Юрась Гецэвіч, загадчык лабараторыі распазнавання і сінтэзу маўлення. Інтэрв'ю для суполкі NLProc.by, частка 1/2

Добры дзень, спадар Юры. Сёння мы сустракаемся, каб Вы расказалі пра свой досвед у камп'ютарнай лінгвістыцы. Раскажыце агульна пра ваш вопыт у гэтай галіне. Калі вы пачалі цікавіцца, на якім узроўні?

Прывітанне, маё імя Юрась Гецэвіч, кандыдат тэхнічных навук, на кафедрах філфака і гумфака працую дацэнтам, у Аб'яднаным Інстытуце праблем інфарматыкі Акадэміі Навук з'яўляюся загадчыкам лабараторыі распазнавання і сінтэзу маўлення.

Да пытання камп'ютарнай лінгвістыкі я ішоў і хутка, і павольна ў то жа самы час. Усё пачалося з таго, што пасля 5 курса ФПМІ БДУ я працаваў пэўны час у адной IT кампаніі, і там мне ўсё меньш і меньш падабалася, бясконцыя php-сайцікі, таксама падыходзіла войска. Але тут узнікла магчымасць пайсці ў магістратуру Акадэміі Навук. З ласкі Госпада я змог за адзін дзень зрабіць усе медспраўкі для паступлення, за ноч вывучыў усе білеты і паступіў. Калі магістратура ішла, было пэўнае алібі, што ў войска не забяруць. Таксама па магістратуры трэба было пісаць магістарскую. У спісе тэм я ўбачыў, што ёсць тэма пра сінтэз беларускага, рускага, украінскага і іншага маўлення - калі я ўбачыў беларускую мову – мяне гэта зацікавіла. Я чалавек веруючы. Лічу, што беларуская мова дадзеная нам Богам на гэтай зямлі. Я прыйшоў да навуковага кіраўніка, гэта быў, дарэчы, Барыс Мяфодзьевіч Лабанаў – доктар тэхнічных навук, які ўсё жыццё займаўся сінтэзам маўлення. Калі ён мне зрабіў кароткую прэзэнтацыю пра працу лабараторыі, я зацікавіўся і пагадзіўся распрацоўваць сінтэзатар беларускай мовы як магістарскую працу.

Першапачаткова я не ведаў, што гэта сфера кам'ютэрнай лінгвістыкі, проста хацелася навучыць машыну размаўляць па-беларуску.

Мая магістарская скончылася на тым, што мы зрабілі маленькі прататып. Ён быў не хуткі, але слова за секунду ён прамаўляў (сінтэз маўлення па тэксце). Гэтага было дастаткова для абароны ў магістратуры.

Пасля гэтага паступіў у аспірантуру. Былі ўжо 2 прычыны: па-першае, ужо спадабалася галіна, па-другое, калі б не паступаў у той год – войска забрала бы. Я вырашыў: хоць грошы невялікія будуць плаціць, але ж ужо лепш на свабодзе буду займацца навукай, чым у войску акопы капаць. Таму вырашыў далей займацца навукай. Мне далі тэму працы - "Алгарытм лінгвістычнай апрацоўкі тэкстаў для сінтэзу маўленняў беларускай і рускай моваў".

Калі паглядзець на агульную схему, то можна заўважыць, што схема сінтэзатара маўлення фактычна пакрывае любую камп'ютарна-лінгвістычную праграму: Sentiment Analysis, Machine translation і іншыя лінгвістычныя працэсары, а можа нават і больш. Таму што з фанетыкай і гукавымі сігналамі, нажаль, у NLProc вельмі мала хто працуе, часамі нават аддаляюць гэту сферу ад камп'ютарнай лінгвістыкі, але гэта не праўда, яна нават большая: гэта не толькі праца з тэкстам, але і з гукавымі (маўленчымі) сігналамі.

Ішоў час. Пастараўся абараніць дысертацыю. Было складана, але атрымалася. Далей, паставілі кіраваць лабараторыяй. За гады выкладання ў беларускі ВНУ з ласкі Госпада ўдалося сабраць штат у 17 чалавек на розную загрузку, на розны занятак у галінах сінтэзу і распазнавання маўлення. Зараз камандай мы ўсе разам удакладняем нейкія часткі сінтэзатара беларускага маўлення, рускага маўлення, якуцкага і так далей. І выпрацоўваем ужо нейкую школу для ВНУ па канкрэтных ведах, якімі чалавек павінен валодаць, каб трапіць і ў нашу галіну, і, у перспектыве, быць шырокім спецыялістам.

Мая праца накіраваная на практыку, я меньш тэарэтык. Гэта пэўны плюс і мінус.

Чаму?

Плюс - мы вельмі часта маем задачы, якія маюць некалькі рашэнняў. Бывае, што ёсць рашэнне, і мы спрабуем яго перапісаць цалкам, па-крокава, каб быць упэўненымі, што да канца валодаем прадуктам.

Так, каб ведаць, што нейкі крок максімальна добра зроблены?

Так, добра і зразумела зроблены. Наша ідэя такая: мы не маем нейкага сур'ёзнага інвестара, хаця бывае, што яны ўзнікаюць адзін раз на 2-3 гады. Напрыклад, сінтэзатар якуцкай мовы мы зрабілі, калі інвестар знайшоўся. Але таксама мы рабілі праект абсалютна бясплатна для слабабачачых дзяцей па ўсталёўцы абноўленых галасоў сінтэзу беларускага маўлення. Таму мы свае рашэнні робім павольней, але фрагменты адчыняем для іншых, каб нам прасцей было вучыць студэнтаў, і, калі цікавяцца, яны да нас прыходзяць – мы адчыняем 30-40%, каб было зразумела, што гэта не закрытая нейкая скрынка, а прылада карысная для адукацыі.

Зараз мы ўсё больш і больш пашыраем наш топік, усё больш займаемся камп'ютарнай лінгвістыкай. Дарэчы, хутка праводзім канферэнцыю па камп'ютарнай лінгвістыцы. Усіх запрашаем.

Вось прыкладна такі шлях і да чаго мы дайшлі за апошнія гады.

Вельмі цікава. Вы казалі, што займаецеся толькі сінтэзам маўлення ў асноўным? Гэта так?

Не, ужо больш, акрамя сінтэзу маўлення, мы займаемся таксама распазнаваннем маўлення. Структуру нашай дзейнасці можна паглядзець на старонцы сайта.

Але ж усё больш і больш прыходзіць думка, што патрэбны і корпусная лінгвістыка, і семантыка, бо не хапае сінтэзатарам маўлення інтэлектуальнасці. Тэксты вельмі складаныя, іх трэба ўмець разбіраць, парсіць. Напрыклад, людзі пішуць са скарачэннямі, з лікамі, значкамі – а мы павінны апрацоўваць гэта для людзей, якія нічога не бачаць ці зараз ня могуць бачыць. Таму трэба зрабіць вельмі глыбокі лінгвістычны аналіз, з дапамогай якога можна будзе вырашыць амаль любую лінгвістычную апрацоўку.

Напрыклад, у мяне пытаюць: “Што вы можаце прапанаваць для Machine Translation?”

Вельмі проста, у Machine Translation вельмі складана перакладаць розныя скарачэнні – а мы разбіраем і тлумачым гэтыя скарачэнні. Напрыклад,

“Самалет ляцеў з хуткасцю 1000 км/г” – вось гэтыя км/г (кіламетры ў гадзіну) мы разгортваем у словы і толькі пасля гэтага аддаем у Machine Translation.

Атрымоўваецца, што вы дадалі яшчэ адзін крок у рашэнне задачы. А вось што вы выкарыстоўваеце – слоўнікі ці што другое?

Ёсць 2 спосабы:

Першы – па слоўніках. Але ж мы сутыкнуліся з тым, што слоўнікаў не хапае, трэба прымяняць Rule-based метад (заснаваны на правілах): разгортка залежыць ад папярэдняй часткі. Напрыклад, 121 *кіламетр* у гадзіну, але 122 *кіламетры* ў гадзіну – калі выкарыстоўваць слоўнікі, то гэта будзе няправільна гучаць, а для славянскіх моваў гэта крытычна.

Хутка выйдзе другая частка нашай размовы, з якой Вы даведаецеся пра тое, якім бокам адносяцца робаты да сінтэзу маўлення, факты з гісторыі камп'ютэрна-лінгвістычнай школы ў Беларусі, а таксама імёны кампаній, якія займаюцца NLProc *у нашай краіне.*

Падчас падрыхтоўкі матэрыялу мы карысталіся сэрвісам праверкі правапісу згаданай лабараторыі.

May 27, 2015 6:11 am by yauhen-info 0 Notes

NLPROC BELARUS SCIENCE ACADEMY MACHINE TRANSLATION

COMPUTATIONAL LINGUISTICS PHONETICS ALLOPHONE SPEECH SYNTHESIS