Консолідація інформації отриманної з мережі інтернет та напів-автоматичне наповнення бази знань

Денис Циганок

Кафедра математичної інформатики, Київський національний університет імені Тараса Шевченка, УКРАЇНА, м.Київ, пр. Глушкова, 4д,
E-mail: denis.tsyganok@gmail.com

How will we acquire structured information from Web? Of course, we can manually enter information to the huge knowledge base – like Wikipedia. But much better to use machine learning algorithms to extract hundred of thousands of facts from unstructured texts and semi-structured sources. We can "macro-read" the web to populate ontologies. It is much easier than solving the full NLP problem.

Ключові слова – база знань, добування інформації, макро-читання, web content mining, information extraction.

I. Вступ

Веб має потенціал бути найбільшим енциклопедичним джерелом у світі, але ми ще далекі від використання цього потенціалу. Цінний науковий та культурний зміст - все змішалося у величезну кількість неструктурованих текстів низької якості та медіа-контенту. Постає питання: чи можемо ми систематично збирати факти з мережі та консолідувати їх у повноцінну машино-орієнтовану базу знань про сутності світу, про їх семантичні властивості та зв'язки один з одним.

Велика кількість, як комерційних проектів, так і наукових розробок в університетах, займається вирішенням цієї проблеми. Найвідоміші з них: freebase.com і trueknowledge.com, які є компіляціями величезної кількості сутністно- та зв'язоково-орієнтованих фактів; проект dbpedia.org, який зусиллями спільноти збирає «трійки» суб'єкт-властивість-об'єкт з Вікіпедії та інших подібних джерел; проект KnowItAll і TextRunner – розробки Вашингтонського університету; Kylin/KOG і Omnivore, метою яких є отримання довільних зв'язків з природно-мовних текстів; проект ReadTheWeb з метою "макро-читання" довільної веб-сторінки; система sig.ma; проект YAGO, який поєднує зв'язками знання з Вікіпедії та WordNet між собою.

Розглянемо найбільш амбіціозний проект – ReadTheWeb, розробку університету Карнегі-Меллон.

II. Приклад реалізації

Центральним об'єктом дослідження у проекті ReadTheWeb є система безперервного навчання природної мови (NELL – never ending language learner). Під «системою безперервного навчання» розуміють комп'ютерну систему, яка постійно працює, та виконує два типи задач: задачу знаходження (добування інформації з текстів Мережі для подальшого

наповнення бази знань, яка складається зі структурованих фактів) та задачу навчання (кожного дня навчатися знаходити краще, ніж напередодні, тобто повернутися до вчорашніх текстових джерел та отримати з них більше інформації з більшою точністю)

Основний принцип, який лежить в основі дослідження – це те що більшість інформації в інтернеті представлена по декілька разів по-різному (один й той самий факт може бути викладений у різній формі). Саме це дозволяє сконцентруватися на “макро-читанні” замість “мікро-читання” і досягти успіху у машинному навчанні.

NELL отримує два види знань: знання про те, які іменникові словосполучення до яких зазначених семантичних категорій відносяться (наприклад, категорії місто, компанія, спортивна команда); знання про те, які пари словосполучень які семантичні зв’язки задовольняють.

Основними блоками системи є: Блок зчитування пов’язаних шаблонів - Coupled Pattern Learner (CPL); Блок SEAL (CSEAL) - модуль, для зчитування напів-структурованих даних; Пов’язаний морфологічний класифікатор [CMC]; Блок навчання правил логічного виводу [RL].

Найважливішим принципом при реалізації такого підходу до навчання є використання підсистем, які роблять некорельовані помилки. Тоді ймовірність того, що вони всі зроблять одну й ту саму помилку – дуже низька. Це дозволяє звести до мінімуму втручання людини у навчання такої системи.

Висновок

Задача макро-читання для заповнення онтології є набагато більш простим завданням ніж вирішення повної задачі «розуміння» природної мови (NLP). Відзначимо також, що мікро-читання буде також важливим, особливо для анотування окремих веб-сторінок та для зчитування інформації, яка з'являлася в інтернеті лише зрідка. Цікавою задачею може бути покращення мікро-читання використовуючи результати макро-читання. Щодо результатів реалізації вище описанної системи, вчені з Карнегі-Мелонн, після запуску через 67 днів отримали 242 000 фактів з оцінкою точності в 74%. Цей результат показує переваги використання різноманітних методів зчитування знань, а також бази знань, яка дозволяє зберігати факти-кандидати та вірогідні факти.

Література

1. G. Weikum, M. Theobald. From Information to Knowledge: Harvesting Entities and Relationships from Web Sources. PODS, 2010.
2. A. Carlson, J. Betteridge, B. Kisiel1, B. Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. Toward an Architecture for Never-Ending Language Learning. AAAI, 2010