

ЗАСТОСУВАННЯ ІНСТРУМЕНТАЛЬНИХ ЗАСОБІВ ПАКЕТУ DKPRO CORE ДЛЯ ВИЗНАЧЕННЯ КЛЮЧОВИХ СЛІВ АНГЛОМОВНОГО ТЕКСТУ.

Яхимович О. В.

Науковий керівник – д.т.н., проф. Бісікало О. В.

Існує велика кількість доступних систем автоматичного виділення ключових слів, розроблених і орієнтованих на обробку природних мов. Ці системи засновані на певних методах визначення ключових слів, які діляться на лінгвістичні та статистичні. Лінгвістичні методи ґрунтуються на значеннях слів, зокрема використовують онтології та семантичні дані про слово. На жаль, ці методи ресурсоємні на ранніх етапах: розробка онтологій, наприклад, вельми трудомісткий процес. З іншого боку, статистичні методи супроводжуються значними обсягами «вербального шуму», який суттєво впливає на якість визначення ключових слів. Тому найбільш перспективними для дослідження, є гібридні методи, для яких швидкість статистичної обробки тексту підсилюється можливостями сучасних лінгвістичних пакетів.

Результати парсерингу природних мов за допомогою сучасних лінгвістичних пакетів дозволяють на доступному програмному рівні оперувати синтаксичними зв'язками між словами окремого речення. Одним з таких пакетів є DKPro Core – це набір програмних компонентів для обробки природної мови, що базується на Apache UIMA framework. Він був побудований з метою підвищення продуктивності дослідників, які працюють з автоматичним аналізом мови. Підхід DKPro Core полягає в тому, що дослідники повинні мати можливість зосередитися на своїх реальних наукових питаннях, а не на розробці технологій.

Запропоновано метод визначення ключових слів, що базується на використанні додаткової інформації про складні залежності між членами англомовного речення. Для функціональної реалізації аналізатора тексту обрано популярний лінгвістичний пакет DKPro Core. Проведені експериментальні дослідження теоретичного обґрунтування методу підтвердили його якісні та кількісні переваги у порівнянні з відомими аналогами. Для англомовного тексту обсягом 1460 слів отримано збільшення повноти визначення ключових слів (на 31,8% за Жаккаром та на 25% за абсолютним значенням) і покращення точності (на 14% за евклідовою і на 5,3% манхетенською відстанями).

Якість отриманих результатів потенційно можна підвищити через окремий аналіз частин мови, оскільки ймовірність релевантності ключового слова, наприклад, іменника і прислівника буде відрізнятися. Окрім цього, варто оцінити збільшення частотних показників для ключових слів шляхом реалізації наявних в DKPro Core компонентів для визначення кореференційних зв'язків.