

МАТЕРІАЛИ НАУКОВИХ ДОСЛІДЖЕНЬ ЗДОБУВАЧІВ, АД'ЮНКТИВ, МАГІСТРІВ, СТУДЕНТІВ, КУРСАНТІВ

Г.В. Карнаух, молодший науковий співробітник Українського мовно-інформаційного фонду Національної академії наук України;

Науковий керівник: кандидат філологічних наук, старший науковий співробітник Українського мовно-інформаційного фонду Національної академії наук України **Л.Л. Шевченко**

АВТОМАТИЧНА ДЕТЕРМІНАЦІЯ ОМОНІМІЧНИХ ОДИНИЦЬ НА МАТЕРІАЛІ ТЕКСТІВ НОРМАТИВНО- ПРАВОВИХ АКТІВ

У світовій лінгвістиці зростає розуміння важливості проведення фундаментального дослідження системних зв'язків у тріаді “інформація – мова – інтелект”. Свідченням цього є значна кількість науково-технічних програм, присвячених вивченню інтелектуальних мовно-інформаційних систем, зв'язків між мовою та мисленням, а також когнітивних процесів, особливо їх лінгвістичних аспектів і т. ін. Зрозуміло, що нові погляди та підходи стосовно шляхів розвитку лінгвістичної науки вимагають від неї вирішення й нових завдань: “наукова проблема в цій галузі лежить у площині побудови формалізованих моделей лінгвістичної компетенції, апробації і верифікації цих моделей та створенні на їхній основі ефективно діючого програмного забезпечення...” (Широков В.А.). Зрештою, потреба автоматизованої обробки текстів викликала

формування цілого науково-технічного напрямку – опрацювання природної мови (Natural Language Processing – NLP).

Серед систем NLP, розроблених в Українському мовно-інформаційному фонді Національної академії наук України (УМІФ НАНУ), широке застосування мають інформаційно-пошукові системи, системи граматичного анотування, індексування текстів, машинного перекладу, семантичного, статистичного аналізу й синтезу мовних одиниць, інструментальні системи укладання словників різного типу (тлумачних, перекладних, орфографічних, етимологічних, фразеологічних, словників словосполучень, частотних тощо), системи менеджменту лінгвістичних корпусів та ін. Останнім часом набуває особливого значення створення лінгвістичних корпусів, які налічують сотні мільйонів і навіть мільярди слововживань і є інструментом фронтального дослідження мови – як у феноменальному, так і в контекстуальному аспектах. При цьому значно загострюється проблема ідентифікації і кваліфікації різного роду лінгвістичних неоднозначностей, і насамперед – зняття граматичної омонімії слів. Від успішного розв’язання останньої проблеми залежить створення ефективного інструментарію автоматизованого маркування корпусів, без якого розвиток корпусної лінгвістики взагалі стає проблематичним.

Питання, пов’язані з вивченням й усуненням граматичної неоднозначності словоформ, досліджено багатьма лінгвістами і на теоретичному, і на практичному рівнях. Теоретичний аспект (здебільшого в контексті загальних проблем мовної неоднозначності, омонімії та полісемії, зумовлених потребами лексикографії, лексикології та граматики) висвітлено в працях О. Потебні, Л. Авксентьева, Ю. Апресяна, О. Ахманової, В. Виноградова, Г. Гнатюк, Т. Грязнухіної Н. Клименко, М. Кочергана, Л. Лисиченко, А. Лучик, М. Муравицької, В. Русанівського,

О. Смирницького, О. Тараненка, Г. Уфімцевої, О. Шипнівської, В. Широкова, Д. Шмельова та ін. Практичний аспект (дослідження особливостей автоматичної ідентифікації текстових слівформ (машинний переклад, автоматичне коригування тексту тощо)) представлено у роботах Т. Аполлонської, С. Білокриницької, Т. Грязнухіної, А. Залізняка, Ю. Зеленкова, Т. Любченко, Ю. Марчука, Т. Молошної, О. Невзорової, Ю. Орехова, Р. Піотровського.

У результаті роботи над розробленням формалізованих методів аналізу й усунення мовних неоднозначностей в УМІФ НАНУ створено програму автоматичної детермінації омонімічних одиниць із використанням статистичних методів. Присвоєння граматичних характеристик слівформам забезпечується Граматичним словником української мови при використанні алгоритму лематизації, унаслідок чого аналізована одиниця набуває набору граматичних параметрів (v, g) , де v – частина мови, а g – граматичне значення. Омонімічна слівформа отримує декілька таких наборів (для кожного значення окремо) у кількості від 1 до n (Кригін М.Ю).

Практичне використання зазначеного програмного засобу показало, що специфіка орфографічного та пунктуаційного оформлення текстів суттєво впливають на результати зняття в ньому граматичної неоднозначності. Стилiстичні особливості текстів при цьому також відіграють неабияку роль. Наприклад, для нормативно-правових актів, з огляду на своєрідність законодавчої мови, можна виробити ряд додаткових функцій, які допоможуть видалити зайву інформацію з граматичного набору омонімічної слівформи. Оскільки граматичний словник досить широко представляє лексику української мови (включаючи розмовну, застарілу, рідковживану та ін.; власні назви нарівні із загальними), більшість слівформ визначаються програмою як омонімічні. Виникає потреба створення лексико-граматичних

фільтрів, формування яких ґрунтується на детальному вивченні особливостей мови зазначених текстів.

Так, до фільтру можна ввести такі леми та граматичні значення: вигуки (*а* – уживається для вираження здогаду, здивування і т. ін.); частки з ремаркою *підсил.* (*і* – уживається для підкреслення, виділення значення слова, перед яким стоїть); усі леми з ремаркою *розм., діал.* (*віда* – звістка, чутка, повідомлення); рідковживані або нехарактерні для мови законодавства слова (*банка* – скляна, металева і т. ін. посудина переважно циліндричної форми; підвищення морського дна, мілина в морі; дошка, яка надає шлюпці або човну поперечної міцності й одночасно служить сидінням для веслярів), серед яких і словоформи, що є власними назвами (*Або* – назва міста у Фінляндії); кличний відмінок однини і множини для іменників чоловічого, жіночого, середнього роду (*адміністрації* – кличний відмінок, множина); особові форми наказового способу (*плати* – відшкодовувати вартість одержаного або використаного), форми 1-ої особи для дієслів (*вищим* – видавати уривчасті, різкі, пронизливі звуки).

Отже, з огляду на специфіку законодавчих текстів, які не описують подій, фактів, наукових понять, почуттів і переживань людей, не характеризують ставлення до певних осіб, явищ дійсності, а владно приписують суб'єктам права конкретну модель поведінки, формулюють вимоги, загальнообов'язкові норми, можна сформулювати певні мовні норми, які, у свою чергу, дозволять розробити відповідні програмні засоби автоматичного опрацювання текстів зазначеного виду, зокрема, створити лексико-граматичні фільтри, що сприятимуть збільшенню кількості правильно ідентифікованих неоднозначних словоформ.