ПРОЈЕКАТ

РАДИОНИЦА О ПРОЦЕНИ КВАЛИТЕТА И ПРЕЧИШЋАВАЊУ ТЕКСТОВА У ОКВИРУ ПРОЈЕКТА „ЕВРОПСКЕ НОВИНЕ“

Александра Тртовац, aleksandra@unilib.bg.ac.rs
Универзитет у Београду, Универзитетска библиотека „Светозар Марковић“, Београд

Универзитетска библиотека „Светозар Марковић“ у Београду један је од 18 пуноправних партнера у пројекту „Европске новине“ у коме учествују најзначајније националне и универзитетске библиотеке Европе, поред организације LIBER и немачке фирме CCS[1], као и 11 придружених партнера. Пројектом ће корисницима Еуропеане бити омогућен приступ до пуног текста, претраживање и проналажење информација на преко 18 милиона дигитализованих страна новина. Пројекат „Европске новине“ започео је 2012. и траје до 2014. године.

У оквиру пројекта, у Универзитетској библиотеци у Београду је 13. и 14. јуна 2013. године одржана радионица о процени квалитета и пречишћавању текстова дигитализованих новина. На радионици је учешће узело педесетак библиотекара и информатичких стручњака из разних делова Европе (од Финске, Естоније, Исланда, Данске, преко Француске, Велике Британије, Немачке, Холандије, Швајцарске, Шпаније, Пољске, Чешке до Словеније, Хрватске, Босне и Херцеговине, Бугарске и Македоније). Учесници радионице су, поред представника партнерских библиотека, биле и колеге из Народне библиотеке Србије, Музеја Николе

1 Content Conversion Specialists

Тесле, Универзитетске библиотеке у Нишу.

После поздравног говора проф. др Александра Јеркова, директора Универзитетске библиотеке „Светозар Марковић“ и основних информација о програму радионице које је изнела Марике Вилмс, главна организаторка радионице и представница LIBER-а, имали смо прилику да чујемо излагања и корисне сугестије колега који у оквиру пројекта раде на развоју апликација, тестирању и анализирању достављених дигитализованих садржаја. Такође, активно смо учествовали у разговорима и дискусијама у малим групама и остављали своје коментаре на тзв. „демократском зиду“ где је Марике Вилмс поставила неколико тема: „Открио сам да...“, „Приметио сам да...“, „Осећам да...“, „Научио сам да...“, „Предложио бих...“

Ханс-Јерг Лидер из Државне библиотеке у Берлину, која је и координатор пројекта, говорио је најпре о технологијама које се користе приликом процене квалитета и пречишћавања новинских текстова – оптичко препознавање карактера (OCR[2]), оптичка сегментација чланака (OLR[3]), препознавање именованих ентитета (NER[4]), препознавање класе странице. Такође, истекао је важност пројекта у смислу метаподатака који се додељују новинским чланцима – поред претраге у пуном тексту, могуће је претраживање и преко метаподатака од којих је посебно занимљиво препознавање текстова који описују фотографије којима су илустровани новински чланци. Такође, Лидер је говорио и о изазовима и потешкоћама на које се наилази приликом пречишћавања текстова старих новинских страница. У првом реду, старе новине су штампане отискивањем неквалитетног мастила, па је оптичко препознавање карактера тиме отежано. Предности за кориснике Европеане, али и за партнере у пројекту су и могућност претраге преко кључних речи и фразног претраживања, претраживање слика и информација о њима, значење текста, укључивање корисника у процес кориговања, али и богаћења садржаја, приступ садржајима и преко апликација за мобилне телефоне...

Клеменс Нојдекер из Националне библиотеке Холандије у Хагу говорио је о процесима пречишћавања текста новина и развоју алата за бољу организацију података које библиотеке достављају. Два партнерске институције раде на овим пословима – Универзитет у Инсбруку и немачка фирма CCS. Кораци који прате процес пречишћавања текста су: одабир новина, постављање услова везаних за ауторска права, унос података у колекцију метаподатака (мастер листа), бинаризација, преименовање докумената и структура фолдера, завршна провера података и метаподатака, OCR/OLR, производња фајла са метаподацима у формату METS[5], и на крају NER. Проблеми који се у овим процесима јављају су углавном везани за различите технике дигитализације и приступ садржају, различите формате докумената, фонтове, језике, писма. Алати који су развијени за што боље пречишћавање текста и организацију података су: BCT – Binarisation and Colour Reduction Tool, FRT – File Rename Tool, FAT – File Analyzer Tool. За оптичко препознавање карактера користи се ABBYY FineReader SDK и софтвер State-of-the-Art OCR, за оптичку сегментацију чланака технологија docWorks, а за NER технологија Stanford CRF-NER која подржава немачки, енглески и холандски језик са додатком француског и литванског за именоване ентитете особа, локација и

2 Optical Character Recognition
3 Optical Layout Recognition
4 Named Entity Recognition

5 *Metadata* Encoding and Transmission Standard

организација.

Штефан Плечахер и Кристијан Клауснер са Универзитета у Салфорду говорили су о потешкоћама приликом препознавања каратера у старим новинама. Процена квалитета је важна због употребе специфичних фонтова која узрокује грешке у текстовима дигиталних верзија новина. Да би се превазишли ови проблеми, потребно је укључити велики број спољних сарадника који би поправке радили на основу штампаног текста.

Након ових предавања и анализа, направљене су мале групе у којима су представници различитих установа имали прилику да се ближе упознају, разговарају и размене искуства везана за организацију послова у пројекту, дигитализацију новина и дају предлоге за унапређење пројекта.

Први дан радионице окончан је посетом Музеју Николе Тесле и заједничком вечером свих учесника.

Другог дана радионица је започета презентацијом Лоте Вилмс из Националне библиотеке Холандије о препознавању именованих ентитета. Она је објаснила да је пројектом предвиђено препознавање именованих ентитета за особе, локације и организације и дала примере за наведене именоване ентитете. У следећем, практичном делу радионице, учесници су, подељени у две групе, пратили презентацију софтвера Named Entity Attestation Tool. Софтвер је једноставан за употребу и омогућава лако уређивање препознатих ентитета, али је мали проблем био тај што учесници радионице не познају холандски језик, па нису били у могућности активније да учествују. За сада је овај алат доступан за препознавање именованих ентитета на немачком, холандском и француском језику.

У току радионице проф. др Цветана Крстев и проф. др Душко Витас, представници Групе за језичке технологије Математичког факултета Универзитета у Београду, договорили су са координатором пројекта, Хансом-Јергом Лидером, да се на обрађеним текстовима српских новина уради и препознавање именованих ентитета за српски језик.

Потом је Клаус Гравенхорст из фирме CCS представио резултате везане за оптичку сегментацију чланака и препознавање метаподатака о њима путем софтвера docWorks. Да би се постигли најбољи резултати потребно је структурирати претходно неструктуриране текстове. Софтвер препознаје речи, колоне са текстом, текстуалне блокове, рубрике, илустрације, рекламе и огласе, табеле и следеће типове страница: насловну страну, садржај, странице са илустрацијама (оне које садрже бар једну илустрацију), као и странице које садрже искључиво рекламе и огласе. Такође, могуће је препознавање наставака чланака (уколико чланак почиње на једној, а завршава се на другој страни). Софтвер docWorks омогућава конверзију метаподатака о препознатим чланцима у формат METS XML.

На ову презентацију надовезао се Гинтер Милбергер са Универзитета у Инсбруку који је најпре говорио о форматима за метаподатке који се користе у пројекту, у првом реду METS и ALTO[6]. У фокусу су били структурални метаподаци и предложено је да структурални елементи буду: подручје наслова, заглавље, огласи и рекламе, илустрације, наслов рубрике, број стране... Такође, предлог је да се чланци разврставају према следећим типовима: вести, кратке вести, прикази књига, позоришних представа, читуље и некролози, рекламе и огласи, пословни огласи, временска прогноза, приче, песме... Учесници у пројекту су подстакнути да

6 Analyzed Layout and Text Object

дају предлоге за структуралне метаподатке имајући у виду потребе корисника, као и да анализирају формални садржај новина.

На овој веома корисној радионици су, осим нових сазнања о току пројекта и развоју апликација и алата, у лепој атмосфери склопљена бројна познанства са библиотекарима и информатичким стручњацима који су дошли из разних крајева Европе. Била је то лепа прилика да се научи нешто ново, али и размене искуства и отворе нове могућности. Као библиотекару Универзитетске библиотеке, изузетно ми је драго што смо, чини ми се, били изузетно успешни како у организацији пословног дела, тако и разноликог забавног садржаја за наше драге госте и колеге.[7]

Примљено: 16. VIII 2013.

7 Презентације и видео материјали са радионице доступни су на порталу пројекта "Европске новине" -http://www.europeana-newspapers.eu/focus-on-newspaper-refinement-quality-assessment-in-belgrade/ и на YouTube каналу Универзитетске библиотеке "Светозар Марковић" - http://www.youtube.com/user/UBSMBeograd