АВТОМАТИЗОВАНЕ ВИДОБУВАННЯ ТЕРМІНОЛОГІЧНИХ ОДИНИЦЬ З НАУКОВО-ТЕХНІЧНИХ ТЕКСТІВ

Борисова Н. В., Решетило С. С.
Національний технічний університет
"Харківський політехнічний інститут",
м. Харків, вул. Пушкінська, 79/2, тел. 707–63–60
e-mail: *borisova_nv@mail.ru, 13-svetik.cat.07@mail.ru*

Вирішення багатьох задач автоматизованої обробки текстів потребує видобування з текстів термінів, тобто слів або словосполучень, що називають поняття певної предметної області. Науковий термін не тільки точно і однозначно визначає чітко окреслене спеціальне поняття будь-якої галузі науки, а й відображає його співвідношення з іншими поняттями в межах предметної області.

Видобування термінів необхідне при вирішенні багатьох задач автоматизованої обробки текстів, а саме машинний переклад, літературно-наукове редагування, видобування знань з наукових текстів, реферування та анотування текстів, складання словників певної предметної області та ін.

Оскільки наявність термінів, понять та їх визначень є особливістю науково-технічних текстів, тому що основною функцією наукового стилю є оформлення, збереження, передача наукової інформації, для виявлення ознак термінів доцільно було б проаналізувати лексико-фразеологічні та дискурсивні особливості цих текстів [1]. Це необхідно для виділення дискурсивних маркерів, які відповідають дискурсійній операції «визначення понять». Саме ці маркери і є ознаками термінів у науково-технічних текстах (табл. 1).

Таблиця 1 – Приклади використання деяких груп дискурсивних маркерів

Опис	**Приклади використання**
1	2
Група «*називатися*»	
дієслово «називатися» у формі третьої особи однини теперішнього часу	***Атмосферою*** *називається зовнішня газова оболонка Землі, що сягає від її поверхні в космічний простір приблизно на 3000 км…* [2]
Група «*бути + називати*»	
дієслово "бути" в формі першої особи множини майбутнього часу	*Угрупованням тут будемо називати досить чітко окреслений та територіально сильно обмежений рівень живого* [2].
Група «*так + званий*»	
дієприкметник «званий» в різних формах	*Небажаним є досягнення так званого «екологічного імперативу» – своєрідної межи або рівня взаємодії суспільства та природи, перевищення якого буде мати катастрофічні наслідки для людства* [2]

Продовження таблиці 1

1	2
Група «*розуміти*»	
дієслово «розуміти» у формі першої особи множини теперішнього часу	*Під біоценозом екологи розуміють історично сформовану сукупність рослин, тварин та мікроорганізмів, що населяє біотоп* [2].
Група «це»	
наявність частки «це»	*Ланцюги живлення – це ряди взаємопов'язаних видів, в яких кожний попередній є об'єктом живлення наступного* [2]
Група «–»	
наявність «–»	*Біоконверсія – біологічна переробка органічних відходів промисловості, сільського й комунального господарства* [2]

Система автоматизованого видобування термінів з науково-технічних текстів певної предметної області виявлятиме терміни саме за дискурсивними маркерами. Для того щоб розробити таку систему можна використати *регулярні вирази* – систему обробки тексту, засновану на спеціальній системі запису зразків (шаблонів, масок), що задають правила пошуку. Зараз регулярні вирази використовуються багатьма текстовими редакторами і утилітами для пошуку та зміни тексту на основі вибраних правил [3].

Алгоритм автоматизованого видобування термінів з науково-технічних текстів з використанням регулярних виразів представлено нижче:

1. Відібрати для аналізу множину текстів предметної області.
2. У текстах предметної області визначити дискурсивні маркери операції «визначення поняття».
3. Розподілити дискурсивні маркери по групах.
4. Задати регулярні вирази для пошуку дискурсивних маркерів, що відповідають дискурсивній операції «визначення поняття».
5. Для кожного тексту з множини здійснити пошук термінів, використовуючи побудовані на кроці 4 регулярні вирази.
6. Вивести список термінів.

Список літератури

1. Баева Н.В. Структурирование и извлечение знаний, представленных в научных текстах / Н.В. Баева, Е.И. Большакова, Н.Э. Васильева // Девятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2004. Труды конференции в 3-х томах. Т. 2. – М.: Физматлит, 2004. – С. 46-54

2. Кучерявий В.П. Екологія. – Львів: Світ, 2001 – 500 с.: [Електронний ресурс]. – Режим доступу: http://eduknigi.com/ekol_view.php?id=1

3. Царьков В.Б. Теория и методика построения регулярных выражений. Проблема самообразования: [Електронний ресурс]. – Режим доступу: http://lipetsk.lug.ru/projects/re/re-building-howto.html