БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИНСТИТУТ ПО ИНФОРМАЦИОННИ И КОМУНИКАЦИОННИ ТЕХНОЛОГИИ

Александър Николаев Попов

Моделиране на лексикалното знание с цел автоматична обработка на естествен език

АВТОРЕФЕРАТ

НА ДИСЕРТАЦИЯ

за присъждане на образователната и научната степен „доктор“

по научна специалност 01.01.12 „Информатика“

Научен ръководител: доцент д-р Кирил Симов

София, 2018

Дисертацията е обсъдена и допусната до защита на разширено заседание на секция „Лингвистично моделиране и обработка на знания“ на ИИКТ-БАН, състояло се на 3 юли 2018 г.

Дисертацията съдържа 147 стр., в които 8 фигури, 23 таблици и 13 стр. литература, включваща 141 заглавия.

Защитата на дисертацията ще се състои на . .2018 г. от : часа в зала на блок на ИИКТ-БАН на открито заседание на научно жури в състав:

1. проф. Мария Нишева, СУ “Кл. Охридски”, ФМИ
2. проф. Иван Койчев, СУ “Кл. Охридски”, ФМИ
3. проф. Людмила Димитрова, ИМИ-БАН
4. проф. Галя Ангелова, ИИКТ-БАН
5. доц. Геннадий Агре, ИИКТ-БАН

Материалите за защитата са на разположение на интересуващите се в стая на ИИКТ-БАН, ул. „Акад. Г. Бончев“, .

Автор: Александър Николаев Попов

Заглавие: Моделиране на лексикалното знание с цел автоматична обработка на естествен език

1 Увод и обща характеристика на дисертационния труд

1.1 Актуалност на темата

Моделирането на лексикалното знание по отношение на естествените езици от дълго време е основна задача за редица дисциплини, изучаващи принципите, чрез които езиковите системи правят възможна човешката комуникация. Компютърната лингвистика и обработката на естествен език (ОЕЕ) следват теоретичната лингвистика в опита ѝ да намери подходящо формално представяне на лексикалното знание. Във все по-голяма степен именно лексиконът бива мислен като средоточие на множество типове информация (като например синтактична структура, семантични валентни рамки, допустими колокации и пр.), използвана от голям брой приложения за ОЕЕ.

Налице са два общи подхода към моделирането на лексикалното знание в ОЕЕ. Един от тях се опитва да формализира лексикона според дадена теоретична рамка. Освен лексикализираните формални граматики, някои символни модели, които се фокусират предимно върху лексикона, са WordNet (Fellbaum, Christiane, 1998), FrameNet (Baker et al., 1998), VerbNet (Schuler, 2005) и др. Изследователски проекти като WordNet (WN) моделират значението в езика през релациите между понятията и цялостната структура на лексикона. Вторият подход е насочен към представяне на лексикалната информация, което е по-размито и основано на вероятностни модели. Изследователски направления като дистрибутивната семантика[1] се опитват да моделират свойствата на езиковите единици въз основа на това кога и колко често те биват използвани. Това е усилие, което включва анализ на големи масиви от данни, на корпуси, които са представителни езикови извадки. При този подход моделите за представяне на значение биват построявани автоматично от данните.

Двете алтернативи, описани горе, явно могат да се допълват взаимно по различни начини. Модели, които успешно обхващат езиковото разнообразие, присъщо на лексикона, могат да помогнат за автоматичното моделиране на езиковата структура на множеста нива на описание. Такова знание от своя

[1] Дистрибутивната семантика изследва сходствата между езиковите единици въз основа на техните разпределения в големи масиви от езикови данни. Основната хипотеза на направлението е, че езикови единици със сходни контексти на употреба (т.е. сходни разпределения) имат сходни значения.

страна би спомогнало да се създават по-качествени приложения, разчитащи на подобен анализ, да се генерират по-експресивни характерни признаци за системи с машинно самообучение, може би дори да се достигне до теоретични пробиви, основани върху анализ на големи данни.

1.2 Посоки на изследване

Дисертацията очертава няколко съществени посоки на изследване, които не са получили достатъчно голямо внимание в наличната за областта литература:

- Обогатяване на символното представяне на значенията в WN чрез извличане на лексикални и семантични релации от други източници на знание.
- Комбиниране на знание от източници със символно представяне на значението (като WN) със знание, построено чрез статистически методи, т.е. получаване на хибридни модели, които съчетават знание, построено едновременно от лингвистичната теория и от текстови масиви.
- Построяване на вероятностни модели на лексикално знание, които включват представяния не само на думи, но и на основни форми, лексикални значения, контексти на употреба, граматическа информация.
- Построяване на такива вероятностни модели като инструмент за оценка на качеството на семантичните мрежи и за оценка на техните изследвани разширения.
- Разработване на невронни архитектури за снемане на лексикалната многозначност (СЛМ), които се справят с набор от проблеми: решаване на задачата за СЛМ на *всички пълнозначни думи* чрез единствен модел; взимане на информирани решения относно думи, които са непознати от учебните данни; извличане на отличителни признаци за вход от различни по вид източници; представяне на контексти в семантично пространство с голяма експресивност; съчетаване на аспекти на лексикалното знание с цел подобряване устойчивостта на представянията.

1.3 Цели на дисертацията

Общите цели на дисертацията са следните:

- Да изследва WN като основа за решаването на задачи при ОЕЕ, включващи лексикален анализ, и при възможност да обогати неговия семантичен модел.

- Да изследва различни автоматични методи за решаването на една от основните задачи за лексикален анализ – снемане на лексикалната многозначност.
- Да разработи дистрибутивни модели на лексикално знание, които да включват и знание, извлечено от теорията.
- Да изследва различните аспекти на лексикалното знание, това как те могат да бъдат кодирани и какво взаимодействие протича между тях (например когато няколко вида задачи за лексикален анализ биват едновременно решавани върху един и същи текст).

2 Дефиниция на задачите

2.1 Определяне частите на речта

Формално под определяне частите на речта (ОЧР) ще се има предвид процедура, която задава етикет $p_i \in P$ на всяка дума в даден текст T. Или с други думи, това е задачата да се намери функция A такава каквато:

$$A(w_i) = p_j$$

където:

- $w_i \in T = (w_1, w_2, ..., w_n)$
- $p_j \in P$

T е текстът, частите на речта на чиито думи трябва да бъдат определени (затова е важно въпросните да бъдат разглеждани в поредност), а P е списъкът с етикети, от които биват избирани частите на речта. Искаме A да ни даде етикет за всяка една от думите в T.

Този труд представя резултати от ОЧР за данни на български език, и по-точно корпуса BulTreeBank (Simov & Osenova, 2004), като използва средно детайлен списък с етикети (153 на брой). В последната част от дисертацията, където задачите за ОЧР и СЛМ биват решавани успоредно, е използван корпусът за английски език SemCor (Miller et al., 1993); за този експеримент е използван опростен списък с 12 етикета за части на речта (Petrov et al., 2011).

2.2 Снемане на лексикалната многозначност

Формално под СЛМ ще се има предвид задачата за намиране на функция A такава каквато[2]:

$$A(w_i) = s^j_{w_i}$$

където:

- $w_i \in T = (w_1, w_2, ..., w_n)$
- $s^j_{w_i} \in (s^1_{w_i}, s^2_{w_i}, ..., s^k_{w_i}) = Senses_D(w_i)$
- $D = (Senses_D(w_1), Senses_D(w_2), ..., Senses_D(w_l))$

Също като при ОЧР, T е текстът, с който работим, но за всяка дума в T, която присъства в речник D, може да има налице k различни съответстващи ѝ значения, като k не е фиксирано положително число. Тоест, при СЛМ няма единен списък с етикети, който се използва за класифицирането на всички думи в текста, а вместо това всяка дума в речника разполага със собствен списък. Това на практика означава, че СЛМ е сбор от класификационни задачи.

Съществуват два широко приети варианта на задачата за СЛМ, което е отразено в популярните състезания Senseval/SemEval. Един от тях е задачата за СЛМ на *лексикална мостра*, при която системата трябва да снеме многозначността на определено подмножество от пълнозначните думи в дадено изречение (обикновено само една от тях). Другият вариант е задачата за СЛМ на *всички думи*, при която всички пълнозначни думи (или тези, които присъстват в речника) са цел за СЛМ. Дисертацията ще се фокусира изцяло върху задачата за СЛМ на всички пълнозначни думи в текста.

Множествата от данни, които използвам, са както следва. За обучение на системи с обучаващ сигнал: SemCor (Miller et al., 1993) е използван за обучение на английски език, съдържа приблизително 360,000 думи от Браун корпуса (Kučera & Francis, 1967), около 226,000 от които имат за етикети значения от WN. За настройване на параметрите на моделите и оценка са използвани следните множества: Senseval-2 (Edmonds & Cotton, 2001), Senseval-3 (Snyder & Palmer, 2004), SemEval-2007 (Pradhan et al., 2007), SemEval-2013

[2]Тази формализация следва относително близко Navigli (2009).

(Navigli et al., 2013), SemEval-2015 (Moro & Navigli, 2015). При СЛМ, основано на знание (СЛМОЗ), моделите биват оценявани върху SemCor на английски и върху BulTreeBank на български (Popov et al., 2014). В един случай за оценка е използвано подмножеството от SemEval-2013.

2.3 Сходство и свързаност между думи

Пресмятането на сходство и свързаност между думи на практика представлява оценка на това колко близки по значение са те. Първата мярка означава доколко два термина означават едно и също, а втората изразява до каква степен термините биха могли да присъстват в един и същ контекст. Обикновено, когато се изграждат златни ресурси, свързани с тази задача, се избира числова скала на близост (напр. от 1 до 10) и анотаторите преценят колко сходни или свързани са членовете на двойките. В дисертацията степента на близост се изчислява от автоматични системи и бива съотнесена спрямо златните ранкирания посредством коефициента на рангова корелация на Спирмън. Следните множества от данни са използвани за оценка: WordSim-353 Relatedness, WordSim-353 Similarity (Agirre et al. (2009) описва двете подмножеста на оригиналното множество WordSim-353) и SimLex-999 (Hill et al., 2015).

3 Основни резултати в областта

3.1 WordNet

Настоящият труд представя анализ, който се опира върху използването на лексикони, представящи списък със значения на думи. За тази цел, като лексикон е избран най-популярният изчислителен ресурс с изброени значения в ОЕЕ – WordNet (Fellbaum, Christiane, 1998). Значенията на думи в WordNet (WN) са организирани в *множества от синоними*, или *синсети*, както ще бъдат наричани оттук нататък. Всяко значение представлява двойка от основна форма (лема) и понятие; понятието е общо между значенията, групирани в синсет и споделящи една и съща част на речта. Значенията за конкретна дума са подредени в синсетите от лексикографи според важността

им. Дисертацията използва като лексикон WordNet 3.0 [3]. За българските значения е използван частичен инвентар, подравнен с оригиналната йерархия на WN и използван за анотацията на синтактичен корпус (Popov et al., 2014). Друг използван ресурс е *eXtended WordNet* (XWN) (Mihalcea & Moldovan, 2001) – проект за анотирането на дефинициите в WN със значения от WN, така че анотираният текст да е достъпен за учебни/оценъчни данни или като някакъв вид обогатяване на WN.

Синсетите и значенията на думи са свързани помежду си чрез множество от лексикално-семантични релации. Тази структура предоставя друга интерпретация на лексикалното значение: лексикалните единици са разположени в мрежа от понятия и така тяхната относителна позиция в нея се превръща в силен индикатор на значение. WN дава достъп до два типа релации: лексикални и семантични. Лексикалните свързват конкретни значения на думи (т.е. значения на конкретни думи, които може би са групирани в синсети с други значения), докато семантичните релации свързват цели синсети (т.е. тези релации се отнасят до всички членове на свързаните синсети). Лексикалните релации включват следните типове: антонимия, отнесеност на прилагателно към дадено съществително (pertainymy), словообразувателни релации. Семантичните типове са: хипернимия, хипонимия, тропонимия, меронимия, холонимия, причинно-следственост, сходство, атрибут, релации от тип "виж още".

3.2 Снемане на лексикалната многозначност: обзор

Методите с машинно самообучение за СЛМ, използващи обучаващ сигнал, целят да обучат класификатори, които да са способни да анотират думи в свързан текст, при условие че възможните значения са налични в лексикален речник. Класификаторите биват обучавани върху анотирани текстови корпуси, в които някои или всички пълнозначни думи са били ръчно определени. Обикновено в задачите за ОЕЕ входът към системите с обучаващ сигнал бива конструиран като множество от характерни признаци. Navigli (2009) представя обзор на няколко такива алгоритми, сред които са *списъци и дървета с правила за взимане на решение*, *наивният класикатор на Бейс*, *учене чрез образци*. Системата, която понастоящем постига най-високите

[3] За повече информация виж `https://wordnet.princeton.edu/documentation/wnstats7wn`

резултати върху задачата за СЛМ, е такава с обучаващ сигнал – It Makes Sense (IMS), използваща метода на опорните вектори и фино настроена стъпка за извличане на характерни признаци (Zhong & Ng, 2010).

Методите за снемане на лексикалната многозначност, основани на знание (СЛМОЗ), представляват семейство от алгоритми за решаване на задачата, които не разчитат толкова на статистическо знание, научено директно от данни, колкото на информация, кодирана в моделиращи лексикона ресурси. Голямо предимство на СЛМОЗ е, че тези алгоритми имат пълно покритие над значенията в използвания лексикон. Някои от най-популярните варианти на СЛМОЗ са групирани под името *методи чрез графи*. При тях обикновено се използва ранкиращ алгоритъм за изчисляване на относителната важност на възлите в семантичния граф. Възлите с най-висок ранк биват подадени като избори за СЛМ. Един от най-популярните методи чрез графи е описан в Agirre & Soroa (2009). Той използва целия граф, за да извършва случайни обхождания по него (с някаква вероятност за спиране на всеки ход), като по този начин обновява теглата на възлите. Формулата за изчисляване на вектора $\mathbf{P}$ (описващ важността на всички възли в графа) е следната:

$$\mathbf{P} = cM\mathbf{P} + (1 - c)\mathbf{v}$$

където M е матрица с вероятности за преход с размери N x N (N е броят на възли в графа), c е заглушаващ параметър (обикновено зададен със стойност 0.85) и $\mathbf{v}$ е стохастичен вектор с размерност N x 1. Втората част от уравнението дава вероятността да бъде извършен случаен скок към който и да е от възлите във $\mathbf{v}$ и да бъде прекъснато обхождането. Алгоритъмът се повтаря, докато не постигне сходимост или в продължение на фиксиран брой пъти.

3.3 Приложения на СЛМ

Navigli (2009) се съсредоточава върху няколко потенциални области на приложение на СЛМ, като например извличане на информация (както в смисъла на "information retrieval" (Stokoe et al., 2003), така и на "information extraction" (Agirre et al., 2015)). СЛМ може да се използва като източник на характерни признаци за модели, извършващи основни задачи от ОЕЕ, като дependентен синтактичен анализ (Agirre et al., 2011), определяне на семантични роли (Zapirain et al., 2013) и др.

Машинният превод се явява естествено поле за приложение на СЛМ. Въпреки това, влиянието на изследванията за СЛМ върху машинния превод на е било особено силно. В експериментите, представени от Carpuat & Wu (2005), използването на СЛМ има негативно влияние върху BLEU резултатите. Авторите използват модел с обучаващ сигнал, за да подават варианти за превод на статистическия преводач, но това влошава резултатите. Други изследвания обаче подхождат към задачата различно и получават положителни резултати: чрез интерпретиране на изборите за СЛМ като "меки" варианти за превод, преосмисляне на СЛМ като задача за превод, максимизиране на дължината на предложените от СЛМ парчета текст, интегриране на СЛМ по-дълбоко в статистическите преводачи и пр. (Chan et al., 2007; Carpuat & Wu, 2007). Simov, Osenova, & Popov (2016a) използват изхода на система за СЛМ, за да подсилят използващия "фактори" статистически автоматичен преводач в контекста на превод от английски на български и обратно. В Simov, Popov, Zlatkov, & Kotuzov (2016) се използва формализма на Минималната рекусивна семантика за подобрено подравняване между текстовете на изходния и целевия език, което след това се използва за последваща преработка чрез правила на резултата от статистическия преводач.

3.4 Невронни мрежи за ОЕЕ

Частта от обзора на литературата, съсредоточена върху невронни езикови модели (НЕМ), стъпва предимно на Popov (2016b), а частта, посветена на СЛМ чрез невронни мрежи (НМ) – на Popov (2018).

3.4.1 Изкуствени невронни мрежи за ОЕЕ

НМ за пряко разпространение (НМПР) са сред най-простите изкуствени НМ. Те могат да се състоят от какъвто и да е брой слоеве между задължителните входен и изходен слой. Многослойните НМПР (често наричани *многослойни перцептрони*) имат допълнителни скрити слоеве, които им позволяват да научават много по-сложни взаимовръзки. Обучението в НМПР може да се осъществи по различни начини, но почти винаги това става чрез метода на *обратно разпространение* на градиентите на грешка. НМПР се характеризират още с изискването връзките между слоевете им да не образуват цикли (Sundermeyer et al., 2015). Това не им позволява да обработват дълги

контексти (изключително важна способност, когато са налице *зависимости с дистантен конституент*, какъвто често е случаят при естествените езици).

Рекурентните невронни мрежи (РНМ) са по-сложен вариант на НМ, които са разработени, за да могат да се справят с подобни дълги времеви серии (Graves, 2012). По-дългите контексти биват обработвани чрез циклични връзки между времевите състояния на скритите слоеве. Този механизъм играе ролята на един вид динамична памет на модела. Това поведение може да бъде формализирано по следния начин. Като при НМПР, за поредица от входни данни x_1, x_2, ..., x_n, мрежата изчислява изходен вектор $\mathrm{y_t}$ по формулата:

$$\mathbf{y_t} = W_{\mathrm{hy}} h_{\mathrm{t}} + b_{\mathrm{y}} \tag{1}$$

където W_{hy} е матрица с теглата на връзките между два слоя (в този случай между скрития и изходния слой), h_t е състоянието на скрития слой за конкретната времева стъпка, а b_y е отместването за изходния слой. Рекурентността е въведена чрез сметката на стойностите за h_t:

$$\mathbf{h_t} = F_{\mathrm{act}}(W_{\mathrm{xh}} x_{\mathrm{t}} + W_{\mathrm{hh}} h_{\mathrm{t\text{-}1}} + b_{\mathrm{h}}) \tag{2}$$

където F_{act} е функцията за активиране в скрития слой, първият член на сумата е входът, умножен по съответните тегла, вторият е векторът за предишното скрито състояние, умножен по съответните тегла, задаващи цикъла в скрития слой, и последният е отместването за скрития слой.

РНМ обаче страдат от проблема с експоненциално нарастващите или изчезващи градиенти на грешката. Колкото по-дълги са контекстите за обработка и съответно колкото по-дълбоки са обработващите ги мрежи, толкова по-големи или близки до нула стават обратно разпространените градиенти. Справянето с втората част от проблема е особено трудна и може да попречи изцяло на една НМ да осъществи обучение. За справянето с този проблем са предложени по-сложни варианти на РНМ, като например *клетките с дълга краткотрайна памет* (КДКП). Състоянието на КДКП бива повлияно от няколко *клапи*: *клапа за забравяне*, *клапа на входа* and *клапа на изхода*. КДКП съдържа четири скрити слоя вътре в себе си, които научават подходящите тегла, нужни за правилното забравяне и обновяване на (части от) състоянието на клетката.

3.4.2 Невронни езикови модели

Най-ранните опити за получаване на смислови представяния чрез НЕМ използват НМПР. Плъзгащият се контекстен прозорец избира думи, които са вложени в споделено пространство. Векторите за думите биват конкатенирани и подадени към скрит слой, който преобразува контекстното представяне на думата до вектор с размерността на пространството на влагане. Накрая функция за получаване на вероятностно разпределение (обичайно *softmax*) съпоставя вектора в пространството на влагане към векторно представяне с размерите на речника, което указва до каква степен е активирана всяка една дума в него.

Mikolov, Chen, et al. (2013) въвеждат *плитките* НМ като способ за обучение на модели върху количества данни, които са непостижимо големи до онзи момент. Тези мрежи са лишени от нелинейните скрити трансформации в многослойните мрежи, между входната и изходната фаза има единствено проектиращ слой. Така времето за обучение с тези *лог-линейни* архитектури бива намалено от седмици до дни, понякога дори часове, върху корпуси от милиарди думи. В цитираната публикация са предложени две архитектури: *CBOW* и *Skip-Gram*.

3.4.3 НЕМ за представяне на понятия, значения и синсети

Описаните методи могат да бъдат приложени за учене на дистрибутивни представяния на единици от всякакъв вид, срещани в естествени поредици. Има няколко подхода за постигане на тази цел по отношение на понятийни единици като синсети: *методи за адаптиране на налични ресурси* (Rothe & Schütze, 2015); автоматично анотиране със значения на големи корпуси и обучаване на НЕМ върху тези данни (Iacobacci et al., 2015); модифициране на плитки НЕМ за успоредно обучаване с думи и възможни синсети (Mancini et al., 2016); генериране на "псевдокорпуси" от идентификатори на синсети и леми чрез случайни обхождания на графи (Goikoetxea et al., 2015).

3.4.4 Невронни мрежи за анотиране на поредици от данни

Една модификация на КДКП, която е много популярна по отношение на решаването на задачи за трансформация на серии по принцип е превръщането им

в *двупосочни* (ДКДКП) Schuster & Paliwal (1997)). Това означава, че входната серия е подадена два пъти на два различни слоя с КДКП – първият получава входа в оригиналната му подредба ($\{w_1, w_2, ..., w_n\}$), а вторият го получава обърнат ($\{w_n, w_{n-1}, ..., w_1\}$). Двете КДКП произвеждат поредица от изходни данни: $\{h_1^{forward}, h_2^{forward}, ..., h_n^{forward}\}$ and $\{h_1^{backward}, h_2^{backward}, ..., h_n^{backward}\}$. Тази модификация на рекурентната архитектура позволява на модела да гледа както напред, така и назад спрямо позицията на думата/елемента, който се опитва да определи.

Някои скорошни трудове, в които ОЧР е третирано като задача за трансформация на серии, са: Wang et al. (2015); Huang et al. (2015); Plank et al. (2016). Всички те използват ДКДКП и векторни представяния на думи като входни характерни признаци и извършват класифициране (чрез softmax или CRF слоеве) спрямо списъка с етикети за част на речта, като постигат върхови или близки до върхови резултати върху данни за различни езици. По отношение на ОЧР за български, Simov & Osenova (2001) описват хибриден подход: РНМ, комбинирана със система от правила, постига "точност от 95.17% за ОЧР и 92.87% за определяне на всички граматически признаци". Simova et al. (2014) докладват точността на няколко тагера за ОЧР върху данните от корпуса BulTreeBank: 95.91% (BLL Tagger), 94.92% (Mate morphology tagger) и 93.12% (TreeTagger).

Съществуват няколко основни подхода за използване на РНМ за СЛМ. Един от тях е да се получи представяне на думата в скрит слой и то да се използва за класификация. Kågebäck & Salomonsson (2016) обучават система с ДКДКП за решаването на задачата за СЛМ на *лексикална мостра* – тя постига върхови резултати, но трябва да научи отделни експертни модели за различните основни форми. Raganato, Bovi, & Navigli (2017) предлагат няколко невронни архитектури за решаване на задачата за СЛМ на *всички думи*. Те също използват тагер с ДКДКП, който обаче се учи да класифицира всички пълнозначни думи по време на едно единствено преминаване през контекста. Друг популярен метод за невронно СЛМ се стреми да представи контекста на употреба на целевата дума като вектор в пространство на влагане и след това да сравни това представяне с предварително изчислени векторни представяния за множеството от възможни значения, свързани с основната форма на целевата дума. Най-популярната система от това семейство е *context2vec* (Melamud et al., 2016).

3.4.5 Успоредно обучение на НМ върху няколко задачи

Под "успоредно обучение върху няколко задачи" се има предвид комбиниране на два или повече обучаващи сигнала, спрямо които НМ оптимизира параметрите си. Могат да бъдат комбинирани различни задачи, напр. ОЧР и синтактичен анализ, или синтактичен анализ и определяне на честотата на думите в контекста на ОЕЕ. По този начин задачите споделят параметри в скритите слоеве на мрежата, стоящи на основния изчислителен път и обновявани отделно спрямо различните обучаващи сигнали. Основната мотивация за такова успоредно обучение е хипотезата, че различните типове анализ разчитат на различни типове структури в езика. Затова взимането предвид на повече от един обучаващ сигнал може да принуди мрежата да научи в скритите си слоеве представяне, което отразява тези различни типове структури. Alonso & Plank (2017) описват изследване на различни комбинации от задачи за ОЕЕ. Те заключават, че успоредното обучение върху няколко задачи не винаги е ефективно и взаимодействието между различните задачи трябва да бъде проучено подробно от гледна точка на теорията на информацията.

4 Рекурентни невронни мрежи за определяне частите на речта

Главата описва невронна архитектура за решаване на задачата за ОЧР, оценена върху корпус с данни на български (Popov, 2016a). Основната цел е да се изследват РНМ, и по-конкретно ДКДКП, като инструмент за анализ на последователности от словоформи – т.е. изречения от езикови корпуси. Допълнителна цел е да бъдат създадени дистрибутивни представяния (т. нар. word embeddings) за български език, а в допълнение към тях и аналогичен модел за представяне на морфологична информация, която е ключова за решаване задачата за ОЧР.

4.1 Векторни представяния на думи и наставки на български

С цел получаване на дистрибутивни представяния на български думи беше събран сравнително балансиран корпус от около 220 милиона думи. За обучението на векторните представяния беше използвана архитектурата Skip-Gram

в Word2Vec[4]; векторите са с размерност 200 позиции. Беше приложена и друга процедура за обучение на дистрибутивни езикови представяния. Този тип представяне цели да обхване информация относно морфологични сегменти. По-малък корпус от 10 милиона думи беше извлечен за целта и предварително обработен, така че да кодира предимно морфологична информация. Думите в текста бяха подменени с части от тях, тук наричани "наставки" и мислени като грубо съответстващи си с морфемите в края на словоформите. Наставките са продукт на проста нормализация на оригиналните думи, т.е. не е извършван истински морфологичен анализ. Размерността на векторните представяния на наставките е зададена със стойност 50.

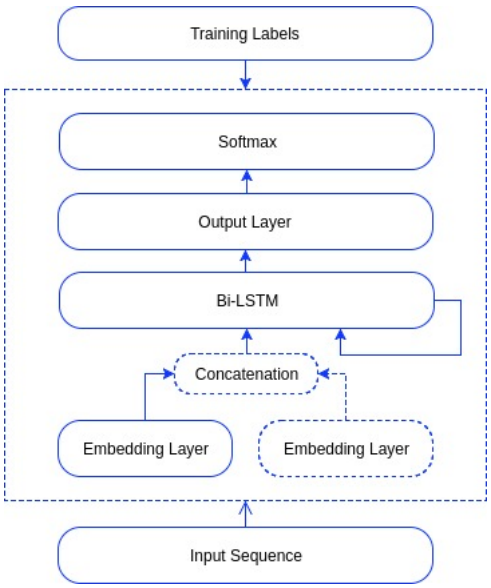
4.2 Многослойна архитектура за ОЧР

Към ОЧР тук е подходено чрез модел с ДКДКП. Състоянията на двата слоя с КДКП биват конкатенирани на всяка времева стъпка и оразмерени чрез линейна трансформация и резултатният вектор бива подаден на класификационен softmax слой, който пресмята вероятностно разпределение върху множеството от възможни етикети за част на речта (ЧР). По време на обучението това разпределение бива сравнявано с вектор, представящ златния етикет чрез бинарна кодировка (единица съответства на правилната категория, навсякъде другаде са записани нули). Архитектурата позволява конкатенирането на различни входове – в този случай векторни представяния на думи и наставки (виж фиг. 4.1).

4.3 Експерименти и резултати

В следващите експерименти са използвани данни за обучение и оценка от корпуса с ЧР BulTreeBank. За обучението на докладваните модели е използвана слабо оптимизирана параметризация на РНМ: един скрит слой с ДКДКП; темпо на учене със стойност 0.3; cross entropy за целева функция; метод на градиентното спускане като тренировъчен алгоритъм; не са използвани техники за регуларизация. Размерът на скрития слой варира слабо при различните експерименти. Размерността на представянията на думи беше определена чрез няколко първоначални експеримента. Две множества от вектори бяха обучени върху корпуса – едно с вектори с размерност 200 и едно с размерност

[4] `https://code.google.com/archive/p/word2vec/`



Фигура 4.1: РНМ за анотиране на поредици от данни: прекъснатите линии означават, че даден компонент е незадължителен. Фигурата е взета от Popov (2017).

600. Резултатите върху задачата за ОЧР дават основания да се предполага, че по-малкият размер е по-подходящ за този контекст (табл. 4.1).

Размер на векторите за думи	Точност
200	77.67%
600	71.89%

Таблица 4.1: Точност при ОЧР в зависимост от размерността на векторите, представящи входните думи (след 10000 учебни итерации)

Представен е още един експеримент, който демонстрира надеждността на подхода чрез представяния на наставки. Невронната мрежа бива захранена само с вектори, представящи наставки, без да използва представяния на думи. Таблица 4.2 показва, че моделът с представяния на наставки се обучава много бързо в началните фази, всъщност почти толкова добре колкото модела, използващ представяния на думи (вж. табл. 4.1).

Размер на векторите за наставки	Точност
50	77.11%
200	78.16%

Таблица 4.2: Точност при ОЧР в зависимост от размерността на векторите, представящи входните наставки (след 10000 учебни итерации)

Таблица 4.3 представя процентната точност за три различни конфигурации на невронната мрежа. Първият модел използва като вход само векторни представяния на думи (с размерност 200). Вторият комбинира на входа

представяния на думи и наставки (чрез конкатенация); заради по-големия вход, той има и по-голям скрит слой (125 неврона на слой в КДКП, сравнено със 100 при първия). Добавянето на векторните представяния на наставките увеличава мощността на тагера за ОЧР. Резултатите валидират използването на архитектурата с РНМ като решение на задачи от ОЕЕ за трансформация на поредици от данни.

Модел	Точност
Вектори за думи (100 неврона)	91.45%
Вектори за думи + наставки (125 неврона)	**94.47**%
Вектори за думи (125 неврона)	91.13%

Таблица 4.3: Точност при ОЧР, когато се използват само векторни представяни на думи и когато последните биват комбинирани с представяния на наставки (след 100000 учебни итерации)

5 Моделиране на лексикалната семантика чрез графи

Тук се изследва хипотезата, че обогатяването на графите със знание (ГСЗ) – чрез допълнителни парадигматични и синтагматични релации[5] – може да бъде осъществено чрез извличане на информация от текстови и от структурирани и полуструктурирани ресурси. Работата, представена тук, е първоначално описана в следните статии: Simov et al. (2015), Simov, Popov, & Osenova (2016b), Simov, Popov, & Osenova (2016a), Simov, Osenova, & Popov (2016b). Като основа на ГСЗ е използвана семантичната мрежа на WN, в която двойки от синсети са свързани чрез семантични релации. Чрез различни техники са извлечени различни множества от релации от разнообразни източници, като чрез задачата за СЛМОЗ са оценени приносите на новите множества с релации, прибавени към основния граф. От теоретична гледна точка, обогатяването на WN със синтагматични релации (т.е. синтактични, контекстуални и т.н.) би трябва да направи семантичната мрежа по-адекватна към моделирането на лексикалното познание, което далеч не се изчерпва с парадигматичните релации, анализирани от лексикалната семантика.

[5]Парадигматичните релации (или отношения) са такива, които са в сила между единици от една и съща категория. Тоест, тези елементи биха могли да бъдат заменени едни с други в контекст, като така ще бъде запазена граматичността на израза, макар и да се измени смисълът му. Синтагматичните релации са отношения между елементи, които са поставени в комбинация в езиков контекст, например в свързано изречение; свързаните чрез синтагматичните релации елементи не могат да бъдат заменяни едни с други по подразбиране. Понятията са популяризирани най-вече от структуралистите в началото на 20-и в.

5.1 Подобряване на СЛМОЗ спрямо данни на български

Първото множество от нови релации, извлечени за целите на СЛМОЗ, не произхожда от корпуса с български данни, а от самия WN. Транзитивната обвивка на хипернимната йерархия бива изчислена и експлицирана във възприетия формат за представяне на релации в ГСЗ (**I**). Например: значението от WN *doctor#1 (doctor%1:18:00::)* (лекар) е хипоним на *medical practitioner#1 (medical_practitioner%1:18:00::)* (специалист по медицина) и хиперним на *surgeon#1 (surgeon%1:18:00::)* (хирург), следователно е добавена директна релация между *surgeon#1 (surgeon%1:18:00::)* и *medical practitioner#1 (medical_practitioner%1:18:00::)*. Този тип логическо експлициране е извършено за всички двойки хиперним-хипоним в мрежата на WN, съдържащи идентификатори на синсети, които се срещат в българския WN.

Второто множество от нови релации между синсети е синтагматично по естество (**S**). За да бъде създадено то, синтактичните анотации в BulTreeBank са използвани заедно с анотациите на значения, когато такива са налични. Този труд се прицелва в следните депендентни релации: *nsubj*, *nmod*, *amod*, *iobj*, *dobj*. Тези релации са преобразувани до WN формата и добавени по този начин към семантичната мрежа. Релациите, извлечени от BulTreeBank, след това са генерализирани към ново множество по същия начин като при транзитивната обвивка на хипернимните релации (**SE**). Извършена е и допълнителна процедура за генерализиране, която взима хипернимите на възлите на съществителните и след това взима всички *техни* хипоними, поставяйки ги в аналогични релации (**SEU**).

Последното множество от нови релации е добито чрез йерархията от семантични полета в WN (Bentivogli et al., 2004). Първо е изпробвана наивна стратегия, при която всички синсети, отнесени към дадено поле, са свързани със специален идентификатор, създаден специално за полето (**D1**). Използвана е и друга стратегия, при която всички синсети в дадено поле са свързани едни с други (**D2**).

Новоизвлечените релации са използвани за построяването на нови ГСЗ, които да бъдат използвани от алгоритъма PageRank. Тъй като някои от новите релации са извлечени от златните данни, част от корпуса BulTreeBank е отделена с цел оценка: от 40-те файла в корпуса, 37 са използвани за извличане на нови релации и 3 – изцяло за оценка. Графите, които не включват релации

от синтактично естество от златните данни, могат да бъдат оценени и върху целия корпус. Таблица 5.1 показва резултатите за множествата от релации, оценени върху всички данни[6]. Таблица 5.2 показва резултатите върху трите файла, които не са използвани за извличане на нови релации. Видно е, че синтактичните релации подобряват точността значително. Генерализираните синтактични релации имат особено голям принос.

Комбинация	Точност
WN	51.6%
WNG	54.2%
WNI	53.7%
WNGI	54.9%
WNGID1	54.9%
WNGID2	55.1%

Таблица 5.1: Резултати върху целия златен корпус

Комбинация	Точност
WN	51.7%
WNG	53.8%
WNI	53.5%
WNGI	53.7%
WNGID1	53.8%
WNGID2	55.0%
WNGIS	56.5%
WNGISE	61.6%
WNGISED1	61.7%
WNGISED2	62.4%
WNGISEUD2	65.6%

Таблица 5.2: Резултати върху част от корпуса (3 файла)

5.2 СЛМОЗ върху английски данни. Анализ на типовете релации

Графите, използвани като основа за сравнение при оценката на различните типове релации от WN, са следните: **WN** (оригиналните релации от WN), **GL** (релациите, получени от корпуса с дефинициите) и **WNG** (комбинацията от двата). В този случай те служат не като долна граница на точност, а по-скоро като горна граница (табл. 5.3).

[6]PageRank беше приложен чрез софтуеъра UKB (http://ixa2.si.ehu.es/ukb/), използван с настройките по подразбиране: контекстен прозорец от 20 думи, класифицирани заедно, след 30 итерации на алгоритъма

ГСЗ	*SemCor*	*BTB*
WN	49.37	52.97
GL	51.66	51.15
WNG	58.97	55.90

Таблица 5.3: Точност върху двата корпуса за оценка, при използване на оригиналните графи (Simov, Popov, & Osenova, 2016a).

За целта на оценката, оригиналните релации от WN са групирани в множества, които съответстват на релационните типове в WN. Следните експерименти дават някакви индикации за полезността на тези различни множества за СЛМОЗ; това е постигнато, като е премерена точността при използване на ГСЗ, които съдържат всяко от тези множества само по себе си. Подмножеството от релации **WN-Hyp** е възприето като най-базовия тип релация и е използвано във всички комбинации в таблица 5.4.

ГСЗ	*SemCor*	*BTB*	ГСЗ	*SemCor*	*BTB*
WN-Hyp	33.52	45.03	WN-Hyp+WN-Mm	33.70	44.81
WN-Hyp+WN-Ant	38.63	48.41	WN-Hyp+WN-Mp	35.67	45.22
WN-Hyp+WN-At	36.97	47.91	WN-Hyp+WN-Ms	33.57	45.31
WN-Hyp+WN-Cls	34.23	46.11	WN-Hyp+WN-Per	39.57	48.19
WN-Hyp+WN-Cs	33.54	44.99	WN-Hyp+WN-Ppl	33.53	45.11
WN-Hyp+WN-Der	39.03	50.63	WN-Hyp+WN-Sa	38.29	48.31
WN-Hyp+WN-Ent	33.30	44.65	WN-Hyp+WN-Sim	42.89	49.28
WN-Hyp+WN-Ins	34.18	45.13	WN-Hyp+WN-Vgp	34.22	46.07

Таблица 5.4: Резултати за различните подмножества от релации в WN, оценени върху SemCor и BTB (Simov, Popov, & Osenova, 2016a).

Резултатите показват, че не всички типове релации допринасят за полезността на ГСЗ, поне що се отнася до СЛМОЗ. Има няколко множества от релации, които или смъкват точността (WN-Ent), или я подобряват почти неуловимо (напр. WN-Cs, WN-Ms, WN-Ppl). Това позволява да се мисли, че някои от оригиналните релации биха могли по принцип да се изключат от ГСЗ, без това да ощети алгоритъма за СЛМОЗ. Подобен анализ е направен върху влиянието, което имат релациите, извлечени чрез логически правила от основните графи WN и WNG. За повече информация относно как са разгънати оригиналните релации, вж. Simov, Popov, & Osenova (2016a). Таблица 5.5 показва, че подобренията, дължащи се на такива релации, извлечени от оригиналната семантична мрежа на WN, са минимални, ако изобщо ги има. Предложен е и анализ на допълнително извлечените по подразбиране релации от eXtended WordNet. Те съставляват множеството **GL**. **GL** е разделено на

GL-A, **GL-N**, **GL-R** и **GL-V**. Таблица 5.6 показва влиянието на отделните подмножества върху точността при СЛМОЗ.

ГСЗ	*SemCor*	*BTB*	ГСЗ	*SemCor*	*BTB*
WN+WN-HypInfer	**53.40**	**53.70**	WNG+WN-HypInfer	58.59	55.20
WN+WN-AntInfer	48.57	**53.05**	WNG+WN-AntInfer	**59.14**	**55.93**
WN+WN-ClsInfer	48.43	**54.62**	WNG+WN-ClsInfer	57.84	**56.14**
WN+WN-Cs1stVInfer	49.32	**56.02**	WNG+WN-Cs1stVInfer	**59.06**	**55.93**
WN+WN-Cs2ndVInfer	**49.39**	**57.28**	WNG+WN-Cs2ndVInfer	58.95	**56.17**
WN+WN-DerNAInfer	48.76	**57.19**	WNG+WN-DerNAInfer	58.49	52.13
WN+WN-DerNNInfer	47.79	**56.74**	WNG+WN-DerNNInfer	58.80	52.86
WN+WN-DerNVInfer	47.73	**55.84**	WNG+WN-DerNVInfer	55.87	52.77
WN+WN-DerVNInfer	48.72	**55.99**	WNG+WN-DerVNInfer	**59.00**	53.56
WN+WN-Ent1stVInfer	49.34	**56.08**	WNG+WN-Ent1stVInfer	**58.98**	52.55
WN+WN-Ent2ndVInfer	49.36	**56.60**	WNG+WN-Ent2ndVInfer	58.92	52.70
WN+WN-InsInfer	49.03	**56.76**	WNG+WN-InsInfer	58.38	52.89

Таблица 5.5: Точност върху SemCor и BTB. Резултатите, които са по-високи от базовите, са потъмнени. (Simov, Popov, & Osenova, 2016a).

ГСЗ	*SemCor*	*BTB*	ГСЗ	*SemCor*	*BTB*
WN+GL-A	52.94	53.08	WN+GL-R	51.76	52.85
WN+GL-N	**57.04**	52.92	WN+GL-V	53.01	**56.01**

Таблица 5.6: Точност при комбинациите от базовите релации от WN и разделени по част на речта подмножества на GL релациите. Най-високите резултати за двата корпуса са потъмнени. (Simov, Popov, & Osenova, 2016a).

Синтактичните анотации също са използвани за генериране на нови семантични релации. SemCor е синтактично анализиран с модулната система за анализ IXA[7]. 49 от документите в корпуса са отделени като данни за оценка, а останалите са използвани като източник на нови релации. Отново, извлечените релации са организирани в нови подмножества, дефинирани от комбинацията между различните части на речта. Така са оформени следните подмножества: **SC-AA**, **SC-AN**, **SC-AV**, **SC-NN**, **SC-NV**, **SC-RA**, **SC-RN**, **SC-RR**, **SC-RV**, **SC-VN**, **SC-VV**.

Таблица 5.7 показва, че повечето от синтактично извлечените подмножества подобряват точността отвъд базовия резултат за WNG. Комбинацията от синтактично извлечени подмножества и базовите релации, която дава най-високите резултати върху SemCor (60.34%), е както следва: **WNG**, **SC-AA**, **SC-AN**, **SC-AV**, **SC-NN**, **SC-NV**, **SC-RA**, **SC-RN**, **SC-RR**, **SC-RV**,

[7] `http://ixa.si.ehu.es/Ixa`

ГСЗ	*SemCor*	*BTB*	ГСЗ	*SemCor*	*BTB*
WNG+SC-AA	**59.20**	**55.93**	WNG+SC-RN	58.94	55.89
WNG+SC-AN	**59.30**	55.89	WNG+SC-RR	**59.07**	**55.93**
WNG+SC-AV	**59.46**	55.78	WNG+SC-RV	**59.43**	52.71
WNG+SC-NN	58.81	**56.21**	WNG+SC-VN	**59.05**	55.55
WNG+SC-NV	**59.31**	**56.21**	WNG+SC-VV	**59.26**	53.78
WNG+SC-RA	**59.52**	**56.18**			

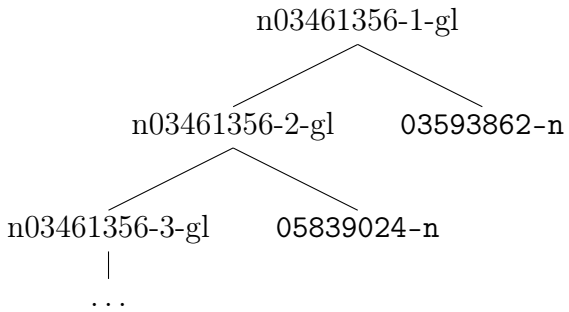
Таблица 5.7: Точност при комбинациите на WNG и синтактично извлечените релации от SemCor

SC-VN, **SC-VV**. Този резултат може да бъде подобрен допълнително чрез добаване на някои от логически извлечените релации от семантичната мрежа на WN. Най-добрата комбинация – **WNG**, **SC-AA**, **SC-AN**, **SC-AV**, **SC-NV**, **SC-RA**, **SC-RR**, **SC-RV**, **SC-VN**, **SC-VV**, **WN-HypInfer**, **WN-AntInfer**, **WN-DerVNInfer**, **WN-Ent1stVInfer**, **WN-Ent2ndVInfer** – дава резултат от 60.70% точност върху SemCor, т.е. с 1.73% повече от базовия резултат с WNG, и 56.39% точност върху BTB.

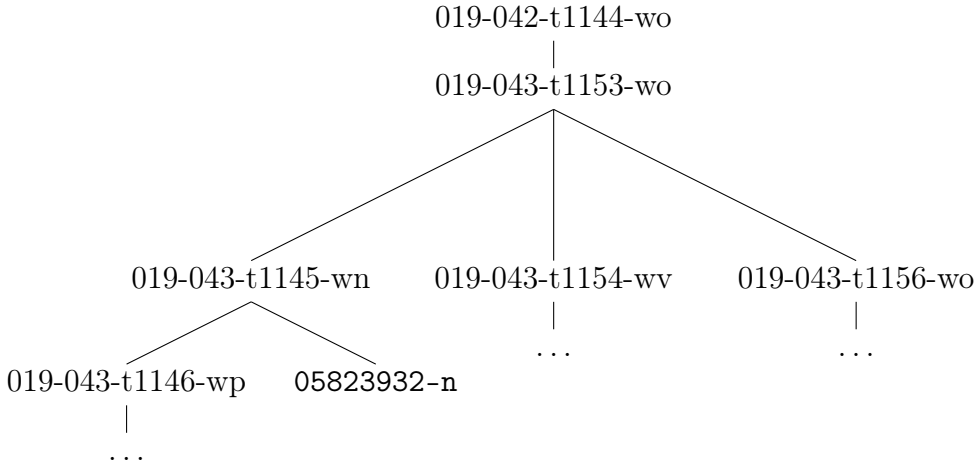
5.3 Допълнителни опити за конструиране на релации

Ресурсът eXtended WordNet, освен че съдържа анотации на пълнозначните думи в дефинициите на значенията, също така предоставя логически форми (т.е. логически представяния от първи ред) на същите тези изречения. От тях е възможно да бъдат построени семантични релации между предикатите и аргументите им в представянето на изречението. Аргументите на релациите са преведени чрез анотациите на значения в корпуса към идентификатори на синсети. Новият граф е наречен **WNGL**.

Друг подход за създаването на нови релации използва информация от SemCor и eXtended WordNet. Изследвани са два различни подхода: да се използва словоредът или синтактичната структура като един вид "свързваща тъкан" на контекстния граф. Първият подход е използван, за да се извадят контексти от XWN, като думите са свързани в бинарни структури, основани върху словореда и свързващи анотациите на синсети с възлите на дърветата. Фиг. 5.1 дава визуално представяне на така конструиран контекст. Новият граф, построен от контекстите, построени въз основа на XWN, тук е наречен **WN30glCon**. При втория подход са използвани синактичните анализи върху изреченията в SemCor – анотациите на синсети отново са свързани със съответстващите им (изкуствено създадени) възли. Коренът на всяко изречение



Фигура 5.1: Дървовидна структура, която онагледява част от един коструиран контекст. Листата на дървото могат да бъдат единствено синсети (Simov, Osenova, & Popov, 2016b).



Фигура 5.2: Горната част на депендентното дърво на изречението, също така свързано с предишното изречение чрез връзка към предходния корен на дърво (Simov, Osenova, & Popov, 2016b). Числата в изкуствените възли са просто индекси към файла, изречението и позицията на думата в SemCor

в корпусните фрагменти е свързан с корена на предходното изречение в съответния фрагмент, като така се създава по-глобален контекст. Фиг. 5.2 дава графична интерпретация на част от такъв контекст. Така полученият граф е наречен **GraphRelSC**.

Таблица 5.8 показва точността на различните комбинации от ГСЗ. Използвани с два различни алгоритмични режима на системата UKB: *статичен (Static)* и *персонализиран (PPRw3w)* PageRank. Резултатите показват, че добавянето на релационна информация от различни ресурси на знание може значително да повиши резултатите на системи за СЛМОЗ.

ГСЗ	*Static*	*PPRw2w*
WN	56.60	56.35
WNG	56.00	57.33
WN + WNG	**59.55**	**62.24**
WNGL	60.46	60.35
WN + WNGL	66.61	67.19
WN + WN30glCon	67.00	66.42
WN + GraphRelSC	67.04	65.97
WN + GraphRelSC + WNGL	68.41	68.51
WN + WN30glCon + GraphRelSC	68.74	68.15
WN + WN30glCon + GraphRelSC + WNGL	**68.77**	68.48
WN + WNG + WN30glCon + GraphRelSC + WNGL	68.39	**68.59**

Таблица 5.8: Точност при СЛМОЗ за различни комбинации от вече съществуващите и новосъздадените графи.

6 Дистрибутивно представяне на думи, основни форми и значения с помощта на лексикални ресурси

В тази глава е описана работа по построяването на дистрибутивни представяния на лексикални единици: думи, основни форми и значения. Описаното се опира върху Simov et al. (2017) и Simov et al. (2018). Основната хипотеза в главата е, че подобрения по топологията на ГСЗ (измерени чрез СЛМОЗ) би следвало да водят и до по-представителни корпуси, генерирани чрез структурата на графите. Върху подобрените корпуси би следвало да могат да бъдат обучавани по-точни модели за векторно представяние (МВП) на лексикона. Чрез изследване на обогатените със синтагматична информация графи, би могло да се достигне до МВП, които комбинират както знание, достъпно от големи корпуси с естествен език, така и такова, което не може да бъде взето директно от тях, защото е имплицитно заложено в теорията, описваща езиковите феномени.

6.1 Векторни представяния от различни графи със знание

За да бъдат комбинирани двата типа знание в МВП – парадигматични релации, кодирани в WN, и синтагматични такива, извлечими от текст, – възприемаме подхода, представен в Goikoetxea et al. (2015) – на генериране на изкуствени поредици от лексикални единици чрез ГСЗ и след това на обучение на МВП върху същите данни. Процедурата по генериране на корпуси,

обучение на МВП и оценяването им е следната:

1. Графът със знание бива конструиран от различни множества от релации.
2. Системата UKB е използвана в режим "обхождане и принтиране" (walkandprint), за да произведе случайни обхождания с различна дължина по продължение структурата на ГСЗ. Системата може да бъде конфигурирана да принтира на всяка стъпка, с определена вероятност, идентификатор на синсет или първата основна форма за посетения синсет.
3. Системата word2vec е използвана, за да се натренира МВП върху изкуствения корпус.
4. Полученият МВП е оценен върху задачата за пресмятане на сходство и свързаност между думи.

Графите, които сме използвали за генериране на учебните корпуси – в допълнение към *WN* и *WNG*, – са описаните по-рано: **WNGL**; **GraphRelSC**; **WN30glCon**; **HypInf**. Таблица 6.1 показва коефициента на рангова корелация на Спирмън за различни МВП спрямо златните данни. Експерименталната оценка показва, че подобрението във СЛМОЗ се отразява в подобрени МВП, при използването на обогатени ГСЗ. Освен това, комбинацията чрез конкатенация на МВП, построен въз основата на ГСЗ (WN+WNglConOne-C15), и МВП, генериран от текст (GoogleNews), постига най-високите резултати върху две от оценъчните множества и се доближава до най-високия резултат върху третото множество.

6.2 Увеличаване на гъстотата на графа със знание чрез филтриране с векторни представяния на граматически роли

Синтактичните релации, които могат да бъдат извлечени от корпус, анотиран от хора със значения на думи, като например SemCor, са една малка част от възможните релации, които са потенциално смислени по отношение на знанието за света и езика. Извличането на релации чрез логически правила над такива "златни" релации може да бъде ефективен подход, както е показано в Simov et al. (2015) и Simov, Osenova, & Popov (2016b). Указаният там метод обаче е присъщо неточен. Заради това, въвеждаме нов подход за добавяне на релации към ГСЗ. Той се основава на идеята, че всички предикати в

Модел за векторно представяне	*WordSim353 Сходство*	*WordSim353 Свързаност*	*SimLex999*
GoogleNews$_{\textbf{Mikolov, Sutskever, et al. (2013)}}$	0.77145	0.61988	0.44196
Dependency$_{\textbf{Levy \& Goldberg (2014)}}$	0.76699	0.46764	0.44730
WN + WNG$_{\textbf{Goikoetxea et al. (2015)}}$	0.78670	0.61316	0.52479
WN + WNG + HypInf *C5*	0.77730	0.54419	0.55192
WN + WNG + HypInf *C15*	0.77205	0.55955	**0.55868**
WN + WNglConOne *C5*	0.77761	0.64747	0.53242
WN + WNglConOne *C15*	0.79659	**0.65548**	0.52632
WN + WNG + WNGL + GrRelSC *C5*	0.79847	0.63587	0.51974
WN + WNG + WNGL + GrRelSC *C15*	**0.81862**	0.61455	0.52350
WN + WNglConOne *C15 +* **GoogleNews**	**0.82684**	**0.70972**	0.54675
WN + WNglConOne *C15 +* **Dependency**	0.80428	0.66570	0.54041

Таблица 6.1: Сравнение на резултатите от различни МВП върху задачата за пресмятане на сходство и свързаност между думи. *C5* and *C15* указват големината на контекстния прозорец, подаван на алгоритъма Skip-Gram. Най-добрите резултати, постигнати чрез единични МВП, са потъмнени. Последните два реда показват резултатите, постигнати чрез комбинация от два модела: такъв, построен въз основа на ГСЗ, и векторите GoogleNews/Dependency.

речника, които могат да имат аргументи, са потенциални възли, от които могат да бъдат разгърнати нови релации, стига да е наличен филтър, който да отделя семантично смислените релации от по-малко вероятните или тривиални такива.

Този филтър трябва да е способен да ни укаже какво представлява прототипният аргумент за определен предикат, така че да сме способни да решим кои аргументи от речника са добри кандидати да влязат в релация с него. Наричаме абстрактните представяния на предикатните аргументи "векторни представяния на граматически роли". В този труд изследваме само граматически роли от типа *subj*, *dobj*, *iobj*. Следващите стъпки описват процедурата по обогатяване на ГСЗ:

1. Построяване на достатъчно голям корпус, който кодира граматически роли заедно със самостойни думи.
2. Обучение на МВП върху корпуса, така че дистрибутивните представяния да са налични както за граматически роли, така и за основни форми и синсети. Граматическите роли за синсети и за самите синсети биват

изчислени чрез пресмятане на средното аритметично на векторите за отделните основни форми към дадения синсет.

3. Използва се речникът на ГСЗ, за да се намерят всички предикати, които са свързани с каквито и да е аргументи в корпуса.
4. За всеки предикат и за всеки тип граматическа роля, предикатите биват поставени в двойка с всички възможни кандидати за аргументи от речника. Оценява се близостта на всеки аргумент с прототипната граматическа роля за предиката.

Успоредното научаване на векторни представяния на граматически роли и основни форми е възможно благодарение на предварителната обработка на учебния корпус. Корпусът WaCkypedia_EN (Baroni et al., 2009) е използван като източник на имплицитно синтактично знание, след като бива лематизиран и парсиран. Аргументите на глаголните предикати биват подменени със специален символ, конструиран по следния начин: ⟨тип_аргумент⟩_⟨основна_форма_на_глагола⟩. Чрез използване на анотациите на частите на речта от текста (RTC), основните форми на отделните токъни са подменени с низовете *основна_фома-част_на_речта*. Генериран е и псевдокорпус (PCWN), за да могат да бъдат заучени и релациите, кодирани в WN. RTC е използван за научаването на векторни представяния сам по себе си, както и заедно с PCWN (чрез обединение с конкатенация) – RTCPCWN. Моделите, които се справят най-добре върху задачата за сходство и свързаност, са оценени и чрез СЛМОЗ.

Граф със знание	SemCor	M13 SemeVal
wn30	51.56	48.41
wn30RTC40	50.32	49.51
wn30RTC45	**52.60**	49.57
wn30RTC47	50.20	48.47
wn30RTC50	50.34	49.63
wn30RTC52	50.58	**51.88**
wn30RTC55	51.05	51.70
wn30RTC57	51.60	51.52

Таблица 6.2: Резултати от СЛМОЗ с релации, ранкирани чрез векторни представяния, получени от истински текст, анотиран с части на речта. Максималното подобрение върху SemCor е **1.04**, а върху M13 SemeVal – **3.47**.

Таблици 6.2 и 6.3 представят основните експерименти. Всяка таблица представя базови нива на точност, получени с ГСЗ, включващи само оригиналните

релации от WN (тук наречени wn30) и точност за обогатените ГСЗ, като всеки от тях съдържа нови релации, построени чрез различни прагове, подадени на филтъра за сходство.

Граф със знание	SemCor	M13 SemeVal
wn30	51.56	48.41
wn30RTCPCWN35	51.88	49.27
wn30RTCPCWN38	53.68	51.39
wn30RTCPCWN40	53.91	**51.45**
wn30RTCPCWN42	**54.33**	50.42
wn30RTCPCWN43	54.08	50.18
wn30RTCPCWN44	52.56	49.93

Таблица 6.3: Резултати от СЛМОЗ с релации, ранкирани чрез векторни представяния, получени от истински текст, анотиран с части на речта, и от псевдо корпус. Максималното подобрение върху SemCor е **2.77**, а върху M13 SemeVal – **3.04**.

Таблица 6.4 представя резултати за ГСЗ, конструирани чрез свити МВП, съдържащи информация само за по-чести синтагматични релации. Нужна е допълнителна работа, за да се установи аналитичен подход за разсъждаване относно тези ефекти. Резултатите обаче сочат недвусмислено, че такъв подход за откриване на смислено ново знание може да бъде използван продуктивно за моделирането на лексикона.

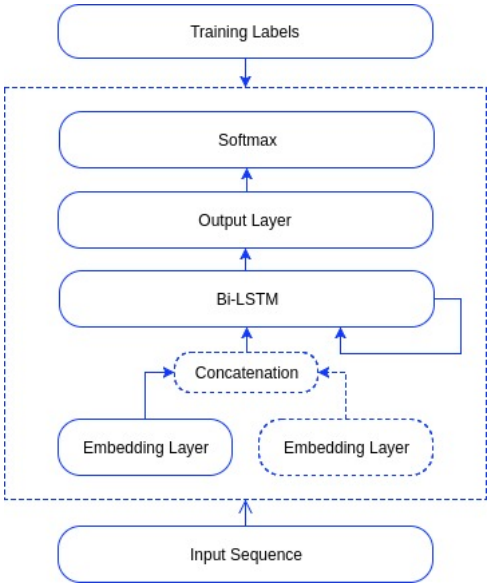
Граф със знание	SemCor	M13 SemeVal
wn30	51.56	48.41
wn30RTCPCWN10-34	**52.35**	51.39
wn30RTCPCWN10-35	50.64	**53.04**
wn30RTCPCWN10-36	50.25	50.72
wn30RTCPCWN10-40	50.49	49.45
wn30RTCPCWN10-45	51.15	49.27
wn30RTCPCWN10-50	51.45	48.29

Таблица 6.4: Резултати от СЛМОЗ с релации, извлечени, след като по-редките граматически роли са премахнати от МВП. Подобрението върху SemCor е **0.79**, а върху M13 SemeVal – **4.62**.

7 Рекурентни невронни мрежи за снемане на лексикалната многозначност

Тази глава стъпва основно върху Popov (2017); въпреки това, някои резултати се основават и на нови експерименти. В нея е до голяма степен преизползвана

РНМ от глава 4, като е адаптирана за целите на СЛМ – друга задача за ОЕЕ, която позволява да бъде моделирана чрез ДКДКП. Изследвани са две различни невронни архитектури, всяка от които интерпретира задачата за СЛМ по различнен начин. За лесната при реферирането, ще наричам първата от двете представени архитектури – *Архитектура A*, а втората – *Архитектура B*. Изследвани са още комбинации от различни МВП като източници на характерни признаци за РНМ, а също така е представен и нов тип, смесен МВП – при него са налице представяния както на основни форми, така и на синсети и контексти на употреба. Този модел разчита изключително много на вече описаните методи за построяване на МВП от ГСЗ.

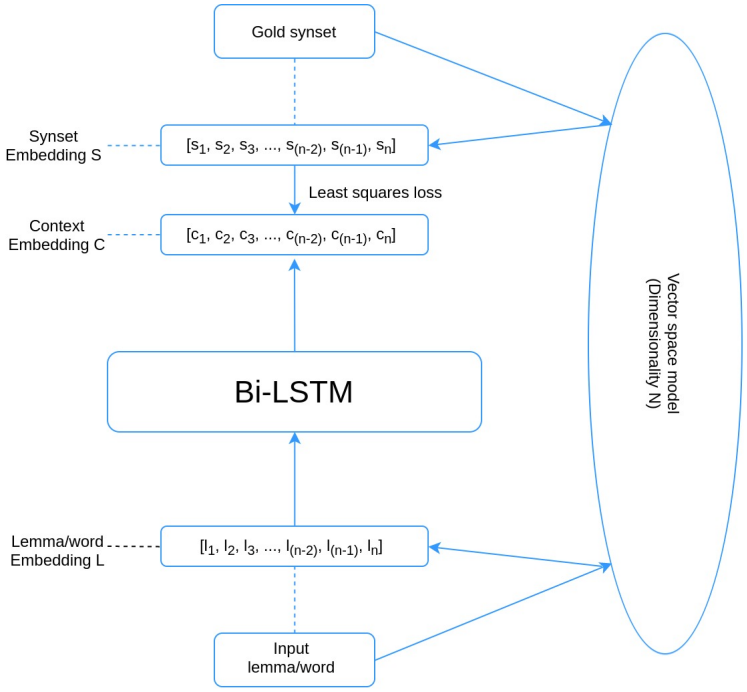


Фигура 7.1: РНМ за СЛМ: прекъснатите линии означават, че даден компонент или връзка не са задължителни (в случай на конкатениране на векторни представяния от два различни източника – напр. на думи от текст и на основни форми от ГСЗ).

Фигура 7.1 представя в графична форма общите принципи на *Архитектура A*. Представянията на отделните думи в скритите слоеве (с размер *2 * hidden_layer_size*), са оразмерени до броя на всички синсети, съответстващи на основните форми, налични в учебните данни. Класификационен softmax слой изчислява разпределението на вероятностите спрямо вектора, описващ лексикона. Крайното решение за СЛМ се взима по отношение на вероятностната маса, концентрирана в позициите на вектора, отредени на онези синсети, асоциирани с основната форма и частта на речта, които биват анализирани на тази стъпка. Когато среща непозната от учебните данни основна форма, системата се обръща към евристика – избира първото значение за конкрет-

ната основна форма. В дисертацията като основен източник на характерни признаци са използвани векторните представяния на думи от Pennington et al. (2014) – GloVe, с размерност 300 позиции. Някои модели ги използват заедно векторни представяния на основни форми, произведени от изкуствени корпуси над WN.

Архитектура B се различава от *Архитектура A* във финалната ѝ фаза, където се представя контекстът и се извършва оптимизирането на мрежата, както и по отношение на това как входните характерни признаци са моделирани по отношение на лексикона. Има две основни разлики: 1) изходният слой вече не отнася скритото представяне на думите към вектор с размер на лексикона; тук той го отнася към вектор, който е равен по размер с входните векторни представяния за отделните думи, напр. 300 позиции; 2) целевата функция, която бива оптимизирана, вече не е *cross entropy*, а сравнение по *метода на най-малките квадрати* между представянето на контекста и векторно представяне на синсета от златната анотация в учебните данни. За визуално представяне, вж. фигура 7.2.



Фигура 7.2: Диаграма, представяща *Архитектура B*.

Този подход разчита на МВП, който съдържа представяния както на входните думи/основни форми, така и на синсетите от златните анотации. Благодарение на това, че се научава да влага контекстите в МВП и да ги доближава максимално до представянията на златните синсети, *Архитек-*

тура B би трябвало да има достъп до множество семантични признаци, които описват правилния отговор. Няколко "смесени" МВП са използвани в експериментите с нея. Методите от по-рано са използвани за обучение на дистрибутивни модели, които комбинират информация за основните форми и синсети в споделено пространство. Изпробвана е и стратегия за увеличаване на покритието на МВП – версия на Уикипедия е сведена до основните форми на думите (лематизация) и конкатенирана към псевдокорпуса.

Обучението и оценката на предложените модели е извършено в Унифицираната рамка за оценка (УРО) на Raganato, Camacho-Collados, & Navigli (2017) – множество от популярни анотирани корпуси за СЛМ, представени в специален единен формат. Корпусът SemCor е използван като източник на обучаващ сигнал. Данните от Senseval-2 са използвани за настройване на параметрите на моделите, а останалата част от множествата данни в УРО – за крайното оценяване. Таблица 7.1 представя параметрите на модела, постигнал най-висока точност върху данните от Senseval-2 (наречен *Модел A1*). Таблица 7.2 представя параметризацията на най-точния модел, обучен с *Архитектура B* (*Модел B1*).

Параметър	**Стойност**
Размер на вект. представяния	300
Вект. представяния - думи	GloVe
Вект. представяния - осн. форми	WN30WN30glConOne
Размер на ДКДКП	2 * 200
Брой ДКДКП	1
Dropout регуларизация	20%
Алгоритъм за оптимизиране	SGD
Темпо на обучение	0.2
Инициализация на ДКДКП	random uniform [-1;1]
Размер на учебните партиди	100
Учебни епохи	100000
Най-добър резултат на епоха №	58100

Таблица 7.1: Параметри на модела, обучен с *Архитектура A* и постигнал най-висока точност върху данните за разработка.

Таблица 7.3 представя сравнение на *Модел A1* и *Модел B1* с някои от системите, оценени в УРО. Също така показва резултати, постигнати от два други модела, които са идентични с *Модел A1*, с разликата, че: *Модел A2* използва само векторните представяния на думи GloVe; *Модел A3* използва друго множество от векторни представяния на основни форми в комбинация

Параметър	**Стойност**
Размер на вект. представяния	300
Вект. представяния - осн. форми	WN30WN30glConOne + WikiLemmatized
Размер на ДКДКП	2 * 400
Брой ДКДКП	1
Dropout регуларизация	0%
Алгоритъм за оптимизиране	SGD
Темпо на обучение	0.2
Инициализация на ДКДКП	random uniform [-1;1]
Размер на учебните партиди	100
Учебни епохи	100000
Най-добър резултат на епоха №	94200

Таблица 7.2: Параметри на модела, обучен с *Архитектура B* и постигнал най-висока точност върху данните за разработка.

с GloVe. МВП на основни форми в *Модел A3* също така се основава на ГСЗ WN30WN30glConOne, но също и на корпуса Wackypedia, обсъден в предишната глава във връзка с представянията на граматическите роли. *Модели A1-3* са под върховите резултати в полето, но не и с голям марж. *Модели A1 и A3* се справят по-добре от *Модел A2* върху някои множества от данни, давайки индикация, че представянията на основни форми може би допринасят с уместно ново знание; но пък върху други дават по-ниска точност.

Накрая, представям някои резултати от експерименти с различни МВП като източници на входни характерни признаци за *Архитектура B*. Таблица 7.4 показва, че дори само конкатенирането на лематизирана версия на Уикипедия към псевдокорпуса води до сериозно подобрение. Има също разлика в представянето на моделите, в зависимост от това как е конструиран псевдокорпусът. Corpus1 (C1) се състои от 100 милиона случайни обхождания върху графа, като до всеки синсет в обхожданията е добавена случайно избрана основна форма от същия синсет; Corpus2 (C2) е направен, като са генерирани 200 милиона случайни обхождания и част от синсетите са директно заменени с основни форми. Така, C1 и C2 са грубо казано с еднакъв размер, но C1 е много по-ефективен при тази оценка. Corpus3 (C3) е направен по същия начин като C1, но броят на случайните обхождания е 200 милиона, т.е. два пъти по-голям е. МВП, които също представят думи/основни форми и синсети в споделено пространство, но са конструирани по различен начин (вж. глава 3), като SW2 и AutoExtend, не се справят по-добре от подхода, предложен тук.

Система	SNE-2	SNE-3	SME-07	SME-13	SME-15	ALL
IMS-s+emb	**72.2**	**70.4**	**62.6**	65.9	71.5	**69.6**
Context2Vec	71.8	69.1	61.3	65.6	**71.9**	69.0
Модел A1	70.4	68.2	57.8	65.3	69.1	67.7
Модел A2	69.6	69.4	59.3	65.0	69.4	67.8
Модел A3	70.1	68.8	56.3	64.2	69.6	67.4
UKB-g*	68.8	66.1	53.0	**68.8**	70.3	67.3
IMS-2010	68.2	67.6	59.1	-	-	-
MFS	65.6	66.0	54.5	63.8	67.1	64.8
IMS-2016	63.4	68.2	57.8	-	-	-
Модел B1	64.7	57.9	47.9	61.9	64.8	61.3
UKB-g	60.6	54.1	42.0	59.0	61.2	57.5

Таблица 7.3: Сравнение на моделите, тренирани с *Архитектури A & B* спрямо други системи, обучени върху SemCor и оценени върху няколко различни множества от данни ("SNE" означава "Senseval", а "SME" – "SemEval"). *IMS-s+emb*, *Context2Vec*, *UKB-g*. UKB-g* и *MFS* са докладвани в Raganato, Camacho-Collados, & Navigli (2017); IMS-2010 е докладван в Zhong & Ng (2010); IMS-2016 (това е конфигурацията IMS + Word2Vec (SemCor)) е докладван в Iacobacci et al. (2016).

Модел за векторно представяне	Точност (SNE-2)
WN30WN30glConOne-C3 + WikiLemmatized	**64.7**
WN30WN30glConOne-C3	63.1
SW2V (600 hidden units)	62.1
WN30WN30glConOne-C1	61.7
SW2V (400 hidden units)	60.2
WN30WN30glConOne-C2	57.4
AutoExtend	53.2

Таблица 7.4: Сравнение между моделите, обучени с *Архитектура B* върху данните от Senseval-2. Векторите SW2V са описани в Mancini et al. (2016); векторите AutoExtend са описани в Rothe & Schütze (2015).

8 Успоредно обучение на РНМ върху няколко задачи

В последната глава, представяща оригинални резултати, е изследвана хипотезата, че анализите на различни задачи, основаващи се на лексикално знание, би следвало да си взаимодействат положително при успоредното обучение на аналитичните модели. Тоест, езиковите структури, които един тип анализ научава от учебните данни, са частично недостъпни за анализатора на друга задача (поради разликите в естеството на учебния сигнал), но в същото време биха могли да ограничат възможните хипотези, с които борави вторият модел. Такова взаимодействие би направило обучението по-стабилно, а лексикалния модел, научен от НМ – по-богат и експресивен. В главата са изследвани две такива комбинации от ралични по вид задачи за лексикален анализ.

8.1 Комбиниране на класификатор за СЛМ и система за научаване на векторни представяния на контексти

Първата идея за обучение върху няколко задачи, изследвана тук, е да се комбинират *Архитектури A and B*. Мотивацията да се споделят параметри между двете е, че макар и да са пригодени за СЛМ, те решават две на практика различни задачи. Промените по имплементацията са минимални: и двата типа връзки между скрит и изходен слой присъстват и резултатите от двете целеви функции са сумирани, преди да бъдат подадени на оптимизатора. Същите параметри са използвани за РНМ като тези в таблица 7.2 от предишната глава. Таблица 8.1 сравнява комбинацията с моделите, решаващи една задача и обучени с аналогични параметри. Резултатите показват, че ученето на няколко задачи наистина помага в този случай. Този модел е по-точен и върху двете задачи от аналозите, обучени върху всяка отделна задача. Резултатите са обнадеждаващи, защото подсказват, че: 1) има голяма степен на взаимна подкрепа между двете задачи; 2) бедността на векторните представяния, добити от ГСЗ (в сравнение с векторите GloVe например), може да бъде преборена чрез подобни конфигурации.

Анализът на поведението на двете под-системи в модела, обучен върху няколко задачи, показва, че клоновете A и B се научават на различни типове информация. За тази цел, три подмножества на златните анотации са извадени от данните за оценка ALL заедно със съответните отговори, дадени от: класификационния модул (A), модула за близост на контекст (B)

	Система	SNE-2	SNE-3	SME-07	SME-13	SME-15	ALL
близост	Модел B1 (една зад.)	64.7	57.9	47.9	61.9	64.8	61.3
	Модел B1 (две зад.)	66.8	60.1	49.2	63.4	67.7	63.3
класификация	Model A1 (една зад.)	70.4	68.2	57.8	65.3	69.1	67.7
	Model A2 (една зад.)	69.6	69.4	59.3	65.0	69.4	67.8
	Model A3 (една зад.)	70.1	68.8	56.3	64.2	69.6	67.4
	Model A4 (една зад.-200)	67.7	66.9	55.8	63.6	68.3	65.9
	Model A4 (една зад.-400)	68.5	67.1	58.2	63.6	67.0	66.2
	Model A4 (две зад.)	68.9	67.8	58.0	63.7	68.4	66.7
	Model A4 (две зад.+dropout)	69.6	68.0	59.1	64.5	70.2	67.5
	MFS	65.6	66.0	54.5	63.8	67.1	64.8

Таблица 8.1: Сравнение между модели, обучени върху една или повече задачи. *Модел A4 (две зад.)* споделя същите принципи и параметри с *B1 (две зад.)* и всъщност е трениран заедно с модел от тип *B*, т.е. *A4 (две зад.)* и *B1 (две зад.)* представляват просто двете отделни пътеки на един и същ модел.

и евристиката за избор на първото значение в WN (C). Таблица 8.2 дава обзор на това колко често един или друг модел е точен. Ако реалността беше такава, че модулите от тип B (близост) просто научават същата информация като модулите A (класификация), би следвало да се очаква, че няма да има примери, в които модулът за близост дава верен отговор, за разлика от модула за класикация, което не се потвърждава.

Комбинация	A!=C!=B	B=C!=A	A=C!=B	Total
A верен	46	256	452	754
B верен	79	598	257	934
C верен	78	598	452	1128
Нито един верен	82	229	241	552
Два (A&B) верни	3	12	15	30

Таблица 8.2: Сравнение на различните модели.

8.2 Комбиниране на ОЧР и СЛМ

Използвана е същата базова ахитектура; този път и двете разклонения от скрития слой към изхода отнасят представянето на контекста към размера на списъка с етикети за съответната задача и изчисляват вероятностното разпределение, използвайки softmax. Учебните данни за моделите отново са от SemCor. Списъкът с етикети за части на речта е преведен към опростения

списък, използван в УРО[8] (Petrov et al., 2011). Корпусът за настройване на параметрите, Senseval-2, е използван за оценка. Таблица 8.3 представя резултатите. Параметриразцията е същата като тази при *Модел A2*. Точностите в таблицата показват, че моделът, обучен върху двете задачи, наистина се справя по-добре както по отношение на СЛМ, така и на ОЧР.

Система	СЛМ (SNE-2)	ОЧР (SNE-2)
Модел A2 (една зад.-WSD)	70.6	-
Модел A2 (една зад.-POS)	-	90.9
Модел A2 (две зад.)	**71.1**	**92.1**

Таблица 8.3: Сравнение между модели, обучени върху една и две задачи (СЛМ и ОЧР). "SNE-2" означава "Senseval-2".

9 Заключение: списък с публикациите, апробация на резултатите и авторска справка

9.1 Списък с публикациите, свързани с дисертацията

1. Popov, A. Neural Network Models for Word Sense Disambiguation: An Overview. Cybernetics and Information Technologies 18.1 (2018): 139-151.
2. Simov, K., Popov, A., Simova, I., & Osenova, P. (2018). Grammatical Role Embeddings for Enhancements of Relation Density in the Princeton Wordnet. In Proceedings of the 9th Global Wordnet Conference.
3. Popov, A. (2017). Word Sense Disambiguation with Recurrent Neural Networks. In Proceedings of the Student Research Workshop Associated with RANLP 2017 (pp. 25-34).
4. Simov, K., Osenova, P., & Popov, A. (2017). Comparison of Word Embeddings from Different Knowledge Graphs. In International Conference on Language, Data and Knowledge (pp. 213-221).
5. Popov, A. (2016b). Neural Network Language Models–an Overview. In The Workshop on Deep Language Processing for Quality Machine Translation (DeepLP4QMT) (p. 20-26).
6. Simov, K., Popov, A., Zlatkov, L., & Kotuzov, N. (2016). Transfer of Deep Linguistic Knowledge in a Hybrid Machine Translation System. In The Workshop on Deep Language Processing for Quality Machine Translation (DeepLP4QMT) (p. 27-33).

[8] `https://github.com/slavpetrov/universal-pos-tags`

7. Simov, K., Osenova, P., & Popov, A. (2016a). Towards Semantic-based Hybrid Machine Translation Between Bulgarian and English. In Proceedings of the 2nd Workshop on Semantics-Driven Machine Translation (SedMT 2016).
8. Simov, K., Osenova, P., & Popov, A. (2016b). Using Context Information for Knowledge-based Word Sense Disambiguation. In International Conference on Articial Intelligence: Methodology, Systems, and Applications (pp. 130-139).
9. Popov, A. (2016a). Deep Learning Architecture for Part-of-speech Tagging with Word and Suffix Embeddings. In International Conference on Artificial Intelligence: Methodology, Systems, and Applications (pp. 68-77).
10. Simov, K., Popov, A., & Osenova, P. (2016a). Knowledge Graph Extension for Word Sense Annotation. In Innovative Approaches and Solutions in Advanced Intelligent Systems (pp. 151-166). Springer.
11. Simov, K., Popov, A., & Osenova, P. (2016b). The Role of the WordNet Relations in the Knowledge-based Word Sense Disambiguation Task. In Proceedings of Eighth Global WordNet Conference (pp. 391-398).
12. Simov, K., Popov, A., & Osenova, P. (2015). Improving Word Sense Disambiguation with Linguistic Knowledge from a Sense Annotated Treebank. In Proceedings of the International Conference Recent Advances in Natural Language Processing (pp. 596-603).
13. Popov, A., Kancheva, S., Manova, S., Radev, I., Simov, K., & Osenova, P. (2014). The Sense Annotation of Bultreebank. Proceedings of TLT13, 127-136.
14. Simova, I., Vasilev, D., Popov, A., Simov, K., & Osenova, P. (2014). Joint Ensemble Model for POS Tagging and Dependency Parsing. In Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-canonical Languages (pp. 15-25).

Цитирания на публикациите, свързани с дисертацията

•Simov et al. (2015) е цитирана в: Hládek, Daniel, et al. "Survey of the word sense disambiguation and challenges for the Slovak language." Computational Intelligence and Informatics (CINTI), 2016 IEEE 17th International Symposium on. IEEE, 2016.

•Simov, Popov, & Osenova (2016b) е цитирана в: Singh, Kuldeep, et al. "Why Reinvent the Wheel: Let's Build Question Answering Systems Together."Proceedings

of the 2018 World Wide Web Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2018.

•Simov, Osenova, & Popov (2016a) е цитирана в: Moussallem, Diego, Matthias Wauer, and Axel-Cyrille Ngonga Ngomo. "Machine Translation Using Semantic Web Technologies: A Survey."arXiv preprint arXiv:1711.09476 (2017).

•Simov, Osenova, & Popov (2016b) е цитирана в: Jelai, Lilyana, et al. "Textual Analysis by using Knowledge-based Word Sense Disambiguation Approach." Journal of Telecommunication, Electronic and Computer Engineering (JTEC) 9.3-3 (2017): 159-162.

Song, Xuebo, "Ontology-based Domain-specific Semantic Similarity Analysis and Applications"(2018). All Dissertations. 2105. `https://tigerprints.clemson.edu/all_dissertations/2105`

•Popov (2016a) е цитирана в: Bhargava, Rupal, Anushka Baoni, and Yashvardhan Sharma. "Composite Sequential Modeling for Identifying Fake Reviews." Journal of Intelligent Systems.

Farrah, Soufiane, Hanane El Manssouri, and Mohammed Ouzzif. "An hybrid approach to improve part of speech tagging system." Intelligent Systems and Computer Vision (ISCV), 2018 International Conference on. IEEE, 2018.

Wagner, Martin. Target Factors for Neural Machine Translation. Diss. Informatics Institute, 2017.

9.2 Апробация на резултатите

Резултати от тезата са представени от автора на следните научни събития: Recent Advances in Natural Language Processing (RANLP). Hissar, Bulgaria 2015; Semantics-Driven Machine Translation Workshop, collocated with the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL: HLT). San Diego, USA 2016; Artificial Intelligence: Methodology, Systems, Applications (AIMSA). Varna, Bulgaria 2016; Workshop on Deep Language Processing for Quality Machine Translation, collocated with the AIMSA conference. Varna, Bulgaria 2016; Student Workshop at RANLP. Varna, Bulgaria 2017; Second International Summer School on Data Science 2017. Split, Croatia (2017); вътрешни семинари на проекта ДемоСем в ИИКТ, БАН, София, България (2017).

9.3 Основни научни и научно-приложни приноси

В това заключение към дисертацията би могло да се каже, че всички задачи, поставени в увода, са успешно изпълнени. Работата по тях доведе до следните **научни приноси**:

1. Направен е подробен обзор на литературата и постигнатите резултати в областта на лексикалното моделиране за целите на ОЕЕ. Особено внимание е обърнато на подходите за свързване на символните и вероятностните подходи за представяне на лексикалното значение. Включен е подробен обзор на методите за лексикално моделиране, използващи многослойни невронни мрежи – една от най-обещаващите в момента посоки на бъдещи изследвания в областта.
2. Показано е изчерпателно, че моделирането на лексикона чрез релационно познание е валиден подход. Демонстрирано е, че в зависимост от целта на моделирането, различни обогатявания на графа с релационно познание помагат по различни начини. Парадигматичните релации в общия случай подобряват моделирането на лексикалното сходство, докато синтагматичните релации помагат при отчитането на близост между лексикалните единици. И двата типа релации са важни от гледна точка на цялостното моделиране, което е нужно за извършването на сложни лексикални анализи, като СЛМ. Тези наблюдения са подплатени с множество експерименти, които проверяват качеството на различните ГСЗ по отношение на различни задачи: СЛМ чрез методи с обучаващ сигнал и такива, основани на знание, както и изчисляване на сходство/близост между думи. Извършен е и сравнителен анализ на релационните множества в WordNet, на тези, извлечени от неговите дефиниции на значения и от синтактично анотирани корпуси. Той показва, че различните релации допринасят по различен начин към лексикалното моделиране, и мотивира по-нататъшни разработки по въпроса за подобряването на ГСЗ.
3. Дисертацията предлага силни доказателства, че кодирането на релационна лексикална информация във вероятностни модели е мощен инструмент за лексикално моделиране. Показано е, че подобряването на наситеността и експресивността на ГСЗ може успешно да бъде преведено под формата на МВП, които от своя страна да бъдат използвани като източник на представяния за машинно самообучаващи се системи, ре-

шаващи разнообразни задачи. Генерирането на учебни данни за такива МВП е постигнато чрез алгоритми за случайно обхождане на графи.

4. Налице са няколко иновативни приноса по отношение представянето на лексикона чрез векторни модели. Първият се състои в генерирането на дистрибутивни представяния за нов тип абстрактен конституент – *граматически роли*, т.е. прототипни синтактични аргументи на глаголни предикати. Методът позволява да се извлече теоретично познание от данни; чрез него е възможно да се отговори на важен за моделирането въпрос: какво представлява адекватен аргумент за конкретен предикат (подходът може да бъде адаптиран и към неглаголни предикати). Експерименталните резултати потвърждават полезността на подхода. Вторият принос от тази група е специфичен подход за обучението на дистрибутивни представяния на основни форми и значения/синсети в споделено пространство – чрез преизползване на техниката за генериране на изкуствени корпуси. Експерименталните резултати показват, че методът е поне толкова успешен, колкото други популярни подходи към същия проблем. В допълнение към тези два приноса, дисертацията също така предлага иновативен подход за кодирането на морфологична информация: т. нар. *вложени представяния на наставки* (suffix embeddings), който представлява метод за лесно представяне на морфологични зависимости. Този подход за моделиране също води до положителни резултати – върху задачата за ОЧР на български език; доколкото съм наясно, това е първото приложение на подобен тип морфологичен анализ към тази конкретна задача.
5. Предложен е нов подход за СЛМ с обучаващ сигнал – който използва РНМ, но вместо директно да класифицира значенията в изходния си слой, се учи да влага контекстите на употреба на думи в споделен МВП, който описва едновременно думи/основни форми и значения/синсети. Дисертацията показа, че по-добри смесени МВП могат да направят моделите, научени с архитектурата, конкурентни на най-добрите системи, макар и целевата им функция да не е директно свързана със СЛМ. Чрез сравнителен анализ е показано, че РНМ, които извършват класификация, и РНМ за влагане на контексти, като разработените тук, научават различни представяния и могат да се допълват взаимно. Подходът за представяне на контексти има предимството, че винаги е способен да вземе самостоятелно решение, без да е нужно да разчита на резервни евристики в случаите на непознати от учебните данни думи.

Всъщност, ако не използва резервна евристика, класификаторът често е по-неточен от моделите за представяне на контекст.

6. Изследвано е успоредното обучение върху няколко задачи, с положителни резултати. Заедно са тренирани модели за: 1) класификация на значения и представяне на контексти в МВП; 2) класификация на значения и ОЧР. И в двата случая експерименталните резултати потвърждават силното взаимодействие между различните аспекти на лексикалното знание. Този принос може да бъде разглеждан като допълнителна мотивация за по-нататъшното изследване на настоящия теоретичен въпрос.

Налице са и следните **научно-приложни приноси**:

1. Разработени са няколко архитектури с РНМ за трансформиране на поредици от данни. По-конкретно, те решават следните задачи, всички от които представляват някакъв аспект на лексикално моделиране: ОЧР, СЛМ и представяне на конктекст на употреба на думи. Резултатите са близки с най-добрите в съответните полета и дават заявка, че с допълнително оптимизиране и подобрения по системите, те могат да достигнат тези върхови резултати.
2. Генерирани са голям брой множества от нови релации между значения на думи (синсети), според конвенционалния формат на WordNet. Множествата с нови релации са оценени спрямо различни множества с данни, върху два различни езика – български и английски. Постигнати са значителни подобрения върху задачата за СЛМОЗ, чрез прибавянето на новите ресурси към ГСЗ. Разработени са различни подходи за извличане на релационно знание от вече съществуващи източници, включително филтриращ подход, който използва векторни представяния на граматически роли с цел да направи изчерпателно търсене на нови релации, позволени от речника.
3. Обучени и оценени са различни МВП. Сред тях присъстват: МВП за представяне на основни форми въз основата на семантични мрежи; смесени МВП, съчетаващи основни форми, синсети и в някои случаи представяния на граматически роли; словоформи и наставки за български език – първите подобни модели, обучени за този език, доколкото ми е известно. Представянията на основни форми постигат върхови резултати върху задачата за изчисляване на сходство/близост между думи, като подобряват резултатите на популярни и широко използвани МВП; също така те допринасят съществени характерни признаци

за задачата за СЛМ с обучаващ сигнал. Смесените МВП от основни форми и синсети се справят по-добре от два популярни такива модела, срещу които са оценени върху задачата за представяне на контексти на употреба на думи.

9.4 Благодарности

Бих искал на първо място и най-вече да благодаря на научния ми ръководител Кирил Симов за многото ценни идеи и за напътствията, които ми е давал през последните четири години. Благодарение на него и на Петя Осенова имах шанса да работя по разнообразни проекти, да посетя ценни обучения и да поддържам приемственост в работата си, която да отразява интересите ми към лексикалната семантика и семантиката по принцип. Много от научните статии, които са в основата на дисертацията, са написани в сътрудничество с тях, както и с други колеги, на които изказвам благодарност.

Искам също да благодаря за финансовата подкрепа по проектите: QTLeap: Quality Translation by Deep Language Engineering Approaches; EUCases - EUropean and National CASE Law and Legislation Linked in Open Data Stack; Дълбоки модели на семантично знание (ДемоСем), финансиран от Фонд научни изследвания.

Библиография

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., & Soroa, A. (2009). A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 19–27).

Agirre, E., Barrena, A., & Soroa, A. (2015). Studying the Wikipedia Hyperlink Graph for Relatedness and Disambiguation. *arXiv preprint arXiv:1503.01655*.

Agirre, E., Bengoetxea, K., Gojenola, K., & Nivre, J. (2011). Improving Dependency Parsing with Semantic Classes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2* (pp. 699–703).

Agirre, E., & Soroa, A. (2009). Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 33–41).

Alonso, H. M., & Plank, B. (2017). When is Multitask Learning Efective? Semantic Sequence Prediction Under Varying Data Conditions. In *15th Conference of the European Chapter of the Association for Computational Linguistics.*

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1* (pp. 86–90).

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky Wide Web: a Collection of Very Large Linguistically PRocessed Web-crawled Corpora. *Language Resources and Evaluation*, *43*(3), 209–226.

Bentivogli, L., Forner, P., Magnini, B., & Pianta, E. (2004). Revising the WordNet Domains Hierarchy: Semantics, Coverage and Balancing. In *Proceedings of the Workshop on Multilingual Linguistic Ressources* (pp. 101–108).

Carpuat, M., & Wu, D. (2005). Word Sense Disambiguation vs. Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 387–394).

Carpuat, M., & Wu, D. (2007). Improving Statistical Machine Translation Using Word Sense Disambiguation. In *EMNLP-CoNLL* (Vol. 7, pp. 61–72).

Chan, Y. S., Ng, H. T., & Chiang, D. (2007). Word Sense Disambiguation Improves Statistical Machine Translation. In *Annual Meeting-Association for*

Computational Linguistics (Vol. 45, p. 33).

Edmonds, P., & Cotton, S. (2001). SENSEVAL-2: Overview. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems* (pp. 1–5).

Fellbaum, Christiane. (1998). *Wordnet*. Wiley Online Library.

Goikoetxea, J., Soroa, A., Agirre, E., & Donostia, B. C. (2015). Random Walks and Neural Network Language Models on Knowledge Bases. In *HLT-NAACL* (pp. 1434–1439).

Graves, A. (2012). Supervised Sequence Labelling. In *Supervised Sequence Labelling with Recurrent Neural Networks* (pp. 5–13). Springer.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. *Computational Linguistics*, *41*(4), 665–695.

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv preprint arXiv:1508.01991*.

Iacobacci, I., Pilehvar, M. T., & Navigli, R. (2015). SensEmbed: Learning Sense Embeddings for Word and Relational Similarity. In *ACL (1)* (pp. 95–105).

Iacobacci, I., Pilehvar, M. T., & Navigli, R. (2016). Embeddings for Word Sense Disambiguation: An Evaluation Study. In *ACL (1)*.

Kågebäck, M., & Salomonsson, H. (2016). Word Sense Disambiguation Using a Bidirectional LSTM. *arXiv preprint arXiv:1606.03568*.

Kučera, H., & Francis, W. N. (1967). *Computational Analysis of Present-day American English*. Dartmouth Publishing Group.

Levy, O., & Goldberg, Y. (2014). Dependency-based Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Vol. 2, pp. 302–308).

Mancini, M., Camacho-Collados, J., Iacobacci, I., & Navigli, R. (2016). Embedding Words and Senses Together via Joint Knowledge-Enhanced Training. *arXiv preprint arXiv:1612.02703*.

Melamud, O., Goldberger, J., & Dagan, I. (2016). context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning* (pp. 51–61).

Mihalcea, R., & Moldovan, D. I. (2001). eXtended WordNet: Progress Report. In *in Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).

Miller, G. A., Leacock, C., Tengi, R., & Bunker, R. T. (1993). A Semantic Concordance. In *Proceedings of the Workshop on Human Language Technology* (pp. 303–308).

Moro, A., & Navigli, R. (2015). Semeval-2015 task 13: Multilingual All-words Sense Disambiguation and Entity Linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (pp. 288–297).

Navigli, R. (2009). Word Sense Disambiguation: A survey. *ACM Computing Surveys (CSUR)*, *41*(2), 10.

Navigli, R., Jurgens, D., & Vannella, D. (2013). Semeval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (Vol. 2, pp. 222–231).

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543).

Petrov, S., Das, D., & McDonald, R. (2011). A Universal Part-of-speech Tagset. *arXiv preprint arXiv:1104.2086*.

Plank, B., Søgaard, A., & Goldberg, Y. (2016). Multilingual Part-of-speech Tagging with Bidirectional Long Short-term Memory Models and Auxiliary Loss. *arXiv preprint arXiv:1604.05529*.

Popov, A. (2016a). Deep Learning Architecture for Part-of-Speech Tagging with Word and Suffix Embeddings. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications* (pp. 68–77).

Popov, A. (2016b). Neural Network Language Models–an Overview. In *The Workshop on Deep Language Processing for Quality Machine Translation (DeepLP4QMT)* (p. 20-26).

Popov, A. (2017). Word Sense Disambiguation with Recurrent Neural Networks. In *Proceedings of the Student Research Workshop Associated with RANLP 2017* (pp. 25–34).

Popov, A. (2018). Neural Network Models for Word Sense Disambiguation: an Overview. *Cybernetics and Information Technologies*, *18*(1), 139–151.

Popov, A., Kancheva, S., Manova, S., Radev, I., Simov, K., & Osenova, P. (2014). The Sense Annotation of BulTreeBank. *Proceedings of TLT13*, 127–136.

Pradhan, S. S., Loper, E., Dligach, D., & Palmer, M. (2007). SemEval-2007 task 17: English Lexical Sample, SRL and All Words. In *Proceedings of the 4th International Workshop on Semantic Evaluations* (pp. 87–92).

Raganato, A., Bovi, C. D., & Navigli, R. (2017). Neural Sequence Learning Models for Word Sense Disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1156–1167).

Raganato, A., Camacho-Collados, J., & Navigli, R. (2017). Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In *Proc. of EACL* (pp. 99–110).

Rothe, S., & Schütze, H. (2015). Autoextend: Extending Word Embeddings to Embeddings for Synsets and Lexemes. *arXiv preprint arXiv:1507.01127*.

Schuler, K. K. (2005). VerbNet: A Broad-coverage, Comprehensive Verb Lexicon.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681.

Simov, K., & Osenova, P. (2001). A Hybrid System for Morphosyntactic Disambiguation in Bulgarian. In *Proceedings of the EuroConference on Recent Advances in Natural Language Processing* (pp. 5–7).

Simov, K., & Osenova, P. (2004). *BTB-TR04: BulTreeBank Morphosyntactic Annotation of Bulgarian Texts* (Tech. Rep.). Technical Report BTB-TR04, Bulgarian Academy of Sciences.

Simov, K., Osenova, P., & Popov, A. (2016a). Towards Semantic-based Hybrid Machine Translation Between Bulgarian and English. In *Proceedings of the 2nd Workshop on Semantics-Driven Machine Translation (SedMT 2016)* (pp. 22–26).

Simov, K., Osenova, P., & Popov, A. (2016b). Using Context Information for Knowledge-based Word Sense Disambiguation. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications* (pp. 130–139).

Simov, K., Osenova, P., & Popov, A. (2017). Comparison of Word Embeddings from Different Knowledge Graphs. In *International Conference on Language, Data and Knowledge* (pp. 213–221).

Simov, K., Popov, A., & Osenova, P. (2015). Improving Word Sense Disambiguation With Linguistic Knowledge from a Sense Annotated Treebank. In *Proceedings of the International Conference Recent Advances in Natural Language Processing* (pp. 596–603).

Simov, K., Popov, A., & Osenova, P. (2016a). Knowledge Graph Extension for Word Sense Annotation. In *Innovative Approaches and Solutions in Advanced Intelligent Systems* (pp. 151–166). Springer.

Simov, K., Popov, A., & Osenova, P. (2016b). The Role of the WordNet Relations in the Knowledge-based Word Sense Disambiguation Task. In *Proceedings of Eighth Global WordNet Conference* (pp. 391–398).

Simov, K., Popov, A., Simova, I., & Osenova, P. (2018). Grammatical Role Embeddings for Enhancements of Relation Density in the Princeton WordNet. In *Proceedings of the 9th Global Wordnet Conference.*

Simov, K., Popov, A., Zlatkov, L., & Kotuzov, N. (2016). Transfer of Deep Linguistic Knowledge in a Hybrid Machine Translation System. In *The Workshop on Deep Language Processing for Quality Machine Translation (DeepLP4QMT)* (p. 27-33).

Simova, I., Vasilev, D., Popov, A., Simov, K., & Osenova, P. (2014). Joint Ensemble Model for POS Tagging and Dependency Parsing. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages* (pp. 15–25).

Snyder, B., & Palmer, M. (2004). The English All-words Task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text.*

Stokoe, C., Oakes, M. P., & Tait, J. (2003). Word Sense Disambiguation in Information Retrieval Revisited. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval* (pp. 159–166).

Sundermeyer, M., Ney, H., & Schlüter, R. (2015). From Feedforward to Recurrent LSTM Neural Networks for Language Modeling. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, *23*(3), 517–529.

Wang, P., Qian, Y., Soong, F. K., He, L., & Zhao, H. (2015). Part-of-speech Tagging with Bidirectional Long Short-term Memory Recurrent Neural Network. *arXiv preprint arXiv:1510.06168*.

Zapirain, B., Agirre, E., Marquez, L., & Surdeanu, M. (2013). Selectional Preferences for Semantic Role Classification. *Computational Linguistics*, *39*(3), 631–663.

Zhong, Z., & Ng, H. T. (2010). It Makes Sense: A Wide-coverage Word Sense Disambiguation System for Free text. In *Proceedings of the ACL 2010 System Demonstrations* (pp. 78–83).

Списък със съкращенията

ГСЗ	граф със знание
ДКДКП	двупосочни клетки с дълга краткотрайна памет
КДКП	клетки с дълга краткотрайна памет
МВП	модели за векторно представяние
НЕМ	невронни езикови модели
НМ	невронни мрежи
НМПР	невронни мрежи за пряко разпространение
ОЕЕ	обработката на естествен език
ОЧР	определяне частите на речта
РНМ	рекурентни невронни мрежи
СЛМ	снемане на лексикалната многозначност
СЛМОЗ	снемане на лексикалната многозначност, основано на знание
УРО	унифицирана рамка за оценка
ЧР	част на речта

IMS	It Makes Sense
WN	WordNet
XWN	eXtended WordNet