УДК 004.855.2

Овсянніков Б.О., магістрант, Полякова Н.П., доцент, науковий керівник,

ДОСЛІДЖЕННЯ ТА РОЗРОБКА МЕТОДІВ СЕМАНТИЧНОГО АНАЛІЗУ ТЕКСТІВ З ВИКОРИСТАННЯМ ПРОДУКЦІЙНОЇ БАЗИ ЗНАНЬ

Запорізька державна інженерна академія, кафедра ПЗАС

З часів так званої “статистичної революції”, яка відбулася в 80-ті та 90-ті годи, велика частина досліджень в галузі природних мов сильно покладалася на машинне навчання. Раніше багато задач в обробці мов виконувалися шляхом ручного кодування правил, що, в цілому, не можна назвати надійним у випадку з природними мовами. Проте в наш час все популярнішим стає таке явище як краудсорсинг, що дозволяє використовувати багате число користувачів Інтернет для отримання від часу доволі таки приголомшливих результатів [1]. Таким чином, рішення які виконують обробку природних мов без використання машинного навчання як і раніше може бути актуальную, якщо створити систему, за допомогою якою певна спільнота могла би займатися редагуванням грамматик. Для побудови систем відповідальних систем та природних мовних інтерфейсів семантичний аналіз став важливою та потужною парадигмою. Семантичні аналізатори відображають природну мову в логічних формах, класичне подання багатьох важливих мовних явищ. Сучасний стан питання полягає в тому, що ми зацікавлені в вивченні семантичних синтаксичних аналізаторів даних, що представляє новий рівень статистичних та обчислювальних проблем [4]. Семантичний синтаксичний аналіз - це багате злиття логічного та статистичного світу, і це поєднання буде мати невід'ємну роль у майбутньому систем усвідомлення природних мов.

Традиційні семантичні парсери мають дві обмеження: вони вимагають анотованих логічних форм як нагляд, і вони працюють в обмежених доменах з невеликою кількістю логічних предикатів. Недавні розробки спрямовані на те, щоб подолати ці обмеження, зменшивши необхідність нагляду або збільшивши кількість логічних предикатів.

На лексичному рівні найважливішою проблемою в семантичному аналізі є відображення природних мовних фраз (наприклад, "відвідування") в логічні предикати (наприклад, "Освіта"). У той час як семантичні аналізатори обмежених доменів здатні вивчати лексику з наочного нагляду, в великих масштабах вони мають недостатнє охоплення. Багато значних ранніх успіхів відбулося в області машинного перекладу, зокрема завдяки роботі в IBM Research, де розроблялися послідовно більш складні статистичні моделі. Ці системи мали змогу скористатися існуючими багатомовними текстовими корпусами, які були створені парламентом Канади та Європейським Союзом внаслідок законів, що вимагають перекладу всіх урядових справ на всі офіційні мови відповідних систем управління. Проте більшість інших систем залежать від корпусів, спеціально розроблених для завдань, що їх виконують ці системи, що було (і часто залишається) основним обмеженням успіху цих систем. Як наслідок, багато досліджень було зроблено в методах більш ефективного навчання з обмеженим обсягом даних.

Останні дослідження все більше зосереджувалися на безконтрольні та напівнаглядові алгоритми навчання. Такі алгоритми здатні вчитися на даних, які не були анотовані вручну з потрібними відповідями, або використовуючи комбінацію анотованих і не анотованих даних. Як правило, це завдання набагато складніше, ніж контрольоване навчання, і зазвичай виробляє менш точні результати для певної кількості вхідних даних.

В цій роботі запропонований алгоритм, який дозволяє розбирати речення послідовно, а також знаходити найбільш вірогідне місце помилки (якщо така мала місце у вхідному висловлюванні). Також особливістю рішення є можливість *індукування правил* на основі існуючих [2]. Ця робота базується на семантичному парсері SEMPRE [3] написаний на мові програмування Java. Окрім цього використовуються власні семантичні функції та перероблена парсингова частина.

Процес семантичного розбору можна умовно поділити на два етапи:

1. Створення семантичного дерева висловлювання на основі грамматики та знаходження взаємозв'язків між окремими частинами.

2. Перетворення дерева у логічні форми за допомогою семантичних функцій.

В якості прикладу використовується розширювана граматика яка переводить формальну англійську в Solidity, мову для написання смарт-контрактів для блокчейну Ethereum. Для внутрішнього представлення семантики використовуються лямбда-вирази, що є домінуючою парадигмою в галузі розуміння природних мов.

Система складається з таких компонентів:

- **Мовний аналізатор.** Виконує первісний аналіз вхідного тексту, токенізує його, обробляє пунктуаційні знаки тощо.
- **Граматика.** Зберігає правила та дозволяє додавати та індукувати нові.
- **Парсер.** Використовуючи правила отримані з граматики та набір семантичних функцій, виконує обидва етапи розбір тексту
- **Семантичні функції.** Набір функцій які відображають фрази у логічні форми

Висновки:

- Були досліджені методи розбору текстів природною мовою.
- Був розроблений алгоритм для семантичного розбору.
- Був досліджений та модифікований процес індукування правил на основі існуючих.

Література

1. Christine Gerber and Martin Krzywdzinski “Brave New World of Work? Crowdwork Distributes Tasks on a Global Scale”. WZB Report 2018
2. Sida I. Wang et al. “Naturalizing a Programming Language via Interactive Learning”, 2017.
3. Jonathan Berant, Percy Liang “Semantic Parsing via Paraphrasing”, 2014
4. Liang (2016) Learning Executable Semantic Parsers for Natural Language Understanding.