УДК 004.6
В. Веселовська, Л. Дмитроца
(Тернопільський національний технічний університет імені Івана Пулюя)

СТАТИСТИЧНИЙ БАГАТОМОВНИЙ ПЕРЕКЛАД ЗАПИТІВ ПРИ ІНФОРМАЦІЙНОМУ ПОШУКУ

UDC 004.6
V. Veselovskaya, L. Dmytrotsa
(Ternopil Ivan Puluj National Technical University, Ukraine)

STATYSTYCHNYY BAHATOMOVNYY PEREKLAD ZAPYTIV PRY INFORMATSIYNOMU POSHUKU

Розглядається можливість покращення релевантної видачі результатів інформаційного пошуку на запити користувача, враховуючи багатомовність вхідних даних розроблюваної системи обробки високошвидкісних потоків текстових даних.

Основною проблемою при частковому перекладі (на рівні запитів) є складність виявлення тематичних зв'язків між змістом пошукових запитів та змістом текстових документів через різне представлення у схожих текстах на різних мовах одної конкретної ситуації чи події. Даючи запит, людина формулює його, керуючись лише своїми представленнями про зміст необхідного документа. Розповсюджені лексичні засоби серед інформаційно-пошукових систем будуються на списках ключових слів. Виражений таким чином семантичний зміст документа обмежується цими списками для різних предметних галузей. Порівнюючи документи, представлені на різних мовах, але з однаковим тематичним змістом, їх схожість виявиться лише у разі співпадіння понять ключових слів для списків на цих мовах.

Одним із варіантів вирішення проблеми є розширення лексичного складу запитів та списків ключових слів або пар ключових слів. Також можливим є залучення алгоритмів статистичного перекладу. Даний тип машинного перекладу широко використовується у великих комерційних організаціях, оскільки потребує великих потужностей для обробки та зберігання мовних пар документів для кожної мови окремо. Ця технологія передбачає існування відкритих онлайн-сервісів – це Google translate та Яндекс-перекладач. Для роботи подібних систем необхідна наявність великих баз паралельних текстів, де зберігаються словосполучення (N-грами) та їх переклади (рис.1).

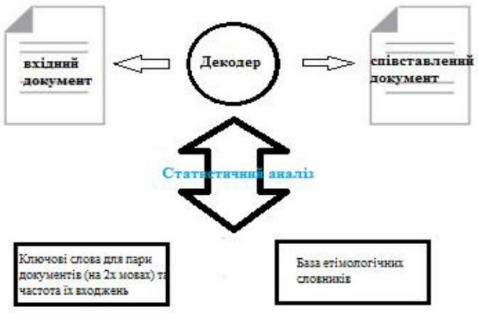

Рис. Блок-схема видачі запитів

Звузивши задачу з повнотекстового перекладу до перекладу лише запитів та видачі результатів на мові оригіналу, можливо суттєво знизити затрати розрахункових потужностей, реалізуючи алгоритми статистичного пошуку. Керуючись правилами спільного походження більшості слів для переліку мов, що входять до однієї мовної групи (слов'янські, романські та ін. мови), можливим є реалізація програмного алгоритму для більш детального виокремлення семантичного навантаження ключових слів запитів та документів на різних мовах. Планується реалізація даного підходу для розроблюваної системи обробки високошвидкісних потоків текстових даних, залучивши до роботи електронні бази етимологічних словників для різних мов (словники, що містять інформацію про фонетичні та семантичні зміни окремих слів та морфем конкретної мови).