Н. О. Ямпольська

У пошуках ефективного використання синтаксису у задачах автоматичного реферування тексту

Автоматичне реферування тексту (АР) – один з напрямків обчислювальної лінгвістики, задачею якого є встановлення головного змісту тексту та створення його стислої форми. Цей напрямок пройшов шлях від суто статистичних методів підрахунку ключовіих слів до складних технологій, що поєднують семантичну, прагматичну, синтаксичну та ін. інформацію.

Процес реферування можна схематично описати наступним чином. У вхідному тексті виділяються ключові слова, базуючись на високій частоті слова у даному тексті і низькій частоті слова у коллекції інших документів. Одночасно враховується позиція речення у тексті (по експерементально встановленних формулах обчислюються найбільш впливові абзаци тексту), наявність у реченні сигнальних слів чи фраз («результатом цієї роботи є…»), та інші більш складні фактори. Після цього кожному реченню припусується коефіцієнт важливості і речення, чий коефіцієнт більше за встановлений поріг, обираються до реферату.

У більшості систем з АР синтаксична інформація використовується головним чином для трансформацій (скорочення, перефразування, та ін.) речення, коли воно вже відібране у реферат. У своїй поточній роботі ми шукаємо шляхи більш широкого використання синтаксичної інформації на етапі виділення важливої інформації у тексті.

Ми працюємо з матеріалом з 90 «ручних» рефератів, що були зроблені волонтерами на базі 30 статтей-новин з New York Times розміром від 1,000 до 2,500 слів. В результаті було отримано 381 унікальних речень (враховуючи те, що деякі речення були обрані відразу 2 або 3 людьми). На сьогоднішній день ми обробили 182 речення.

Після обробки речення, а саме встановлення їх синтаксичної структури та визначення тих синтаксичних елементів, що повторюються найчастіше, ми отримали попередні результати, які мають бути перевірені на більшому матеріалі. Ми ввели поняття *щільності* синтаксичних елементів, що враховує відносну кількість елементу в рефераті і входному тексті. Водночас, нами були підраховані розподіл важливих речень у вхідному тексті (напр., відмічено, що речення різних рефератів на один текст відрізняються один від одного позицією на 1 чи 2 речення).

Роботу планується продовжити на більшому матеріалі та з залученням синтаксичного аналізатору Connexor.