ЕВРОПЕЙСКИ СЪЮЗ
ЕВРОПЕЙСКИ ФОНД ЗА
РЕГИОНАЛНО РАЗВИТИЕ

ОПЕРАТИВНА ПРОГРАМА
НАУКА И ОБРАЗОВАНИЕ ЗА
ИНТЕЛИГЕНТЕН РАСТЕЖ

Стратегии за подобряване на покритието на българския уърднет BTB-WordNet

Зара Кънчева
Изкуствен интелект и езикови технологии
ИИКТ – БАН

ЦЕНТЪР ЗА ВЪРХОВИ ПОСТИЖЕНИЯ ПО
ИНФОРМАТИКА И ИНФОРМАЦИОННИ И
КОМУНИКАЦИОННИ ТЕХНОЛОГИИ

План

- ✓ Увод
- ✓ BulTreeBank-WordNet
- ✓ Подобряване на покритието на BTB-WN
- ✓ Заключение
- ✓ Използвана литература

Увод

- ✓ Научен проект № 4: Езикови технологии и технологии, базирани на съдържание, за приложения над големи данни
- ✓ Задача № 4.1: Езикови технологии и технологии, базирани върху съдържание, за машинен превод

Увод (1)

✓ Цели и задачи:

- развиване на елементи от ресурса BTB-WordNet за съпоставяне към формални онтологии (представяне на знание с множества от категории и свойства и връзки между тях) и за използване на методи за автоматично генериране на данни за обучение на векторно представяне на езикови данни (*word embedding*)
- свързване на BTB-WordNet с различни онтологии и мрежи от знание (*knowledge graph* – мрежа от обекти от реалния свят – обекти, събития, ситуации или понятия с връзки помежду си)
- експерименти с трениране на векторно представяне на езикови данни
- проверка и редакция на съществуващите синонимни множества в BTB-WN и обогатяване с нови значения от сравнение с българската Уикипедия

BulTreeBank-WordNet

- BTB-WN е лексикален ресурс с характеристики на тълковен и синонимен речник, обогатен с енциклопедично знание (Osenova and Simov, 2018)
- сливане на лингвистично и енциклопедично знание (BabelNet (семантична мрежа, свързана с Wikipedia и Princeton WordNet), FrameNet (семантичен ресурс, свързан с PWN), Uby (PWN, GermaNet, Wiktionary, Wikipedia, FrameNet, VerbNet), VerbNet (синтактично-семантична мрежа, свързана с PWN и FrameNet), обединяването на PWN и Wikipedia и на plWordNet и PWN)
- подходящ за редица задачи при обработката на естествен език: сваляне на многозначност, извличане на релации, извличане и парсиране на именовани същности и на многокомпонентни думи, машинен превод и други

ИКТ

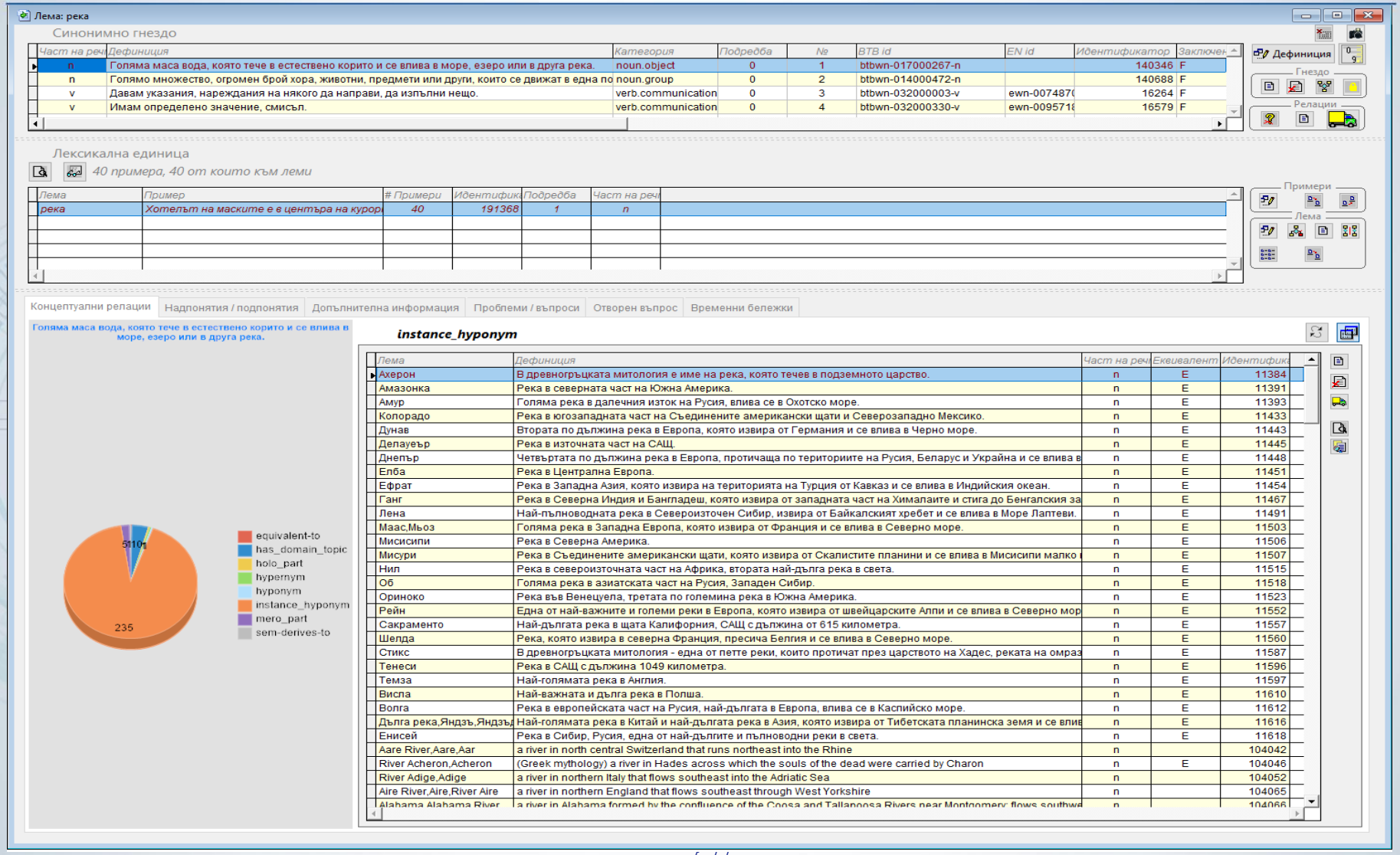
Проект BG05M2OP001-1.001-0003 „Център за върхови постижения по Информатика и информационни и комуникационни технологии“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

BulTreeBank-WordNet (1)

- понятията са организирани в синонимни множества (synsets) – групи от значения, подредени в йерархия, с различни релации (еквивалентност, хиперонимия, хипонимия и други)
- повече от 26 000 синсета в BTB-WN
- части на речта – съществителни, прилагателни и числителни имена, наречия, глаголи
- ръчен превод на Core WordNet от Princeton WordNet (Fellbaum, 1998) – 5000 базови понятия
- обогатяване със значения от българската трибанка BulTreeBank, честотен списък, българските Уикиречник и Уикипедия
- по-ранно разширяване на BTB-WN с данни от Уикипедия – 15% увеличение (Simov et al., 2019)

Проект BG05M2OP001-1.001-0003 „Център за върхови постижения по Информатика и информационни и комуникационни технологии“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

BulTreeBank-WordNet (2)



Подобряване на покритието на BTB-WN

- ✓ Уикипедия:
- богат и постоянно разширяващ се източник на информация
- 277 076 статии на български език
- понятия от разнообразни сфери – специализирани термини и общоупотребима лексика
- възможност за тематичен подход – обогатяване със значения от определена област, защото Уикипедия предлага разглеждане на понятия, разпределени в собствени категории
- лесно наблюдение на йерархията на категориите и избор на нива, от които да се извлекат понятия
- и двата ресурса съдържат понятия и отделни случаи на понятия – Уикипедия има голям обем от имена на хора, географски обекти и събития (спортни състезания, концерти, войни и прочие), които не са застъпени в BTB-WN

Проект BG05M2OP001-1.001-0003 „Център за върхови постижения по Информатика и информационни и комуникационни технологии“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

Подобряване на покритието на BTB-WN (2)

Страници в категория „Военни звания“

Показани са 71 от общо 71 страници в тази категория.

- Военно звание

*

- Военни звания през Втората световна война
- Пагон
- Сигнифер

А

- Ага
- Адмирал
- Адютант
- Армейски генерал
- Атаман

Б

- Багатур
- Бей
- Бейлербей
- Бригаден генерал

В

- Велик войвода
- Войвода
- Войник

Г

- Гардемарин
- Генерал
- Генерал от авиацията
- Генерал от артилерията
- Генерал от артилерията (Руска империя)
- Генерал от кавалерията
- Генерал от свързочните войски
- Генерал-адмирал
- Генерал-аншеф
- Генерал-лейтенант
- Генерал-майор
- Генерал-полковник
- Генералисимус
- Главнокомандващ
- Гросадмирал

Д

- Декурион (военен командир)
- Джигит

Е

- Ефрейтор

К

- Канонир
- Кадет
- Капитан
- Капудан паша
- Комбриг
- Кондотиер
- Курсант

Л

- Лейтенант

М

- Майор
- Мирмиран
- Младши лейтенант
- Младши сержант
- Офицерски кандидат

П

- Паша
- Подполковник
- Полковник
- Портупей-юнкер
- Прапоршчик

Р

- Редник

С

- Сержант
- Старши лейтенант
- Старши сержант
- Старшина
- Субалтерн-офицер

Ф

- Фелдмаршал
- Флигел-адютант
- Флотилен адмирал

Х

- Хауптман

Ц

- Центурион

Ч

- Чорбаджия

Ш

- Шогун

ИКТ

Проект BG05M2OP001-1.001-0003 „Център за върхови постижения по Информатика и информационни и комуникационни технологии“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

Заключение

- ✓ Ползи от разширяването на BTB-WN с понятия от Уикипедия:
- обогатяване на BTB-WN със синоними на съществуващи понятия, с нови понятия от разнообразни области и отделни случаи на понятия
- по-богати данни за векторно представяне
- по-качествена семантична анотация на текст

ИКТ

Използвана литература

- P. Osenova, K. Simov, The Data-driven Bulgarian WordNet: BTBWN, Cognitive Studies | Études cognitives 18 (2018)
- C. Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, 1998
- Rudnicka, E. K., Piasecki, M. T., Piotrowski, T., Łukasz Grabowski, and Bond, F. (2017). Mapping WordNets from the perspective of inter-lingual equivalence. Cognitive Studies — Etudes cognitives , 17(1373):1–17
- McCrae, J. P. (2018). Mapping WordNet Instances to Wikipedia. In Proceedings of Ninth Global WordNet Conference, pages 62–69. The Global WordNet Association
- Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. (2012). Uby – a large-scale unified lexical-semantic resource based on lmf. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 580–590, Avignon, France, April. Association for Computational Linguistics
- Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence, 193:217–250
- Kiril Simov, Petya Osenova, Laska Laskova, Ivajlo Radev, and Zara Kancheva. 2019. Aligning the Bulgarian BTBWordNet with the Bulgarian Wikipedia. In Proceedings of the 10th Global WordNet Conference, 290-297
- Collin Baker. 2008. FrameNet, present and future. In Jonathan Webster, Nancy Ide, and Alex Chengyu Fang, editors, The First International Conference on Global Interoperability for Language Resources, Hong Kong. City University, City University
- Karin Kipper-Schuler. 2005. VerbNet: A broad-coverage, comprehensive verb lexicon. Ph.D. thesis, University of Pennsylvania

Проект BG05M2OP001-1.001-0003 „Център за върхови постижения по Информатика и информационни и комуникационни технологии“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

Благодаря за вниманието!

Проект BG05M2OP001-1.001-0003 „Център за върхови постижения по Информатика и информационни и комуникационни технологии“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.