoritmilor dezvoltați în proiectele P2, P3 și P4 și adnotări manuale și semiautomate la corpusul voce/text

Deadline: Noiembrie 2020
Responsabili: Andrei Scutelnicu, Serban Boghiu, Felix Cristian Pericică

1. Rezumatul etapei

În această etapă a proiectului complex ReTeRom consorțiul și-a propus să consolideze și apoi să exploateze rezultatele acumulate în primii doi ani, cu obiectivele: existența unui Portal pregătit a primi și prelucra resurse bilingve românești, dezvoltarea în continuare a unei colecții de resurse care să corespundă formatului agreat de consorțiu, perfecționarea lanțurilor de prelucrări lingvistice și sonore, atât asupra componentelor textuale cât și vocale ale resurselor bimodale, care să permită alinieri între componentele vocale și textuale, recunoașterea cu minimum de erori a vocii, generarea expresivă a vocii și antamarea de aplicații bazate pe aceste tehnologii.

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

2. Rezumatul activității

Activitatea 3.2 a avut ca obiectiv creșterea în dimensiune a corpusului bimodal acumulat în Portalul CoBiLiRo până la sfârșitul etapei precedente și realizarea automată a alinierilor prin utilizarea tehnologiilor rezultate în proiectele partenerilor.

3. Descrierea științifică și tehnică

Tehnologiile folosite în cadrul acestei etape au fost cele dezvoltate de către partenerii din proiect și anume, în cadrul proiectului TEPROLIN - au fost realizate o serie de servicii care realizează diverse prelucrări asupra resurselor de tip text (mai multe detalii în rapoartele A1.5 și A1.6) și în cadrul proiectului TADARAV - care a avut ca obiectiv principal realizarea și dezvoltarea unui serviciu de aliniere automată a unei resurse de tipul text-voce (descris în rapoartele A1.13 și A2.11).

Prin urmare, în cadrul platformei CoBiLiRo, odată cu adăugarea unei resurse, fișierul text poate fi trimis la endpoint-ul de procesare TEPROLIN.

TEPROLIN oferă în acest moment peste 15 flow-uri de procesare a textului pentru limba română:

- text-normalization
- diacritics-restoration
- word-hyphenation
- word-stress-identification
- word-phonetic-transcription
- numeral-rewriting
- abbreviation-rewriting
- sentence-splitting
- tokenization
- pos-tagging
- lemmatization
- named-entity-recognition
- biomedical-named-entity-recognition
- chunking
- dependency-parsing

Utilizatorul poate selecta anumite *flow-uri* prin care fișierele text ale resursei urmează să treacă, rezultatele procesării urmând să fie salvate la nivel de resursă. Un exemplu de rezultat este prezentat în Anexa 1. Aceste rezultate, codificate ca fișiere json, pot fi atașate componentelor textuale ale resurselor, la cererea utilizatorilor.

Folosind serviciile oferite de proiectul TADARAV [1], au fost trecute prin procesare următoarele colecții: "*Alma Mater"*, "*Ghici Cine (Mai) Vine La Cina?"*, "*Romania 100 Iasul in arcul timpului"*. Folosind aceste servicii, au rezultat două tipuri de procesări:

- Un prim proces de aliniere dintre fișierul audio și text (Figura 1), în care sunt marcate timpii de început și sfârșit al fiecărui cuvânt.

```
Alma-Mater-Iassiensi-24-Facultatea-de-Drept:
 a(0.36,18.69)
 stimați(20.4,20.85)
 telespectatori(20.85,21.51)
 bine(21.51,21.78)
 v-am(21.78,21.96)
 regăsit(21.96,22.38)
 la(22.38,22.5)
 o(22.5,22.59)
 nouă(22.59,22.92)
 emisiune(22.95,23.49)
 alma(23.52,23.79)
 mater(23.79,24.15)
 probabil(24.96,25.32)
 că(25.32,25.5)
 mulți(25.5,25.83)
 dintre(25.83,26.07)
 dumneavoastră(26.07,26.46)
 s-au(26.46,26.64)
 obișnuit(26.64,27.21)
 a(27.87,27.96)
 trecut(27.96,28.26)
 aproape(28.26,28.56)
 o(28.56,28.59)
 jumătate(28.59,29.01)
 de(29.01,29.16)
 an(29.16,29.28)
 de(29.28,29.37)
 când(29.37,29.73)
```

Figura 1 - Fragment rezultat în urma procesului de aliniere dintre fișierul audio și text

- Cel de-al doilea tip de procesare în care au fost marcate *timestamp*-uri aferente secvențelor de cuvinte. Formatul este după cum urmează: câte un fișier text pentru fiecare resursă audio. Pe fiecare linie este câte o propoziție (cu *timestamps* la nivel de cuvânt într-o variantă; fără *timestamps* în cealaltă variantă - după cum se poate observa în Figura 2). În cadrul lunilor de implementare rămase aferente proiectului se va realiza prelucrarea acestui rezultat oferit de serviciul TADARAV, pentru a aduce acest rezultat la formatul CoBiLiRo convenit (livrabilul A3.1).

```
<s> pentru orice român pentru orice străin care vine </s> (Alma-Mater-Iassiensis-1-Aniversarile-Universitatii_00)
<s> instituţii care dau </s> (Alma-Mater-Iassiensis-1-Aniversarile-Universitatii_01)
<s> oraşului nostru este vorba de mitropolie </s> (Alma-Mater-Iassiensis-1-Aniversarile-Universitatii_02)
<s> vorba de palatul administrativ </s> (Alma-Mater-Iassiensis-1-Aniversarile-Universitatii_03)
<s> palatul culturii </s> (Alma-Mater-Iassiensis-1-Aniversarile-Universitatii_04)
<s> este vorba </s> (Alma-Mater-Iassiensis-1-Aniversarile-Universitatii_05)
<s> universitatea din iaşi </s> (Alma-Mater-Iassiensis-1-Aniversarile-Universitatii_06)
<s> începe un nou ciclu de emisiuni la tvr iaşi </s> (Alma-Mater-Iassiensis-1-Aniversarile-Universitatii_07)
```

Figura 2 - Secvență de fișier rezultat în urma procesării cu serviciile oferite de proiectul TADARAV

Datorită unor probleme tehnice întâmpinate pe serverul care găzduiește Portalul CoBiLiRo, au fost pierdute din baza de date o serie de înregistrări ale vechilor resurse. Din acest motiv a trebui să implementăm un job care să parcurgă recursiv folderele unde erau ținute back-up-urile anumitor resurse, pentru a reintroduce automat acele resurse în baza de date, astfel încât

ele să fie vizibile din nou pe platformă. Jobul creat caută toate fișierele audio, text și XML aflate la nivel de resursă. Pentru a recupera metadatele parcurse, a trebuit parsat fișierul XML [2].

În exemplele de cod din Figura 3, sunt introduse id-urile resurselor care există pe hard disk, dar care nu figurează actualmente cu înregistrări în baza de date. Prin urmare, este parcurs fiecare folder cu id-ul specific, recreată resursa și apoi salvată în baza de date.

```
var resourcesPathsAndIds :IEnumerable<(resourceId, folderPath)> = missingResourceIds.Select(
    resourceId :string =>  new
    {
        resourceId = resourceId,
        folderPath = _hostingEnvironment.WebRootPath + Path.DirectorySeparatorChar + "Uploads" + Path.DirectorySeparatorChar + resourceId
    }
    );

foreach (var item :(resourceId, folderPath) in resourcesPathsAndIds)
{
    var xmlFolderPath :string = item.folderPath + Path.DirectorySeparatorChar;
    var xmlFilePath :string = Directory
        .GetFiles(xmlFolderPath, searchPattern: "*.xml*", SearchOption.AllDirectories).FirstOrDefault();
    if (xmlFilePath != null)
    {
        var xmlFile :string = System.IO.File.ReadAllText(xmlFilePath);
        resources.Add(item: new Resource
        {
            ResourceId = Guid.Parse(item.resourceId),
            ApplicationUserId = user.Id,
            ApplicationUser = user,
            UploadPath = "",
            UploadedDateTime = DateTime.Now,
            Collection = GetAttribute(xmlFile, attribute: "Collection") ?? "",
            Title = GetAttribute(xmlFile, attribute: "Title") ?? "",
            Ident = GetAttribute(xmlFile, attribute: "Ident") ?? "",
            LongTitle = GetAttribute(xmlFile, attribute: "LongTitle") ?? "",
            Description = GetAttribute(xmlFile, attribute: "Description") ?? "",
            Keywords = GetAttribute(xmlFile, attribute: "Keywords") ?? "",
            Language = GetAttribute(xmlFile, attribute: "Language") ?? "",
            MetadataCreator = GetAttribute(xmlFile, attribute: "MetadataCreator") ?? "",
            AnnotationCreator = GetAttribute(xmlFile, attribute: "AnnotationCreator") ?? "",
            AnnotationLevel = GetAttribute(xmlFile, attribute: "AnnotationLevel") ?? "",
            SpeechCreator = GetAttribute(xmlFile, attribute: "SpeechCreator") ?? "",
            AcusticMedia = GetAttribute(xmlFile, attribute: "AcusticMedia") ?? "",
            Duration = GetAttribute(xmlFile, attribute: "Duration") ?? "",
            SamplingFrequency = GetAttribute(xmlFile, attribute: "SamplingFrequency") ?? "",
            Rezolution = GetAttribute(xmlFile, attribute: "Rezolution") ?? "",
            RecordDate = GetAttribute(xmlFile, attribute: "RecordDate") ?? "",
            RecordTime = GetAttribute(xmlFile, attribute: "RecordTime") ?? "",
            Equipment = GetAttribute(xmlFile, attribute: "Equipment") ?? "",
            Broadcast = GetAttribute(xmlFile, attribute: "Broadcast") ?? "",
            SpeechFileType = GetAttribute(xmlFile, attribute: "SpeechFileType") ?? "",
            SpeechFile = GetAttribute(xmlFile, attribute: "SpeechFile") ?? "",
            SpeechSegmentation = GetAttribute(xmlFile, attribute: "SpeechSegmentation") ?? "",
            SpeakerName = GetAttribute(xmlFile, attribute: "SpeakerName") ?? "",
            SpeakerGender = GetAttribute(xmlFile, attribute: "SpeakerGender") ?? "",
            SpeakerAge = GetAttribute(xmlFile, attribute: "SpeakerAge") ?? "",
            SpeakerAccent = GetAttribute(xmlFile, attribute: "SpeakerAccent") ?? "",
            TextCreator = GetAttribute(xmlFile, attribute: "TextCreator") ?? "",
            TextFormat = GetAttribute(xmlFile, attribute: "TextFormat") ?? "",
            Distribution = GetAttribute(xmlFile, attribute: "Distribution")
        });
    }
}
```

```
29 references | Serban Paul Boghiu, 14 days ago | 1 author, 1 change | 0 exceptions
private string GetAttribute(string xmlFile, string attribute)
{
    var searchKey :string = attribute + "=";
    var keyPosition :int = xmlFile.IndexOf(searchKey);
    var subText :string = xmlFile.Substring(keyPosition + searchKey.Length + 1);
    var attributeLength :int = xmlFile.Substring(keyPosition + searchKey.Length + 1).IndexOf(@"""");
    var attrValue :string = subText.Substring(startIndex: 0, attributeLength);
    return attrValue;
}
```

Figura 3: Codul de restaure a resurselor pierdute din Portal

4. Concluzii

În această activitate au fost depuse eforturi de includere în corpus a resurselor identificate în activitatea A3.1, precum și de realizare automată a alinierilor dintre componentele sonore și textuale.

O parte din aceste resurse au fost introduse pe platformă atât automat folosind scripturi implementate în cadrul acestei activități, cât și manual, de către utilizatori prin încărcarea fișierelor audio, text și completarea în interfață cu informațiile aferente. În momentul de față pe server exista aproximativ 70 de mii de fișiere audio cu perechile textuale respective, ce conțin transcrieri ale înregistrărilor sonore, precum și fișierele corespunzătoare cu metadate.

Pe lângă fișierele existente și prezentate și în livrabilul A3.1, urmează să se încarce pe platformă o serie de resurse audio ce au fost obținute în ultimele săptămâni și la care se lucrează la partea de transcriere a textelor și la alinieri.

Aceste resurse noi sunt în curs de integrare în platforma online. Până la finalul proiectului, aprilie 2021, platforma va include toate resursele bimodale completate cu metadate.

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

Bibliografie

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

[2] Pistol. I., Pădurariu C., Boghiu Ș., Scutelnicu A., Raport Activitate A2.2: *Raport asupra soluțiilor de armonizare a reprezentărilor colecțiilor existente text /vorbire (metadate și adnotări)*.

Anexa 1: Extras dintr-un rezultat al unui proces TEPROLIN


```
{
    "teprolin-conf": {
        "abbreviation-rewriting": "expander-utcluj",
        "biomedical-named-entity-recognition": "bioner-icia",
        "chunking": "ttl-icia",
        "dependency-parsing": "nlp-cube-adobe",
        "diacritics-restoration": "diac-restore-icia",
        "lemmatization": "ttl-icia",
        "named-entity-recognition": "ner-icia",
        "numeral-rewriting": "expander-utcluj",
        "pos-tagging": "ttl-icia",
        "sentence-splitting": "ttl-icia",
        "text-normalization": "tnorm-icia",
        "tokenization": "ttl-icia",
        "word-hyphenation": "tts-utcluj",
        "word-phonetic-transcription": "tts-utcluj",
        "word-stress-identification": "tts-utcluj"
    },
    "teprolin-result": {
        "sentences": [
            "Fiscul va face verificări la firmele indicate de CNSP, iar pe zona de dezvoltare va acorda granturi, precum cele pentru primării.",
            "Diabetul zaharat este un sindrom caracterizat prin valori crescute ale concentrației glucozei în sânge (hiperglicemie) și dezechilibrarea metabolismului."
        ],
        "text": "Fiscul va face verificări la firmele indicate de CNSP, iar pe zona de dezvoltare va acorda granturi, precum cele pentru primării.\nDiabetul zaharat este un sindrom caracterizat prin valori crescute ale concentrației glucozei în sânge (hiperglicemie) și dezechilibrarea metabolismului.",
        "tokenized": [
            [
                {
                    "_bner": "",
                    "_chunk": "Np#1",
                    "_ctg": "NSRY",
                    "_deprel": "nsubj",
                    "_expand": "",
                    "_head": 3,
                    "_id": 1,
                    "_lemma": "fisc",
                    "_msd": "Ncmsry",
                    "_ner": "ORG",
                    "_phon": "f.i.s.k.u.l",
                    "_syll": "f'is.cul",
                    "_wordform": "Fiscul",
                    "upos": "NOUN",
                    "ner": "ORG"
```

```
                },
                {
                    "_bner": "",
                    "_chunk": "Vp#1",
                    "_ctg": "VA3S",
                    "_deprel": "aux",
                    "_expand": "",
                    "_head": 3,
                    "_id": 2,
                    "_lemma": "vrea",
                    "_msd": "Va--3s",
                    "_ner": "",
                    "_phon": "v.a",
                    "_syll": "va",
                    "_wordform": "va",
                    "upos": "AUX",
                    "ner": ""
                },
            ...
            ]
        ]
    }
}
```