Вибір текстів за тематикою для побудови лінгвістичної моделі мови в системах розпізнавання злитого мовлення

Пилипенко В.В.

Міжнародний науково-навчальний центр
інформаційних технологій та систем, Київ
valery_pylypenko@mail.ru

This paper presents a automatic continuous speech recognition system with statistical language model. To choose documents for language model it is proposed to download large amount of texts from Internet and to use the modified K-means cluster algorithm to select useful documents. First experts form small document corpus with known subject. Agglomeration algorithm allows to build begin clusters for K-means algorithm. The threshold to reject far documents is introduces.

1. Вступ

Лінгвістична модель мови є необхідним етапом у побудові автоматичної системи розпізнавання злитого мовлення (АСРЗМ), оскільки будь-яка сучасна АСРЗМ має бути настроєна на словник та тему повідомлення, яке розпізнається. Сучасні алгоритми АСРЗМ використовують статистичні методи для побудови лінгвістичної моделі за текстами, які найбільш схожі до розпізнаваних мовленнєвих повідомлень. Загалом, можуть бути задіяні будь-які тексти, але необхідно вибрати саме ті, що містять корисну інформацію за схожою тематикою.

2. Формування корпусу текстів за допомогою мережі Internet

Internet містить безліч інформації, яку можна використати для формування корпусів текстів. Однак слід задати алгоритм знаходження корисної інформації.

Розглядалися три стратегії пошуку інформації:

1. Використання сайтів, що містять тексти за необхідною тематикою у достатньому обсязі. Наприклад, сайт Верховної Ради України містить усі стенограми засідань депутатів і в [1] ці тексти використовувались для побудови лінгвістичної моделі мови.
2. Стратегія, яка передбачає вказівку експерта на сайти, які можуть містити корисну інформацію.
3. Задається словник системи розпізнавання. Для кожного слова формується запит до загальної пошукової системи (наприклад, Goggle, Yandex або інші). Пошукова система формує список адрес сайтів, де вказане необхідне слово. Після цього завантажуються усі тексти, що містять вказане слово.

Досвід завантаження для першої стратегії показує, що достатньо 100 МБ текстів для побудови надійної лінгвістичної моделі. Друга стратегія приводить до значних обсягів корпусу (більше 1000 МБ), який містить текстовий матеріал, що не відноситься до тематики АСРМ. Третя стратегія дозволяє завантажити цільовий текстовий матеріал середнього обсягу (близько 500 МБ), але під час використання такого методу можуть бути порушені статистичні залежності природної мови.

Пропонується використати наступну стратегію завантаження текстового матеріалу:

1. За допомогою експертів сформувати невеликий корпус текстів із відомою заданою тематикою;
2. Завантажувати всі доступні сайти Internet із текстами досліджуваної мови;
3. Відібрати корисні тексти за тематикою за допомогою модифікованих алгоритмів кластеризації документів.

3. Метрика TF-IDF

Тематика тексту визначається словами, які використовуються для передачі смислу, але різні слова мають бути враховані у різному ступені. Простіший погляд на частоту вживання слів звичайно модифікується з метою зменшення ролі загальнопоширених слів.

Метрика TF-IDF [2] обчислюється для кожного слова зі словника документу

$$TFIDF_i = \frac{n_i}{\sum_k n_k} * \ln\left(\frac{|D|}{|t_i \in D_i|}\right)$$

де перший множник TF_i є частотність i-го слова у даному документі, а другий множник IDF_i враховує інверсію частотності появи i-го слова у всіх розглядуваних документах.

4. Формування вектора ключів

Для обчислення відстані між документами потрібно задати словник, для якого формується метрика TF-

IDF. Такий словник називається вектором ключів. Вичерпну інформацію про відстань несе увесь словник для всіх розглядуваних документів, але обсяг обчислень стає дуже великим. Слід відзначити, що більша частина усього словника не є корисною для обчислень метрики — загальнопоширені та рідковживані слова не несуть ніякої корисної інформації.

Для формування вектора ключів обчислюються коефіцієнти TF-IDF для всіх розглядуваних документів, а всі слова документу впорядковуються за цими коефіцієнтами. Перші кілька десятків слів (приблизно 100) додаються до словника ключів. Потім видаляються найбільш частотні (з частотою появи більше ніж 0.001) та найменш частотні слова (з частотою появи менш ніж 0.00001). Достатньо хорошим вважається вектор довжиною від 1000 до 10000 слів.

5. Початкова агломераційна кластеризація документів

Для сформованого вектора ключів можна обчислювати відстань між документами різними способами, наприклад, за допомогою евклідової відстані, манхеттенової відстані, відстані Чебишева та ін.

Найпоширеніша косинусна міра дозволяє обчислювати відстань між документами D_i та D_l:

$$sim(D_i, D_l) = \frac{\sum_j w_{ij} w_{lj}}{\sqrt{\sum_j w_{ij}^2} \sqrt{\sum_j w_{lj}^2}}$$

де w_{ij} – вага j-го слова в і-м документі за метрикою TF-IDF.

Спочатку обчислюється матриця відстаней між усіма документами. Потім агломераційний алгоритм кластеризації об'єднує два документи з максимальною відстанню та переобчислює матрицю відстаней. Алгоритм зупиняється при досяганні деякого порогу в відстанях або при наявності необхідної кількості кластерів.

З метою запобігання занадто великих кластерів використовують поправочний коефіцієнт при обчисленні косинусної міри, залежний від обсягу кластера, наприклад, квадратний корінь або логарифм від кількості слів у кластері.

Агломераційний алгоритм потребує N^2 пам'яті та $O(N^2)$ обчислень, де N — кількість документів. Цей алгоритм використовується для кластеризації невеликої кількості документів, наприклад, для обчислення початкових кластерів для алгоритму К-середніх.

Таким чином формується декілька кластерів документів, які можна розглядати як однорідні за тематикою.

6. Модифікація алгоритму К-середніх для відбору корисних текстів із великих корпусів документів

Алгоритм К-середніх потребує початкового завдання К центрів кластерів. Для цього використовується результат роботи агломераційного алгоритму кластеризації на відносно невеликому обсязі документів.

Алгоритм сканує всі документи з метою виявлення, до якого кластера належить поточний документ. Уводиться поріг відстані, який дозволяє не враховувати в подальшому розгляді документи з далекою відстанню від усіх центрів кластерів. Після сканування обчислюються нові центри кластерів і виконується декілька ітерацій алгоритму.

Після закінчення роботи алгоритму кластери містять тільки корисні документи за заданою тематикою.

7. Подальші дослідження

Відібрані тексти використовуються для побудови статистичної лінгвістичної моделі в системі розпізнавання злитого мовлення, що покращує точність розпізнавання.

Крім алгоритму К-середніх можуть використовуватися більш сучасні алгоритми кластеризації, наприклад CLOPE [3,4].

8. Література

[1] Пилипенко В., Робейко В. Експериментальна система автоматизованого стенографування українського мовлення. // Обробка сигналів і зображень та розпізнавання образів: Дев'ята Всеукраїнська міжнародна конференція. – Київ, 2008. –С. 73-76.

[2] http://ru.wikipedia.org/wiki/TF-IDF

[3] Yang, Y., Guan, H., You. J. CLOPE: A fast and Effective Clustering Algorithm for Transactional Data In Proc. of SIGKDD'02, July 23-26, 2002, Edmonton, Alberta, Canada.

[4] Wang, K., Xu, C.. Liu, B. Clustering transactions using large items. In Proc. CIKM'99, Kansas, Missouri, 1999.