UDK: 004.9:811.16’374
811.16'322

i-Librarian – бесплатна онлајн библиотека за грађане Европе

Диман Карагиозов (diman@tetra.com), Tetracom Consulting
Анелиа Белогај (anelia@tetracom.com), Tetracom Consulting
Дан Кристеа (dcristea@info.uaic.ro), Универзитет Александру Јоан Куза
Светла Коева (svetla@dcl.bas.bg), Институт за бугарски језик,
Бугарска академија наука
Мађеј Огродничук (maciej.ogrodniczuk@ipipan.waw.pl),
Институт за рачунарство, Пољска академија наука
Поливиос Раксис (raxis@atlantisresearch.gr), Atlantis Consulting
Емила Стојанов (emil@tetracom.com), Tetracom Consulting
Кристина Вертан (cristina.vertan@uni-hamburg.de), Универзитет у Хамбургу

превод с енглеског **Јелена Бајић**

Апстракт
Појавом веба (WWW) као главног извора за дистрибуцију садржаја, уследила је поплава информација. Велики обим и разноврсност садржаја захтевају приступ који ће редефинисати начин на који се информације анализирају. Квантитативни метод обраде информација, који се ослања на алате за управљање садржајем омогућава структурну анализу. Изазов са којим се суочавамо је да узнапредујемо од процеса објављивања података како пристижу, до степена разумевања који одређује вредност садржаја.
Решење које представљамо инкорпорира технологије за обраду природних језика у процес управљања вишејезичним садржајима на вебу. i-Librarian је веб локација направљена уз помоћ софтверске платформе отвореног кода ATLAS. ATLAS лингвистичком платформом употпуњује компоненту за управљање садржајем, која је у виду софтвера као сервиса, а користи се за изградњу и вођење веб локација код којих је садржај у првом плану, као и за управљање њима. Ова платформа обогаћује садржај тих веб локација индикативним детаљима и умањује потребу да уредници документа класификују ручно, пошто се категоризација садржаја обавља аутоматски. Платформа подржава шест европских језика: бугарски, немачки, грчки, енглески, пољски и румунски.

i-Librarian је бесплатна онлајн библиотека која помаже ауторима, студентима, младим истраживачима, научницима, библиотекарима и руководиоцима да лако стварају, организују и објављују различите врсте докумената. Она омогућава корисницима да одржавају свој лични радни простор тако да у њему чувају, са другима деле и објављују различите врсте докумената који се аутоматски категоризују, резимирају и анотирају према важним речима, изразима и именима. Такође омогућава лако проналажење сличних докумената на разним језицима.

Кључне речи
Дигитална библиотека, ланци обраде језика, UIMA, систем управљања садржајем, вишејезичност, машинско превођење, резимирање, класификација

Увод

Појавом веба, настала је револуција у начину на који се рукује садржајем и на који се он пласира. То за последицу има чињеницу да су дигитални садржаји на различитим језицима постали широко доступни на Интернету, а њихова велика бројност и језичка разноврсност пружају прилику за усвајање нових метода и алата за стварање и дистрибуцију садржаја. Иако су недавно постигнута значајна побољшања у области управљања садржајима на вебу, још увек постоји све већа потреба за онлајн сервисима из домена садржаја у које су инкорпориране језичке технологије.

Постојећа софтверска решења и услуге, као што су Google Docs, Slingshot и Amazon примењују неке од лингвистичких механизама којима се посвећује пажња у платформи о којој ће бити речи. Најпопуларнији системи отвореног кода за управљање вишејезичним садржајима на мрежи (Joomla, Joom!Fish, TYPO3, Drupal)[1] нуде могућности за управљање вишејезичним садржајима које су на ниском нивоу, обезбеђујући механизме за изградњу вишејезичних веб локација. Доступне услуге су, међутим, уско оријентисане на задовољавање потреба веома специфичних циљних група, због чега остаје незадовољена све већа потреба за свеобухватним решењем за управљање вишејезичним садржајима у којем се посвећује пажња проблемима насталим због увећавања породице језика који се говоре у Европској унији.

Представићемо платформу отвореног кода за управљање садржајима и вишејезичну библиотеку коју та платформа покреће, као доказ да поменути концепт функционише. Овај рад има за циљ да докаже да људи, који читају веб локације које покреће наша платформа за управљање вишејезичним садржајима, могу лако да проналазе документе сређене уз помоћ аутоматске класификације, као и садржаје који зависе од контекста, лоцирају сличне документе у обимној збирци вишејезичних података и добију кратке резимее на различитим језицима који помажу корисницима да јасније него икада пре препознају информације од кључне важности.

1. Онлајн библиотека

i-Librarian је бесплатна онлајн библиотека која помаже ауторима, студентима, младим истраживачима, научницима, библиотекарима и руководиоцима да лако стварају, организују и објављују различите врсте докумената и да их деле са другима без икаквих трошкова.

i-Librarian умањује потребу да уредници документа класификују ручно, пошто се класификација издања, вести, докумената обавља

1 http://www.joomla.org/, http://www.joomfish.net/, http://typo3.org/ http://drupal.org/

аутоматски и пружа алате за резимирање докумената и њихових превода. И не само то, корисници могу лако да пронађу најважније текстове у великим збиркама докумената и добију бољи преглед публикација, захваљујући детаљнијим информацијама заснованим на анотирању текста према важним именима, датумима, нумеричким изразима и значајним фразама.

Систем i-Librarian је развијен у оквиру пројекта ATLAS (Applied Technology for Language-Aided CMS (Примењена технологија за језички потпомогнут систем управљања садржајем)) који финансира Европска комисија кроз програм Програм за подршку спровођења закона и стратегија у области информационо - комуникационих технологија који је део Оквирног програма за конкурентност и иновативност (ICT Policy Support Programme - ICT PSP; уговор о додели финансијских средстава бр. 250467)[2]. Његов главни циљ је да олакша развој вишејезичних садржаја на мрежи и управљање њима, нарочито стварање и одржавање вишејезичних веб локација, као и одржавање њихових различитих верзија. Једно од главних достигнућа пројекта је библиотека i-Librarian која има користи од интеграције језичких технологија у управљање вишејезичним садржајима.

Основна функција система је описана, из угла различитих типова корисника, у одељцима који следе.

1.1. Угао читалаца

Руководиоци, чланови академске заједнице, људи који путују, људи који воле да читају:

1. Читалац поставља неколико дигиталних књига у свој лични радни простор. Књиге се у оквиру ове услуге обрађују, организују у одговарајуће предметне категорије, резимирају и анотирају према важним речима и изразима. Читалац потом може да приступи књигама и да их чита са било ког места на свету путем прелистача или мобилног уређаја (iPhone-a, уређаја који користе оперативни систем Android, итд.). Осим тога, могуће је дискутовати о омиљеним књигама са другим корисницима система i-Librarian, и ако се читаоцима нарочито допадне нека књига, могу да траже сличне књиге без обзира на ком су језику.
2. Руководилац има пословни састанак са потенцијалном муштеријом ван канцеларије. Жели да покаже клијенту важан документ, али документ се не налази у меморији његовог преносивог уређаја. Приступа свом налогу у систему i-Librarian и лако проналази жељени документ, јер су сви документи категоризовани, а доступна је и проширена верзија њихових резимеа.

1.2. Угао аутора

Студенти, истраживачи, аналитичари, консултанти:

1. Студент пише истраживачки рад и потребно му је да брзо одабере и прочита најбитније текстове из велике збирке извештаја, вести и научних публикација. Поставља архивску датотеку са свим документима у свој радни простор у оквиру система i-Librarian који резимира документе и врши екстракцију важних речи, израза и имена. Пошто прочита резимее и одломке текстова, студент одлучује које документе вреди детаљније прегледати, а које може да одбаци. Осим тога, може лако да се креће кроз свој радни простор, зато што i-Librarian аутоматски сортира постављене документе у одговарајуће предметне категорије и међусобно повезује документе на основу одломака. На крају студент може да

2 http://www.ATLASproject.eu

објави завршени рад или у делу библиотеке i-Librarian доступном јавности или на постојећим веб локацијама.

2. Истраживач објављује рад у систему i-Librarian. Сервис аутоматски анотира рад према важним речима, изразима и именима и преводи анотације на неколико језика. Други истраживач који се бави истом облашћу, али говори неки други језик, проналази тај документ у систему i-Librarian, користећи вишејезичну претрагу на основу критеријума „пронађи слично“ и контактира аутора. Тако истраживачи могу да размене знања о одређеној теми и да одлуче да сарађују.
3. Научник држи говор на научном форуму о тренутним резултатима свог истраживања. После предавања, научнику бива упућено питање из публике и он мора да поткрепи одговор податком објављеним у раду написаном за једну конференцију. Приступа свом налогу у оквиру система i-Librarian и проналази рад смештен у лист „Конференције“ на категоризационом стаблу. Као корисник сервиса i-Librarian, он је имао могућност да на њега поставља све документе, радове и истраживачке публикације и да их организује и сређује користећи нарочиту функционалност разврставања у кластере.

1.3. Библиотекаров угао

1. Дигитално издање треба да буде представљено на привлачан начин да би привукло пажњу читалаца. Књижаре користе сервис i-Publisher за обраду дигиталних садржаја и обогаћивање постојећих информација за дигитална издања књига. Уз библиографске информације као што су аутор, наслов и датум објављивања, свако издање има и резиме који генерише i-Publisher, да би читаоци могли да добију кратак преглед садржаја књиге. Читалац добија најчешће коришћене именичке фразе, имена, линкове и датуме који се тичу књиге. Ако кликне, на пример, на неку фразу, проналази листу књига у којима се помиње та фраза. Читалац добија листу дигиталних књига сличних оној коју тренутно прегледа. Корист: Садржај добија на вредности. Систем води корисника до садржаја који имају везе са оним за шта је првобитно заинтересован. Осим тога, корисник са лакоћом прави избор књига. Књижаре ће профитирати од повећане продаје књига, пошто ће корисници лакше проналазити релевантне информације, т.ј. књиге на сасвим прецизно одређене теме. Такође ће имати користи од продаје великог броја примерака појединих књига, чија ће продаја порасти захваљујући чињеници да систем предлаже сличне документе, што ће омогућити читаоцима да проналазе и купују више књига на омиљене теме.
2. Библиотека објављује велике количине информација, као што су публикације, књиге, чланци и билтени итд. које треба да буду анотиране, категоризоване и свакодневно доступне на мрежи. Библиотека интегрише алате за ископавања из текстова у постојећи софтверски систем. Члан тима обучава модел за категоризацију дигиталних садржаја уз помоћ ручно категоризованих података или интегрише већ обучен модел. Резултат тога је да ће новододати садржаји бити аутоматски категоризовани у складу са тим моделом. Новододати садржаји су обогаћени аутоматски компилираним анотацијама, као што је екстракција најчешће

коришћених именичких фраза у тексту, датума, линкова, именованих ентитета и детаљни резиме састављен од извода из самог текста. Такође, анотације су машински преведене на језике на којима је веб локација доступна. Корист: Аутоматска категоризација. Мања потреба да уредници документа класификују ручно, пошто систем аутоматски сугерише категорије за делове садржаја. Додатне информације које се објављују на веб локацији, пружају корисницима бољи преглед публикација. Такође, предложена листа сличних докумената може бити веома корисна за проналажење релевантних информација на неку тему.

2. Стварање библиотеке

Веб апликација, i-Librarian је развијена ради демонстрације могућности новог система за управљање садржајем названог ATLAS (Ogrodniczuk i Karagiozov 2011). Пошто представља систем управљања садржајем (CMS - content management system), ATLAS је коришћен за конфигурисање модела података апликације i-Librarian, њеног изгледа, регистрације и профила корисника и одржавање изолованог приватног корисничког простора. И не само то, систем за управљање садржајем, ATLAS омогућава напредно уређивање садржаја, подесиве токове рада са одобравањем садржаја и гранулиран систем права приступа, флексибилно конфигурисање изгледа, велики избор предефинисаних тема и модела садржаја.

Графичка корисничка сумећа лака за коришћење је изграђена сврх језгра система управљања садржајем ATLAS. Заснована на ZK-у[3], графичка корисничка сумећа (GUI - graphical user interface), ATLAS је садржајна апликација која се примењује на Интернету, функционише у различитим прелистачима, а такође је подесиве величине и безбедна.

Као апликација развијена у потпуности уз помоћ ATLAS-а, i-Librarian користи технологије које омогућавају интелигентну обраду података и стварају додатну вредност коју није могуће обезбедити на други начин. У оквиру листе карактеристика која следи, укратко су описане најважније технике и алгоритми из ATLAS-а који се користе за прављење интелигентних веб апликација:

- Индексирање и претраживање читавог текста – модерни систем за управљање садржајем дозвољава дизајнерима информација да структуришу садржаје и односе између делова садржаја на динамичан начин, а затим и да праве упите за претрагу читавог текста у свим деловима садржаја. Најчешће коришћена библиотека претаживача читавог текста која се интегрише у систем за управљање садржајем је Apache Lucene или алати засновани на Lucene, као што је Apache Solr.
- Идентификација важних, „индикативних“ речи и израза – именице (и именичке фразе) се по традицији дефинишу као “особе, места, ствари и идеје”. Amazon прво дефинише термин „статистички невероватне фразе“ као „најкарактеристичније фразе у тексту одређене књиге… у односу на све књиге (у збирци)”. Главна додатна вредност коју добија систем за управљање садржајем је презентовање читаоцу главних појмова и идеја присутних у делу садржаја.
- Идентификација именованих ентитета – именовани ентитети су именичке фразе из којих је додатно отклоњена вишезначност и категоризоване су према значењу и функцији у тексту.

3 http://www.zkoss.org/

Екстраховани именовани ентитети се користе за одговарање на шест питања (ко, шта, зашто, где, када и како) и за проналажење сличног садржаја. Популарни сервиси за екстракцију именованих ентитета су OpenCalais, Stanford CoreNLP и OpenNLP.

- Груписање сличних делова садржаја – филтрирање, контрола и одржавање релација између делова садржаја одузима време и захтева напор дизајнера информација и снабдевача садржајем. Стога су систему за управљање садржајем потребне опције као што су „још оваквих садржаја“, „препорука за читање“ и „погледајте и следеће“. Према хипотези сврставања у кластере („Документи из истог кластера се понашају слично по питању релевантности у односу на информационе потребе“.), најважније карактеристике делова садржаја су скоро исте код сличних делова садржаја из једне те исте групе.
- Аутоматско додељивање етикета деловима садржаја – тагирање садржаја (додељивање кључних речи) олакшава његово претраживање и проналажење информација; међутим, процес се обавља ручно, што захтева улагање великог напора. Прављење таксономије и додељивање етикета су две технике које се могу обављати полуаутоматски уз помоћ рачунара, док се контрола и исправљање грешака раде ручно.
- Превођење уз помоћ рачунара за вишејезичне веб апликације, машинско превођење, као поље истраживања доживљава бум, нуди нове могућности, али је слабо интегрисано у процес управљања садржајем. Са друге стране, потражња за вишејезичним веб локацијама се убрзано повећава. Механизми за машинско превођење помажу снабдевачима садржајем при изради прве верзије превода текстуалних материјала. Такође су од помоћи корисницима веб апликација да превазиђу језичке баријере. Услуге машинског превођења тренутно пружају Moses, Google Translate, Bing Translator.

3. Иза дигиталних полица

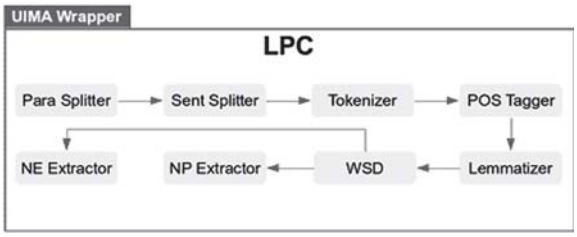
У овом поглављу чланка се излажу детаљи о архитектури ATLAS-а, као и о компоненти обраде природних језика, која је у основи карактеристика којима се одликују интелигентне веб апликације. Колико ми знамо, тренутно не постоји систем управљања садржајима који на транспарентан начин интегрише алате за обраду природних језика и пружа сумеђу за примену тих алата која је лака за коришћење. Циљ система управљања садржајем, ATLAS је да омогући интеграцију постојећих разнородних алата за обраду природних језика (NLP – Natural Language Processing) у процес управљања садржајима.

3.1. Ланци обраде језика

Текстуалне информације обично нису структурисане; међутим, људи имају способност да их обрађују и да проналазе најважније информације. За разлику од људи, рачунари не могу да врше такву анализу – они су програмирани да извршавају низ задатака са циљем да открију главне концепте и међусобне релације у тексту. Секвенцијални задаци, названи „ланац обраде језика“ (LPC - Language Processing Chain) се састоје од атомских алата за обраду природних језика који у текст додају анотације ниског нивоа, чинећи га тако структурисаним. Анотације ниског нивоа се користе за екстракцију важних речи и израза и именованих ентитета у каснијој фази обраде. Осим тога, на анотације ниског нивоа примењујемо статистичке алгоритме да бисмо пронашли најзначајније карактеристи-

ке анализираног текста.

Пример ланца обраде језика се састоји од следећих алата за обраду природних језика: токенизатора (дели сирови текст на токене) → разделника на параграфе (дели текст на параграфе) → разделника на реченице (дели параграфе на реченице) → тагера врста речи (означава сваки токен одговарајућом етикетом Врста речи) → лематизатора (одређује лему за сваки токен) → отклањања вишезначности у значењу речи (отклањања вишезначност сваког токена и додељује му јединствено значење) → екстрактора именичких фраза (означава именичку фразу у тексту) → екстрактора именованих ентитета (означава именоване ентитете у тексту) (видети слику 1).



Слика 1. Пример ланца обраде језика.

Да би се постигла оптимална прецизност ланца обраде језика, комбинујемо статистичке алате за обраду природних језика са правилима специфичним за сваки појединачни језик. На пример, ланац обраде језика за енглески језик се састоји од следећих компоненти које се извршавају у низу:

1. разделника на параграфе и анотатора URL-а/email-а – засновани на скупу регуларних израза.
2. разделника на реченице, токенизатора и тагера врста речи базираних на OpenNLP-у.
3. лематизатора – користи алат за морфолошку анализу из система RASP (Robust Accurate Statistical Parsing) (Робусни прецизни статистички парсинг) (в. 2).
4. екстрактора именичких фраза – граматика и структура именичке фразе у енглеском језику су описане кроз сет од 14 правила у формату подкомпоненте ParseEst.
5. пепознавача именованих ентитета – заснован на седам модела именованих ентитета из OpenNLP-а за препознавање датума, израза којима се означава време, локација, израза који се односе на новац, имена организација, процената и личних имена.

3.2. Захтеви и архитектура ATLAS-а

Главни захтеви за систем за управљање садржајем који се не тичу његових функција су:

- брзина одговора – класични сценарио захтев-одговор треба да се одвија најбрже могуће.
- подесивост величина – систем за управљање садржајем треба да подешава величину хоризонтално и вертикално да би се обезбедио максимално ефикасан рад.
- могућност одржавања – систем за управљање садржајем обилује значајним и мање значајним функцијама које се често преклапају и/или су комплементарне. Одржавање оваквог система није тривијалан задатак, стога је потребно да архитектура подржава овај процес у највишем могућем степену.
- међусобно уклапање – сумеђа између система за управљање садржајем и других система треба да буде што стандардија. То ће омогућити њихову надоградњу у будућности и интеграцију екстерних функција.

Интеграција било каквих модула за обраду језика не сме да угрози испуњавање ни једног од ова четири веома значајна захтева. Сваком од наведених захтева који се не тичу функција система за управљање садржајем се приступа

на следећи начин:

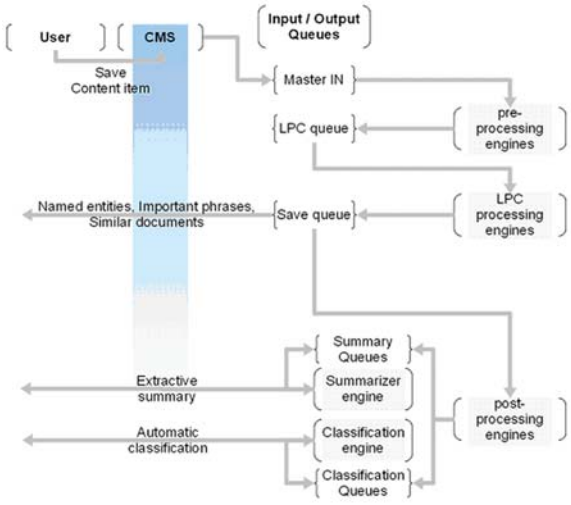
- Брзина одговора. Обављање задатака обраде природних језика је обично споро. Њихова укупна ефикасност зависи од ефикасности атомских алата за обраду природних језика и дужине текста који се уноси. Разлог зашто ланац обраде језика не може бити представљен класичним ланцем типа захтев-одговор се налази у чињеници да је немогуће предвидети време потребно да се добије одговор. Стога користимо асинхрони канал за комуникацију између компонената - система управљања садржајем и ланца обраде језика. Систем управљања садржајем асинхроно шаље поруку, идентификујући документ и пружајући његов садржај механизму ланца обраде језика и информише корисника да је обрада захтева у току. Док механизам ланца обраде језика обрађује поруку, кориснику се приказује одговарајућа информација о статусу задатка. Резултати задатка постају доступни у систему управљања садржајем када порука коначно буде обрађена.
- Механизам ланаца обраде језика заснован на оквиру OSGi. Оквир OSGi је систем модула и платформа услуга за Јаву који примењује комплетни и динамични модел компоненти. Апликације и компоненте се могу даљински покретати, заустављати и ажурирати, а да при том није потребно поновно покретање система. Облик примене оквира OSGi, Equinox је изабран да буде окосница предложене архитектуре механизма ланца обраде језика. Наша архитектура је састављена од три главне компоненте:
  - Ред чекања при размени порука. Сумеђа за програмирање апликација (API – Application Porgramming Interface) Јава услуге за размену порука (JMS - Java Messaging Service) је посреднички софтвер оријентисан на поруке који служи за слање порука између два или више клијената. Он омогућава да комуникација између различитих компоненти дистрибуиране апликације буде лабаво спрегнута, поуздана и асинхрона. Примену агента за пренос порука између система управљања садржајем и различитих компоненти ланца обраде језика смо базирали на Apache ActiveMQ.
  - Атомски анотатор. Атомски анотатор је одговоран за почетни сет анотација потребних за обављање задатака вишег нивоа у области обраде природних језика. Анотатор вади поруку из реда и прослеђује обраду:
    - претпроцесору. Ова компонента препознаје тип вишенаменских прикључака за Интернет пошту (mime-type) у који спада садржина порука, ако је потребно, врши екстракцију текста, открива на ком језику је текст написан и шаље интерну поруку за природнојезичку обраду;
    - процесору за природнојезичку обраду. Ова компонента обезбеђује основне анотације у тексту поруке. Слично као OSGi за Јаву, неструктуриране апликације за управљање информацијама (UIMA - Unstructured Information Management Applications)

омогућују да се комплексне апликације за обраду природних језика раставе на компоненте. Сваки атомски алат за обраду природних језика је смештен унутар примитивног механизма; секвенцирање примитивних механизама врши обједињени механизам. Неструктуриране апликације за управљање информацијама нису компатибилне са оквиром OSGi и зато смо обједињени механизам UIMA сместили у OSGi компоненту (природнојезички процесор), а тиме смо га учинили доступним осталим компонентама у инсталацији.

- постпроцесор. Ова компонента се позива када су анотације спремне и меморише анотације, даје извештај о учинку, информише систем за управљање садржајем да су анотације доступне и позива компоненте вишег нивоа, категоризатор и алат за резимирање.

- Категоризатор и алат за резимирање. Категоризатор и алат за резимирање имају исту унутрашњу архитектуру, зато је само алат за резимирање описан детаљно. Компонента вади поруку из реда, учитава потребне реченице и токене за тражени документ, покреће механизам за резимирање (реализација LexRank-a или екстерни алат OpenText Summarizer), прави резиме документа и шаље резиме у ред ради даље обраде (сачуван у складишту података).
- Подесивост величине. Коришћењем реда чекања при размени порука у архитектури механизма ланца обраде језика, омогућава се прилично мало хоризонтално подешавање величине једноставно инсталирањем нових инстанци механизма ланца обраде језика. Типичним корисницима библиотеке i-Librarian енглески није матерњи језик, те се очекује да ће користити i-Librarian у двојезичном окружењу – на језику који је њима матерњи и енглеском. Тако је очекивани број докумената на енглеском много већи него докумената на другим језицима. У овом случају, има смисла поставити неколико ланаца обраде језика на енглеском који раде истовремено, да би се време које претходи обради поруке свело на најмању могућу меру.

Описана архитектура је приказана на дијаграму на слици 2.



Слика 2. Најважније компоненте архитектуре ATLAS-а и комуникациони канали између њих.

У блоку на дијаграму стоји:

- Корисник. Корисник покреће ланце обраде језика тако што обави неку активност управљања садржајем на веб локацији (додавање или ажурирање садржаја).
- Систем за управљање садржајем. Систем за управљање садржајем "комуницира" са механизмима ланаца обраде језика путем реда чекања при размени порука, преко добро дефинисане сумеђе за програмирање апликација (API). Тренутно је доступан само API заснован на оквиру OSGi.
- Улазни и излазни редови чекања. Асинхрона комуникација између компонената се остварује помоћу реализације Јава услуге за размену порука. Порука се шаље у улазни ред чекања; једна компонента вади поруку из реда, трансформише је и шаље у други ред чекања. Компонента ланца обраде језика и систем за управљање садржајем примењују усмеривач порука, алат за превођење порука, мрежни пролаз за поруке, пројектне обрасце вођене догађајима, намењене потрошачима и конкурентске обрасце намењене предузећима (Hohpe & Woolf, 2003)
- Механизми за претпроцесирање. Ова компонента омогућава откривање типа вишенаменских прикључака за Интернет пошту, екстракцију текста, идентификацију језика и пречишћавање текста.
- Механизми за обраду ланаца обраде језика. Компонента смешта ланац обраде језика за дати језик.
- Механизми за постпроцесирање. Компоненте складиште анотације у складишту података.
- Механизми за резимирање и категоризацију. Ове компоненте резимирају документ и праве листу категорија које су применљиве на њега. Архитектура механизама дозвољава интеграцију више алгоритама за резимирање и алата за категоризацију.

Пошто се базира на спецификацији OSGi, архитектура ATLAS-а се може лако проширити да би могла да подржи више језика (тренутно су бугарски, енглески, немачки, грчки, пољски и румунски доступни у облику ланаца обраде језика) и више врста анотација, као што су ланци кореференције, као и дубљу интеграцију са Wordnet-има, да би се постигло боље семантичко разумевање текстуалних садржаја.

4. Језичке технологије за вишејезичне збирке

Вишејезични библиотечки садржај се обрађује у ланцима обраде језика који се разликују у зависности од језика,и нуде исти сет активности у домену обраде за све подржане језике (откривање границе између реченица, токенизацију, лематизацију, етикетирање врста речи, одређивање именичких фраза, препознавање именованих ентитета). Иако се техничке компоненте које се примењују у обради разликују од језика до језика, овај приступ нуди заједничку основу за обраду језика, а његове резултате могу несметано користити напредне језичке компоненте (за класификацију докумената на основу садржаја, статистичко машинско превођење, резимирање на основу фраза), као и за директну визуализацију.

4.1. Аутоматска категоризација

Класификација докумената је задатак који подразумева да се документ сврста у једну или више категорија или класа. Аутоматизовање тог процеса је од велике важности за савремене апликације; стога су током година развијени разни методи.

Методи за аутоматску класификацију се могу неформално поделити у две групе: статистички алгоритми и структурални алго-

ритми. Примери статистичких алгоритама су регресија и наивни бајесовски. Структурални алгоритми се могу даље поделити на алгоритме засноване на правилима (стабла одлучивања, правила извођења) и алгоритме засноване на растојању (kNN, Centroid), као и неуронске алгоритме.

Класификација према једном обележју се врши са циљем учења од групе докумената који су повезани са једним обележјем (класом) l из скупа обележја L. У класификацији према више обележја, сваки документ може бити повезан са више обележја из скупа L. Ако L садржи тачно два обележја, проблем учења се зове „бинарна класификација“, а ако L садржи више од два обележја, проблем се зове „вишекласни“.

4.1.1. Примењени алгоритми

Наш систем има модул за аутоматску категоризацију докумената према више обележја и више класа. Алгоритми који су тренутно укључени у модул су наивни бајесовски, алгоритам заснован на релативној ентропији и центроидни са класама“ (class-featured centroid). Изведени експерименти су показали да поменути алгоритми обезбеђују разуман ниво прецизности при класификовању и да су много бржи него комплекснији методи, као што су машине са потпорним векторима (Support Vector Machines). Горенаведена листа, међутим, није коначна, пошто стално експериментишемо са новим методима и стратегијама класификације који ће бити укључени у касније верзије система.

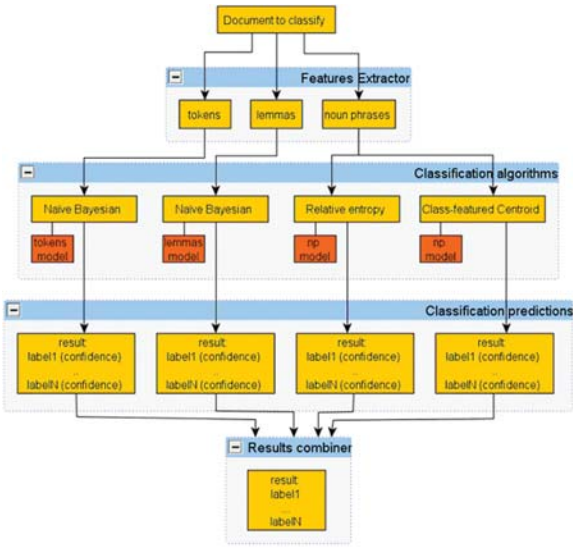
4.1.2. Решавање задатка класификације

Задатак класификације се састоји од две основне фазе: обучавања и класификовања. У фази обучавања се обрађује група обележених докумената ради прављења модела. У кораку који подразумева класификацију се користи модел да би се документима који нису етикетирани доделила једна или више етикета.

Да би се направио модел, модул представља сваки документ као скуп својстава. Та својства се касније користе за прављење модела за различите класе. У зависности од алгоритма који се користи за класификацију, током обраде се може применити метод за смањење броја својстава.

Ланац обраде језика обрађује сваки документ и обезбеђује приступ различитим врстама својстава – токенима, лемама, именичким фразама, водећим токенима. Тиме се омогућава успостављање једног алгоритма који ради са различитим врстама својстава. Осим тога, модул за категоризацију може да примењује неколико алгоритама истовремено. Резултати добијени од различитих класификатора се комбинују, а резултат класификације се одређује већином гласова.

На дијаграму на слици 3 су описани главни кораци од којих се састоји задатак класификације докумената:



Слика 3: Главне фазе процеса класификације докумената.

У горенаведеном примеру, постоје четири класификатора – наивни бајесовски са токе-

нима у простору својстава, наивни бајесовски са лемама у простору својстава, релативна ентропија са именичким фразама у простору својстава и центроидни са класама са именичким фразама у простору својстава. Класификациони модул региструје те алгоритме као OSGi услуге, према конфигурационим подешавањима. Екстракцију својстава новог документа врши оквир ланаца обраде језика, а потом се те особине прослеђују одговарајућем класификатору. Сваки класификатор користи свој модел за предвиђање скупа обележја. Коначно се сви резултати комбинују и класификација се презентује кориснику.

4.2. Машинско превођење

Машинско превођење је кључна компонента ATLAS-a – WebCMS-a и биће уграђена у систем i-Librarian за потребе „превода за асимилацију". Развој механизма преставља посебан изазов, пошто превођење треба примењивати у различитим доменима и на текстове различитих жанрова. Осим тога, већина језичких парова који се разматрају, припадају групи језика за које је на располагању мање ресурса и за које је двојезични материјал за обучавање и тестирање ограничен.

Механизам за машинско превођење је интегрисан у платформу ATLAS на два различита начина:

У системима i-Librarian и EuDocLib (видети пододељак 6.1), механизам за машинско превођење обезбеђује превод за асимилацију, што значи да ће корисник који проналази документе на разним језицима користити механизам да би добио назнаке о садржају докумената и одлучио да ли жели да их сачува. Ако се сматра да је превод прихватљив, биће сачуван у бази података.

Интеграција механизма за машинско превођење у систем за управљањем садржајима на вебу уопште, а посебно у систем ATLAS, доноси два главна изазова из угла корисника:

1. Корисник може проналазити документе из различитих домена. Прилагодљивост домена је веома значајан проблем у машинском превођењу, а нарочито што се тиче метода заснованих на корпусу. Лоша лексичка покривеност и лажно отклањање вишезначности су главни проблеми при превођењу докумената из домена обуке.
2. Корисник може проналазити документе из различитих временских периода. Како се језик временом мења, алати језичких технологија развијени за савремене језике не функционишу, или функционишу уз већи проценат грешака када се примењују на дијахроне документе.

Тренутно доступна технологија не омогућава да се обезбеди систем за превођење који би био независан од домена и језичких варијација и применљив на неколико хетерогених језичких парова. Стога наш приступ предвиђа систем упућивања корисника, да би доступност и предвиђено функционисање система били транспарентни у сваком тренутку.

Имајући у виду чињеницу да се платформа ATLAS користи за језике који припадају различитим језичким породицама, као и да би требало да механизам подржава најмање неколико домена, приступ типа интерлингуа није одговарајући. Изградња система трансфера за све парове језика такође захтева много времена и не омогућава да се платформа лако преноси на друге језике. С обзиром на корисничке и системске захтеве, парадигме машинског превођења засноване на корпусу су једине које треба узети у обзир. У наставку ћемо описати експерименте које смо извршили да бисмо одредили најбољи приступ који ћемо применити.

Статистичко машинско превођење (SMT - Statistical Machine Translation) је парадигма која се најчешће користи када је циљ система превод за асимилацију. Систем статистичког машинског превођења Moses (Koehn et. al 2007)

је не само механизам за превођење, већ дозвољава и развој и коришћење преводилачких и језичких модела узимајући у обзир варијације неколико параметара. Извршили смо неколико експеримената да бисмо утврдили да ли:

- су уобичајене вредности параметара које се користе у кампањама евалуације погодне за парове језика у којима оба језика имају богату морфологију,
- корак који подразумева подешавање а који захтева много времена, доводи до значајних побољшања,
- модели у којима се као фактор узима врста речи значајно побољшавају квалитет резултата.

Експерименте смо извршили на свим паровима језика које су чинили немачки, румунски и енглески, користећи вредности параметара примењене у кампањи евалуације са Радионице о статистичком машинском превођењу (WMT - Workshop on Statistical Machine Translation) из 2010. године. Као корпусе за обучавање смо користили JRC-Acquis као и корпус ROGER, ручно поравнати корпус који обухвата одређени домен (Gavrila & Vertan 2011).

Осим тога, упоредили смо резултате са системом за машинско превођење заснованом на примерима (EBMT - Example Based MT) описаном у (Gavrila 2011). То је систем који функционише независно од језика, на нивоу низова у који се уграђују лингвистичке информације из извора-инпута. Извршени су следећи експерименти:

- SMT: поређење резултата нашег система са резултатима изнетим у релевантним истраживачким радовима,
- SMT у односу на EBMT примењено на документ Acquis Communautaire[4] ,
- SMT у односу EBMT примењено на ROGER,
- ROGER као тест корпус за SMT тренирано на документу Acquis Communautaire.
- Експерименти су довели до следећих закључака:
- Чак и уз иста подешавања Mosesa, могу се добити различити резултати BLEU[5] (Papieni et. al 2002) пошто:
  - се подаци за тестирање могу разликовати, *и*
  - број референтних превода варира.
- Само један референтни превод проузрокује лошије BLEU резултате;
- BLEU и TER[6] нису увек у корелацији, т.ј. BLEU расте, а TER је или лошији или остаје непромењен. То може бити показатељ да се BLEU ослања само на вокабулар и н-граме, док TER подражава „нешто мало синтаксе".
- Чак и уз наведене мане, применом система статистичког машинског превођења, уз класична подешавања из кампање евалуације, добијају се резултати слични онима који се описују у литератури, као што се може видети на слици 4.
- Подешавање захтева изузетно много времена, а побољшања су минимална.
- Модели у којима се као фактор узима врста речи, незнатно побољшавају ре-

4 Acquis Communautaire је скуп закона, правних аката и судских одлука који чине главнину закона Европске уније. JRC-Acquis представља збирку паралелних текстова на 22 језика коју је из тог извора сачинио Заједнички истраживачки центар (Joint Research Centre) Европске комисије.

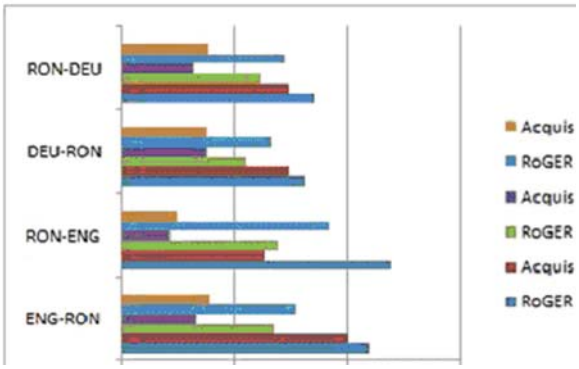
5 BLEU (Bilingual Evaluation Understudy - Двојезични систем за евалуацију) је алгоритам за оцењивање квалитета текста који је машински преведен са једног на други природни језик. Квалитет се дефинише као поклапање између резултата машинског превођења и превода који изради човек: „машински превод је бољи што је ближи професионалном преводу који изради човек"

6 Проценат грешака у преводу је мера за грешке при машинском превођењу којом се означава број измена које је потребно извршити да би се излазни резултат система преточио у референтни превод.

зултате евалуације.

Када је реч о величини корпуса, што представља веома важно питање када се ради са језицима за које постоји мањи број ресурса, наши експерименти су показали следеће:

- Обучавање на мањем корпусу, као што је ROGER (неколико хиљада реченица) не води до веома лоших резултата, под условом да подаци који се тестирају припадају истом домену.
- Ефикасност у великој мери зависи од прецизности поравнавања речи.



Слика 4. Евалуација система EBMT и SMT

Следи резиме поређења статистичког система за машинско превођење и система заснованог на примерима:

- Ако су реченица или делови улаза идентични као делови садржани у корпусу за обучавање, EBMT (Somers 1999) има бољи учинак, пошто се аутоматски проналазе одговарајуће јединице циљног језика.
- Резултати евалуације су лошији у случају EBMT-a.
- Међутим, после ручне евалуације приближно 100 реченица, показало се да је квалитет превода неколико реченица бољи у случају EBMT-a.
- На нивоу ниски у статистичком систему за машинско превођење не постоји могућност уградње лингвистичких информација из језика изворника, што може бити релевантно за прављење превода. EBMT има ту могућност.
- Различити домени у обучавању и тестирању података могу да умање учинак система због великог броја речи које не припадају домену.

У случају механизма за машинско превођење у систему ATLAS, одлучили смо се за хибридну архитектуру, у којој се на нивоу речи комбинују машинско превођење засновано на примерима и статистичко машинско превођење (неће бити коришћена синтаксичка дрвета). За статистичку компоненту машинског превођења се користе модели у којима се као фактор узимају врсте речи или домен, као у (Niehues & Waibel 2010), да би се обезбедила прилагодљивост домена. Оригинални приступ нашег система се огледа у интеракцији механизма за машинско превођење са другим модулима система описаним у делу текста који следи.

Модул за категоризацију документа додељује сваком документу један или више домена. Систем администратор има могућност да за сваки домен сачува информације о доступности одговарајућег конкретног корпуса за обучавање. Ако за одговарајући домен не постоји конкретан обучени модел, корисник добија упозорење да постоји могућност да превод неће бити адекватан, што се тиче лексичке покривености.

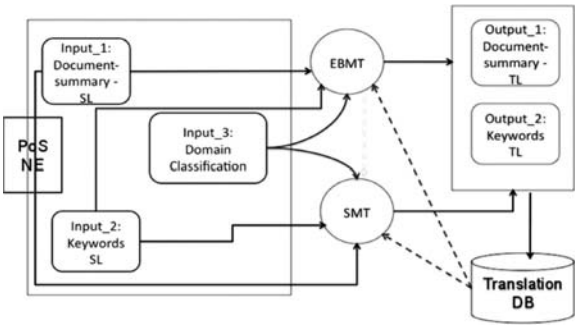
Излаз модула за резимирање се обрађује тако да се изостављају елипса и анафора, а лексички материјал се прилагођава корпусу за обучавање.

Модул за екстракцију информација пружа информације о метаподацима докумената, укључујући и време када је документ објављен. За документе објављене пре 1900. године не дајемо превод, уз објашњење кориснику да у недостатку корпуса за обучавање, превод може бити нетачан.

Систем администратор може у сваком тренутку да измени ограничења, која се тичу доме-

на и времена настанка докумената, када постане доступан адекватан модел за обучавање.

Описана архитектура је представљена на слици 5.



Слика 5. Архитектура система за механизам машинског превођења ATLAS

4.3. Резимирање текста

Сврха резимирања докумената у апликацијама ATLAS је да заинтересованом читаоцу представи резиме чланка или књиге. Читалац може да користи могућности претраживања да на Интернету прегледа информације на теме које га занимају и када пронађе неки текст за који се заинтересује, пожели да брзо прегледа садржај документа који на први поглед изгледа занимљиво. Може се десити да је текст написан на језику који не разуме. Комбинацијом алата за резимирање и модула за превођење, таквом кориснику се може пружити резиме на његовом језику.

Заинтересовани смо за две врсте резимеа: резимеи кратких текстова (кратке приче или чланци, не дужи од неколико страница) и резимеи дужих текстова (на пример, романа). Јасно је да је у случају краћих текстова могуће назначити дужину резимеа као одређени проценат изворног текста, док у случају дужих текстова то више није могуће зато што би варијације у дужини добијеног резимеа биле сувише велике. Стога се у ATLAS-у користе две сасвим различите стратегије за добијање резимеа: за краће текстове, стратегија се базира на идентификацији структуре дискурса, па се резимеи формирају од извода из текста, док се за дуже текстове стратегија базира на екстракцији релевантних информација, па се резимеи базирани на шаблонима добијају генерисањем новог текста. У наставку, описујемо само филозофију резимирања коришћену у пројекту за кратке текстове.

Резимирање кратких текстова у ATLAS-у има користи од целог ланца обраде, а притом додаје неколико других модула на крају ланца. Почетне фазе које се примењују на документ који се резимира су: идентификација параграфа и граница реченица, дељење реченица на клаузе, токенизација, тагирање врста речи и лематизација, препознавање именованих ентитета, плитко парсирање за идентификацију именичких фраза и разрешење анафора. До овог тренутка, изрази којима се исказује референција (нарочито заменице, али и други именички изрази и именовани ентитети) се препознају и везују за своје антецеденте. Ови ланци кореференције помажу да се идентификује најверодостојнија структура дискурса. Структура дискурса, у облику дрвета, гради се поступно, уз примену стратегије beam-search[7] да би се ограничила експоненцијална експлозија генерисаних структура. У сваком тренутку током парсирања, из таласа N дрвета (која се називају „дрва у развоју“) се задржавају она која највише обећавају у датом тренутку, а остала се одбацују. Потом се врши парсирање следеће реченице, и генеришу се сва могућа мања дрвета (која се називају „помоћна дрвета“), вођена маркерима дискурса садржаним у реченици. Након тога се сва та помоћна дрвета комбинују на све дозвољене начине са сваким од N дрвета у развоју,

7 Beam search је хеуристички алгоритам за претрагу који пролази кроз граф проширујући чвор који највише обећава у ограниченом скупу.

прикључивањем на десној граници (Cristea & Webber, 1997) и супституцијом (ове две операције комбиновања дрвета инспирисане су граматикама са прикључивањем дрвета (Tree Adjoining Grammar - TAG[8]). У случају стабала у развоју, која настају из тог процеса, а већа су од првобитних са јединицама из још једне реченице, израчунати резултати се придружују на основу другачије хеуристике. Затим се читава шума насталих стабала у развоју рангира на основу тих резултата и најбоља N дрвета се чувају за следећи корак.

На крају ове процедуре, треба да остане низ коначних стабала, а на крају оно које је најбоље рангирано се предлаже да баш оно представља структуру дискурса улазног текста. Сваки резиме који се базира на неком проценту текста би онда једноставно могао да се извуче из дрвета дискурса. Сви ти резимеи су кохерентни и заменице не могу да промаше своје антеценденте. Осим тога, нити које придружене јединицама дискурса (Cristea, 2009) дозвољавају генерисање резимеа који се фокусирају на одређене ентитете, чак и ако су ти ентитети од мањег значаја у тексту и не би се нашли у општем резимеу.

5. Утисци корисника библиотеке

Тренутно потенцијални корисници врше евалуацију система i-Librarian. Циљ је да се изврши процена прихваћености онлајн услуге, применом индикатора који мере степен задовољства корисника које проистиче из њиховог искуства у коришћењу услуге. Индикатори оцењују параметре невезане за функције система i-Librarian, као што су:

- Прилагођеност потребама корисника и задовољство корисника, јасноћа одговора и лакоћа коришћења;
- адекватност и комплетност понуђених података и функција;
- утицај на извесне активности корисника и степен извршавања уобичајених задатака.

Примарних корисника система i-Librarian има три врсте (т.ј. постоје три корисничке групе), и то:

1. ГК1 – студенти и научници: стварање личне библиотеке, којој се може приступити преко Интернета, формулисање аутоматски генерисаних вишејезичних извода из текста и резимеа докумената, итд.
2. ГК2 – аутори, млади научници и истраживачи: неометано управљање различитим документима на разним језицима, дељење преведених извода из текстова и резимеа радова, чланака, итд са другима.
3. ГК3 – обични корисници Интернета, умерено искусни у коришћењу веба: стварање личне дигиталне библиотеке, којој се може приступити преко Интернета, објављивање и превод извода, итд.

Охрабрујемо све кориснике да испробају услугу путем интернета. Основни сценарио је понуђен као помоћни алат, употпуњен вежбањем са предлозима о различитим задацима и корацима од којих се састоје активности. Главни инструмент за прикупљање реакција корисника је интерактивни електронски упитник доступан онлајн на адреси: http://ue.ATLASproject.eu/ … само изаберите групу корисника којој припадате и пратите упутства на екрану!

6. Друге библиотеке докумената

Осим система i-Librarian, уз примену језичког и техничког оквира ATLAS су припремљене још две библиотеке докумената – EUDocLib и PLDocLib, овога пута, углавном за потребе демонстрације њиховог функционисања, пошто широј јавности није дозвољено уношење измена. Обе библиотеке нуде могућност претраге која је лингвистички подржана.

8 Граматике са прикључивањем дрвета су донекле сличне граматикама независним од контекста (контекстно-слободним граматикама), али основна јединица преписивања је дрво, а не симбол.

Слика 6. Сумеђа система i-Librarian

6.1. EUDocLib

Услуга EUDocLib[9] представља репозиторијум докумената Европске уније из збирке EUR-LEX у који јавност има приступ. Тај репозиторијум омогућава лакши приступ релевантним документима на језику корисника и обезбеђује:

- аутоматски категоризован, резимиран и анотиран садржај, са важним именичким фразама и именованим ентитетима,
- боље кретање кроз садржај (захваљујући, на пример, листама сличних докумената) засновано на међусобно повезаним анотацијама текста,
- машински преведене изводе из докумената и њихово коришћење за категоризацију и груписање докумената.

Тренутно се на веб локацији налази 140 хиљада докумената (182 милиона токена).

9 http://eudoclib.ATLASproject.eu/

Слика 7. Сумеђа библиотеке EUDocLib: резултати претраживања

6.2. PLDocLib

Пољска варијанта библиотеке EUDocLib[10] је веб локација чије функционисање обезбеђује ланац обраде језика који омогућава претрагу и прегледање око хиљаду правних аката пољског парламента (Сејма) аутоматски анотираних уз помоћ сета алата за пољски језик интегрисаних у ATLAS, а то су:

- Morfeusz – морфолошки анализатор,
- Pantera – Брилов (Brill) тагер за пољски заснован на правилима,
- Spejd – механизам за плитко парсирање уз помоћ каскадних граматика,
- Алат plNER – статистички препознавач именованих ентитета који се базира на условним случајним пољима (CRF)[11].

На основу анотација, веб апликација за сваки документ обезбеђује скуп препознатих именованих ентитета, значајних именичких фраза (у кластерима формираним на основу њихове сличности и значаја), као и листу сличних докумената. За потребе презентације се генеришу основни облици (у смислу речничке одреднице) вишечланих речи, а користе се ручно додељене категорије.

10 http://www.ATLASproject.eu.pl/

11 Условна случајна поља (CRF – Conditional Random Fields) представљају класу статистичког метода моделирања која се често примењује у препознавању шаблона и машинском учењу, где се користе за структурисано предвиђање.

Слика 8. Сумећа PLDocLib: резултати претраживања и пример документа са темама и сличним документима

Закључци

Обиље знања нам дозвољава да проширимо примену алата заснованих на обради природних језика развијених у истраживачком окружењу. Комбинација управљања садржајима на вебу и најсавременијих језичких технологија, омогућава читаоцу да превазиђе језичку баријеру, да пронађе најрелевантније информације у великим збиркама података и да све те информације буду сређене на одређени начин. Наменски развијен систем гласања максимизира искоришћеност различитих алгоритама за категоризацију. За резимирање кратких и дужих текстова се користе два различита приступа, а њихов превод обезбеђује најсавременији хибридни систем за машинско превођење.

Језички оквир ATLAS ће бити објављен као софтвер отвореног кода. Ланци обраде језика за бугарски, грчки, румунски, пољски и немачки су у потпуности имплементирани почетком 2012. године.

Очекујемо да ће та платформа постати основа за будући развој алата за дубинску анализу које ће моћи да генеришу резимее у којима су сажете информације као и моделе за обуку за системе за доношење одлука.

Литература

Cristea Dan and Bonnie Lynn Webber. 1997. Expectations in Incremental Discourse Processing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid.

Cristea Dan. 2009. Motivations and implications of veins theory: a discussion of discourse cohesion. In *International Journal of Speech Technology*, 12(2/3), 83–94.

Gavrila Monica. 2011. Constrained recombination in an example-based machine translation system. EAMT-2011: *the 15th Annual Conference of the European Association for Machine Translation*, Eds. Vincent Vondeghinste, Mikel L. Forcada, and Heidi Depraetere, 193–200, Leuven, Belgium: EAMT.

Gavrila Monica and Cristina Vertan. 2011. Training data in statistical machine translation – the more, the better? In *Proceedings of the RANLP-2011* Conference, Hissar, Bulgaria.

Hohpe Gregor and Bobby Woolf. 2003. *Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions*. Addison-Wesley Professional.

Koehn Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, 2007. Moses: *Open Source Toolkit for Statistical Machine Translation*, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

Niehues Jan and Alex Waibel, 2010. Domain Adaptation in Statistical Machine Translation using Factored Translation Models, *Proceedings of EAMT 2010*, Saint-Raphael.

Ogrodniczuk Maciej and Diman Karagiozov. 2011. ATLAS – The Multilingual Language Processing Platform. *Procesamiento del Lenguaje Natural*, vol. 47, 241–248.

Papineni Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In ACL-2002: *40th Annual meeting of the Association for Computational Linguistics,* 311–318.

Somers Harold. 1999. Review Article: Example-based Machine Translation. *Machine Translation*, 14(2):113-157.