СОЦИАЛЕН ПОДХОД КЪМ СЕМАНТИЧНОТО АНОТИРАНЕ НА Е-КНИГИ

Ясен Кипров, Иван Койчев, Борис Крайчев

Факултет по Математика и Информатика,
Софийски университет Св. Климент Охридски”
koychev@fmi.uni-sofia.bg

***Резюме**: Уеб 3.0 цели създаването на общ интерфейс за публикуването на различни бази от знания и свързването им чрез общи онтологии. Това създава предпоставки за изграждане на ново поколение интелигентни системи. Тези технологии дават възможност и на е-книгите, да станат „по-умни”. Целта на представената работа е създаването на инструменти, които да улеснят вмъкването на явна семантика в е-книгите. Статията прави обзор на съществуващите семантики технологии и софтуерни инструменти за семантично анотиране на текст. Представя и прототип на система, който позволява семантично анотиране на книги. В системата са интегрирани платформите Insemtives и KIM. Разработени са съответни интерфейси: за автор - позволява полуавтоматично семантично обогатяване на книгата и съответната база от знания; а за читателя – да използва вече вкараната семантика и възможност за създаването на допълнително персонално семантично обогатяване, което може да бъде и споделено.*

***Ключови думи**: Е-книги, Семантичен Уеб*

1. Увод

През последното десетилетие уеб 2.0 и социалните мрежи като среда, дават възможност на хората да променят изцяло нормите в общуването. От друга страна зараждането на “версия 3.0” (*web 3.0*) на мрежата спомага за създаването на общ интерфейс за данни, които различни системи използват в интернет. Публикуването на различни бази от знания и свързването им чрез общи онтологии улеснява създаването на интелигентни системи. Тези технологии дават възможност е-книгите да станат „по-умни” [1] чрез вкарване на семантични технологии в тях. Въпреки динамичното развитие на семантичните и социални технологии, няма утвърдена система, която да помага, в процеса на работа, на авторите на статии и книги. Представената система подпомага процеса на създаване и възприемане на образователни, публицистични и научни текстове. Използвайки съществуващите платформи, тя съчетава възможностите на семантични бази от знания, текстов анализ и силата на колективната интелигентност на автори и читатели.

От гледна точка на *авторите*, основната задача на системата ще бъде да предоставя фактологични знания за термините, с които боравят, докато пишат.

Това ще им даде възможност да добавят интересни факти в текста, както и да добавят факти в общата база от знания.

За *читателите* системата ще показва разпознатите термини, заедно с фактите, свързани с тях, успоредно с текста. Читателите също ще имат възможност да маркират именувани обекти и термини, които не са разпознати от системата, като по този начин допринасят за качеството на документа. Освен представянето на знания, друга полза от структурираните знания за всеки документ, е възможността за разширено търсене.

Реализираният прототип включваща следните компоненти:

- Семантична база данни, пазеща фактологична информация, както и връзките между именовани обекти и документите, които ги споменават. Тя ще бъде обща за всички автори и ще обединява фактите, които те добавят в процес на работа.
- Система за текстов анализ, с вход – изречение и изход – изречението с добавени метаданни за споменатите в него обекти. Тази част ще служи за разпознаване на термини в реално време, като добавя връзки към семантичната база от знания към части от текста.
- Общ публичен интерфейс, предоставящ достъп на различни приложения до семантичната база и системата за текстов анализ. Този слой ще използва утвърдени интернет протоколи за да свържа останалите части на системата. В него се концентрират права за достъп, паралелизиране на процесите и други.
- Прототипи на два алтернативни потребителски интерфейса за авторите, а именно приставка на Microsoft Word и уеб сайт. Чрез тях авторите ще имат възможност да разглеждат фактите от базата данни, както и да редактират автоматично създадените анотации.

Настоялата статия прави обзор на съществуващите стандарти и формати за семантични данни и известни системи за текстов анализ. Описва прототипа на системата за семантично анотиране.

2. Обзор на семантичното анотиране

В тази глава ще направим анализ на съществуващи алгоритми и системи, близки до системата, която предлагаме. Ще завършим със сравнителен анализ на решенията, които използваме, както и техни алтернативи.

За **разпознаване на именувани обекти** се използват следните подходи:

- **Подходи, базирани на правила:** При тези алгоритмите зависимостите се описват от експерти лингвисти. Целта на правилата е на базата на локални признаци да определят дали дадена дума или фраза в текста изразява именуван обект. Най-често се базират на процесори на регулярни изрази,

като например в системите UIMA [6] и GATE [11]. Такива системи имат склонност да работят добре върху предоставения корпус, но да не се справят с нови документи.

- **Машинно самообучение**. Най-често се използват алгоритми от вида обучение с учител, т.е. които „учат” на класификатори от наблюдения.
- **Хибридни системи.** В много случаи системите разчитат на предварителна обработка с правила, както и на допълнителна обработка след машинната класификация. Среди за разработка като GATE [7] и UIMA дават възможност за лесно подреждане на различни видове алгоритми в поточни линии.

Общото между тях е целта да се намерят тенденции и зависимости в корпуса при споменаването на именувани обекти в текста. На базата на предоставените правила, положителни и отрицателни примери, за всяка дума или фраза в текста се определя дали принадлежи към някой от дефинирани за съответната задача типове именувани обекти [5]. Въпреки голямото разнообразие на изказа, съществуват алгоритми за текстов анализ, които постигат добри резултати в голяма част от случаите.

Процесът на **семантично анотиране** се състои в добавяне на семантична информация към анотациите. Въпреки бързото развитие през последните две десетилетия в областта на обработката на естествени езици, добавянето на семантични знания започва да се развива едва в последните години. Причина за развитието на областта е най-вече налагането на стандарти и развитието на глобалната семантична мрежа (web 3.0). За да бъде възможно семантично анотиране на именувани обекти са необходими:

- **Онтология** определяща класовете именовани обекти.
- **Уникални идентификатори** за всеки обект от базата знания;
- База знания с описания на обектите.

При конструиране на семантична система за разпознаване на обекти първо, речниците с познати имена се заменят с речници, използващи онтологията и базата от знания, така че да могат да се свържат разпознатите имена с конкретни обекти в базата. След внимателно подбиране на речниците, системите за правила също се обогатяват със знания [2]. Така, към методите за обработка на регулярни изрази над анотации, се добавя знанието за йерархията на класовете в онтологията.

През последните няколко години семантичният уеб набира скорост, като все повече научни постижения намират реално приложение в бизнеса. Тук ще направим **обзор на съществуващите значими системи за анализ и семантично анотиране на текст**. Общото между системите, разглеждани в тази глава е главната цел, която те си поставят: *Да извличат семантична*

информация от големи количества документи, така че хора и програми да могат да ги използват по най-добрия начин.

OpenCalais [9] е продукт на Томсън Ройтерс, чиято основна цел е класификация и семантично анотиране на новинарски статии. Тъй като OpenCalais е комерсиален продукт, малко неща са известни за алгоритмите и структурата на системата.

AlchemyAPI [10] е система, подобна на OpenCalais, която покрива широк спектър от задачи, свързани с обработка на естествени езици. Инструментите, предоставени от AlchemyAPI позволяват разпознаване на език, намиране на именувани обекти, намиране на ключови думи, разрешаване на двусмислия в именуваните обекти, разпознаване на мнения и отношения и други.

Factiva [4], подобно на OpenCalais системата е разработвана от Dow Jones и доскоро от Reuters. Служи за анотиране и класификация на новини, като дава възможност на крайния потребител да използва разширени методи за търсене, базирани на семантиката, добавена към статиите. Особеност на тази система, е че семантичната информация се добавя почти изцяло от хора.

S-cream [3] е платформа за полуавтоматично анотиране. Разпознаване на именувани обекти е базирано на правила, извлечени от съществуващи анотации. Използва се алгоритъм, който разпознава съществителни собствени имена при предварителната обработка на текста. От тях се създават хипотези, базирани на лингвистични образци и наличната онтология.

KIM (Knowledge and Information Management) [8] платформата, разработена от Онтотекст, използва алгоритми за текстов анализ, подобно на OpenCalais и AlchemyAPI. Основните разлики между KIM и другите предложени системи са две:

- KIM не се предлага като завършен продукт, а авторите наблягат на адаптации на алгоритмите и семантичните знания за различните цели.
- KIM предоставя разнообразни, предварително определени, интерфейси за откриване на документи.

Архитектурно, KIM може да се раздели на 3 модула:

- **Извличане на информация –** използва се GATE за разпознаване на именувани обекти, но и има възможност за добавяне и на други процеси към обработка на текстовата.
- **Семантична база от знания** - за база от знания, както и за връзки между документите и споменатите в тях факти, се използва семантичното хранилище OWLIM.
- **Пълно-текстово търсене** - използва утвърдената система LUCENE.

KIM използва PROTON [12] **онтологията от високо ниво**, което позволява лесното добавяне на специфични онтологии за различните задачи.

Съвместима е с онтологиите на DbPedia [16], GeoNames [17], и други допълнителни бази от знания.

За крайните потребители, KIM предлага разнообразни опции за търсене на документи. чрез уеб базирано приложение, което позволява лесно добавяне на нови интерфейси. Няма да се спираме подробно на тях, тъй като за задачата ни са по-интересни програмните интерфейси.

След интеграцията на платформата Insemtives [14], KIM предоставя **достъп чрез уеб услуги до съществуващото API**. Различни уеб услуги имат достъп до пълната функционалност на KIM, което позволява на потребителите да използват функциите на KIM в свои приложения.

Таблица 1 прави сравнение по основните характеристики на представените по-горе системи, които са в контекста на поставената задача.

Таблица 1. Сравнение на описаните системи.

	KIM	KIM + Insemtives	Alchemy API	OpenCalais	Factiva	S-cream
безплатна	да*	да	да*	да*	не	да
с отворен код	не	да, без KIM	не	не	не	не
стабилност на работа	да	не	да	да	да	не
анотиране с машинно обучение	не	не	да	да	***	да
анотиране с правила	да	да	не	не	***	да
ръчно анотиране	не	**	не	не	да	да
опции за промяна на алгоритмите	да	да	не	не	***	да
извличане на релации	не	не	да	да	да	не
фокус			новини	новини	финансови новини	
вход за документи	файлове	уеб услуги, файлове	уеб услуга	уеб услуга, уеб сайт	RSS и други	****
изход за документи	уеб услуга, уеб сайт	уеб услуга, уеб сайт	уеб услуга	уеб услуга, уеб сайт	уеб сайт	****

** за некомерсиални цели*

*** Insemtives проекта има основна цел да се създават ръчно анотации, чрез приложения, използващи платформата*

**** Factiva използва неизвестен базов алгоритъм за текстов анализ с ниска точност*

***** системата е само експериментална и не се поддържа или разработва*

Основната причина да използваме платформите Insemtives и KIM е лесната адаптация на алгоритмите използвани за текстов анализ. Отворената архитектура на платформата Insemtives от друга страна предоставя възможности за добавяне на необходимата функционалност. Като зрялост и качество на анотациите, OpenCalais и AlchemyAPI са по-добри, но са комерсиални системи, насочени към конкретни приложни. За целите на предложената система е необходима система, която е адаптируема към различни области, с възможности за лесно добавяне на алгоритми за обработка на естествен език според различните задачи.

В повечето системи семантичното анотиране се разглежда като отделен процес от създаването на документа. Не съществува утвърдена система, създаваща семантични анотации в помощ на автора, докато пише. Системи като FACTIVA дават възможност за сътрудничество между потребителите, но на инфраструктурно ниво за конкретния набор от документи. В предложената система, сътрудничеството между автори и читатели се издига на ниво допълване на базата от знания и полуавтоматичното семантично анотиране.

3. Система за полуавтоматично анотиране

Предложената система има две основни задачи – да помогне на авторите и читателите в процеса на писане и четене на книги и да обогатява знанията си за света. В следващите глави ще демонстрираме реализация на основните компоненти в системата, както и на алгоритъма за подреждане.

Основната задача на експерименталната система е да предоставя знания на авторите и читателите на книги и статии, използвайки семантична база данни. Проблемът с определянето на знания, които биха били интересни за потребителя е много сериозен и няма точно решение. Причина за това е социалният му характер. Нашият подход интегрира два различни метода: *семантично анотиране и социално подреждане на фактите*. Семантичното анотиране има за цел да прихване обекти от текста, за които системата има данни. Следва определяне на важността на фактите, чрез иновативен подход. Вярваме, че този алгоритъм може да се приложи и от системи за семантично търсене, с цел по-полезно подреждане на резултатите.

Системата се състои от два основни компонента – сървър и интерфейси. Интерфейсите, които ще предложим са под формата на приставка за MS Word, развивайки съществуващият прототипа за контекстно извличане на информация [15] и на уеб приложение. Сървърът представлява база от знания, заедно с механизми и инфраструктура за нейната поддръжка.

В нашата работа ще разглеждаме два типа потребители: *автори и читатели*. Тъй като възможностите за читателите са подмножество на тези за авторите, ще опишем първо как системата работи за авторите. На фигура 1 са

показани основните стъпки при работата на автора: 1) **Писане на текст** в текстов редактор; 2) **Семантично анотиране;** 3) **Попълване на данни.**

Фигура 1. *Работа на автора.*

3.1. Семантично анотиране

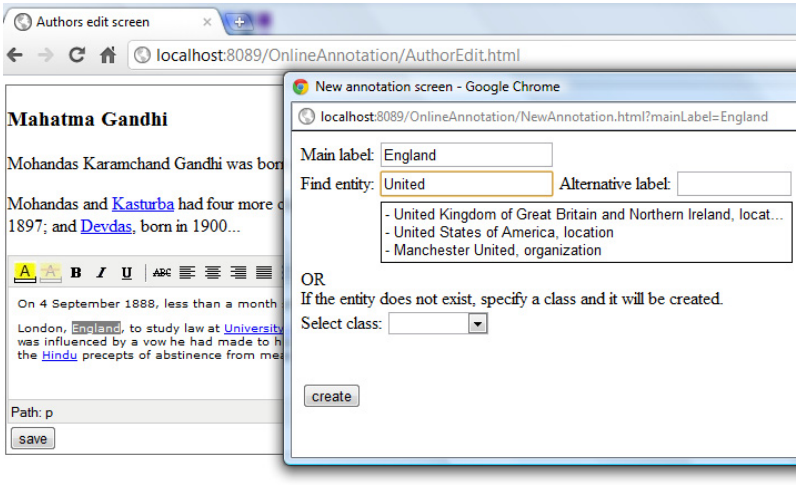
При завършване на изречение, системата взема изречението и извършва семантично анотиране. Това се случва в реално време и отнема не повече от секунда. Авторът вижда на момента кои термини в текста са разпознати от системата, като може да ги поправя, изтрива или да добавя нови анотации. Заради процеса на поправяне, анотирането се нарича полуавтоматично.

За семантично анотиране използваме съществуващата езикова поточна линия на GATE, която е вградена в системата KIM.

Тъй като очакванията са системата, с помощта на авторите и читателите, да създава много точни анотации, една възможност е класификационен алгоритъм да бъде обучен чрез първоначално създадените със системата документи. Така при постоянно усъвършенстване, алгоритъмът на практика ще използва натрупаното знание от всички предходни документи.

Авторите и читателите са приканвани да свържат добавените от тях анотации към съществуващи инстанции в базата данни. Така по естествен начин базата се обогатява с имена, неизвестни досега. Тези имена влизат в

семантично обогатените списъци и по този начин биват разпознавани при следващи анотации. Вижда се, че с малко усилия от страна на всеки автор (или читател), ще се създават големи количества полезни факти и ще се попълват знанията на системата.



Фигура 2. *Уеб интерфейс за ръчно създаване на анотации в документа.*

3.2. Представяне на данните

При създаване на анотации, системата се уведомява, с цел да бъдат намерени необходимите факти, свързани с тях. Това също става в реално време, в специално създаден за целта прозорец.

Фактите се представят като списък от тройки със субект, предикат и обект. Списъкът се запълва, като за всеки намерен именувани обект се показват най-често използваните факти. При желание от страна на автора, могат да бъдат показани повече. При избор на някой от фактите, той се асоциира с анотацията и се добавя към данните за документа. Така при разглеждане на документа, данните асоциирани в него ще бъдат показвани в списък отстрани, като ще остава ясно кой факт за коя анотация се отнася.

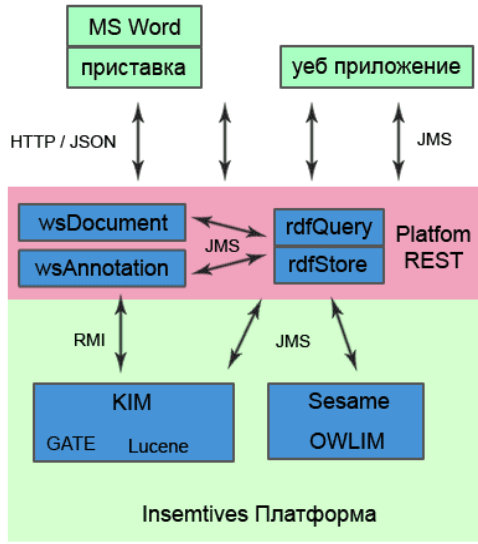
3.3. Читателски интерфейс

Предложената система цели да подобри сътрудничеството между авторите, но също така да включи и читателите в процеса на добавяне на знания. Създаването на wiki-подобна структура със сложни групи и права е извън пределите на тази работа. Нещо, което досега не е правено на практика обаче, е масова колаборация за попълване на база от знания в процеса на

създаване на документи. Теоретично, масовото създаване на база от знания е описано в статията на Ричардсън и Домингос [14], а задачата на проекта Insemtives е свързана с генериране на данни за вече съществуващи ресурси. Добавянето на знания е пряко свързано с писането на книги и статии, затова представеният тук подход се очаква да е успешен. Целта е да се представи един мета-слой от знания за всеки документ, който може да бъде актуализиран и поправян от авторите и от всеки читател. Така всеки читател, който знае нещо интересно по темата, ще може да го добави като факт. Способността, да разглеждаме фактите като обекти, позволява да се добавят например знания кой потребител е добавил факта, има ли потребител, който смята факта за неверен и т.н. Подредбата на факта основана на тяхното социално използване, позволява естествено пресяване на важните факти. Така лесно може да се дефинира „достоверен” факт, като факт, който е или добавен от достоверен автор, или има много висок рейтинг.

3.4. Програмна реализация

За да се създаде една работеща система, изпълняваща описаните функции в пълен размер е изключително трудна задача. Затова разработихме прототип, демонстриращ основните компоненти, без да претендираме за завършеност.



***Фигура 3**. Архитектура на приложението*

На фигура 4 е показана цялостната **архитектура** на предложеното приложение. Както вече отбелязахме, то се базира на платформата Insemtives, описана по-рано. Макар да изглежда, че се използва голяма част от платформата, всъщност основните усилия бяха насочени към интегриране на KIM сървъра. Модулите KIM и OWLIM са изцяло вътрешни за системата, като комуникацията към тях става по показания начин чрез съобщения.

4. Заключение

Настоящата статия, прави сравнителен анализ на значимите публично достъпни системи, за семантично анотиране. От това изложение ясно се вижда нуждата от система, която интегрира създаването на семантични знания с процеса на създаване на документи. Предложената система, запълваща тази празнина, базирана на отворен код и лесно може лесно да се развива за различни приложения. Възможно бъдещото развитие на системата включва: усъвършенстване на потребителският интерфейс, за да могат да се използват от практикуващи писатели; свързваме с други бази от знания, например, интегриране с FactForge [18] и др.

Благодарности: Това изследване е подпомогнато от проект „Умна Книга”, финансиран от Фонд „НИ“ към МОМН, договор № D002-111/15.12.2008 г.

Литература

1. Koychev I., Dicheva D. and Nikolov, R.: SmartBook: Semantics Inside. Serdica Journal of Computing. (4), 2010, p. 263-278.
2. Kiryakov, Atanas, et al. Semantic annotation, indexing, and retrieval. 2004, Journal of Web Semantics.
3. Handschuh, S., Staab, S. and Ciravegna, F. S-CREAM - Semi-automatic CREAtion of Metadata. Springer-Verlag, 2002. EKAW '02. pp. 358-372.
4. Dow Jones. Factiva. [Online] http://www.dowjones.com/factiva/index.asp.
5. Nadeau, D. and Sekine, S. A Survey of Named Entity Recognition and Classification.. 2007, Lingvisticae Investigationes, pp. 3-26.
6. Ferrucci, D. and Lally, A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. 2004, Natural Language Engineering, pp. 327-348.
7. Bontcheva, K., et. al. Evolving GATE to Meet New Challenges in Language Engineering. 2004, Natural Language Engineering, pp. 349-373.
8. Popov, B., et al. KIM – a semantic platform for information extraction and retrieval.. 2004, Natural Language Engineering, pp. 375 - 392.
9. Thomson Reuters. OpenCalais. [Online] http://www.opencalais.com/.
10. Alchemy API. [Online] http://www.alchemyapi.com/.

11. Cunningham, H., et al. GATE: A framework and graphical development environment for robust NLP tools and applications. 2002. 40th Anniversary Meeting of the Association for Computational Linguistics.
12. Terziev, Ivan, Kiryakov, Atanas and Manov, Dimitar. D1.8.1. Base upper-level ontology (BULO) Guidance. Ontotext, 2008. SEKT Deliverable.
13. Richardson, M. and Domingos, P. Building large knowledge bases by mass collaboration.: ACM New York, 2003. K-CAP '03. pp. 129-137.
14. Nozhchev, Marin, et al. D3.2.2. Semantic Content Management Platform (final version). 2011. Insemtives Deliverable.
15. Chenkova, E. and Koychev, I. A Just-In-Time Information Retrieval Agent that Provides Context Aware Support of Text Creation.. Sofia : ,2009. S3T'09.
16. http://dbpedia.org/About
17. http://www.geonames.org/
18. http://factforge.net/

A SOCIAL APROACH TO SEMANTIC ANNOTATION OF E-BOOKS

Yasen Kiprov, Ivan Koychev and Boris Kraychev

Faculty of Mathematics and Informatics, University of Sofia "St. Kliment Ohridski"
koychev@fmi.uni-sofia.bg

Abstract: *Web 3.0 aims to create a common interface for publishing different knowledge bases and connecting them through common ontology, which creates prerequisites for building a new generation of intelligent systems. These technologies provide the opportunity also for e-books to become "smarter." The present work aims to create a integrated set of tools which facilitate the incorporation of explicit semantics in e-books. The article reviews the existing semantics technologies and software tools for semantic annotation of text. Further it describes a prototype system that allows semantic annotation of e-books. The system integrates Insemtives and KIM platforms.*