УДК 004.82

Д.С. Сергеєв, А.В. Хіміч

ВИЗНАЧЕННЯ КАТЕГОРІЇ «ЗНАННЯ» ТА ЇЇ ВИКОРИСТАННЯ В ІНФОРМАЦІЙНИХ ПРИРОДНО-МОВНИХ ТЕХНОЛОГІЯХ

Анотація: Стаття присвячена особливостям організації структури природно-мовних текстів, проблемам, що виникають на шляху їх обробки та методам вирішення цих проблем. Ключовим елементом дослідження постає природно-мовне повідомлення, якому часто характерні складність структури та сильна зв’язність. Виділення знань з такого повідомлення пропонується здійснювати на основі системи синтаксичних правил та визначення семантичних зв’язків. Також розглядаються питання зіставлення різних баз знань при синтезі та аналізі повідомлення і зовнішнього контексту як необхідної складової при опрацюванні сильно зв'язаного тексту.

Ключові слова: природно-мовний текст, база знань, лінгвістичний процесор, базова семантико-синтаксична структура, квант знань.

Вступ

Дана стаття присвячена проблемі упорядкування природно-мовних (ПМ) знань, тобто зіставлення первинного вигляду представлення знання – природно-мовного тексту – і структурованої інформації, придатної для автоматичного опрацювання.

Сучасні комп'ютерні системи, які мають потужні процесори і здатні обробляти величезні масиви даних, ефективно працюють лише з такою інформацією, структуру якої можливо формалізувати. Людські ж знання у загальному випадку представлені у природно-мовному вигляді, як словесний опис. Звісно, більшість спеціалізованих наукових знань можна описати формулами, алгоритмами, законами тощо, які спираються на конкретні терміни і визначення – але цей вигляд представлення знань є результатом цілеспрямованої роботи спеціалістів, тобто ручної праці.

Основна функція природно-мовних повідомлень – передача повідомлення, сформованого в процесі мислення, з найменшими витратами ресурсів. Як правило, повідомлення складається з багатьох пов'язаних частин – тверджень, фактів тощо; відповідний текст також містить ці частини, але у скороченому вигляді. Цього достатньо для розуміння

повідомлення людиною, але викликає певні складності при його аналізі – адже в одному неперервному реченні, абзаці тощо може міститися багато окремих частин – квантів знань – тісно пов'язаних між собою.

Постановка задачі

В першу чергу необхідно визначити сам об'єкт роботи. Оскільки загальноприйнятого наукового визначення терміну «знання» у ІТ немає – скористаємось тим визначенням, яке використовуємо у власному дослідженні.

Квант знань – відображення ситуації зорового рівня – об'єкт, його дія та їх повне атрибутивне оточення, що потрапляють на центральну ямку ока. У тексті ситуація представлення у вигляді БССС – окремої структури, яка може описати довільну ситуацію [1, 2]. Відповідно, текст – це сукупність квантів знань і зв'язків між ними, стиснені для економії ресурсів на комунікації.

Мета дослідження – виділення знань з тексту, тобто перетворення тексту у структуру пов'язаних квантів знань, якими можемо оперувати як окремими об'єктами. Таким чином, об'єктом дослідження є структура ПМ інформації, а предметом – сутності (кванти знань) у ПМ тексті та зв'язки (відношення) між ними.

Аналіз аналогів

Існує багато систем, які ставлять за мету створення на основі тексту структурованого масиву даних. Умовно можемо розділити їх на два типи: системи автоматичної обробки тексту (системи перевірки орфографії, парсери тощо), та системи збереження знань (такі як семантичні мережі та онтології). Найбільш повно ці можливості представлені у системах аналізу тексту (*text mining* [3]), зокрема елементарні мережі суміжності (*Co-occurrence networks - CON* [4]) і їх більш складні модифікації. Особливості, що виділяють їх серед інших систем, це:

– врахування контексту фрагменту тексту – абзацу, тексту (на відміну від парсерів, де кожне речення опрацьовується незалежно);

– отримання в результаті зв'язної структури даних (на відміну від систем перевірки орфографії, де кінцевим результатом обробки тексту є теж текст);

– повністю автоматична обробка тексту (не онтології, що заповнюються вручну або з втручанням оператора).

CON не є ідеальним інструментом для вирішення цієї задачі. Час-тково це зумовлено самою концепцією *CON*, частково особливостями їх реалізації у реальних системах:

– складність структури, що генерується на основі тексту: вигляд структури сильно залежить від конкретного фрагменту тексту;

– сильно зв'язані дані: з отриманої структури важко виділити окремі незалежні кванти знань;

– складність автоматизації: часто для коректної роботи парсера CON використовуються стохастичні методи (які вносять похибки, що накопичуються, на кожному етапі) або втручання оператора (що перекреслює усі переваги повної автоматизації).

Особливості запропонованого підходу

Характерною особливістю існуючих підходів і систем є презумпція первинності тексту – тобто, текст розглядається як окремий, завершений об'єкт, який має завершену структуру даних і містить у собі усю необхідну інформацію для її розуміння. Пропонуємо іншу концепцію. По-перше, стверджуємо, що текст відображає частину дійсності, тобто текст спирається на знання, які автор вважає загальновідомими. По-друге, стверджуємо, що текст складається з окремих квантів знань, пов'язаних через відношення, а не з цільної структури знань. Це дозволяє припускати наявність у тексті пропусків (інформація, яка має бути у читача для повного розуміння тексту), і допускати часткове заповнення структури знань тексту.

Оскільки структура БССС заснована на особливостях нейрофізіології людини, ролі слів у тексті визначаються в першу чергу структурою БССС, а не правилами граматики. Таким чином, алгоритми аналізу апріорі матимуть скінченну складність, а результат — передбачувану форму. Зазначимо, що ці кванти знань не можуть використовуватися самостійно, отже їх необхідно розкрити перш ніж використовувати. Тобто, при декомпозиції необхідно записувати усі можливі варіанти, оскільки текст може мати кілька варіантів тлумачення, і його «основне» розуміння може змінюватись при розширенні накопичених знань – бази знань (БЗ). Для уникнення подібних ситуацій, в даній статті обмежуємо контекст невеликим фрагментом тексту – таким як речення або абзац. Розглянемо загальний алгоритм та приклад роботи запропонованої системи.

 ISSN 1560-8956

Алгоритм виділення знань з тексту

Першими можемо виділити у тексті прості синтаксичні зв'язки *Obj/Subj/Attr* – наприклад, однозначний збіг граматичних форм слів. В отриманій таким чином БЗ вже є достатньо знань для вирішення простих задач, скажімо для визначення деяких *Obj* через характерні для них *Atrr* та *Mov* (дії). Зауважимо, що отримана таким чином БЗ нагадує *CON*, але є суттєва відмінність: усі її елементи пов'язані семантично – а отже, отримуємо саме знання і їх можна використовувати.

Основною метрикою такої БЗ є точність, яка має сягати 100%, навіть якщо для цього на початкових етапах доведеться відкинути значну частину знань. Очевидно, точність завжди буде менша через недосконаліть правил/неповну БЗ, але таке жорстке обмеження дозволяє ліквідувати штучну похибку у знаннях. Це компенсується збереженням високої точності при наповненні бази; крім того, ті ж самі вихідні тексти можна буде обробляти більш повно після кожного розширення бази граматичних правил та семантичних зв'язків. Розглянемо ці процеси детальніше.

При введенні нових синтаксичних правил необхідно оновлювати всю базу – одночасно або поетапно; при отриманні нових семантичних зв'язків необхідно оновлювати лише ті фрагменти знань, які з ними пов'язані. Оскільки наразі існують досить ефективні системи аналізу (парсингу) текстів, часта зміна бази граматичних правил не передбачається, але таку можливість виключати не можна – а отже, враховувати її необхідно ще на ранніх етапах проектування. Зауважимо, що такий підхід загалом відповідає роботі мозку людини при отриманні нових знань.

Одразу зазначимо також потенціал для автоматичної обробки знань: по-перше, зі збільшенням БЗ тільки за рахунок комбінування існуючих квантів знань можливо виділяти велику кількість нових тверджень, що будуть істинними у рамках цієї бази (тобто, моделюємо виведення нових знань після навчання); по-друге, у отриманій таким чином базі можливо простежити будь-яке твердження від елементарних фактів і до складних ланцюжків пов'язаних знань, а отже – автоматично знайти протиріччя і розбіжності на будь-якому рівні.

Приклад роботи

Розглянемо роботу такої системи на прикладі статей про арифметичні операції з Вікіпедії.

Введемо два тексти: (1) «*Додавання — бінарна арифметична операція*» та (2) «*Віднімання — двомісна математична операція*».

Виділимо очевидні *Obj*, *Subj*, *Attr* за збігом граматичних ознак.

Отримуємо:

(1) Obj «операція» Attr «арифметична», «бінарна»

(2) Obj «операція» Attr «математична», «двомісна»

Ці вирази є квантами знань, і з ними стає можливо працювати як зі знан-нями – наприклад, вибирати усі *Attr* для *Obj* «операція», або визначати, чи існує в БЗ БССС (квант знань, тобто факт), який містить даний *Obj* («операція») та даний *Attr* (наприклад, «бінарний» – так, «унарний» - ні)

Спробуємо розширити базу граматичних правил. Визначимо правило: конструкція «—» еквівалентна дієслову «є». Зазначимо, що це правило може помилково опрацювати інші випадки використання тире «—», і це призведе до неправильних записів у БЗ; водночас, кількість можливих граматичних правил щодо використання тире обмежена, і при їх правильній обробці система буде завжди коректно опрацьовувати правильно побудовані тексти (такі, які не містять граматичних помилок).

Після повторного опрацювання тексту БЗ доповнюється виразами:

(3) Obj «додавання» Mov «є»

(4) Obj «віднімання» Mov «є»

Зазначимо, що в такому вигляді ці вирази не мають сенсу, але ми все одно зберігаємо їх в БЗ.

Перевіримо роботу системи при додаванні семантичного правила. Визначимо направлене елементарне відношення «*S1* є *S2*», де *S1*,*S2* – структури БССС. Тепер на основі виразів (1)-(4) можемо заповнити дані про відношення:

(5) [Obj «додавання» Mov «є»] «є» [Obj «операція» Attr «арифметична», «бінарна»]

(6) [Obj «віднімання» Mov «є»] «є» [Obj «операція» Attr «математична», «двомісна»]

Ці твердження вже можна використовувати не лише як кванти знань (факти), а й як окремі фрагменти знань, або твердження. Скажімо, «чи пов'язаний *Obj* «додавання» з *Obj* «операція» через відношення «*S1* є *S2*»» - або, простіше кажучи, «чи є додавання операцією».

Тим не менш, для повної передачі знань у цьому тексті необхідний і зовнішній контекст. Так, встановлення для галузі знань «математика», до

якої належить текст, синонімії між «двомісна операція» та «бінарна операція» дозволяє визначити спільну другу частину у твердженнях (5),(6). Ця синонімія ніяк не вказана у оригінальних фрагментах тексту, але вона входить до базових понять галузі «математика», і автори оригінальних фрагментів спираються на них як на вже відомі знання [5]. Зауважимо, що для читача, який не знайомий з основами галузі «математика», не буде зрозуміло, що ця синонімія існує, і він не зможе об'єднати «додавання» та «віднімання» як «бінарна/двомісна операція»; а отже, і автоматична система не буде здатна зробити цього за тих самих умов

Висновки

У статті запропоновано підхід до визначення знань у природномовних текстах на основі структури БССС, що відображає квант знань – ситуацію зорового рівня. Цей підхід дозволяє вирішити проблему складності структури знань природно-мовного тексту і надає основу для розробки лінгвістичного процесора – набору правил розбору тексту по БССС.

Оскільки кожен квант знань, отриманий за запропонованим алгоритмом, має чітко визначену структуру, отримані знання не є сильно зв'язаними. Таким чином, їх можна безпосередньо використовувати у система обробки знань – експертних системах, системах комп'ютерної логіки та прийняття рішень.

Для підтримання високої точності роботи системи пропонується встановлювати максимальний рівень точності, що негативно впливає на повноту БЗ. Ця проблема вирішується накопиченням великого обсягу знань і повної бази граматичних правил, але це не є обов'язковим, оскільки алгоритм однаково добре працює на текстах різного обсягу.

Працездатність запропонованого алгоритму і окреслені проблеми показано на прикладі обробки реального тексту. Результати обробки тексту збігаються з теоретичними побудовами.

Список використаних джерел

1. Кисленко, Ю.И. От мысли к знанию (нейрофизиологические основания) / Ю.И. Кисленко.— Київ : Український літопис, 2008.

2. Kyslenko Y. Cognitive architecture of speech activity and modelling thereof / Y. Kyslenko, D. Sergeiev. // Biologically Inspired Cognitive Architectures. – 2015. – №12. – p.134–143.

3. Bretonnel Cohen K. Getting Started in Text Mining / K. Bretonnel Cohen, L. Hunter. // PLoS Comput Biol. – 2008. – №4.

4. Lund K. Producing high-dimensional semantic spaces from lexical co-occurrence / K. Lund, C. Burgess. // Behavior Research Methods, Instruments, & Computers. – 1996. – №28. – p. 203–208.

5. Сергеєв Д.С. Оптимізація використання природно-мовних баз знань шляхом тематичної декомпозиції / Д. Сергеєв // ЕлІТ-2016 , (Львів-Чинадієво, 27-30.08.2016 р.) – Львів: Львів. нац. ун-т ім. І. Франка, ф-т електроніки, 2016. – С. 25-28.