СПЕЦИАЛИЗИРАН НАУЧЕН СЪВЕТ ПО ИНФОРМАТИКА И МАТЕМАТИЧЕСКО МОДЕЛИРАНЕ ПРИ ВАК

Светлин Иванов Наков

АВТОМАТИЧНО ИЗВЛИЧАНЕ НА ФАЛШИВИ ПРИЯТЕЛИ ОТ ПАРАЛЕЛЕН ДВУЕЗИЧЕН КОРПУС

АВТОРЕФЕРАТ

на

ДИСЕРТАЦИЯ

за присъждане на образователна и научна степен “Доктор”
по научна специалност: 01.01.12 “Информатика”

Научен ръководител:
ст. н. с. II ст. д-р Елена Паскалева

София, 2009

Дисертационният труд е обсъден на заседание на катедра "Информационни технологии" към Факултета по математика и информатика на Софийски университет "Св. Климент Охридски" на 30.09.2009 г. и насочен за защита пред СНС по информатика и математическо моделиране по специалност 01.01.12 (Информатика).

Публичната защита на дисертационния труд ще се състои на 12.04.2010 г. от 16:00 часа в мултимедийната зала на ИМИ към БАН на открито заседание на СНС по информатика и математическо моделиране.

Материалите по защитата са на разположение в библиотеката на ИМИ към БАН (София, ул. акад. Георги Бончев, бл. 8).

Пълният обем на дисертацията е 142 страници, от които заглавната страница, съдържанието, използваната литература и приложенията са 13 страници. Дисертацията се състои от увод, пет глави, заключение, авторска справка, приложения и списък с цитираната литература.

Използваната литература включва 99 заглавия (85 на английски език, 6 на български език, 8 на руски език и 4 електронни публикации). Списъкът от публикации на докторанта по същината на дисертацията включва 7 заглавия.

Въведение

В компютърната лингвистика когнатите са двойки думи от различни езици, които звучат или се изписват по подобен начин и имат еднакво или сходно значение, например думата *вода* в български и в руски език. Фалшивите приятели, подобно на когнатите, са двойки думи, които звучат или се изписват по подобен начин, но за разлика от когнатите, имат различни значения. Например българската дума *майка* на руски език означава *потник*, а руската дума *быстрота* означава *бързина*, а не *бистрота*. Фалшивите приятели са от особен интерес за преводачите, а и за всеки изучаващ чужд език, защото лесно могат да породят грешна аналогия с близка по звучене дума в другия език.

В настоящата дисертация са изследвани когнатите и фалшивите приятели между български и руски език и са предложени алгоритми за тяхното извличане. Разработени са нови алгоритми за измерване на ортографска и семантична близост в едноезикова и междуезикова плоскост и са демонстрирани приложенията им при решаването на различни задачи от компютърната лингвистика: извличане на синоними, различаване между когнати и фалшиви приятели и подравняване на думи. Предложен е двустъпков метод за автоматично извличане на фалшиви приятели от паралелен двуезичен корпус: на първата стъпка се търсят двойки думи език със сходно изписване, след което всяка от двойките се категоризира като когнати или като фалшиви приятели на базата на измерване на междуезиковата семантична близост с използване на уеб като корпус и прилагане на статистически техники, отчитащи броя срещания на двойките думи – заедно в съответни изречения и поотделно.

Актуалност на темата

Темата за когнатите и фалшивите приятели в български и в руски език е интересна и актуална както от лингвистична гледна точка, така и като проблем на компютърната лингвистика. За лингвистиката съставянето на пълни речници на когнатите и фалшивите приятели е проблем, който вълнува изследователи от десетки години, тъй като такива речници помагат на изучаващите езика да избягват грешки под влияние на друг език, който владеят по-добре. За компютърната (или известна още като *изчислителна*) лингвистика автоматичното извличане на когнати и разграничаването им от фалшивите приятели има съществено приложение при решаването на редица задачи като подравняване на думи, машинен превод, автоматично построяване на речници, автоматично резюмиране на текст, представяне и моделиране на знания, избор на правилното значение измежду няколко възможни, разширяване на заявки за търсене и при много други задачи.

Изключително важна задача за компютърната лингвистика е моделирането на семантиката в езика и построяването на лексикални семантични ресурси (речници, лексикални онтологии и семантични мрежи). Семантични ресурси и инструменти се използват при решаването на много задачи от обработката на естествен език: резюмиране на текст, категоризиране на текст, извличане на информация от текст, отговаряне на въпроси, извличане на знания, извличане на факти, разрешаване на многозначност, перифразиране на текст и идентификация на перифразиран текст, генериране на текст, подравняване на думи, машинен превод, семантичен анализ и други. Интересът към изграждането и използването на семантични ресурси и инструменти за извличане и анализиране на семантични връзки се доказва и със значителния брой научни публикации, свързани с тях. Например,

само публикациите, свързани с WordNet – един от основните семантични ресурси за английски език – към август 2009 г. са над 10 000 (според Google Scholar). По тази причина изследванията и разработването на нови алгоритми и инструменти за измерване на едноезикова и междуезикова семантична близост, както и проучването и разработването на подходи и инструменти за извличане на когнати и фалшиви приятели допринасят съществено за компютърната лингвистика и обработката на естествен език.

Болшинството лексикални ресурси, достъпни в Интернет, и изследванията, свързани със семантиката, когнатите и фалшивите приятели, се отнасят за западноевропейските езици и най-вече за английски език. Езиковите ресурси, достъпни в електронен вид, свързани с източноевропейските езици, са крайно недостатъчни за нуждите на съвременната компютърна лингвистика, поради което създаването на инструменти и ресурси за тях е актуална тема.

Семантичната близост и подходите, алгоритмите и инструментите за нейното измерване привличат интереса на голям брой учени както от областта на компютърната лингвистика и обработката на естествен език, така и от областта на биоинформатиката и генетиката. Доказателство за това са многото научни публикации, свързани с измерване на езикова семантична близост, моделиране и използване на лексикални онтологии и семантични мрежи, както и моделиране на генна онтология и семантична близост между части от гени. Само публикациите, свързани с езикова семантична близост, към август 2009 г. са над 5 000 (според Google Scholar). По тази причина развитието на съществуващите и създаването на нови подходи, алгоритми и инструменти за измерване на семантична близост води до пряк напредък при решаването на голям брой задачи. Това прави изследванията в областта на езиковата семантична близост изключително актуална тема.

Цели и задачи дисертацията

В настоящата дисертация са поставени следните основни цели:

- **разработване на алгоритъм за измерване на междуезикова семантична близост;**
- **разработване на алгоритъм за извличане на фалшиви приятели от паралелен двуезичен корпус.**

Във връзка с основните цели се поставят следните конкретни задачи:

- да се проектира, опише, реализира, тества и изследва алгоритъм за измерване на ортографска близост между двойка българска и руска думи, съобразно лингвистичните особености на двата езика;
- да се проектира, опише, реализира, тества и изследва алгоритъм за измерване на семантична близост между думи от един и същ език;
- да се проектира, опише, реализира, тества и изследва алгоритъм за измерване на семантична близост между думи от различни езици;
- да се проектира, опише, реализира, тества и изследва алгоритъм за различаване между когнати и фалшиви приятели;
- да се проектира, опише, реализира, тества и изследва алгоритъм за извличане на фалшиви приятели от паралелен двуезичен корпус.

Решението на поставените задачи е разработено в контекста на конкретна двойка езици – български и руски. Изборът е мотивиран от следните причини:

- фалшивите приятели между български и руски език са обект на дългогодишни изследвания в лингвистиката, но досега не са били изследвани от компютърни лингвисти;
- българският и руският са родствени езици и между тях има голям брой когнати и фалшиви приятели, което прави задачата интересна.

Глава 1. Когнати и фалшиви приятели, ортографска и семантична близост – обзор

В обзорната първа глава са дефинирани понятията *когнати* и *фалшиви приятели,* както и някои техни подвидове в зависимост от принципите на класификацията им. Разгледани са и по-важните съществуващи алгоритми за измерване на ортографска и семантична близост между двойки думи, както и алгоритми за търсене и извличане на фалшиви приятели от текстови корпуси.

Определение за когнат

Лингвистите дефинират *когнатите* като думи с общ исторически произход (в един и същ език или в различни езици). Според тях когнатите имат общ корен, наследен от дума-предшественик от езика, от който произхождат.

Изследователите от компютърната лингвистика използват опростена дефиниция на термина *когнати*, като игнорират историческия произход на двойката думи и взимат предвид единствено тяхното изписване и семантика. Те дефинират *когнатите* като ортографски близки думи от различни езици, които са превод една на друга или имат близки значения.

В настоящата дисертация се възприема именно тази последна дефиниция от компютърната лингвистика, тъй като обект на изследване са ортографските и семантичните сходства между думите, без оглед на произхода им:

> **Дефиниция:** *Когнати* наричаме двойка думи от различни езици, които имат сходно изписване и са близки по значение.

Истински и фалшиви когнати

Когнати (англ. *cognates,* фр. *vrais amis)* са двойки думи от различни езици, които се изписват или звучат сходно и имат близки значения. Често се наричат още *истински приятели* или *истински когнати*. Според дефиницията на [Simard & колектив, 1993] "*когнатите са двойки думи от различни езици, които споделят очевидни фонологични, ортографски и семантични сходства и поради това често се използват като превод една на друга*".

Примери за когнати с идентично изписване:

- българската дума *вода* и руската дума *вода*
- френската дума *train* и английската дума *train*

Примери за когнати със сходно изписване:

- българската дума *слънце* и руската дума *солнце*

- испанската дума *banco* и английската дума *bank*

Фалшиви приятели (англ. *false friends,* фр. *faux amis*) са двойки думи от различни езици, които имат сходно изписване или звучене заради което често се възприемат като близки, но имат различно значение.

Примери за фалшиви приятели, които имат идентично изписване (за примерите с кирилица и латиница – идентично след транскрибиране):

- българската дума *позор* и чешката дума *pozor*, която на български означава *внимание*
- българската дума *март* и английската дума *mart*, която на български означава *търговски център*
- българската дума *прост* и немската дума *Prost*, която на български означава *наздраве*

Примери за фалшиви приятели, които имат сходно, но не идентично изписване:

- българската дума *бистрота* и руската дума *быстрота*, която означава на български *бързина*
- българската дума *дрян* и руската дума *дрянь*, която означава на български *вехтории*; *глупости*; *дреболия*

Следва дефиниция на термина *фалшиви приятели*, която ще бъде приета и използвана в цялата дисертация:

> **Дефиниция:** *Фалшиви приятели* (*фалшиви когнати*) наричаме двойка думи от различни езици, които имат сходно изписване и различни значения.

Частични когнати

Пълни когнати (едновременно съвпадане или прилика на буквения състав и на значенията – преводи на думата) са най-често еднозначните думи (с един превод в речника). При повече значения (а следователно и преводи) може да се получи *частичен когнат*, т.е. съвпадане или прилика на буквения низ на думата с един от многото й преводи.

Например руската дума *машина* се превежда на български с 3 думи – *машина, лека кола, двигател*, само една от които съвпада с нейното изписване на руски. Това ни позволява да я определим като *частичен когнат* на българското *машина*. Ако в обратната посока разгледаме пък българската дума *бас* и нейните руски преводи (*бас* и *пари*, в значение – *мъжки глас* и *облог*), само един от които съвпада с нея, виждаме, че тя е частичен когнат на руското *бас*.

Определението за *частични когнати (partial cognates)* на [Frunza & Inkpen, 2006], изхожда не от речниковите значения на двойката думи, а от възможната им употреба в речта: "*частични когнати са двойки думи, които имат близко изписване и означават едно и също в някои контексти, но се различават по значение в други контексти*". Следва дефиниция на термина *частични когнати*, която ще бъде приета и използвана в настоящата дисертация:

> **Дефиниция:** *Частични когнати* наричаме двойка думи от различни езици, които имат сходно изписване и в някои случаи означават едно и също (т.е. се използват като преводни съответствия), а в други случаи се различават

по значение.

Ортографска близост

Понятието ортографска близост между две думи означава близост между буквения им състав. Пълната близост означава съвпадане на двата буквени низа, т.е. идентичност (като *emotion* за френски и английски, независимо от различния им звуков състав при произнасяне).

Формалното измерване на ортографска близост е свързано със съпоставянето на число в интервала [0; 1] за двойка думи от един език (или от различни езици, но изписани с една и съща азбука), такова, че то е по-голямо, когато думите съвпадат побуквено в по-голямата си част и по-малко, когато думите имат по-малко съответстващи си букви. При пълна идентичност в изписването съпоставената мярка за ортографска близост трябва да е 1, а при пълно несъвпадение на буквите в състава на двете думи мярката трябва да е 0. При частично съвпадение на една или няколко последователности от букви мярката трябва да съпоставя число между 0 и 1. Следва дефиниция на термина *ортографска близост*, която ще бъде приета и използвана в настоящата дисертация:

Дефиниция: *Ортографска близост* между две думи от различни езици наричаме степента на сходство в изписването на двете думи, измерена като число в интервала [0; 1].

Семантична близост

Семантична близост между двойка думи се дефинира като сходство между множествата от техните значение. Естественото възприятие на значенията на думите в човешкото съзнание почти винаги е в състояние да даде отговор на въпроса дали дадена двойка думи са по-близки семантично от друга двойка думи. Например, лесно може да се прецени дали *китаец* и *пералня* или *саксия* и *цвете* са по-близки семантично. Това означава, че формалната мярка за семантична близост трябва да съпоставя по-ниска степен на близост за *китаец* и *пералня*, отколкото за *саксия* и *цвете*.

Формалното измерване на семантична близост изисква съпоставка на някакво число (мярка), което отразява степента на близост между множествата от значенията на думите. Това число трябва да е най-голямо при синоними, по-малко при други семантично свързани думи (например при хипоними) и най-малко или нула, ако между думите няма никаква или почти никаква смислова връзка.

Усредненатa човешка оценка за семантичната близост между двойка думи може да бъде измерена като двойката думи бъде оценена по някаква скала (например с числа от 0 до 4) от внимателно подбрана представителна извадка на хората владеещи съответния език, включваща участници от различни възрастови групи, пол, географско разположение, образование, професия, етническа принадлежност и социален статус, след което оценките бъдат усреднени и съотнесени към интервала [0; 1]. Следва дефиниция на термина *семантична близост*, базирана на концепцията за усредняване на човешката оценка, която използвана в настоящата дисертация:

Дефиниция: *Семантична близост* между две думи наричаме степента на

сходство между множествата от техните значения и тя може да бъде формално измерена с число в интервала [0; 1] чрез *усреднена човешка оценка*.

Мярката за семантична близост между две думи от различни езици би трябвало да има голяма степен на корелация с оценката за семантична близост, дадена от човек, който владее на добро ниво двата езика. Както човек, владеещ български и руски език би могъл да обясни, че *китаец* и *стиральная машина* (пералня) семантично са по-далечни, отколкото *цвете* и *цветочный горшок* (саксия), така и мярката за близост трябва да съпостави съответно по-малка и по-голяма стойност за двойките думи от примера. Въз основа на тези разсъждания може да бъде дадена дефиниция за термина *междуезикова семантична близост*, която е използвана в настоящата дисертация:

Дефиниция: *Междуезикова семантична близост* между две думи от различни езици наричаме степента на сходство между множествата от техните значения, и тя може да бъде измерена с число между 0 и 1 чрез усреднени човешки оценки.

Предишни разработки в проблемната област

Като част от обзора в първа глава са проучени съществуващите алгоритми за измерване на ортографска близост, измерване на семантична близост и извличане на фалшиви приятели и е направено заключение, че те не могат да бъдат приложени директно за български и руски език, поради което е необходимо да бъдат специално адаптирани или да бъдат предложени нови алгоритми.

Глава 2. Измерване на модифицирана ортографска близост между български и руски език

Във втора глава е описано разработването на алгоритъм MMEDR, който измерва близост в изписването или звученето на двойка думи – българска и руска. Алгоритъмът съобразява типичните изменения на морфеми и фонеми при преминаването им от руски към български език. Използва се разбирането, че освен думи с пълна ортографска идентичност и думи с различно изписване също могат да се възприемат като близки, когато съдържат еднаква или близка стема като например думите ***афектир**ахме* (в български) и ***аффектир**овались* (в руски) и думите ***уч**а* (в български) и ***уч**иться* (в руски).

Разработеният алгоритъм, измерва до каква степен дадена българска и дадена руска дума се възприемат като близки от човек, който владее двата езика. Тази мярка е различна от традиционната ортографска близост, която взима предвид единствено изписването на думите, и е различна от традиционната фонетична близост, която оценява единствено сходство в звуците. Тя е наречена *модифицирана мярка за ортографска близост* поради липса на по-подходящ термин.

Описание на алгоритъма

Алгоритъмът MMEDR извършва няколко стъпки. Първо прилага лематизация и привежда българската и руската дума към основната им форма. След това транслитерира руската дума до българската азбука (например заменя буквата "*ы*" с "*и*"). Следва прилагането на лингвистично мотивирани правила за трансфор-

мация на n-грами. Те се отнасят най-вече за окончанията на думите. Например прилагателното "*вечный*" се трансформира във "*вечен*", а глаголът "*ходить*" в "*ходя*". Накрая се измерва нормализирано разстояние на Левенщайн (известно още като NED /normalized edit distance/ и MEDR /minimum edit distance ratio/) с тегла при замяната на букви. Теглото е необходимо, защото замените на букви не са еднакво вероятни. Например замяната на "*у*" с "*ъ*" се случва често за разлика от замяната на "*а*" с буквата "*щ*", която на практика никога не се случва. Следва описание на алгоритъма MMEDR в стъпки:

***Алгоритъм:* MMEDR.**
Вход: българска дума w_{bg}; руска дума w_{ru}. *Изход:* мярка за ортографска близост между думите $MMEDR(w_{bg}, w_{ru})$.
Стъпки: Стъпка 1. Лематизация на българската дума w_{bg} (незадължителна стъпка). Стъпка 2. Лематизация на руската дума w_{ru} (незадължителна стъпка). Стъпка 3. Трансформация на окончанието на руската дума w_{ru} (незадължителна стъпка). Стъпка 4. Транслитерация на руската дума w_{ru}. Стъпка 5. Премахване на някои двойни съгласни от w_{bg} и w_{ru}. Стъпка 6. Изчисляване на модифицирано разстояние по Левенщайн (MED) с тегла при замяна на букви. Стъпка 7. Нормализиране и изчисляване на MMEDR.

Експерименти и резултати

Като тестови данни са използвани думи от романа на руския писател А. Беляев "Владетелят на света" [Беляев, 1940a] и съответния му превод на български език [Беляев, 1940b]. От двата текста са извадени съответно първите 200 различни български думи и руски думи и между всяка от получените 40 000 двойки думи е измерена близостта по алгоритъм MMEDR. Измежду тях 163 двойки думи са оценени от лингвист като възприемани като сходни, а останалите 39 837 са оценени като различни. Алгоритъмът MMEDR подрежда извлечените 40 000 двойки думи в намаляващ ред спрямо изчислената близост. Очакването е на първите 163 позиции да застанат думите, които се възприемат като сходни, а на останалите позиции – останалите думи. Полученият списък се оценява с мярката за оценка на алгоритми за извличане на информация *усреднена интерполирана точност в 11 точки (11 point interpolated average precision)*.

Алгоритъм MMEDR е сравнен с класическите мерки за ортографска близост LCSR (longest common subsequence ration) и MEDR (minimal edit distance ratio), както и със случайната наредба (означена в резултатите като RAND и служеща като долна граница). Получени са следните резултати:

Алгоритъм	**Усреднена интерполирана точност в 11 точки**
RAND	0,31%
LCSR	69,06%

MEDR	72,30%
MMEDR	**90,58%**

Глава 3. Измерване на семантична близост

В трета глава е описан и изследван алгоритъм за автоматично измерване на семантична близост между двойки думи от различни езици (български и руски) чрез използване на уеб като корпус. Алгоритъмът е предложен в два варианта: алгоритъм SemSim за измерване на едноезикова семантична близост и алгоритъм CrossSim за измерване на междуезикова семантична близост.

Описание на алгоритмите

Предложените алгоритми изпълняват заявки в уеб търсеща машина и анализират върнатите като резултат отрязъци от текстове. От тях извличат т. нар. *локален контекст* на всяка анализирана дума (думите в непосредствена близост около нея), тъй като той съдържа думи, които са смислово свързани с нея [Hearst, 1991]. Колко думи наляво и надясно се взимат е параметър, наречен *размер на контекста*. По извлечените локални контексти за всяка дума построяват *контекстен честотен семантичен вектор*, който представлява множество от двойки {дума, брой срещания}. Заради флексивността на българския и руския език се прилага лематизация и векторите съдържат само основни форми на думи. За получаване на мярката за семантичната близост между двойката разглеждани думи се измерва косинус между техните честотни семантични вектори. Когато разглежданите думи са от различни езици, техните контексти (които също са на различни езици) се сравняват като предварително единият контекст се превежда на другия език чрез речник. Алгоритмите може да се ползват за измерване на семантична близост не само между думи, но и между фрази.

Пример: направено е търсене в Google на думата "*рокля*" на български език. Резултатът би могъл да изглежда по следния начин:

Изборът на булчинска **рокля** « Всичко за сватбата
Сега е времето за най-забавната задача от списъка за организацията на сватбата - избирането на вашата булчинска **рокля**. Това е важно, защото показва изцяло ...

Покупка на подходяща бална **рокля** :: Advise
Обикновено покупката на подходяща бална **рокля** започва с едно (не)разумно пазаруване. Но това не означава да пазаруваме само заради цената на дадена **рокля**...

Дълга **рокля** - 8-ми Март - секс магазин
Дълга **рокля** със сребърни отблясъци. Универсален размер (45-80 кг.). Ръчно пране в хладка вода ... Комплект бродирана **рокля** и прашки. Най - гледани продукти: ...

...

Думите от локалния уеб контекст на "*рокля*" на фигурата са оградени в правоъгълник. Ако по аналогичен начин бъде направено търсене и за думата "*блуза*", могат да бъдат извлечени следните семантични честотни контекстни вектори:

рокля	
рокля	422
сватбен	262
бална	202
булчински	167
вечерен	94
черен	84
купувам	72
ваш	56
червен	37
...	...

блуза	
блуза	461
дамски	386
женска	345
вълнен	205
памучен	183
поръчвам	176
класически	188
магазин	98
фирма	12
...	...

Сходството между тях може да бъде измерено чрез измерване на косинус между вектори, като множеството от всички различни думи бъде използвано за техни размерности, а за координати се ползват броят срещания на всяка от думите.

Размер на контекста

Един от параметрите на алгоритъма за измерване на семантична близост чрез уеб контексти е размерът на контекста. Това е цяло положително число, което указва колко думи вляво и вдясно от търсената дума да бъдат считани за част от нейния локален контекст. Това може да е една дума, две, три или повече думи.

Лематизация

Българският и руският език са силно флексивни, т.е. използват окончания, за да изразяват различни форми на една и съща дума. Думите получават различно окончание в зависимост от род, число, падеж, глаголна форма, членуване и др. *Лематизация* означава преминаване към основната форма на думата и помага да се уеднаквят различните форми на една и съща дума, срещани в локалния контекст, извлечен от уеб.

За лематизация се използват богати граматични речници на българския и руския език (с над 960 000 словоформи за български и с над 1 390 000 словоформи за руски език). Всяка дума, която попада в контекста на търсената дума, се заменя с нейната основна форма (лема). Когато за една дума има няколко леми (например думата *гори* с леми *гора* и *горя*), се взимат всичките ѝ леми, защото няма как да се определи точно коя от тях е била в локалния контекст.

Лематизация на заявката

Техниката *лематизация на заявката* работи по следния начин: вместо да се изпълнява единично търсене в уеб търсеща машина за дадена дума, се изпълнява поредица търсения за всяка от словоформите на дадената дума и от резултати за всяко от търсенията се извлича по-богат семантичен контекст.

TF.IDF претегляне

При извличане на информация (information retrieval) често пъти се прилага т. нар. TF.IDF претегляне на честотите на отделните думи. Числото TF.IDF (term frequency – inverse document frequency) е статистическа мярка, която измерва колко е важна определена дума за даден документ от даден корпус с документи. Важността на думата се увеличава пропорционално спрямо броя на срещанията ѝ в документа, но намалява пропорционално на броя документи, които я съдържат.

Обратен контекст

Техниката на *обратния контекст* се основава на идеята, че ако две думи са семантично близки, то първата трябва да се среща често в локалния контекст на втората и същевременно втората трябва да се среща често в локалния контекст на първата. Например в семантичния контекст на думата *картина* често се срещат думи като *художник*, *галерия* и *изкуство*, но и паразитни думи като *поръчвам*, *новини* и *сайт*. Ако бъде извършено търсене за първите три думи, ще се установи, че *картина* се среща често в техните семантични уеб контексти. Ако, обаче, бъде извършено търсене за последните три думи, ще се установи, че в техните семантични уеб контексти *картина* почти не се среща.

Нека с F(x, y) е означен броят срещания на думата y в семантичния уеб контекст на думата x. Нека е дадена думата a и са извлечени думите w_i от нейния локален уеб контекст заедно с броя им срещания F(a, w_i). Нека за всяка дума w_i бъде извлечено числото F(w_i, a) – броят срещания на думата a в уеб контекста на w_i (наречено е *обратен контекст*). Получава се вектор на взаимните срещания на думата a с всички думи от нейния семантичен уеб контекст. Той е съставен от думите w_i с честоти min(F(a, w_i), F(w_i, a)). Полученият вектор съдържа минималният брой взаимни срещания на дадена дума с всяка друга дума от нейния уеб контекст и съдържа по-точна семантична информация в сравнение с чистия уеб контекст. При изчисляването на вектора на взаимните срещания е добре да се игнорират думи, които се срещат прекалено малко на брой пъти (примерно по-малко от 10), защото това може да е случайно. С промяна на този параметър (*праг на честотата*) може да се влияе върху точността на резултатите.

Обогатяване на контекста

Техниката *обогатяване на контекста*, наричана още *използване на индиректен контекст*, предложена от [Hagiwara & колектив, 2007], препоръчва към семантичния контекст на дадена дума да се добавят думите от семантичните контексти на всички често срещани думи от нейния контекст. По този начин семантичният контекст на думата се разширява с още думи, които оригинално не присъстват в него, но са свързани смислово с тази дума.

При обогатяване на контекста, е добре да се игнорират думи, които се срещат в него прекалено малко на брой пъти (примерно по-малко от 10), защото тяхната поява може да е случайна. С промяна на този параметър (*праг на честотата*) може да се влияе върху точността на резултатите.

Алгоритъм WebExtract

Следва описание в стъпки на алгоритъма WebExtract за извличане на контекстен честотен семантичен вектор за дадена дума от уеб:

Алгоритъм: **WebExtract.**

Вход: дума *w*; език *L*.

Изход: контекстен честотен семантичен вектор F_w.

Настройки: размер на контекста *size*; използване на лематизация (да/не).

Стъпки:

Стъпка 1. Изпраща се заявка към Google за търсене на думата *w* при език *L*. Извличат се максималният брой резултати от търсенето (до 1 000), всеки от които се състои от заглавие и отрязък от текст. Съвкупността от тези заглавия и отрязъци формира мултимножество *R* от символни последователности (текстове).

Стъпка 2. В мултимножеството *R* се заменят се всички главни букви с малки.

Стъпка 3. Символните последователности $r \in R$ се разделят на списъци от думи. За разделител се използват всички символи, които не са букви от азбуката на езика *L* (за български и руски това са символите, които не са букви от кирилицата). Резултатът е мултимножество от списъци с думи *S*.

Стъпка 4. От всеки от списъците с думи $s \in S$ се премахват думите с дължина по-малка от 3 символа.

Стъпка 5. От всеки от списъците с думи $s \in S$ се премахват всички функционални думи за езика *L* (предлози, местоимения, съюзи, частици, междуметия и някои наречия).

Стъпка 6. Започва се от мултимножество от думи $C = \varnothing$. Преминава се през всички списъци от думи $s \in S$ и се търси в тях думата *w*. При всяко срещане на *w* в *s* се взимат до *size* на брой думи вляво и вдясно от намерената дума и се прибавят към *C*, като се позволява всяка дума от *s* да участва в *C* най-много веднъж.

Стъпка 7. Ако лематизацията е включена, се прилага лематизация на всички думи от мултимножеството *C*. Лематизацията се извършва чрез замяна на всяка дума от *C* с нейната основна форма (лема). Ако дадена дума има няколко леми, се замества с множеството от тях.

Стъпка 8. Построява се векторът F_w като в него се включват всички думи от мултимножеството *C* заедно с броя им срещания в *C*.

Алгоритъм SemSim

Следва описание в стъпки на алгоритъма SemSim за измерване на семантична близост от уеб:

Алгоритъм: **SemSim.**

Вход: едноезични думи w_1 и w_2.

Изход: мярка за семантична близост между думите $SemSim(w_1, w_2)$.

Настройки: език, размер на контекста; използване на лематизация (да/не); лематизация на заявката (да/не); използване на TF.IDF претегляне (да/не); използване на обратен контекст (да/не); обогатяване на контекста (да/не); праг на честотата при обратен или обогатен контекст.

Стъпки:

Стъпка 1. По алгоритъма WebExtract се извлича контекстен честотен семантичен вектор F_1 за думата w_1 от откъсите текст в първите 1 000 резултата от търсене на w_1 в Google по зададения размер на контекста.

Ако е включена настройката "използване на лематизация", в алгоритъма WebExtract се използва лематизация при извличане на думите от уеб контекста.

Ако е включена настройката "лематизация на заявката", векторът F_1 се извлича не само от думата *w*, а като обединение на контекстните честотни семантични вектори на всички думи, които споделят една и съща лема с думата *w*.

Ако е включена настройката "използване на обратен контекст", се прилага техниката *обратен контекст* върху вектора F_1 със зададения *праг на честотата*.

Ако е включена настройката "използване на обогатен контекст", се прилага техниката *обогатяване на контекста* върху вектора F_1 със зададения *праг на честотата*.

Ако е включена настройката "използване на TF.IDF претегляне", се прилага техниката *TF.IDF претегляне* върху вектора F_1.

Стъпка 2. Аналогично на стъпка 1 се извлича контекстен честотен семантичен вектор F_2 за думата w_2.

Стъпка 3. Изчислява се мярката за едноезикова семантична близост SemSim като косинус между векторите F_1 и F_2.

Измерване на междуезикова семантична близост

Алгоритъмът CrossSim измерва междуезикова семантична близост по следния начин: Първо се извличат контекстните честотни семантични вектори V_{bg} и V_{ru} от уеб за двете думи (единият вектор е на български език, а другият – на руски). След това думите от V_{bg} се превеждат на руски и тогава се сравняват двата вектора. За целта се използва преводен речник (глосарий), който съдържа двойки съответствия между българска и руска дума. При многозначност се взимат предвид всички значения. Използва се лематизация за уеднаквяване на различните словоформи на думите. Следва описание в стъпки на алгоритъма за измерване на междуезикова семантична близост от уеб, наречен CrossSim:

Алгоритъм: **CrossSim.**

Вход: българска дума w_{bg} и руска дума w_{ru}.

Изход: мярка за семантична близост между думите $CrossSim(w_{bg}, w_{ru})$.

Настройки: размер на контекста; използване на лематизация (да/не); лематизация на заявката (да/не); използване на TF.IDF претегляне (да/не); използване на обратен контекст (да/не); обогатяване на контекста (да/не); праг на честотата при обратен контекст.

Стъпки:

Стъпка 1. Чрез алгоритъма WebExtract се извлича контекстен честотен семантичен вектор F_{bg} за българската дума w_{bg} от откъсите текст в първите 1 000 резултата от търсене на w_{bg} на български език в

Google по зададения размер на контекста.

Ако е включена настройката "използване на лематизация", в алгоритъма WebExtract се използва лематизация при извличане на думите от уеб контекста.

Ако е включена настройката "лематизация на заявката", векторът F_{bg} се извличат не само от думата *w*, а като обединение на контекстните честотни семантични вектори на всички думи, които споделят една и съща лема с думата *w*.

Ако е включена настройката "използване на обратен контекст", се прилага техниката *обратен контекст* върху вектора F_{bg}.

Ако е включена настройката "използване на обогатен контекст", се прилага техниката *обогатяване на контекста* върху вектора F_{bg} със зададения *праг на честотата*.

Ако е включена настройката "използване на TF.IDF претегляне", се прилага техниката *TF.IDF претегляне* върху вектора F_{bg}.

От вектора F_{bg} се извлича вектор G_{bg}, който съдържа честотите на всички български думи в преводния речник *G*.

Стъпка 2. Аналогично на стъпка 1 се извлича контекстен честотен семантичен вектор F_{ru} за руската дума w_{ru} от откъсите текст в първите 1 000 резултата от търсене на w_{ru} на руски език в Google. От вектора F_{ru} се извлича вектор G_{ru}, който съдържа честотите на всички руски думи в преводния речник *G*.

Стъпка 3. Изчислява се мярката за междуезикова семантична близост CrossSim като косинус между векторите G_{bg} и G_{ru}.

Семантична близост чрез съвместно срещане с множество думи

Идеята произхожда от [Fung & Yee, 1998], но е адаптирана да използва уеб като корпус. Нека е даден речник (глосарий) G, който се състои от 300 често срещани български и руски думи, превод една на друга ([Fung & Yee, 1998] ги наричат *seed words*). Дадени са българска дума w_{bg} и руска дума w_{ru}, между които трябва да се измери семантична близост. Построява се български вектор V_{bg} и руски вектор V_{ru} с размерности 300, така че всяка тяхна координата съответства на дадена двойка думи (g_{bg}, g_{ru}) от глосария G. Стойността в координатата за g_{bg} във V_{bg} се изчислява като общия брой съвместни срещания на w_{bg} и g_{bg} в уеб, където g_{bg} непосредствено предхожда или непосредствено следва w_{bg}. Това число се извлича чрез търсене в уеб търсеща машина (например Google) със заявки "g_{bg} w_{bg}" и "w_{bg} g_{bg}" и сумиране на броя резултати, които търсещата машина декларира, че е намерила. Аналогично се изчисляват стойностите на вектора V_{ru}. Накрая семантичната близост се изчислява като косинус между векторите V_{bg} и V_{ru}.

Експерименти и резултати

Като тестови данни е използвана адаптация на списъка от 30 двойки думи, предложени от Милер и Чарлз [Miller & Charles, 1991]. В оригинал тези думи са на английски език и представляват внимателно подбрани двойки съществителни имена, за всяка от които е дадена човешка оценка на семантичната близост от 51 души в скала от 0 до 4, след което оценката е усреднена. При адаптацията първата дума е преведена на български език, а втората – на руски:

#	Дума 1	Дума 2	Български пре-вод на дума 1	Руски превод на дума 2	Семантична близост оценена от човек (по Милер и Чарлз)
1	car	automobile	автомобил	автомобиль	3,92
2	gem	jewel	скъпоценен камък	драгоценность	3,84
3	journey	voyage	пътешествие	путешествие	3,84
4	boy	lad	момче	мальчик	3,76
5	coast	shore	крайбрежие	побережье	3,70
6	asylum	madhouse	психиатрия	сумасшедший дом	3,61
7	magician	wizard	магьосник	волшебник	3,50
8	midday	noon	пладне	полдень	3,42
9	furnace	stove	пещ	печь	3,11
10	food	fruit	храна	фрукт	3,08
11	bird	cock	птица	петух	3,05
12	bird	crane	птица	журавль	2,97
13	tool	implement	инструмент	орудие	2,95
14	brother	monk	брат	монах	2,82
15	lad	brother	момче	брат	1,66
16	crane	implement	жерав	орудие	1,68
17	journey	car	пътуване	автомобиль	1,16
18	monk	oracle	калугер	оракул	1,10
19	cemetery	woodland	гробища	лесистая местность	0,95
20	food	rooster	храна	петух	0,89
21	coast	hill	крайбрежие	холм	0,87
22	forest	graveyard	гора	кладбище	0,84
23	shore	woodland	бряг	лесистая местность	0,63
24	monk	slave	калугер	раб	0,55
25	coast	forest	крайбрежие	лес	0,42
26	lad	wizard	момче	волшебник	0,42
27	chord	smile	корда	улыбка	0,13
28	glass	magician	стъкло	маг	0,11
29	rooster	voyage	петел	путешествие	0,08
30	noon	string	пладне	нитка	0,08

Върху 30-те български и руски думи и фрази от таблицата са проведени серия експерименти за оценяване на семантичната им близост чрез изпълнение на описаните алгоритми при различни техни параметри (с различни размери на контекста, с и без лематизация, с различна големина на речниците, с и без прилагане на TF.IDF, с и без използване на обратен контекст, с и без обогатяване на контекста и при различни стойности на минималната честота на срещане на думите). Следвайки учени като [Resnik, 1995] и [Jiang & Conrath, 1997] получените резултати от предложените алгоритми за автоматичното измерване на междуезикова семантична близост са сравнени с човешката оценка чрез изчисление на *коефициента на корелация на Пирсън*.

Използвани ресурси

- Online уеб търсеща машина Google
- Граматичен речник на българския език. Използва се за лематизация на български думи. Съдържа 963 339 словоформи и 73 113 леми.
- Граматичен речник на руския език. Използва се за лематизация на руски думи. Съдържа 1 390 613 словоформи и 66 101 леми.
- Списък с функционални думи (598 български и 507 руски думи – предлози, местоимения, съюзи, частици, междуметия и наречия).
- Кратък българо-руски речник – 4 562 двойки преводни думи.
- Подробен българо-руски речник – 59 582 двойки преводни съответствия.

Влияние на размера на контекста и големината на речника

Алгоритми, участващи в експериментите:

- **CrossSim-1, CrossSim-2, ..., CrossSim-20** – основният алгоритъм CrossSim с краткия българо-руски речник, с размер на контекста 1, 2, 3, …, 20 думи, без използване на обратен или обогатен контекст, с прилагане на лематизация, без лематизация на заявката.
- **CrossSim-MAX** – модификация на CrossSim-1 алгоритъма с безкрайно голям размер на контекста.
- **CrossSim-1-BIG, CrossSim-2-BIG, ..., CrossSim-20-BIG** – модификации на CrossSim-1 алгоритъма с размер на контекста съответно 1, 2, 3, ... 20 думи и с използване на подробния българо-руски речник.
- **CrossSim-MAX-BIG** – модификация на CrossSim-1 алгоритъма с безкрайно голям размер на контекста и с използване на подробния българо-руски речник.

Получени резултати:

Алгоритъм	Корелация
CrossSim-1	0,6598
CrossSim-2	0,6864
CrossSim-3	0,7043
CrossSim-4	0,7143
CrossSim-5	0,7166
CrossSim-6	0,7197
CrossSim-7	0,7213
CrossSim-8	0,7215
CrossSim-9	0,7226
CrossSim-10	**0,7245**
CrossSim-11	0,7238
CrossSim-12	0,7227

Алгоритъм	Корелация
CrossSim-1-BIG	0,6178
CrossSim-2-BIG	0,6202
CrossSim-3-BIG	0,6210
CrossSim-4-BIG	0,6235
CrossSim-5-BIG	0,6257
CrossSim-6-BIG	0,6282
CrossSim-7-BIG	0,6303
CrossSim-8-BIG	0,6302
CrossSim-9-BIG	0,6312
CrossSim-10-BIG	0,6317
CrossSim-11-BIG	0,6319
CrossSim-12-BIG	0,6320

CrossSim-13	0,7219		CrossSim-13-BIG	0,6330
CrossSim-14	0,7222		**CrossSim-14-BIG**	**0,6334**
CrossSim-15	0,7224		CrossSim-15-BIG	0,6334
CrossSim-16	0,7223		CrossSim-16-BIG	0,6334
CrossSim-17	0,7223		CrossSim-17-BIG	0,6334
CrossSim-18	0,7223		CrossSim-18-BIG	0,6334
CrossSim-19	0,7223		CrossSim-19-BIG	0,6334
CrossSim-20	0,7223		CrossSim-20-BIG	0,6334
CrossSim-MAX	0,6598		CrossSim-MAX-BIG	0,6320

Влияние на останалите параметри

За да се оцени влиянието на останалите параметри върху алгоритъма CrossSim са извършени следните експерименти:

- **RAND** – случайна близост, зададена за всички двойки думи (baseline).
- **CrossSim** – основният алгоритъм CrossSim с размер на контекста 10 думи, с краткия българо-руски речник, без използване на обратен или обогатен контекст, с прилагане на лематизация, без лематизация на заявката и без TF.IDF претегляне.
- **CrossSim-NO-STOP-WORDS** – модификация на CrossSim алгоритъма без премахване на функционалните думи (предлози, местоимения и т.н.).
- **CrossSim-NO-LEMMA** – модификация на CrossSim алгоритъма без прилагане на лематизация.
- **CrossSim+QUERY-LEM** – модификация на CrossSim алгоритъма с прилагане на лематизация на заявката.
- **CrossSim+TF.IDF** – модификация на CrossSim алгоритъма с използване на TF.IDF претегляне.
- **CrossSim-REV-0**, **CrossSim-REV-10**, **...**, **CrossSim-REV-50** – модификации на CrossSim алгоритъма с използване на обратен контекст с прагове на честота съответно 0, 10, 20, 30, 40 и 50.
- **CrossSim-IND-0**, **CrossSim-IND-10**, **...**, **CrossSim-IND-50** – модификация на CrossSim алгоритъма с използване на обогатен контекст с прагове на честота съответно 0, 10, 20, 30, 40 и 50.
- **FUNG-YEE** – алгоритъма за измерване на семантична близост чрез съвместно срещане с множество думи (използвани са 300 най-често срещани преводни съответствия).

Получени са следните резултати:

Алгоритъм	**Праг 0**	**Праг 10**	**Праг 20**	**Праг 30**	**Праг 40**	**Праг 50**
RAND	0,0339	-	-	-	-	-
CrossSim	**0,7245**	-	-	-	-	-

CrossSim-NO-STOP-WORDS	0,5893	-	-	-	-	-
CrossSim-NO-LEMMA	0,6818	-	-	-	-	-
CrossSim+QUERY-LEM	0,6970	-	-	-	-	-
CrossSim+TF.IDF	0,7216	-	-	-	-	-
CrossSim-REV	0,5982	0,5804	0,5719	0,5703	0,5725	0,5640
CrossSim-IND	0,4738	0,5047	0,5203	0,6161	0,6504	0,6589
FUNG-YEE	0,5043	-	-	-	-	-

От получените резултати може да бъде направен изводът, че за измерване на междуезикова семантична близост най-добре работи алгоритъмът CrossSim. Всички описани техники за неговото подобряване работят по-лошо.

Глава 4. Приложения на алгоритмите за измерване на семантична близост

В четвърта глава са демонстрирани практически приложения на алгоритмите за измерване на семантична близост при решаването на няколко актуални задачи на компютърната лингвистика: различаване между когнати и фалшиви приятели, извличане на синоними и подравняване на думи.

Различаване между когнати и фалшиви приятели

Алгоритъмът за различаване между когнати и фалшиви приятели директно ползва алгоритъма CrossSim. Като вход се задават две думи от различни езици, които притежават достатъчно ортографско, фонетично или друго сходство, за да се възприемат като близки от хора, владеещи тези езици. Алгоритъмът измерва семантичната близост между двете думи и връща дали те са по-вероятно фалшиви приятели или когнати (пълни или частични).

За експериментите, е подготвен списък от 200 двойки думи (българска и руска), съществителни имена, записани в основната им форма, от които 100 са фалшиви приятели и 100 са когнати. По няколко алгоритъма, които се съпоставят, е изчислена семантична близост между всяка двойка от разглежданите 200 двойки българска и руска думи. В резултат се получават списъци от 200 двойки думи, подредени по изчислената мярка за семантична близост от съответния алгоритъм. Полученият резултат се оценява чрез стандартната мярка усреднена интерполирана точност в 11 точки. Следва описание на участвалите в експериментите алгоритми:

- **RAND** – случайна подредба на двойките думи – база за сравнение с другите алгоритми (baseline).
- **MEDR** – алгоритъм за измерване на ортографска близост MEDR (minimum edit distance ratio).
- **LCSR** – алгоритъм за измерване на ортографска близост LCSR (longest common subsequence ratio).
- **CrossSim** – основният алгоритъм CrossSim с краткия българо-руски речник, с размер на контекста 10 думи, без използване на обратен или обогатен контекст, с прилагане на лематизация, без лематизация на заявката. Размерът

на контекста е 10 думи, защото при предишни експерименти този размер е дал най-добри резултати.

- **CrossSim+BIG** – модификация на CrossSim алгоритъма с използване на подробния българо-руски речник и с размер на контекста 14 думи. Размерът на контекста е 14 думи, защото такъв размер е дал най-добри резултати при предишни експерименти с подробния българо-руски речник.
- **CrossSim+QUERY-LEM** – модификация на CrossSim алгоритъма с прилагане на лематизация на заявката.
- **CrossSim+TF.IDF** – модификация на CrossSim алгоритъма с използване на TF.IDF претегляне.
- **CrossSim+REV** – модификация на CrossSim алгоритъма с използване на обратен контекст и праг на честотата 0.
- **CrossSim+REV+TF.IDF** – модификация на CrossSim алгоритъма с прилагане на обратен контекст с праг на честотата 0 срещания и TF.IDF претегляне.
- **CrossSim+IND** – модификация на CrossSim алгоритъма с използване на техниката "обогатяване на контекста" при прагове на честотата 50, 40, 30, 20 и 10 срещания.
- **CrossSim+BIG+REV** – модификация на CrossSim алгоритъма с използване на подробния българо-руски речник с размер на контекста 14 думи и с прилагане на обратен контекст.
- **CrossSim+BIG+REV+TF.IDF** – модификация на CrossSim алгоритъма с използване на подробния българо-руски речник с размер на контекста 14 думи, с прилагане на обратен контекст и TF.IDF претегляне.
- **FUNG-YEE** – адаптация на алгоритъма на [Fung & Yee, 1998] за измерване на семантична близост чрез съвместно срещане с множество думи.

Получени са следните резултати:

Алгоритъм	**Точност (11pt average precision)**
RAND	50,00%
MEDR	53,84%
LCSR	53,84%
FUNG-YEE	65,50%
CrossSim+IND-10	90,88%
CrossSim+IND-40	91,37%
CrossSim+IND-20	91,56%
CrossSim+IND-50	91,81%
CrossSim+IND-30	92,00%
CrossSim+QUERY-LEM	94,11%
CrossSim+TF.IDF	94,25%
CrossSim	95,24%

CrossSim+BIG+REV	95,40%
CrossSim+BIG+REV+TF.IDF	95,43%
CrossSim+REV	95,78%
CrossSim+REV+TF.IDF	95,79%
CrossSim+BIG	**96,17%**

Извличане на синоними чрез измерване на семантична близост

Предложен е алгоритъм за извличане на синоними, който използва директно алгоритъма SemSim за измерване на едноезикова семантична близост чрез използване на уеб като корпус. За оценяване на ефективността на SemSim при решаването на тази задача са проведени серия експерименти. При всеки от тях е използван списък от 94 думи от терминологията на изобразителното изкуство в руски език, приготвени ръчно от лингвист:

абрис, адгезия, алмаз, алтарь, амулет, асфальт, беломорит, битум, бородки, ваятель, вермильон, возрождение, вохрение, выжигание, высветление, гематит, диамант, жезл, жертвенник, закрепление, инсигнии, искусствоведение, искусствознание, кентавр, киноварь, ковроделие, ковроткачество, колье, контур, кровавик, ландшафт, модельер, модельщик, модерн, мозаичист, мозаичник, муштабель, мягчители, нюанс, ожерелье, оливин, орлец, оседание, основа, основание, осыпание, отпечаток, оттенок, оттиск, пакля, палка, пейзаж, пенька, перидот, пирография, плагиат, плагиатор, плакатист, плакатчик, пластификаторы, подрамник, подрамок, полирование, полировка, прилипание, пробойники, проектант, проектировщик, разжижители, растворители, регалии, резец, ренессанс, ритм, ритмичность, родонит, селенит, сецессион, скипетр, скульптор, талисман, тонирование, тонировка, фиксаж, фиксация, фиксирование, центавр, шлифование, шлифовка, штихель, экспрессивность, экспрессия, эстетизм, эстетство

Сред всички възможни двойки думи измежду приготвените 94 (които са 4 371 на брой) съществуват точно 50 двойки синоними, които се очаква да бъдат открити от предложения алгоритъм, т.е. да бъдат изведени на челни позиции в списъка, който се получава като резултат. Задачата се разглежда като класически проблем на извличането на информация, в която се търсят всички синоними измежду списък с двойки думи. Извършени са експерименти с различни модификации на алгоритъма SemSim. При всеки от експериментите се измерва семантичната близост между всяка двойка думи от списъка и като резултат се получава списък с 4 371 двойки думи, подредени в намаляващ ред по съпоставената им от съответния алгоритъм семантична близост. Оценяването на резултатите е извършено чрез *усреднена интерполирана точност в 11 точки* и чрез изчисляване на броя намерени синоними сред първите N резултата. Следва описание на проведените експерименти:

- **RAND** – връща двойките в случаен ред. Използва се като база за сравнение с другите алгоритми (baseline).
- **SemSim-1, SemSim-2, …, SemSim-10** – основният алгоритъм SemSim с размер на контекста съответно 1, 2, ..., 10 думи, с краткия българо-руски

речник, без използване на обратен или обогатен контекст, с прилагане на лематизация, без лематизация на заявката.

- **SemSim-1+TFIDF, SemSim-2+TFIDF, ..., SemSim-6+TFIDF** – модификация на SemSim с TF.IDF претегляне и размер на контекста съответно 1, 2, ..., 6 думи.
- **SemSim-1+QUERY-LEM, SemSim-2+QUERY-LEM, ..., SemSim-5+QUERY-LEM** – модификация на SemSim алгоритъма с лематизация на заявката и размер на контекста съответно 1, 2, 3, 4 и 5 думи.
- **SemSim+REV2**, **SemSim+REV3**, **...**, **SemSim+REV7** – модификация на SemSim алгоритъма с използване на техниката "обратен контекст" с прагове на честотата съответно 2, 3, 4, 5, 6 и 7.

Получени са следните резултати:

Алгоритъм	**11pt avg. precision**		**Алгоритъм**	**11pt avg. precision**		**Алгоритъм**	**11pt avg. precision**
SemSim-1	63,47%		SemSim-1+TF.IDF	66,95%		SemSim-1+ QUERY-LEM	63,68%
SemSim-2	63,33%		**SemSim-2+TF.IDF**	**67,09%**		SemSim-2+ QUERY-LEM	63,51%
SemSim-3	59,80%		SemSim-3+TF.IDF	63,64%		SemSim-3+ QUERY-LEM	59,92%
SemSim-4	57,60%		SemSim-4+TF.IDF	60,65%		SemSim-4+ QUERY-LEM	57,84%
SemSim-5	56,16%		SemSim-5+TF.IDF	58,12%		SemSim-5+ QUERY-LEM	56,23%
SemSim-6	55,42%		SemSim-6+TF.IDF	55,47%			
SemSim-7	54,83%						
SemSim-8	53,47%						
SemSim-9	52,79%						
SemSim-10	52,29%						

Резултатите от експериментите с използване на обратен контекст SemSim+REV2, SemSim+REV3, ..., SemSim+REV7 не могат да се оценят с мярката усреднена интерполирана точност в 11 точки, защото тези алгоритми връщат нулева стойност за една голяма част от двойките думи, поради което тяхната наредба е неопределена. По тази причина е предложен друг начин за сравнение на резултатите – по брой намерени синоними сред първите 1, 5, 10, 20, 30, 40 50, 100, 200 резултата и общ брой открити синоними сред всички резултати:

Алгоритъм	**1**	**5**	**10**	**20**	**30**	**40**	**50**	**100**	**200**	**Max**
RAND	0	0,1	0,1	0,2	0,3	0,4	0,6	1,1	2,3	**50**
SemSim-1	1	4	**8**	**16**	20	25	29	43	**47**	**50**
SemSim-2+TFIDF	1	**5**	**8**	**16**	**23**	27	30	**44**	**47**	**50**

SemSim+REV2	1	4	**8**	**16**	21	27	32	42	43	46
SemSim+REV3	1	4	**8**	**16**	20	**28**	32	41	42	46
SemSim+REV4	1	4	**8**	15	20	**28**	**33**	41	42	45
SemSim+REV5	1	4	**8**	15	20	**28**	**33**	40	41	42
SemSim+REV6	1	4	**8**	15	22	**28**	32	39	40	42
SemSim+REV7	1	4	**8**	15	21	27	30	37	39	40

Най-добре работещия алгоритъм SemSim-2+TF.IDF извежда съответно повече от половината синоними сред първите 40 резултата, 88% от всички синоними сред първите 100 резултата и 94% от всички синоними сред първите 200 резултата. Следователно, ако се търсят не непременно всички синоними, а е необходимо намирането на възможно по-голям брой по-рано, този алгоритъм ще свърши отлична работа. Въпреки това, е видно, че при търсене на синоними алгоритъмът SemSim не работи достатъчно добре и е удачно да се комбинира с други методи.

Подобряване подравняването на думи

Задачата за подравняване на думи (word alignment) е изключително важна за съвременния статистически машинен превод. Системите за статистически машинен превод (statistical machine translation, SMT) използват паралелен двуезичен корпус, за да научат автоматично вероятни преводни съответствия между думи и фрази и след това превеждат чрез търсещ алгоритъм (decoder), който използва вече заучените фрази, за да намери най-вероятното преводно съответствие на даденото за превод изречение [Lee, 2007]. Ако една система за статистически машинен превод използва по време на своето обучение по-добро подравняване на думите в изреченията от учебния паралелен корпус, това води до построяване на по-качествен модел на превода и съответно до по-добър превод.

Задачата за *подравняване на думи* (*word alignment*) може да се формализира като задача за намиране на най-точното съответствие между думите в съответните изречения на паралелен двуезичен корпус, такова, че между дадена двойка думи съществува връзка тогава и само тогава, когато съответните думи са превод една на друга в конкретното изречение.

Описание на алгоритъма

Подравняването на думи в разглеждания алгоритъм се извършва чрез комбинация между алгоритъма MMEDR за измерване на модифицирана ортографска близост и алгоритъма CrossSim за измерване на междуезикова семантична близост чрез използване на уеб като корпус. Приложена е адаптация на алгоритъма *конкурентно свързване (competitive linking)*, описан в [Melamed, 2000]. Алгоритъмът за подравняване на думи работи по следния начин: по дадени две съответни изречения от паралелния корпус първо се извличат всички думи, съдържащи се във всяко от тях. След това се премахват функционалните думи (предлози, местоимения, съюзи, частици, междуметия и някои наречия), а останалите се комбинират всяка с всяка и така се получават възможните съответствия между думите от двете изречения. Между всяка от получените двойки думи се измерва близостта (ортографска, семантична, комбинирана или друга) и двойките думи се подреждат спрямо нея в намаляващ ред. След това започва свързването на думите. Първо се свързват първата двойка думи (с най-голяма близост), след нея втората двойка и т.н. Всяко такова свързване се извършва само ако и двете

двойки думи не са били свързани с никоя друга дума до момента. Свързването приключва, когато думите, които са останали свободни, свършат или близостта между тях достигне някаква предварително дефинирана долна граница θ. Двойки думи с мярка на близост под долната граница не се свързват, защото такова свързване най-вероятно е некоректно. В резултат алгоритъмът установява връзки между някои двойки думи от съответните преводни изречения от паралелния текст. Някои думи остават умишлено несвързани, например функционалните думи, както и думи, за които не е намерено достатъчно близко съответствие в преводното изречение.

Коректността на полученото подравняване се оценява чрез мярката AER (alignment error rate), въведена от [Och & Ney, 2003].

Експерименти и резултати

За експериментите е използван българо-руски паралелен корпус, подравнен на ниво изречение (5 827 съответни изречения, от които 4 827 са използвани като учебни и 1 000 като тестови). За оценката на резултатите с мярката AER като образец за правилно подравняване на думите от съответните изречения в текста (gold standard) е използвано подравняване, извършено ръчно от лингвист. Следва детайлно описание на всеки от проведените експерименти:

- **CL-LCSR** – думите в съответните изречения от паралелния корпус се подравняват по метода *конкурентно свързване* чрез използване на традиционната мярка за ортографска близост LCSR.
- **CL-MEDR** – алгоритъмът подравнява думи чрез конкурентно свързване, точно както CL-LCSR, но използва мярката за ортографска близост MEDR.
- **CL-MMEDR** – алгоритъмът подравнява думи чрез конкурентно свързване, точно както CL-LCSR, но вместо LCSR използва модифицираната мярка за ортографска близост MMEDR.
- **CL-CrossSim** – алгоритъмът подравнява думи чрез конкурентно свързване, използвайки като мярка за близост алгоритъма CrossSim с размер на контекста 3 думи, с краткия българо-руски речник, без използване на обратен или обогатен контекст, с прилагане на лематизация, без лематизация на заявката и без TF.IDF претегляне.
- **CL-CUT** – алгоритъмът подравнява думи чрез конкурентно свързване като комбинира мерките за близост MMEDR и CrossSim чрез отсичане по следния начин: при дадени две думи w_{bg} и w_{ru} и граница на отсичане α = 0,62 близостта *s* се изчислява като *s* = 1 при $MMEDR(w_{bg}, w_{ru}) > \alpha$ и $s = CrossSim(w_{bg}, w_{ru})$ в противен случай. Най-подходящата стойност на α се установява като се пробват всички възможности измежду {0,01; 0,02; …, 0,99} върху тренировъчен набор данни.
- **CL-AVG** – алгоритъмът подравнява думи чрез конкурентно свързване като ползва като мярка за близост между двойките думи средното аритметично от стойностите на мерките MMEDR и CrossSim.
- **CL-MAX** – алгоритъмът подравнява думи чрез конкурентно свързване като ползва като мярка за близост по-голямата от стойностите на мерките MMEDR и CrossSim.

Най-подходящата стойност за границата θ за отсичане при конкурентно свързване за всеки от експериментите е установена чрез изследване на всички стойности θ

є {0,01; 0,02; …; 0,99} и избиране на тази от тях, която работи най-добре за алгоритъма от съответния експеримент върху учебен корпус.

Получени са следните резултати:

Алгоритъм	Alignment Error Rate (AER)
CL-LCSR	0,176
CL-MEDR	0,177
CL-MMEDR	0,161
CL-CrossSim	**0,122**
CL-CUT	0,150
CL-AVG	0,124
CL-MAX	0,180

Глава 5. Алгоритъм за извличане на фалшиви приятели от паралелен двуезичен корпус

В пета глава е разработен алгоритъм FFExtract за извличане на фалшиви приятели от паралелен двуезичен корпус на български и руски език, подравнен на ниво изречение. По даден паралелен текст алгоритъмът трябва да извлича от него двойките думи, които представляват фалшиви приятели, т.е. се възприемат като близки поради ортографско или фонетично сходство, но имат напълно различни значения. Обект на изследване са единствено фалшивите приятели и различаването между пълни когнати и частични когнати е извън обхвата на поставената задача. Извършените експерименти се отнасят за извличането на фалшиви приятели между български и руски език, но описаните методи са приложими и за други двойки езици (след известна адаптация).

Описание на алгоритъма

Алгоритъмът FFExtract работи на две стъпки: на първата извлича кандидатите за когнати и фалшиви приятели чрез алгоритъма MMEDR, а на втората прилага няколко техники за отделяне на фалшивите приятели от когнатите. Едната техника използва статистически наблюдения за срещанията на думите в паралелните изречения от текста заедно и поотделно и след това изчислява вероятността двете думи да са фалшиви приятели чрез няколко формули. Другата техника прилага алгоритъма за измерване на семантична близост CrossSim.

Намиране на кандидат когнати / фалшиви приятели

На първата стъпка от предложения алгоритъм се извличат всички двойки думи от паралелния текст, които се възприемат като близки от хора, владеещи двата езика, и съответно са кандидати за когнати или фалшиви приятели. За целта от дадения паралелен текст се извличат всички български и всички руски думи и за всяка двойка българска и руска дума { w_{bg}, w_{ru} } се измерва модифицираната ортографска близост по алгоритъма MMEDR (описан във втора глава). Като кандидати за когнати и фалшиви приятели се разглеждат всички двойки българска и руска дума, които преминават над определена граница на близост α. Понеже българският и руският език са силно флексивни, в тях род, число, определителен

член и падеж могат да се изразяват чрез окончания, които формират различни словоформи на една и съща дума. Алгоритъмът счита различните словоформи на една и съща дума за различни и така всяка от тях може да участва в двойките кандидати за когнати и фалшиви приятели.

Различаване между когнати и фалшиви приятели: статистически подход

Използваният статистически подход за различаване между когнати и фалшиви приятели в паралелен текст, подравнен на ниво изречение, е базиран на наблюдения за срещанията на анализираните думи самостоятелно в текста и съвместните им срещания в съответни изречения. Използван е фактът, че в паралелен текст се наблюдава тенденция когнатите да се срещат често в съответни преводни изречения, докато такава не е в сила за фалшивите приятели [Nakov P. & Pacovski, 2006]. За да бъде формализирана тази идея се въвеждат следните означения:

- $S_{bg}(w_{bg})$ – брой български изречения от паралелния текст, съдържащи българската дума w_{bg}.
- $S_{ru}(w_{ru})$ – брой руски изречения, съдържащи руската дума w_{ru}.
- $S_{bg\&ru}(w_{bg}, w_{ru})$ – брой съответни преводни изречения, съдържащи съответно думата w_{bg} в българското изречение и думата w_{ru} в руското изречение.

Използвани са следните формули за моделиране на вероятността две думи да имат сходно значение и следователно да са когнати:

$$F_6(w_{bg}, w_{ru}) = \frac{S_{bg\&ru}(w_{bg}, w_{ru}) + 1}{\max\left(\frac{1 + S_{bg}(w_{bg})}{1 + S_{ru}(w_{ru})}, \frac{1 + S_{ru}(w_{ru})}{1 + S_{bg}(w_{bg})}\right)}$$

$$F_1(w_{bg}, w_{ru}) = \frac{(S_{bg\&ru}(w_{bg}, w_{ru}) + 1)^2}{(S_{bg}(w_{bg}) + 1)(S_{ru}(w_{ru}) + 1)}$$

$$F_2(w_{bg}, w_{ru}) = \frac{(S_{bg\&ru}(w_{bg}, w_{ru}) + 1)^2}{(S_{bg}(w_{bg}) - S_{bg\&ru}(w_{bg}, w_{ru}) + 1)(S_{ru}(w_{ru}) - S_{bg\&ru}(w_{bg}, w_{ru}) + 1)}$$

Различаване между когнати и фалшиви приятели: семантичен подход

Предложеният семантичен подход за различаване между фалшиви приятели и когнати е базиран на алгоритъма CrossSim за измерване на междуезикова семантична близост между двойка думи чрез контексти, извлечени от уеб. Основната идея, която се използва, е че ако две думи са когнати, то те би следвало да са в по-голяма степен семантично близки, отколкото, ако са фалшиви приятели.

Различаване между когнати и фалшиви приятели: комбиниран подход

Комбинираният подход комбинира трите статистически формули F_1, F_2 и F_6 с оценката за семантична близост CrossSim чрез просто сумиране, както и с тегла при сумирането.

Алгоритъм FFExtract

Следва постъпково описание на алгоритъма FFExtract:

Алгоритъм: **FFExtract.**

Вход: паралелен българо-руски текст { t_{bg}, t_{ru} }, състоящ се от изреченията s^1_{bg} … s^n_{bg} ∈ t_{bg} и s^1_{ru} … s^m_{ru} ∈ t_{ru}.

Изход: списък R от наредени множества с двойки думи и техните оценки { w_{bg}, w_{ru}, $F_1(w_{bg}, w_{ru})$, $F_2(w_{bg}, w_{ru})$, $F_6(w_{bg}, w_{ru})$, $CrossSim(w_{bg}, w_{ru})$ }.

Настройки: граница α (0 < α < 1); използване на лематизация (да / не).

Стъпки:

Стъпка 1. Извличане на кандидатите за когнати и фалшиви приятели.

Стъпка 1.1. Извлича се множеството W_{bg} от всички български думи от изреченията на българския текст s^1_{bg} … s^n_{bg}.

Стъпка 1.2. Извлича се множеството W_{ru} от всички руски думи от изреченията на руския текст s^1_{ru} … s^m_{ru}.

Стъпка 1.3. Построява се множеството C от кандидати за когнати и фалшиви приятели, което се състои от всички двойки думи { w_{bg} ∈ W_{bg}, w_{ru} ∈ W_{ru} }, за които $MMEDR(w_{bg}, w_{ru}) \geq \alpha$.

Стъпка 2. Различаване между когнати и фалшиви приятели.

Стъпка 2.1. За всички думи w_{bg} ∈ W_{bg} и w_{ru} ∈ W_{ru} по множествата съответни изречения s^1_{bg} … s^n_{bg} ∈ t_{bg} и s^1_{ru} … s^m_{ru} ∈ t_{ru} се изчисляват стойностите $S_{bg}(w_{bg})$, $S_{ru}(w_{ru})$ и $S_{bg\&ru}(w_{bg}, w_{ru})$. В случай, че е включено използването на лематизация, се броят срещанията не на конкретните думи w_{bg} и w_{ru}, а на техните леми (една или няколко основни граматични форми).

Стъпка 2.2. Първоначално R е празен списък.

Стъпка 2.3. За всяка двойка думи { w_{bg}, w_{ru} } ∈ C се изпълнява следното:

Изчислява се $CrossSim(w_{bg}, w_{ru})$ при размер на контекста 10 думи, с използване на лематизация, с краткия преводен речник и без използване на техниките "TF.IDF претегляне", "обратен контекст", "обогатяване на контекста" и "лематизация на заявката".

Изчисляват се стойностите $F_1(w_{bg}, w_{ru})$, $F_2(w_{bg}, w_{ru})$ и $F_6(w_{bg}, w_{ru})$.

Към списъка R се добавя наредената множество { w_{bg}, w_{ru}, $F_1(w_{bg}, w_{ru})$, $F_2(w_{bg}, w_{ru})$, $F_6(w_{bg}, w_{ru})$, $CrossSim(w_{bg}, w_{ru})$ }.

Експерименти и резултати

Като входни данни е използвана извадка от руската книга "*Властелин мира*" от Александър Беляев и преводът ѝ на български език. Чрез прилагане на описания двустъпков алгоритъм FFExtract на първата стъпка са отделени от паралелния текст като кандидати за когнати и фалшиви приятели всички двойки думи, които притежават модифицирана ортографска близост над 90%, а на втората стъпка списъкът с двойките думи е подреден в нарастващ ред по съответно измерената близост (комбинация от оценките F_1, F_2 и F_6 и CrossSim). Оценката на резултатите е направена чрез изчисляване на интерполирана усреднена точност в 11 точки в получения нареден списък (както е направено в изследването на [Bergsma & Kondrak, 2007]). Следва описание на проведените експерименти:

- **ASC** – двойките думи подредени по азбучен ред. Използва се като база за сравнение с другите алгоритми (baseline)
- **PAR** – статистическият алгоритъм с формула F_6.
- **PAR+L** – алгоритъмът PAR, модифициран да използва лематизация.
- **F1** – алгоритъмът PAR с използване на формулата F_1.
- **F1+L** – алгоритъмът PAR с използване на формулата F_1 и с лематизация.
- **F2** – алгоритъмът PAR с използване на формулата F_2.
- **F2+L** – алгоритъмът PAR с използване на формулата F_2 и с лематизация.
- **CrossSim+L** – алгоритъмът CrossSim с размер на контекста 10 думи, с използване на лематизация и без използване на техниките "TF.IDF претегляне", "обратен контекст", "обогатяване на контекста" и "лематизация на заявката".
- **CrossSim+L+PAR+L** – алгоритъмът CrossSim+L комбиниран с PAR+L чрез сумиране на стойностите на CrossSim+L и PAR+L.
- **CrossSim+L+F1+L** – алгоритъмът CrossSim+L комбиниран с F1+L чрез сумиране на стойностите на CrossSim+L и F1+L.
- **CrossSim+L+F2+L** – алгоритъмът CrossSim+L комбиниран с F2+L чрез сумиране на стойностите на CrossSim+L и F2+L.
- **1.5*(CrossSim+L)+F1+L** – алгоритъмът CrossSim+L комбиниран с F1+L чрез претеглено сумиране на стойностите 1.5 * (CrossSim+L) и F1+L.
- **1.5*(CrossSim+L)+F2+L** – алгоритъмът CrossSim+L комбиниран с F2+L чрез претеглено сумиране на стойностите 1.5 * (CrossSim+L) и F2+L.
- **CrossSim+L+1.5*(F1+L)** – алгоритъмът CrossSim+L комбиниран с F1+L чрез претеглено сумиране на стойностите CrossSim+L и 1.5 * (F1+L).
- **CrossSim+L+1.5*(F2+L)** – алгоритъмът CrossSim+L комбиниран с F2+L чрез претеглено сумиране на стойностите CrossSim+L и 1.5 * (F2+L).

Получени са следните резултати:

Алгоритъм	**Усреднена точност в 11 точки**
ASC	4,17%
F2	38,60%
F1	39,50%
PAR	43,81%
PAR+L	53,20%
CrossSim+L+PAR+L	61,28%
CrossSim+L	63,68%
F1+L	63,98%
F2+L	66,82%
CrossSim+L+1.5*(F2+L)	74,34%
1.5*(CrossSim+L)+F1+L	75,07%
CrossSim+L+1.5*(F1+L)	75,46%

CrossSim+L+F2+L	76,15%
CrossSim+L+F1+L	77,50%
1.5*(CrossSim+L)+F2+L	**77,64%**

Заключение

Основните цели и задачи на дисертационния труд налагат извършването на изследвания, свързани с извличането на семантична близост, когнати и фалшиви приятели. По време на работата по дисертацията са направени голям обем експерименти и са разработени нови алгоритми. Предложени са нов алгоритъм MMEDR за измерване на ортографска близост между български и руски език и алгоритми SemSim и CrossSim за извличане на едноезична и междуезикова семантична близост от уеб. Демонстрирано е приложение на разработените алгоритми при решаването на различни задачи на компютърната лингвистика. Разработен е алгоритъм FFExtract за извличане на фалшиви приятели от паралелен българо-руски корпус. Получени са убедителни резултати по основните цели и е разработено решение на всички поставени конкретни задачи. Резултатите са публикувани на престижни научни конференции и в авторитетни научни издания.

Използвана литература

1. [Adamson & Boreham, 1974] Adamson G., Boreham J. "The Use of an Association Measure Based on Character Structure to Identify Semantically Related Pairs of Words and Document Titles". *Information Storage and Retrieval*, volume 10, pages 253-260, 1974
2. [Al-Onaizan & колектив, 1999] Al-Onaizan Y., Curin J., Jahr M., Knight K., Lafferty J., Melamed D., Och F., Purdy D., Smith N., Yarowsky D. "Statistical Machine Translation". *Technical Report, Johns Hopkins University Summer Workshop*, Baltimore, MD, United States, 1999
3. [Atanassova & колектив, 2003] Atanassova I., Nakov S., Nakov P. "ArtsSemNet: From Bilingual Dictionary to Bilingual Semantic Network". *Proceedings of the Workshop on Balkan Language Resources and Tools, 1st Balkan Conference in Informatics*, Thessaloniki, Greece, 2003
4. [Bergsma & Kondrak, 2007] Bergsma S., Kondrak G. "Alignment-Based Discriminative String Similarity". *Proceedings of the ACL '2007*, pages 656-663, Prague, Czech Republic, 2007
5. [BgRu.net, 2007] Online Bulgarian-Russian and Russian Bulgarian dictionary – http://www.bgru.net/intr/dictionary/ (посетен през април 2007)
6. [BGTree, 2009] Български синонимен речник on-line, http://www.bgtree.net/dictionary/ (посетен през април, 2009)
7. [Bickford & Tuggy, 2002] Bickford A., Tuggy D. "Electronic Glossary of Linguistic Terms (with Equivalent Terms in Spanish)", http://www.sil.org/mexico/ling/glosario/E005ai-Glossary.htm, Version 0.6, *Summer Institute of Linguistics*, Mexico, April 2002 (посетен през март 2007)
8. [Bollegala & колектив, 2007] Bollegala D., Matsuo Y., Ishizuka M. "Measuring Semantic Similarity between Words Using Web Search Engines". *Proceedings of the 16th International World Wide Web Conference (WWW2007)*, pages 757-766, Banff, Canada, 2007
9. [Brew & McKelvie, 1996] Brew C., McKelvie D. "Word-Pair Extraction for Lexicography". *Proceedings of the 2nd International Conference on New Methods in Language Processing*, pages 45–55, Ankara, Turkey, 1996
10. [Brown & колектив, 1993] Brown P., Della Pietra S., Della Pietra V., Mercer R. "The Mathematics of Statistical Machine Translation: Parameter Estimation". *Computational Linguistics*, volume 19, issue 2, pages 263-311, 1993
11. [Budanitsky & Hirst, 2006] Budanitsky A., Hirst G. ”Evaluating WordNet-based Measures of Lexical Semantic Relatedness”. *Computational Linguistics*, volume 32, issue 1, pages 13-47, MIT Press, USA, 2006

12. [Buncic, 2000] Buncic D., "Псевдо-аналогонимия. "Ложные друзья переводчика как единица сопоставительной лексикологии". *XXIX межвузовская научно-методическая конференция преподавателей и аспирантов*, Санкт-Петербург, Русия, 2000
13. [Chen & колектив., 2006] Chen H., Lin M., Wei Y. "Novel Association Measures Using Web Search with Double Checking". *Proceedings of the COLING-ACL 2006*, pages 1009-1016, Sydney, Australia, 2006
14. [Cilibrasi & Vitanyi, 2007] Cilibrasi R., Vitanyi P. "The Google Similarity Distance". *IEEE Transactions on Knowledge and Data Engineering*, volume 19(3), pages 370-383, USA, 2007
15. [Curran & Moens, 2002] Curran J., Moens M. "Improvements in Automatic Thesaurus Extraction". *Proceedings of the Workshop on Unsupervised Lexical Acquisition, SIGLEX 2002*, pages 59-67, Philadelphia, PA, USA, 2002
16. [Dijkstra & колектив, 1999] Dijkstra T., Grainger J., van Heuven V. "Recognition of Cognates and Interlingual Homographs: The Neglected Role of Phonology". *Journal of Memory and Language*, ISSN 0749-596X, issue 41, pages 496-518, London, 1999
17. [Fellbaum, 1998] Fellbaum C. (editor) "WordNet: An Electronic Lexical Database". The MIT Press, Cambridge, MA, USA, 1998
18. [Frunza & Inkpen, 2006] Frunza O., Inkpen D. "Semi-Supervised Learning of Partial Cognates using Bilingual Bootstrapping". *Proceedings of COLING-ACL,* pages 433-440, Sydney, Australia, 2006
19. [Fung & Yee, 1998] Fung P., Yee L. "An IR Approach for Translating New Words from Nonparallel, Comparable Texts". *Proceedings of COLING-ACL*, volime 1, pages 414-420, Montreal, Canada, 1998
20. [Fung, 1998] Fung P. "A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora". *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA '98)*, pages 1-16, Springer, 1998
21. [Gabrilovich & Markovitch, 2007] Gabrilovich E., Markovitch S. "Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis". *Proceedings of IJCAI-07*, pages 1606-1611, Hyderabad, India, 2007
22. [Gale & Church, 1993] Gale W., Church K. "A Program for Aligning Sentences in Bilingual Corpora". *Computational Linguistics*, volume 19, issue 1, pages 75-102, 1993
23. [Google, 2007] Google online search engine – http://www.google.com (използван в периода април 2007 – март 2009)
24. [Guy, 1994] Guy J. "An Algorithm for Identifying Cognates in Bilingual Wordlists and Its Applicability to Machine Translation". *Journal of Quantitative Linguistics*, volume 1, issue 1, pages 35-42, 1994
25. [Hagiwara & колектив, 2007] Hagiwara M., Ogawa Y., Toyama K. "Effectiveness of Indirect Dependency for Automatic Synonym Acquisition". *Proceedings of CoSMo 2007 Workshop, held in conjuction with CONTEXT 2007*, Roskilde, Denmark, 2007.
26. [Harris, 1985] Harris Z. "Distributional Structure". In: Katz J. (editor), *The Philosophy of Linguistics*, Oxford University Press, pages 26-47, New York, NY, USA, 1985
27. [Hearst, 1991] Hearst M. "Noun Homograph Disambiguation Using Local Context in Large Text Corpora". *Proceedings of the 7th Annual Conference of the University of Waterloo Centre for the New OED and Text Research*, pages 1-22, Oxford, England, 1991
28. [Inkpen & колектив, 2005] Inkpen D., Frunza O., Kondrak G. "Automatic Identification of Cognates and False Friends in French and English". *Proceedings of RANLP 2005*, pages 251-257, Borovets, Bulgaria, 2005
29. [Inkpen, 2007] Inkpen D. "Near-Synonym Choice in an Intelligent Thesaurus". *Proceedings of NAACL-HLT 2007*, New York, NY, USA, 2007
30. [Iosif & Potamianos, 2007] Iosif E., Potamianos A. "Unsupervised Semantic Similarity Computation using Web Search Engines". *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (WI-2007)*, Silicon Valley, USA, 2007

31. [Jiang & Conrath, 1997] Jiang J., Conrath D. "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy". *Proceedings of the 10th International Conference Research on Computational Linguistics (ROCLING X)*, Taiwan, 1997
32. [Kilgarriff & Grefenstette, 2003] Kilgarriff A., Grefenstette G. "Introduction to the Special Issue on the Web as Corpus". *Computational Linguistics*, volume 29, issue 3, pages 333-347, 2003
33. [Koehn & Knight, 2002] Koehn P., Knight K. "Learning a Translation Lexicon from Monolingual Corpora". *Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, pages 9-16, Philadelphia, PA, USA, 2002
34. [Kœssler & Derocquigny, 1928] Kœssler M., Derocquigny J. “Les Faux Amis Ou Les Pièges Du Vocabulaire Anglais”, Paris, France, 1928
35. [Kondrak & Dorr, 2004] Kondrak G., Dorr B. "Identification of Confusable Drug Names: A New Approach and Evaluation Methodology". *Proceedings of 20th International Conference on Computational Linguistics (COLING 2004)*, pages 952-958, Geneva, Switzerland, 2004
36. [Kondrak & Sherif, 2006] Kondrak G., Sherif T. "Evaluation of Several Phonetic Similarity Algorithms on the Task of Cognate Identification". *Proceedings of the COLING-ACL Workshop on Linguistic Distances*, pages 37-44, Sydney, Australia, 2006
37. [Kondrak & колектив, 2003] Kondrak G., Marcu D., Knight K. "Cognates Can Improve Statistical Translation Models", *Proceedings of HLT-NAACL 2003*, companion volume, pages 46-48, Edmonton, Canada, 2003
38. [Kondrak, 2000] Kondrak G. "A New Algorithm for the Alignment of Phonetic Sequences". *Proceedings of NAACL/ANLP 2000*, pages 288-295, Seattle, WA, USA, 2000
39. [Kondrak, 2001] Kondrak G. "Identifying Cognates by Phonetic and Semantic Similarity". *Proceedings of NAACL 2001*, pages 103-110, Pittsburgh, PA, USA, 2001
40. [Kondrak, 2003] Kondrak G. "Identifying Complex Sound Correspondences in Bilingual Wordlists". *Proceedings of CICLING 2003*, pages 432-443, Mexico City, Mexico, 2003
41. [Kondrak, 2004] Kondrak G. "Combining Evidence in Cognates Identification". *Proceedings of the 17th Canadian Conference on Artificial Intelligence*, pages 44-59, London, 2004
42. [Lee, 1999] Lee L. "Measures of Distributional Similarity". *Proceedings of ACL'99*, pages 25-32, College Park, MD, USA, 1999
43. [Lee, 2007] Lee G. "Statistical Machine Translation". *Lecture Notes from the Course "Advances in Human Language Technology (EECS703A)"*, Intelligent Software Lab, Pohang University of Science and Technology (POSTECH), Hyoja-Dong, Korea, 2007
44. [Levenshtein, 1965] Levenshtein V. "Binary Codes Capable of Correcting Deletions, Insertions and Reversals". *Doklady Akademii Nauk SSSR*, volume 163, issue 4, pages 845-848, Moscow, Russia, 1965
45. [Lin, 1998] Lin D. "Automatic Retrieval and Clustering of Similar Words". *Proceedings of COLING-ACL '98*, pages 768-774, Montreal, Canada, 1998
46. [Mann & Yarowski, 2001] Mann G., Yarowsky D. "Multipath Translation Lexicon Induction via Bridge Languages. *Proceedings of NAACL 2001*, pages 151-158, Pittsburgh, PA, USA, 2001
47. [Manning & колектив, 2008] Manning C., Prabhakar R., Schütze H. "Introduction to Information Retrieval". *Cambridge University Press*, ISBN 0521865719, New York, USA, 2008
48. [Marzal & Vidal, 1993] Marzal A., Vidal E. "Computation of Normalized Edit Distance and Applications". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 15, issue 9, pages 926-932, USA, 1993
49. [McDonald, 2000] McDonald S. "Environmental Determinants of Lexical Processing Effort". *Ph.D. Thesis*, University of Edinburgh, United Kingdom, 2000
50. [McEnery & Oakes, 1995] McEnery T., Oakes M. "Cognate Extraction in the CRATER Project: Methods and Assessment". *Proceedings of From Texts to Tags: Issues in Multilingual Language Analysis*, pages 77-86, Dublin, Ireland, 1995
51. [Melamed, 1999] Melamed D. "Bitext Maps and Alignment via Pattern Recognition". *Computational Linguistics*, ISSN:0891-2017, volume 25, issue 1, pages 107-130, 1999

52. [Melamed, 2000] Melamed D. "Models of Translational Equivalence among Words". *Computational Linguistics*, volume 26, issue 2, pages 221-249, 2000
53. [Miller & Charles, 1991] Miller G., Charles W., "Contextual Correlates of Semantic Similarity". *Language and Cognitive Processes*, volume 6, pages 1-28, 1991
54. [Mitkov & колектив, 2007] Mitkov R., Pekar V., Blagoev D., Mulloni A. "Methods for Extracting and Classifying Pairs of Cognates and False Friends". *Machine Translation*, volume 21, issue 1, pages 29-53, Springer Netherlands, 2007
55. [Mulloni & Pekar, 2006] Mulloni A., Pekar V. "Automatic Detection of Orthographic Cues for Cognate Recognition". *Proceedings of LREC-06*, pages 2387-2390, Genoa, Italy, 2006
56. [Mulloni & колектив, 2007] Mulloni A., Pekar V., Mitkov R., Blagoev D. "Semantic Evidence for Automatic Identification of Cognates". *Proceedings of the 1st International Workshop on Acquisition and Management of Multilingual Lexicons*, pages 49-54, Borovets, Bulgaria, 2007
57. [Nakov & колектив, 2007a] Nakov P., Nakov S., Paskaleva E. "Improved Word Alignments Using the Web as a Corpus". *Proceedings of RANLP 2007*, pages 400-405, Bulgaria, 2007
58. [Nakov & колектив, 2007b] Nakov S., Nakov P., Paskaleva E. "Cognate or False Friend? Ask the Web!". *Proceedings of the 1st International Workshop on Acquisition and Management of Multilingual Lexicons, part of RANLP 2009*, pages 55-62, Borovets, Bulgaria, 2007
59. [Nakov & колектив, 2009a] Nakov S., Nakov P., Paskaleva E. "Unsupervised Extraction of False Friends from Parallel Bi-Texts Using the Web as a Corpus". *Proceedings of RANLP 2009*, pages 292-298, Borovets, Bulgaria, 2009
60. [Nakov & колектив, 2009b] Nakov S., Paskaleva E., Nakov P. "A Knowledge-Rich Approach to Measuring the Similarity between Bulgarian and Russian Words", *Workshop on Multilingual Resources, Technologies and Evaluation for Central and Eastern European Languages held in conjuction with RANLP 2009*, Borovets, Bulgaria, 2009
61. [Nakov P. & Pacovski, 2006] Nakov P., Pacovski V. "Acquiring False Friends from Parallel Corpora: Application to South Slavonic Languages". *Readings in Multilinguality, Selected Papers from Young Researchers in BIS-21*++, pages 87-94, INCOMA, Shoumen, Bulgaria, 2006
62. [Nakov, 2008a] Nakov S. "Automatic Acquisition of Synonyms Using the Web as a Corpus". *Proceedings of the 3rd Annual South-East European Doctoral Student Conference (DSC 2008)*, volume 2, pages 216-229, Thessaloniki, Greece, 2008
63. [Nakov, 2009] Nakov S. "Automatic Identification of False Friends in Parallel Corpora: Statistical and Semantic Approach". *Serdica Journal of Computing*, volume 3, pages 133-158, 2009
64. [Och & Ney, 2003] Och F., Ney H. "A Systematic Comparison of Various Statistical Alignment Models". *Computational Linguistics*, volume 29, issue 1, pages 19-51, 2003
65. [Paskaleva & Mihov, 1997] Paskaleva E., Mihov S. "Second Language Acquisition from Aligned Corpora". *Proceedings of the International Conference "Language Technology and Language Teaching"*, pages 43-52, Groningen, The Netherlands, 1997
66. [Paskaleva, 2002] Paskaleva E. "Processing Bulgarian and Russian Resources in Unified Format". *Proceedings of the 8th International Scientific Symposium MAPRIAL*, pages 185-194, Veliko Tarnovo, Bulgaria, 2002
67. [Plas & Tiedemann, 2006] Plas L., Tiedemann J. "Finding Synonyms Using Automatic Word Alignment and Measures of Distributional Similarity". *Proceedings of COLING-ACL 2006*, Sydney, Australia, 2006
68. [Rapp, 1995] Rapp R. "Identifying Word Translations in Non-Parallel Texts". *Proceedings of ACL '95,* pages 320-322, Cambridge, MA, United States, 1995
69. [Rapp, 1999] Rapp R. "Automatic Identification of Word Translations from Unrelated English and German Corpora". *Proceedings of ACL '99,* pages 519-526, College Park, MD, USA, 1999
70. [Resnik, 1995] Resnik Ph. "Using Information Content to Evaluate Semantic Similarity in a Taxonomy". *Proceedings of IJCAI-95*, pages 448-453, Montreal, Canada, 1995
71. [Rodgers & Nicewander, 1988] Rodgers J., Nicewander W. "Thirteen Ways to Look at the Correlation Coefficient". *The American Statistician*, volume 42, issue 1, pages 59-66, 1988
72. [Russel, 1918] Russel R. "U.S. Patent 1,261,167", Pittsburgh, PA, USA, 1918

73. [Sahami & Hailman, 2006] Sahami M., Heilman T. "A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets". *Proceedings of 15th International World Wide Web Conference (WWW2006)*, pages 377-386, Edinburgh, Scotland, 2006
74. [Sanchez & Moreno, 2005] Sanchez D., Moreno A. "Automatic Discovery of Synonyms and Lexicalizations from the Web". *Artificial Intelligence Research and Development*, vol. 131, 2005
75. [Shao & Ng, 2004] Shao L., Ng T. "Mining New Word Translations from Comparable Corpora". *Proceedings of COLING 2004,* pages 618-624, Geneva, Switzerland, 2004
76. [Simard & колектив, 1993] Simard M., Foster G., Isabelle P. "Using Cognates to Align Sentences in Bilingual Corpora". *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 67–81, Montreal, Canada, 1993
77. [Smucker & колектив, 2007] Smucker M., Allan J., Carterette B. "A Comparison of Statistical Significance Tests for Information Retrieval Evaluation". *Proceedings of ACM-CIKM 2007*, pages 623-632, Lisboa, Portugal, 2007
78. [Sparck-Jones, 1972] Sparck-Jones K. “A Statistical Interpretation of Term Specificity and its Application in Retrieval”. *Journal of Documentation*, volume 28, pages 11-21, 1972
79. [Taskar & колектив, 2005] Taskar B., Lacoste-Julien S., Klein D. "A Discriminative Matching Approach to Word Alignment". *Proceedings of HLT/EMNLP 2005*, pages 73-80, Vancouver, Canada, 2005
80. [Tiedemann, 1999] Tiedemann J. "Automatic Construction of Weighted String Similarity Measures". *Proceedings of the SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 213-219, College Park, MD, USA, 1999
81. [Tiedemann, 2004] Tiedemann J. "Word to Word Alignment Strategies". *Proceedings of COLING 2004*, pages 212-218, Geneva, Switzerland, 2004
82. [Tufis & колектив, 2006] Tufis D., Ion R., Ceausu A., Stefanescu D. "Improved Lexical Alignment by Combining Multiple Reified Alignments". *Proceedings of EACL 2006*, pages 153-160, Trento, Italy, 2006
83. [Vogel & колектив, 1996] Vogel S., Ney H., Tillmann C. "HMM-based Word Alignment in Statistical Translation". *Proceedings of COLING-96*, pp. 836-841, Copenhagen, Denmark, 1996
84. [Vossen, 1998] Vossen P. (editor): "EuroWordNet: A Multilingual Database with Lexical Semantic Networks", Dordrecht, Netherlands, Kluwer, 1998
85. [Wagner & Fischer, 1974] Wagner R., Fischer M., "The String-to-String Correction Problem", *Journal of the ACM*, volume 21, issue 1, pages 168-173, New York, NY, USA, 1974
86. [Weeds, 2003] Weeds J. “Measures and Applications of Lexical Distributional Similarity”. *Ph.D. Thesis, University of Sussex*, United Kingdom, 2003
87. [Zobel & Dart, 1996] Zobel J., Dart P., "Phonetic String Matching: Lessons from Information Retrieval". *Proceedings of ACM-SIGIR '96*, pages 166–172, Zurich, Switzerland, 1996
88. [Адамчик, 2008] Адамчик Н. В. "Самый полный курс русского языка", издательство. ”Харвест”, 2008
89. [Акуленко, 1969] Акуленко В. "О ложных друзьях переводчика". *Англо-русский и русско-английский словарь ложных друзей переводчика*, сост. В. В. Акуленко, С. Ю. Комиссарчик, Р. В. Погорелова, В. Л. Юхт, Москва, Русия, 1969
90. [Беляев, 1940a] Беляев А. "Властелин мира" (1940 г.), *издателство "Оникс 21 век"*, 2005 г., ISBN 5-329-01356-9, http://lib.ru/RUFANT/BELAEW/lordwrld.txt (посетен през април, 2009)
91. [Беляев, 1940b] Беляев А. "Владетелят на света" (1940 г.), превод от руски А. Траянов, *изд. „Народна младеж“*, 1977 г., http://www.chitanka.info/lib/text/2130 (посетен април, 2009)
92. [Бернщайн, 1986] Бернщайн, С. Б. "Българо-руски речник". *Издателство "Русски език"*, Москва, Русия, 1986
93. [Влахов & Тагамлицка, 1985] "Руско-български речник" под редакцията на С. Влахов и Г. Тагамлицка, *издателство "Наука и изкуство"*, София, 1985
94. [Зализняк, 1977] Зализняк А. "Граматический словарь русского языка", *издательство "Русский язык"*, Москва, Русия, 1977

95. [Зинкевич, 2001] Зинкевич А. "Учебный болгарско-русский словарь ложных лексических параллелей", Минск, 2001
96. [Колесников, 1995] Колесников Н. "Семонимические словари. I. Словарь паронимов русского языка. II. Словарь антонимов русского языка", Ростов на Дон, Русия, 1995
97. [Наков, 2008b] Наков С. "Измерване на междуезикова семантична близост чрез търсене в Google". *Доклади от 5ma международна конференция "Езикът: феномен без граници"*, ISBN 978-954-9685-43-5, страници 238-242, Варна, България, 2008
98. [Новиков & колектив, 1987] Новиков Л., Иванов В., Кедайтене Е., Тихонов А. "Современный русский язык. Теоретический курс. Лексикология", Москва, Русия, 1987
99. [Чукалов, 1986] Чукалов С. К. "Руско-български речник", *издателство "Русски език"*, Москва, Русия, 1986

Авторска справка

В настоящата секция са описани приносите и публикациите, свързани с дисертацията и кратки биографични данни за автора.

Научно-приложни приноси

1. Разработени и изследвани са алгоритми SemSim и CrossSim за измерване на едноезична и междуезикова семантична близост между двойка думи чрез използване на уеб като корпус и анализиране на резултатите от серия справки в уеб търсеща машина. Резултатите са докладвани на научна конференция "Езикът – феномен без граници" [Наков, 2008b] и представляват съществена част от публикации на престижни международни конференции и специализирани научни списания: [Nakov & колектив, 2007a], [Nakov & колектив, 2007b], [Nakov, 2008a], [Nakov, 2009] и [Nakov & колектив, 2009a].
2. Разработен и изследван е нов, алгоритъм MMEDR за подобрено измерване на ортографска близост между двойка българска и руска думи, който съобразява лингвистични особености на тези езици. Алгоритъмът подобрява точността на традиционните ортографски мерки за близост LCSR и MEDR с над 18%. Резултатите са докладвани на престижна научна конференция по обработка на естествен език (RANLP 2009) в специализирана секция (workshop) за междуезикови ресурси, технологии и оценяване за езиците от Централна и Източна Европа [Nakov & колектив, 2009b].
3. Разработен и изследван е алгоритъм за различаване между когнати и фалшиви приятели, постигащ висока усреднена интерполирана точност (над 96%). Резултатите са докладвани на престижна научна конференция по обработка на естествен език (RANLP 2007) в секция (workshop) за съставяне и управление на многоезични речници [Nakov & колектив, 2007b].
4. Разработен и изследван е алгоритъм за извличане на фалшиви приятели от паралелен двуезичен корпус, който комбинира статистически и семантични признаци за различаване между когнати и фалшиви приятели. Алгоритъмът постига усреднена интерполирана точност от над 78%, което е значително по-добър резултат от известните до момента алгоритми за решаване на тази задача. Резултатите са публикувани в научно списание Serdica Journal of Computing [Nakov, 2009] и на престижна научна конференция по обработка на естествен език RANLP 2009 [Nakov & колектив, 2009a].
5. Разработен и изследван е алгоритъм за автоматично извличане на синоними от текстови корпуси, базиран на измерване на семантична близост чрез използ-

ване на търсеща машина като източник на семантична информация. Резултатите са публикувани на научна конференция DSC-2008 [Nakov, 2008a].

6. Разработен и изследван е алгоритъм за подобряване на подравняването на думи чрез използване на уеб като корпус. Резултатите са докладвани на престижна научна конференция по обработка на естествен език RANLP 2007 [Nakov & колектив, 2007b].

Приложни приноси

1. Описаният във втора глава алгоритъм MMEDR (за български и руски език) е имплементиран като част от публично достъпния инструмент с отворен код TECFF (Toolkit for Extraction of Cognates and False Friends): http://code.google.com/p/cognates-and-false-friends-tools/.
2. Описаните в трета глава алгоритми SemSim (за български, руски и английски език) и CrossSim (за български, руски и английски език) са имплементирани като част от инструмента TECFF, публично достъпен от http://code.google.com/p/cognates-and-false-friends-tools/.
3. Алгоритъмът FFExtract, описан пета глава, е имплементиран като част от инструмента TECFF, публично достъпен от http://code.google.com/p/cognates-and-false-friends-tools/.

Публикации по дисертацията

1. Nakov P., Nakov S., Paskaleva E. "Improved Word Alignments Using the Web as a Corpus", *Proceedings of International Conference "Recent Advances in Natural Language Processing" (RANLP 2007)*, pages 400-405, Borovets, Bulgaria, 2007
2. Nakov S., Nakov P., Paskaleva E. "Cognate or False Friend? Ask the Web!", *Proceedings of the 1st International Workshop on Acquisition and Management of Multilingual Lexicons, held in conjunction with RANLP 2007*, pages 55–62, Borovets, Bulgaria, 2007

   Статията е цитирана в следните публикации, които не са свързани с авторите:

   - Mitkov R., Pekar V., Blagoev D., Mulloni A. "Methods for Extracting and Classifying Pairs of Cognates and False Friends". *Machine Translation*, volume 21, issue 1, pages 29-53, Springer Netherlands, 2007
   - Uzun L., Salihoglu U. "Cognates and False Cognates: Compiling a Corpus and Testing How They are Translated by Computer Programs". *Poznań Studies in Contemporary Linguistics*, volume 45, issue 4, Versita, Warsaw, 2009
3. Nakov S. "Automatic Acquisition of Synonyms Using the Web as a Corpus". *Proceedings of the 3rd Annual South-East European Doctoral Student Conference (DSC 2008)*, Volume 2, pages 216-229, Thessaloniki, Greece, 2008
4. Наков С. "Измерване на междуезикова семантична близост чрез търсене в Google". *Доклади от 5ma международна конференция "Езикът: феномен без граници"*, ISBN 978-954-9685-43-5, страници 238-242, Варна, България, 2008
5. Nakov S. "Automatic Identification of False Friends in Parallel Corpora: Statistical and Semantic Approach", *Serdica Journal of Computing*, issue 3, pages 133-158, 2009
6. Nakov S., Nakov P., Paskaleva E. "Unsupervised Extraction of False Friends from Parallel Bi-Texts Using the Web as a Corpus", *Proceedings of International*

Conference "Recent Advances in Natural Language Processing" (RANLP 2009), pages 292-298, Borovets, Bulgaria, 2009

7. Nakov S., Paskaleva E., Nakov P. "A Knowledge-Rich Approach to Measuring the Similarity between Bulgarian and Russian Words", *Workshop on Multilingual Resources, Technologies and Evaluation for Central and Eastern European Languages held in conjuction with RANLP 2009*, Borovets, Bulgaria, 2009

Кратка автобиография

Светлин Наков е роден през 1980 г. в гр. Велико Търново. Завършва местната математическа гимназия с Национална диплома за отличен успех и високи резултати в областта на информатиката през 1999 г. Като ученик е победител в десетки конкурси и състезания по програмиране и е носител на четири медала от международни олимпиади. Приет е за студент във ФМИ на СУ без изпит от олимпиадата по информатика. Завършва бакалавърска степен на обучение във ФМИ през 2003 г. и магистърска степен също във ФМИ по специалност "Разпределени системи и мобилни технологии" през 2005 г.

Светлин е автор на 5 книги по програмиране (някои с обем над 1000 страници) и на десетки научни и приложни публикации. Лектор е в над 50 семинара и технологични конференции. Като хоноруван преподавател е организирал и провел над 20 университетски и извънуниверситетски курса за обучение в областта на информационните технологии. Учредител и председател е на Българска асоциация на разработчиците на софтуер (БАРС).

През 2004 г. Светлин е награден от Президента на България с наградата "Джон Атанасов" за принос към развитието на информационните технологии и информационното общество. Получава още няколко десетки отличия и награди за заслуги в областта на информационните технологии и е обявен за почетен гражданин на Велико Търново през 2005 г.

Работи като софтуерен инженер, ръководител на екипи и проекти, консултант, преподавател, технологичен предприемач и съдружник в няколко успешни фирми.

Научните му интереси са в областта на компютърната лингвистика, разработката на софтуер и обучението по програмиране и информационни технологии.