2. Pakhnenko V. V. Differential equations. Text-book [Текст] : підр. / V.V. Pakhnenko, Ye. O. Shkvar ; NAU. —К. : NAU, 2002. – 104 p.
3. Karupu O. W. Elements of theory of functions of complex variable. Lectures.[Текст] : навч. посібник / O.W. Karupu ; NAU. – K.: NAU, 2002. – 68 p.
4. Karupu O.W. Operational calculus. Lectures.[Текст] : навч. посібник / O.W. Karupu ; NAU. – K.: NAU, 2003. – 52 p.
5. Oleshko T.A., Pakhnenko V.V., Trofymenko V.I. Elements of mathematical statistics. Methodical guide [Текст] : навч. посібник / T.A. Oleshko, V.V.Pakhnenko, V.I. Trofymenko ; NAU. – K.: NAU, 2003. – 72 p.
6. Higher mathematics. Part 1: Manual [Текст] : навч. посібник / V.P. Denisiuk, L.I. Grishina, O.V. Karupu, T.A. Oleshko, V.V. Pakhnenko, V.K. Repeta ; NAU. – K.: NAU, 2006. – 268 p.
7. Higher mathematics. Part 3: Manual [Текст] : навч. посібник / V.P. Denisiuk, L.I. Grishina, O.V. Karupu, T.A. Oleshko, V.V. Pakhnenko, V.K. Repeta.; NAU. – K.: NAU, 2006. – 232 p.
8. Higher mathematics. Part 2: Manual [Текст] : навч. посібник / V.P. Denisiuk, V.G. Demydko, V.K. Repeta; NAU.– K.: NAU, 2009. – 248 p.
9. Higher mathematics. Part 4: Manual[Текст] : навч. посібник / V.P. Denisiuk, L.I. Grishina, O.V. Karupu , T.A. Oleshko, V.V. Pakhnenko, V.K. Repeta ; NAU. – K.: NAU 2012.-250p.(сдана в печать).

Виконано аналіз сучасних технологій обробки природномовних текстів. Показано, що в системах контролю знань необхідним є використання систем типу «текст-питання-відповідь», та запропоновано концептуальні елементи створення таких систем

Ключові слова: обробка природномовних текстів, формування тестів

Выполнен анализ современных технологий обработки естественноязычных текстов. Показано, что в системах контроля знаний необходимым является использование систем типа "текст-вопрос-ответ", и предложены концептуальные элементы создания таких систем

Ключевые слова: обработка естественноязычных текстов, формирование тестов

The analysis of modern technologies for natural text processing is executed. It is shown that in the control knowledge systems it is necessary the use of the systems as a "text-question-answering", and the conceptual elements of such systems creation are offered

Key words: natural language processing, test creating

УДК 004.91

АВТОМАТИЧНЕ ФОРМУВАННЯ ТЕСТІВ ЯК ОДНА ІЗ ЗАДАЧ ОБРОБКИ ПРИРОДНОМОВНИХ ТЕКСТІВ

А.В. Мороз
Аспірант
Кафедра управління проектами
Черкаський державний технологічний університет
бул. Шевченка, 460, м. Черкаси, Україна, 18006
Контактний тел.: 063-848-41-24
E-mail: morozanvi@gmail.com

1. Вступ

Технології NLP (natural language processing) на сьогодні є найважливішим інструментарієм створення штучного інтелекту. Саме NLP дозволяє розв'язувати задачі розпізнавання та синтезу мови, машинного перекладу, «видобування» інформації з текстів, інформаційного пошуку, аналізу висловлювань та створення систем типу «питання-відповідь» (QAS-question answering systems). Лавиноподібний ріст інформації в світі та обмежені можливості людського мозку щодо її обробки та інтерпретації роблять NLP однією із найактуальніших інтелектуальних технологій. В основі переважної більшості методів-елементів NLP лежать моделі математичної лінгвістики та технології математичної статистики. Останні дозволяють здійснювати частотний аналіз інформаційного контенту та будувати його ієрархію. Розбиття на елементи, аналіз та синтез текстів виконується із застосуванням математичної лінгвістики. Далі розглянемо основні напрямки аналізу текстової інформації та створення QAS-систем як основних елементів, які необхідно використовувати у навчальному процесі.

2. Аналіз парадигм NLP в Україні та світі

Аналіз наукових літературних джерел свідчить про присутність декількох найбільш потужних наукових шкіл у вказаному напрямку. В Україні значна кількість релевантних досліджень виконується під керівництвом академіка НАНУ О.В. Палагіна [1, 2], зокрема:

– розробка концептуально-семантичних відношень між категоріями, концептами та примітивами;

– пропозиція застосування комбінованого підходу до розпізнавання синтактико-семантичних відношень, потрійного аналізу неоднозначностей та алгоритму співвідношення анафоричних зв'язків у природномовних текстах;

– розробка абстрактної моделі мовно-онтологічної картини світу, що базується на лексикографічній базі даних природної мови та семантико-синтаксичних відношеннях між мовними одиницями;

– підхід до адаптації відомих описів концептуальних графів у застосуванні до візуалізації формалізованого представлення природномовних висловів для флективних мов;

– аналіз складу та архітектури онтологій як основи семантичного Weby, в якому особлива увага приділена поняттю баз знань як центрального компонента інтелектуальної інформаційної системи та Інтернету;

– класифікація тестових документів;

– розробка методологічного, онтологічного і логічного аспектів проектування знанняорієнтованих інформаційних систем, функціонування яких опирається на автоматизацію процесів «видобування» і формалізації змісту природномовних об'єктів з наступною обробкою формалізованого представлення цього змісту логіко-семантичними методами з орієнтацією на конкретну предметну область;

– застосовність байєсівських процедур для розпізнавання текстів;

Значний обсяг досліджень із вказаної тематики проводиться в Лабораторії аналізу інформаційних ресурсів Науково-дослідного обчислювального центру МДУ ім. М.В. Ломоносова. Вченими лабораторії та їх колегами розроблені методи [3, 4]:

– автоматизованого аналізу текстової інформації на основі поєднання лінгвістичного аналізу та застосування методів машинного навчання;

– формування лінгвістичних онтологій великого розміру;

– інтеграції різнорідних структурованих і слабко структурованих даних для корпоративних та Інтернет-систем;

– класифікація документів, реферування, кластеризація пулу новин, пошук відповідей на питання.

За кордоном дослідження проблематики є значно ширшим. Про це свідчить хоча б той факт, що на запит «Natural language processing» Google видає 14900000 результатів, а на запит «Обработка естественного языка» – 341000. На сайті англомовної Вікіпедії вказано [5], що до задач NLP належать: автоматичне резюмування, визначення кореференції, аналіз бесід, машинний переклад, морфологічна сегментація, розпізнавання ідентифікованих сутностей, генерація природномовних текстів із інформації з баз даних, розуміння природної мови, оптичне розпізнавання текстів, розпізнавання частин мови, побудова графічних структур речень, генерація відповідей на питання, визначення відношень між елементами тексту, розпізнавання емоцій в тексті, сегментація мови і тексту, сегментація слів, вибір слова із синонімічного ряду, інформаційний пошук, «видобування» інформації.

Порівнюючи публікації російських та українських вчених із закордонними, можна зробити висновок про домінуючу теоретичну складову у вітчизняних дослідженнях та практичну спрямованість результатів закордонних вчених [6]. Проблема обробки природної мови, природномовних текстів є слабко структурованою, важко формалізованою, тому очевидним є необхідність створення систем простого локального використання із відкритою архітектурою, що дозволить здійснювати їх модифікацію та вдосконалення. Водночас, зауважимо, що в основі таких систем лежать словники, тезауруси та онтології, створення яких є трудомістким процесом. Без обмеження загальності розглянемо декілька поширених типових NLP-систем типу «питання-відповідь».

3. QAS-системи в навчальному процесі

Для працівників навчальних закладів особливий інтерес становлять системи, що дозволять об'єктивізувати процес взаємодії викладача та осіб, що навчаються. До таких систем належать автоматизовані системи навчання і контролю знань та системи типу «питання-відповідь», елементи яких активно використовуються в Інтернеті. Зауважимо, що на запит «Question answering system» Google видав 14-700000 посилань, а на запит «Вопросно-ответные системы» всього 47600 посилань, що свідчить про недостатню увагу до таких систем вітчизняних вчених. У той же час очевидно, що релевантність та пертинентність відповідей на запити залишається на низькому рівні. І пошук потрібної інформації з вже виданої пошуковими системами є складним і тривалим процесом.

Розглянемо деякі найбільш поширені парадигми NLP, що мають програмну реалізацію. Розроблена спеціалістами система Apache UIMA (архітектура управління неструктурованою інформацією) [7] є програмною інфраструктурою, призначеною для аналізу значних інформаційних масивів та «видобування» з них знань. В цій системі реалізовано модульний підхід до аналізу текстів, який, в принципі, може вважатись класичним. Зокрема, алгоритм аналізу тексту має такі кроки: визначення мови тексту; встановлення границь речень; пошук іменованих входжень (імена,назви тощо). Використання Apache UIMA приводить до збагачення текстової інформації шляхом використання онтологій та словників, що дозволяє здійснювати тлумачення термінів та понять, а також використовувати синоніми та різні ідентифікатори одного і того ж елемента тексту.

Відомо [8], що переважна більшість підходів до аналізу текстів базується на синтаксичному аналізі та статистичній обробці інформації, що дозволяє визначати семантичну близькість слів у тексті.

Відповідні системи раціонально використовувати для аналізу значних обсягів однотипної інформації, для побудови QAS-систем вони є мало застосовними. Для одержання найбільш точних відповідей на питання в Інтернеті рекомендується використання когнітивного словника WordNet [9, 10], базовими одиницями якого є синсети – структури, що визначають значення слів. Використання WordNet разом з іншими словниками дозволяє розв'язувати складніші задачі обробки текстів, оскільки здійснюється більш об'ємне бачення проблемної ситуації чи предметної області.

4. Елементи концепції інтегрованих систем автоматичного контролю знань

Розробка перших мов програмування дала поштовх створенню експертних систем, які з часом зайняли важливу нішу в процесах автоматизації діяльності людини. На сьогодні розвиток мережі Інтернет та дистанційної освіти є причиною виведення на передній план задач розробки автоматизованих систем навчання і контролю знань. Проблема підвищення ефективності цього процесу є настільки багатогранною, що десятки тисяч публікацій та сотні розроблених систем залишають враження мінімальної дослідженості предметної області. Майже в кожному навчальному закладі використовуються відповідні системи, алгоритми розробки яких та функціонування зводиться до декількох простих кроків: викладач формує базу питань та відповідей; особа, що навчається, відповідає на певну сукупність таких питань; за деяким алгоритмом визначається оцінка. Недоліком такого процесу є значний суб'єктивізм при формуванні питань та витрати часу викладача на створення бази даних.

Одним із способів оптимізації процесів контролю знань пропонуємо застосування інтелектуальних елементів Text Mining [11] як технології «видобування» знань із текстів. Відомо біля десяти напрямків – складових Text Mining. Для нашої задачі релевантним є напрямок, пов'язаний із створенням систем типу «питання-відповідь» [12]. Системи QAS почали створюватись із 60-х років минулого століття і були природномовними оболонками для експертних систем. Сучасні QAS системи призначені для пошуку відповідей на питання в документах з використанням технологій обробки природних мов.

Водночас за межами реалізації ідей та принципів функціонування QAS систем залишились проблеми формування і можливих відповідей. Виконуючи аналіз ефективності автоматизованих систем контролю знань, робимо висновок про доцільність більш глибокої і змістовної автоматизації такого процесу. Зокрема, пропонуємо створення систем типу «текст-питання-відповіді» (TQAS – text-question-answering system). Така система працюватиме за алгоритмом:

1. Обробка (T) навчального електронного тексту (препроцесінг).
2. Формування (Q) множини питань до тексту.
3. Формування (A) множини відповідей до тексту.
4. Проведення контролю знань та оцінювання.

Зауважимо, що виконання усіх чотирьох кроків повинно здійснюватись автоматично без участі людини. Очевидно, що у повному обсязі їх реалізація, тобто обробка текстів будь-якої тематики, складності та формування різнотипних питань, надто складна задача, враховуючи різноманіття природної мови. Певна складність додається у випадку обробки технічних текстів з формулами, графіками та таблицями. Але, як відомо, працююча складна система є результатом працюючих простих систем. Тому виконаємо аналіз аспектів, які супроводжують реалізацію наведених вище трьох кроків.

Обробка навчального тексту як задача складатиметься із декількох компонент

$$T = < S_l, C, R, S_s >,$$

де S_l – словник, який включатиме як загальну, так і спеціальну лексику; C – структурні схеми, у відповідності до яких далі будуть формуватись питання; R – алгоритм редукції складних речень до таких, що відповідатимуть схемам C; S_s – алгоритм визначення статистичних параметрів елементів тексту.

Формування множини питань до тексту як процес має такі кортежі:

$$Q = < Q_f, G, K >,$$

де Q_f – структурні форми питань; G – відображення структурних схем C на структурні форми Q_f; K – класифікація типів питань.

Відносно новою є задача формування можливих відповідей на питання. Відомо, що контроль знань у залежності від типів питань може бути як відкритим, так і закритим. У першому випадку питання не мають варіантів можливих відповідей і особа, що навчається, пропонує свою відповідь на питання. У другому випадку варіанти можливих відповідей задані й особа, що навчається, обирає один або декілька з них. Якщо контроль знань відбувається на основі питань із відомими варіантами відповідей, то виникає задача їх раціонального формування. Очевидно, що тоді одна або декілька відповідей є правильними. Декілька інших відповідей у певній метриці мають бути близькими до правильних відповідей, інколи одна або більше відповідей є максимально далекими від правильної відповіді або взагалі не мають відношення до даної предметної області. Маємо

$$A = < N, A_r, A_n > = < N, h(T,Q), f(N, A_r, d, q) >,$$

де N – процедура визначення кількості потенційних відповідей; A_r – процедура визначення правильної відповіді, $A_r = h(T,Q)$; A_n – процедура формування неправильних відповідей, $A_n = f(N, A_r, d, q)$, d – метрика, q – її порогове значення.

Формування питань до тексту та потенційних відповідей дозволить значно спростити процес розробки тестових питань, скоротити час на їх електронну обробку та урізноманітнити склад множини питань для кожного навчального курсу та предметної області. Зауважимо, що на перших етапах експлуатації TQAS системи потребуватимуть настройки викладачем або особою, що приймає рішення.

5. Висновки

Вважаючи за потрібне подальший розвиток QAS-систем, ми пропонуємо, щоб такі системи дозволяли не лише здійснювати ефективний пошук відповіді на питання, але і для будь-якого наукового чи навчального тексту виконувати його статистичний аналіз, виділяти концептуальні структури, автоматично формувати питання чи генерувати відповіді на них як правильні, так і неправильні. Такі системи названо TQAS-системами (text question answering system).

Подальші дослідження полягатимуть у конструктивній реалізації запропонованої парадигми, актуальність чого є безсумнівною, особливо у зв'язку із розвитком Інтернету та дистанційного навчання. Частково її елементи зможуть бути використаними при аналізі природномовних текстів як Інтернет-контенту.

Література

1. Палагин, А. В. Знание-ориентированные информационные системы с обработкой естественно-языковых объектов: основы методологии и архитектурно-структурная организация [Текст] / А. В. Палагин, С. Л. Кривой, Н. Г. Петренко // Управляющие системы и машины. – 2009. – № 3. – С. 42-55.
2. Палагин, А. В. Концептуальные графы и семантические сети в системах обработки естественно-языковой информации [Текст] / А. В. Палагин, С. Л. Кривой, Н. Г. Петренко // Математичні машини і системи. – 2009. – № 3. – С. 67-79.
3. Лукашевич, Н. В. Описание понятийной системы русского языка в виде тезаурусноорганизованной семантической сети : труды междунар. конф. «Знания-Диалог-Решение» / Н. В. Лукашевич, Б. В. Добров. – Спб. – 2001. – Т. 2. – С. 438-444.
4. Добров, Б.В. Онтологии для автоматической об работки текстов: описание понятий и лексических значений : труды междунар. конф. «Компьютерная лингвистика и интеллектуальные технологии», 31 мая - 4 июня 2006 г., Бекасово / Н. В. Лукашевич, Б. В. Добров. – М.: Изд-во РГГУ, 2006. – С. 138-142.
5. Natural language processing [електронний ресурс]. – Режим доступу : http://en.wikipedia.org/wiki/Natural_language_processing.
6. Bates, M. Models of natural language understanding : proceedings of the National Academy of Sciences of the United States of America / M. Bates. – 1995. – Vol. 92, No. 22. P. 9977-9982.
7. Машинная обработка естественных языков: Apache UIMA. – Режим доступу : http://habrahabr.ru/blogs/sw/56461/.
8. Вопросно-ответные системы и обработка знаний – нестатистический подход. – Режим доступу : http://rykov-qa2.narod.ru/.
9. Сухоногов, А. М. Разработка русского WordNet : труды 6-ой Всероссийской научн. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», Пущино / А. М. Сухоногов, С. А. Яблонский. – Пущино, Россия, 2004. – С. 234-237.
10. WordNet – A lexical database for English. – Режим доступу: http://wordnet.princeton.edu/.
11. Feldman, R. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data / R. Feldman, J. Sanger. – Cambridge: Cambridge University Press, 2006.
12. Hirschman, L. Natural Language Question Answering. The View from Here. Natural Language Engineering, 7:4 / L. Hirschman, R. Gaizauskas. – Cambridge University Press, 2001. – P. 275-300.