Модель формалізації смислового значення елементів факту речення англійської мови

Ніна Хайрова

д. т. н., професор, професор кафедри інтелектуальних та комп'ютерних систем, Національний технічний університет «Харківський політехнічний інститут», Україна, E-mail: nina_khajrova@yahoo.com

Тетяна Гайденко

студентка факультету соціально-гуманітарних технологій, Національний технічний університет «Харківський політехнічний інститут», Україна, E-mail: gaudenko.tanya@gmail.com

This work is related to the search and processing of facts by using the logical-linguistic model in the English technical documentation. We consider the fact in the form of a triplet: Subject>Predicate>Object with the Predicate representing relations and the Object and Subject pointing out two entities. The logical-linguistic model is based on the use of the grammatical and semantic features of words in English sentences.

Ключові слова — логіко-лінгвістична модель, автоматичне вилучення фактів, технічна документація, триплет факту, граматичні категорії.

I. Вступ

Сьогодні одним із перспективних напрямків обробки текстової інформації є отримання інформації з тексту. Вилучення інформації (Information extraction) – це задача автоматичної побудови структурованих даних з неструктурованих або слабкоструктурованих машинозчитуваних документів. У більшості випадків ця діяльність стосується зчитування текстів природної мови шляхом його обробки (NLP). Вилучення інформації є інтелектуальним процесом і не може бути вирішеним повною мірою існуючими здебільшого статистичними інформаційними методами. Вирішення цього завдання вимагає залучення апарату штучного інтелекту і евристичних методів обробки інформації.

Наразі розроблено велику кількість методів обробки текстів, що дозволяють вирішувати задачі вилучення фактів у вузьких предметних галузях, однак залишаються не розв'язаними задачі машинного розуміння тексту, що дозволяють вилучати факти різних типів із різноманітних галузей.

У статті ми пропонуємо логіко-лінгвістичну модель, яка дозволяє витягувати різні факти із текстів технічної документації англійської мови.

II. Огляд літературних джерел

Через стрімке збільшення кількості неструктурованих матеріалів, зокрема в Інтернеті, усе більше зростає роль такої інформаційної процедури, як вилучення інформації. Типова задача вилучення інформації – просканувати набір документів, написаних природною мовою, і наповнити базу даних виділеною корисною інформацією. Але сучасні підходи вилучення інформації у більшості випадків використовують методи обробки природної мови, спрямовані лише на дуже обмежений набір тем (питань, проблем) – часто лише на одну тему. При цьому традиційно використовувані підходи: пошук за шаблоном, пошук опорного елемента, пошук за онтологією тощо – мають як свої переваги, так і недоліки. Усі вони працюють тільки на заздалегідь визначених предметних галузях, що обмежують тематику досліджуваних текстів, і вимагають добре структурованих текстів (патентів, бібліографічних описів, авторефератів тощо). У той же час переважна частина текстової інформації, яка подається в комп'ютерних мережах, – це не структуровані і слабкоструктуровані тексти різної тематичної спрямованості.

Зберігання даних у неструктурованій формі без будь-якої певної схеми даних є найпоширенішим способом подачі інформації [6]. Вилучення фактів із такої інформації може стати додатковим потужним джерелом для різних видів завдань, наприклад, генерація онтологій із корпусу тексту, написаного природною мовою, розвиток інтелектуальної питально-відповідної системи, бізнес-аналізу, бізнес-аналітики тощо [4].

Часто автоматичне вилучення інформації з тексту пов'язано з проблемою спрощення тексту. Вирішення цієї проблеми має на меті спростити структуроване подання інформації, представленої у вільному тексті. Загальна мета полягає в тому, щоб створити більш легко машинозчитуваний текст для опрацювання речень [7].

Типові підзадачі вилучення інформації включають:

- вилучення іменованих сутностей (named entities recognition);
- екстракцію визначення відносин між суб'єктами (relationship extraction) [5].

Вилучення іменованих сутностей має на увазі визначення імен сутностей (для людей і організацій), географічних назв, тимчасових виразів і деяких типів чисельних виразів, які використовують існуючі знання про домен або інформацію, отриману з інших речень. Як правило, завдання розпізнавання включає в себе призначення унікального ідентифікатора для вилученої сутності. Більш простим завданням є визначення сутності, що виявляє осіб, які заздалегідь не мають будь-яких знань про екземпляри сутностей [3, 5].

Вирішення цього завдання пов’язано з розв’язанням анафор і кореференцій (coreference resolution) – виявлення кореферентних і анафоричних зв'язків між текстовими структурами.

Другий етап завдання – це вилучення фактів, визначення відносин між сутностями, що є більш складним і менш розробленим [7, 1].

Використання традиційних статистичних методів для отримання інформації є досить неефективним. Для цього є кілька причин.

1. Статистичні методи розцінюють документи як невпорядкований «мішок слів», що є типовим для задач пошуку інформації та завдань класифікації тексту. Але таке спрощене уявлення тексту не використовує обробку природної мови і втрачає багато знань, пов'язаних з граматикою мови, синтаксисом та семантикою.

2. У більшості випадків, розглядаючи вилучення фактів, ми маємо на увазі вилучення з речення, а не з корпусу текстів. Це засновано на ідеї розгляду факту у вигляді триплета: Subject → Predicate → Object, з предикатом, що передає дієслово, об'єктом і суб'єктом, які є іменниками. Іменник представляє учасника дії в реченні.

3. Інша причина низької ефективності використання статистичних методів під час вилучення фактів полягає в синонімії та неоднозначності мовних одиниць. Це призводить до втрати важливих даних, що містяться в документах, коли суб'єкт, об'єкт, і предикат представлені різними словами (іноді різними частинами мови) [4].

Одним із основних сучасних методів вирішення задачі вилучення фактів є використання автоматичного навчання з учителем (Supervised learning) – одного з найпоширеніших розділів машинного навчання. Тут виділяють безліч об'єктів (ситуацій) і безліч можливих відповідей (відгуків, реакцій). Існує деяка залежність між відповідями і об'єктами, але вона невідома. Відома тільки кінцева сукупність прецедентів – пар «об'єкт, відповідь», що називається навчальною вибіркою. На основі цих даних потрібно відновити залежність, тобто побудувати алгоритм, здатний для будь-якого об'єкта видати досить точну відповідь. Під учителем розуміють саму вибірку або того, хто вказав на заданих об'єктах правильні відповіді [5].

Навчання з частковим залученням учителя, напівавтоматичне навчання або часткове навчання (англ. Semi-supervised learning) – спосіб машинного навчання, різновид навчання з учителем, яке також використовує немарковані дані для тренування. Зазвичай існує невелика кількість розмічених даних і велика кількість нерозмічених даних.

Найбільш відомою системою вилучення фактів є TextRunner (the TextRunner system), яка витягує доброякісну інформацію із речень у розширюваному і загальному вигляді. Замість того, щоб вимагати точне визначення відносин з вхідного потоку даних, TextRunner вивчає відносини, класи і об'єкти з його корпусу, використовуючи стосунково-незалежну модель екстракції [2].

III.Модель формалізації смислового значення елементів факту

Нами розроблена модель, яка формалізує смислове значення елементів факту за допомогою явного визначення синтаксичних і граматичних характеристик слів у реченні. При цьому для вилучення суб’єкта та об’єкта факту описують формалізовані граматичні характеристики партиципантів речень. Такими характеристиками в англійській мові є:

- конкретний прийменник після дієслова;
- присвійний відмінок іменника чи займенника;
- розташування іменника щодо дієслова в реченні;
- будь-яка форма дієслова "to be";
- форма основного дієслова в обороті;
- використання модальних дієслів;
- наявність заперечення.

У ході роботи за допомогою програми було визначено частоту всіх видів модальних дієслів у різних текстах. Особливістю технічної документації, яка розглядається у цьому дослідженні, є наявність модальних дієслів. У дослідженні проаналізовані такі модальні дієслова, які об'єднані в смислові групи:

- перша група включає дієслова, які часто зустрічаються, такі як: ***can, may, must***;
- друга група включає дієслова, які рідко використовуються: ***should, might, could, need***;
- остання група включає тільки одне модальне дієслово, яке практично не трапляється в технічній документації – ***would***.

У ході дослідження було розглянуто, яким чином модальні дієслова впливають на стиль технічної документації і як часто вони вживаються. На основі отриманих результатів предметні змінні, подані у статті [4], були доповнені.

Для формалізації і явного уявлення засобами поверхневої структури суб'єкта і об'єкта триплета факту Subject → Predicate → Object названого реченням англійської мови, виділені і описані предметними змінними наступні кінцеві множини синтаксичних і морфологічних категорій

$$
\begin{gathered}
z^{to} \vee z^{by} \vee z^{with} \vee z^{about} \vee z^{of} \vee z^{on} \vee z^{at} \vee \\
z^{in} \vee z^{out} = 1, \\
y^{ap} \vee y^{aps} \vee y^{out} = 1, \\
x^{f} \vee x^{l} \vee x^{kos} = 1, \\
m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee \\
m^{was} \vee m^{were} \vee m^{be} \vee m^{out} = 1, \\
p^{III} \vee p^{ed} \vee p^{I} \vee p^{ing} \vee p^{II} = 1, \\
f^{can} \vee f^{may} \vee f^{must} \vee f^{should} \vee f^{might} \vee f^{could} \\
\vee f^{need} \vee f^{would} \vee f^{out} = 1, \\
n^{not} \vee n^{out} = 1,
\end{gathered}
\qquad (1)
$$

де z – предметна змінна, яка визначає синтаксичні характеристики наявності (to, by, with, about, of, on, at, in) або відсутності (out) прийменника в англійській фразі;
y – предметна змінна, яка визначає наявність (ap, aps) або відсутність (out) апострофа в кінці слова;
x – предметна змінна, яка визначає позицію іменника перед (f), після (l) основного дієслова або після непрямого доповнення (kos);
m – предметна змінна, яка визначає існування будь-якої форми дієслова "to be" (is, are, havb, hasb, hadb, was, were, be, out);
p – предметна змінна, яка визначає форму основного дієслова (III, ed, I, ing, II);
f – предметна змінна, яка визначає форму / наявність модального дієслова (can, may, must, should, might, could, need, would, out);
n – предметна змінна визначає наявність заперечення (not) або його відсутність (out).

IV. Опис проведеного експериментального дослідження

На основі дослідження 50 файлів (середнім обсягом 340 кілобайт) інструкцій користувачів мобільних телефонів виділено три структурного типу фактів, що зустрічаються в інструкціях: 1) Subject → Predicate; 2) Object → Predicate; 3) триплет факту Subject → Predicate → Object.

Проаналізувавши перший і третій типи фактів, на основі введених в (1) предметних змінних було визначено предикат, що описує граматичні категорії іменника, що називає суб'єкт дії:

$$\gamma_1^{(z,y,x,m,p,f,n)} = y^{out}(f^{can} \vee f^{out})\, n^{out}(p^{I} \vee p^{ed} \vee p^{II})\,(x^{f} m^{out} z^{out} \vee x^{l}(m^{is} \vee m^{are})\, z^{by}). \quad (2)$$

Наприклад, в такому реченні представлений перший тип факту Subject → Predicate:

1) Optical fiber cables and sfp modules that are not clean can cause signal errors frames not being received bouncing routes and permanently affected performance.

Ми можемо виділити дієслово «cause», яке відноситься до фактів причини чогось. Потім ми можемо визначити групу іменника "optical fiber cables and sfp modules" як суб'єкт факту. Граматичні та синтаксичні особливості головного іменника цієї групи відповідають рівнянню:

$$\gamma^1{}_1(z, y, x, m, p, f, n) = z^{out} \wedge y^{out} \wedge x^{f} \wedge m^{out} \wedge p^{1} \wedge f^{can} \wedge n^{out}.$$

Наприклад,

2) Sgsn change is used for a context released in the old sgsn after an inter sgsn routing area update and after an intra sgsn inter system change.

Ми можемо виділити дієслово «released», яке відноситься до фактів випуску або реалізації. Потім ми можемо визначити іменник «a context» як суб'єкт факту. Граматичні та синтаксичні особливості іменника відповідають наступній рівнянню:

$$\gamma^2{}_1(z, y, x, m, p, f, n) = z^{in} \wedge y^{out} \wedge x^{f} \wedge m^{out} \wedge p^{ed} \wedge f^{out} \wedge n^{out}.$$

У такому реченні представлений другий тип факту Subject→Predicate→Object:

3) The subscriber is not allowed to use the apn in the visitor public land mobile network vplmn.

Ми можемо виділити дієслово «use», яке відноситься до фактів використання та застосування. Потім ми можемо визначити іменник "the subscriber" як суб'єкт факту. Граматичні та синтаксичні особливості іменника відповідають рівнянню:

$$\gamma^3{}_1(z, y, x, m, p, f, n) = z^{out} \wedge y^{out} \wedge x^{f} \wedge m^{is} \wedge p^{1} \wedge f^{out} \wedge n^{not}.$$

Потім ми можемо визначити іменник "the apn" в якості об'єкта факту. Граматичні та синтаксичні особливості іменника відповідають рівнянню:

$$\gamma^4{}_1(z, y, x, m, p, f, n) = z^{out} \wedge y^{out} \wedge x^{l} \wedge m^{is} \wedge p^{1} \wedge f^{out} \wedge n^{not}.$$

Приклад

4) If there istopinformation in the ps_monitor log file the command will use the processes from the ps information.

Ми можемо виділити дієслово «use», яке відноситься до факту використання та застосування. Потім ми можемо визначити іменник "the command" як суб'єкт факту. Граматичні та синтаксичні особливості іменника відповідають наступній рівності:

$$\gamma^5{}_1(z, y, x, m, p, f, n) = z^{out} \wedge y^{out} \wedge x^{f} \wedge m^{out} \wedge p^{1} \wedge f^{out} \wedge n^{out}.$$

Потім ми можемо визначити іменник "the processes" в якості об'єкта факту. Граматичні та синтаксичні особливості іменника відповідають наступній рівнянню:

$$\gamma^6{}_1(z, y, x, m, p, f, n) = z^{out} \wedge y^{out} \wedge x^{l} \wedge m^{out} \wedge p^{1} \wedge f^{out} \wedge n^{out}.$$

Проаналізувавши другий тип факту, на основі введених в (1) предметних змінних було визначено предикат, що описує граматичні категорії іменника, що називає об'єкт дії:

$$\gamma_2^{(z,y,x,m,p,f,n)} = y^{out}(n^{out} \vee n^{not})\,(f^{out} \vee f^{can} \vee f^{may} \vee f^{must} \vee f^{need})\,(z^{out} x^{l} m^{out} (p^{I} \vee p^{II} \vee p^{ed}) \vee x^{f}(z^{out} \vee z^{by})\,(m^{is} \vee m^{are})\,(p^{III} \vee p^{ed}). \quad (3)$$

Наприклад, в такому реченні представлений перший тип факту Object→ Predicate:

7) Optical fiber cable cleaning kits such as the cletop reel type cleaner pn cletop rl a realm fieldmaster pn realm fp 6 or 3m kits pn 4144 are also good solutions do not use acetone as a cleaning solvent on the fiber optical surfaces.

Ми можемо виділити дієслово «use», яке відноситься до фактів використання та застосування. Потім ми можемо визначити іменник "acetone" як об'єкт факту. Граматичні та синтаксичні особливості іменника відповідають рівнянню:

$$\gamma^1{}_2(z, y, x, m, p, f, n) = \\ = z^{out} \wedge y^{out} \wedge x^{l} \wedge m^{out} \wedge p^{l} \wedge f^{out} \wedge n^{not}.$$

Приклад:

8) The following example shows how to use 3gdt for the default apn.

Ми можемо виділити дієслово «use», яке відноситься до фактів використання та застосування. Потім ми можемо визначити іменник "3gdt" в якості об'єкта факту. Граматичні й синтаксичні особливості іменника відповідають рівнянню:

$$\gamma^2{}_2(z, y, x, m, p, f, n) = \\ = z^{out} \wedge y^{out} \wedge x^{l} \wedge m^{out} \wedge p^{1} \wedge f^{out} \wedge n^{out}.$$

Висновок

Основним результатом нашого дослідження є розроблена логіко-лінгвістична модель вилучення фактів із тестів технічної документації англійської мови. Модель дозволяє витягувати кілька типів фактів з різних англійських речень. Програмна реалізація моделі дозволяє фахівцям автоматично отримувати ці факти з різних типів текстів. У ході дослідження експерименти показали, що використання моделі підвищило ефективність вилучення фактів з неструктурованих текстів.

Створена модель витягу фактів є лише експериментальним прототипом, у майбутньому ця модель буде частиною підсистеми, яка використовуватиметься для отримання інформації з неструктурованих англійських текстів.

Література

1. Agichtein, E., Gravano, L. Snowball: Extracting Relations from Large Plaintext Collections. In: Proceedings of the 5th ACM International Conference on Digital Libraries, 85–94, San Antonio, Texas, (2000)

2. Etzioni, O., Banko, M., Soderland, S., Weld, D. Open information extraction from the Web. In: Communications of the ACM, 68-74, (2008).

3. Fader, S., Soderland, O.: Etzioni Identifying relations for open information extraction. In: Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, 1535 – 1545 (2011)

4. Хайрова Н., Логіко-лінгвістична модель генерації фактів із текстових потоків інформаційно-корпоративної системи/ Ніна Хайрова. Наталя Шаронова, Аджит Пратап Сінгх Гаутам// International Journal Informatiom theories & application – 2015. vol. 22. № 2. – P 142-152.

5. Mooney, R. J., Bunescu R. Mining Knowledge from Text Using Information Extraction. In: Newsletter. ACM SIGKDD Explorations Newsletter - Natural language processing and text mining, vol.7, issue 1, 3–10 (2005)

6. Sint, R. , Schaffert, S., Stroka, S., Ferstl, R. Combining Unstructured, Fully Structured and Semi-Structured Information in Semantic Wikis. In: Proceedings of the 4th Semantic Wiki WorkShop (SemWiki) at the 6th European Semantic Web Conference, ESWC (2009)

7. Yahya, M., Whang, E. S., Gupta R., Halevy A. ReNoun: Fact Extraction for Nominal Attributes. In: Proceedings of the Conference on Empirical Methods in Natural Language (EMNLP), 325 – 335 (2014)

8. Вопросно-ответные системы: развитие и перспективы: ежемесячный научно-технический сборник/ В. А. Лапшин. – М.: ВИНИТИ, 2012. – 32 с.

9. Извлечение объектов и фактов из текстов в Яндексе: Лекция для Малого ШАДа // https://habrahabr.ru/company/yandex/blog/205198/, 01.11.2016.