

МОЛБА
ЗА ОДОБРАВАЊЕ ТЕМЕ МАСТЕР РАДА

Молим да се одобри израда мастер рада под насловом:

„Проширивање српског *WordNet*-а у медицинском домену применом великих језичких модела”

Значај теме и области:

Тема припада области обраде природног језика (*NLP*), подобласти вештачке интелигенције. *WordNet* представља лексичко-семантичку мрежу која организује речи у синсетове и описује њихове семантичке односе. Српски *WordNet* (*SrpWN*) представља значајан језички ресурс за српски језик, али је његова покривеност специјализованим доменима, посебно медицинским, ограничена. Последњих година велики језички модели показали су значајан потенцијал у задацима превођења и изградње лексичких ресурса. Овим радом би се истражила могућност њихове примене за проширивање *SrpWN*-а медицинским терминима, чиме би се унапредили ресурси за обраду медицинских текстова на српском језику.

Специфични циљ рада:

Циљ рада је проширивање базе *SrpWN* у медицинском домену додавањем недостајућих синсетова, литерала, дефиниција и примера употребе из *Princeton WordNet*-а. Стратегије примене великих језичких модела биће поређене дуж две димензије: избора модела (*Claude*, *GPT*, уз *Google Translate* као основу за поређење) и стратегије промптовања (*zero-shot*, *few-shot*, *chain-of-thought*, промптови обogaћени контекстом синсета, генерисање више кандидата превода). Најуспешнија комбинација, утврђена на мањем узорку, биће примењена на цео скуп. Квалитет превода биће испитан аутоматски (слагање између модела) и ручно: аутор ће на узорку од 100–200 синсетова оценити тачност литерала, дефиниције и примера употребе (тачно/прихватљиво/погрешно/не могу да проценим). Практична вредност ресурса биће испитана класификацијом медицинских текстова на српском по специјалностима (нпр. кардиологија, неурологија), у поређењу са оригиналним *SrpWN*-ом.

Остале битне информације:

У оквиру рада биће формиран корпус од неколико стотина до хиљаду медицинских текстова на српском језику, прикупљених са јавно доступних здравствених портала (нпр. *doktor.rs*, *stetoskop.info*, *B92 Здравље*). Проширени *WordNet* ресурс, изворни код и експериментална подешавања биће јавно доступни путем *GitHub* репозиторијума ради обезбеђивања репродуктивности резултата.

Јана Милутиновић, 1065/2024, Информатика
(име и презиме студ., бр. инд., ознака програма и модула)

Сагласан ментор **Јелена Граовац**

(својеручни потпис студента)

(својеручни потпис ментора)

(датум подношења молбе)

Чланови комисије

1. Јована Ковачевић
2. Сташа Вујичић Станковић

Катедра за рачунарство и информатику је сагласна са предложеном темом.

(шеф катедре)

(датум одобравања молбе)