УДК 004.415 (045)

Бадьоріна Л.М.

КІЛЬКІСНА ОЦІНКА ЗМІСТОВОЇ БЛИЗЬКОСТІ ДОВІЛЬНИХ ТА ЕТАЛОННИХ ТЕКСТОВИХ ВИЗНАЧЕНЬ У КОМП'ЮТЕРНИХ СИСТЕМАХ ТЕСТУВАННЯ ЗНАНЬ

Інститут комп'ютерних технологій
Національного авіаційного університету

Запропонована методика кількісного оцінювання відповідності текстових визначень на завдання відкритого типу у тестувальних системах.

Вступ

Інтеграційні процеси, впровадження засобів телекомунікації, комп'ютерізація людської діяльності становить безліч нових завдань в науковій області, яка знаходиться між напрямками комп'ютерних технологій і лінгвістикою[1]. Розвиток сучасних інформаційних технологій в освітній сфері створює необхідність автоматизованого контролю знань студентів. Велике значення для автоматизованих систем освітнього призначення мають моделі оцінки відповідей не у вигляді обраних варіантів, а у виді вільного тексту довільної довжини з урахуванням слів-синонімів. Впровадження прогресивних форм навчання створює необхідність переходу комп'ютерного тестування знань студентів. Оцінювання знань студентів можливе лише шляхом порівняльного аналізу тексту відповіді з еталонним текстом та визначення їх релевантності. Теоретико - множинна модель на основі синонімії термінів предметної області дозволяє встановити відповідність між еталонним та фактичним визначенням, які представлені у виді тексту довільної довжини з використанням слів-синонімів [3]. Оцінка правильності текстової відповіді заснована на методі абсолютного збігу відповіді з одним з еталонів. Оскільки визначення терміна формується через систему базових понять (термів), кожне з яких має своє визначення, пропонується для обчислення показника релевантності відповідей на завдання відкритого типу використати кількісні показники синонімії термів предметної області.

Для оцінки релевантності еталонного визначення і відповіді необхідно: встановити відповідність між термами еталонного визначення і відповіді шляхом їх нормалізації.

Постановка задачі

Сучасні наукові дослідження лінгвістів [2], надають нам ряд підходів, що дозволяють формалізувати представлення кожної лінгвістичної одиниці в зручному для машинної обробки виді. Слова мови можна представити у виді конкатенації окремих фрагментів (основа слова, префікси, суфікси і т.д.). Використовуючи морфемний аналіз слова, тобто виділяючи всі значущі частини слова, які є в даному слові. Синтаксичні форми одного і того слова – представлені однією базовою формою – термом [4].

Мова предметної області, як сукупність лінгвістичних одиниць являє собою множину термів тематичного словника даної предметної області: $T = \{ t_i \}$.

Тезаурус предметної області складається з множини термінів, кожен термін колекції в загальному випадку може мати не одне формулювання. Це зв'язано з тим, що той самий навчальний матеріал по конкретній предметній області викладений одночасно в декількох підручниках різними авторами, що дають конкретним термінам колекції свої формулювання, а тому має експертне наповнення тезаурусів як своєрідних баз знань. Кожен експерт дає своє формулювання того або іншого терміна, що у такий спосіб формує свій опис входжень термів T в формулювання терміна. Кожен термін виражений набором значень місць термів

(порядкових номерів) в формулюванні конкретного терміна для даного експерта. Сукупність таких наборів являє собою багатомірний простір термів-термінів, кваліфікованих по числу експертів. Для рішення цієї задачі необхідно: установити взаємно однозначну відповідність між термами еталонного визначення і відповіді шляхом їхньої нормалізації.

Обробка текстових визначень з урахуванням синонімії понять

Припустимо, що висловлення розглядаються як сукупність термів. Таким чином, еталонне визначення розглядається як сукупність базових термів, а відповідь, як сукупність термів t, для кожного з яких потрібно знайти відповідний базовий терм e. Пошук відповідності базового терму і терму відповіді припускає визначення функції $e = \varphi(t)$ й обчислення величини синонімічної відповідності $k = \theta(e,t)$. Таким чином, пара $< e,k >$ дозволить характеризувати терм t стосовно терму-эталона e. Це означає нормалізацію термів відповіді до базових термів [5].

Нехай A - множина термів еталонного визначення, B - множина термів відповіді.

Тоді опис еталонного визначення і відповіді має вигляд:

$$A = \{e_1, e_2, ..., e_i, 1 \le i \le N\},$$
$$B = \{t_1, t_2, ..., t_i, 1 \le i \le M\}$$

де N - кількість термів еталонного визначення; M - кількість термів відповіді.

Для обчислення відповідності між термами еталонного визначення і відповіді необхідно характеризувати терми t стосовно терм-еталонів e.З цією метою пропонується провести нормалізацію термів відповіді до термів еталонного визначення.

Результатом нормалізації є одне з наступних співвідношень між множинами A і B:

1. $A = B$ - відповідь збігається з еталонним визначенням.

2. $A \subset B$ - відповідь містить всі терми з еталонного визначення та надлишкові терми.

3. $B \subset A$ - відповідь частково відповідає еталонному визначенню, відповіді відсутні деякі базові терми.

4. $A \cap B \neq \varnothing$ - відповідь має однакові терми з еталонним визначенням..

5. $A \cap B = \varnothing$ - відповідь не відповідає еталонному визначенню.

В процесі нормалізації між окремими термами еталонного визначення і відповіді можуть виникнути наступні ситуації:

1. *Одному терму еталонного визначення відповідає тільки одних базовий терм відповіді.*

Це можна представити як биєктивне відображення між множинами A і B:

$$\alpha : B \to A, a_i = \alpha(b_j), a_i \in A, b_j \in B.$$

В цьому випадку між термами еталонного визначення і відповіді установлений взаємно однозначний зв'язок. До такого виду повинні бути приведені всі відносини між термами еталонного визначення і відповіді.

2.*Одному терму еталонного визначення відповідає декілька різних термів відповіді.*

В цьому випадку існують пересічні множини $\{a_i, b_j\} \cap \{a_i, b_b\} \neq \varnothing$. Кожне з цих множин характеризується функцією θ, що позначає показник синонімії термів $a_i, b_j : k_m = \theta(a_i, b_j)$.

Для досягнення поставленої мети необхідно видалити з розгляду одну з пересічних множин за наступним правилом:

Якщо $\theta(a_i, b_j) \succ \theta(a_i, b_b)$, то видалити множину $\{a_i, b_b\}$, тому що терм b_j є найбільш близьким синонімом для терму a_i, і використовувати для подальшої обробки показник k_h, що характеризує числову величину синонімії терму c_i і b_j.

Якщо $\theta(a_i, b_j) = \theta(a_i, b_b)$, то видалити кожну з множин $\{a_i, b_j\}$ або $\{a_i, b_b\}$,

тому що терми b_j і b_b є **однаково** близькими синонімами для терму a_i, і використовувати для подальшої обробки показник k, що характеризує числову величину синонімії терму c_i і залишкові терми фактичного визначення.

Якщо $\theta(a_i,b_j) \prec \theta(a_i,b_b)$, то видалити множину $\{a_i,b_j\}$, тому що терм b_b є найбільше близьким синонімом для терму a_i, і використовувати для подальшої обробки показник k_a, що характеризує числову величину синонімії терму c_i, b_b.

3. *Деяким різним термам еталонного визначення відповідає той самий терм фактичного визначення.*

В цьому випадку встановлений взаємно однозначний зв'язок. До такого виду повинні бути приведені всі відносини між термами еталонного визначення і відповіді.

4. *Одному терму еталонного визначення відповідає декілька різних термів відповіді.*

В цьому випадку існують пересічні множини $\{a_i,b_j\} \cap \{a_i,b_b\} \neq \varnothing$. Кожна з цих множин характеризується функцією θ, що позначає показник синонімії термів $a_i,b_j : k_m = \theta(a_i,b_j)$.

Для досягнення поставленої мети необхідно видалити з розгляду одну з пересічних множин за наступним правилом:

Якщо $\theta(a_i,b_j) \succ \theta(a_i,b_b)$,

то видалити множину $\{a_i,b_b\}$, тому що терм b_j є найбільше близьким синонімом для терму a_i, і використовувати для подальшої обробки показник k_h, що характеризує числову величину синонімії терму c_i і b_j.

Якщо $\theta(a_i,b_j) = \theta(a_i,b_b)$, то видалити кожну з множин $\{a_i,b_j\}$ або $\{a_i,b_b\}$, тому що терми b_j і b_b є однаково близькими синонімами для терму a_i, і використовувати для подальшої обробки показник k, що характеризує числову величину синонімії терму c_i і ті, що залишились терми фактичного визначення.

Якщо $\theta(a_i,b_j) \prec \theta(a_i,b_b)$, то видалити множину $\{a_i,b_j\}$, тому що терм b_b є найбільш близьким синонімом для терму a_i, і використовувати для подальшої обробки показник k_a, що характеризує числову величину синонімії терму c_i, b_b.

5. *Деяким різним термам еталонного визначення відповідає один терм фактичного визначення.*

В цьому випадку існують пересічні множини $\{a_a,b_j\} \cap \{a_i,b_j\} \neq \varnothing$. Для досягнення поставленої мети необхідно видалити з розгляду одне з пересічних множин за наступним правилом:

Якщо $\theta(a_a,b_j) \succ \theta(a_i,b_j)$, то варто видалити множину $\{a_i,b_j\}$ і використовувати для подальшої обробки показник k_h, що характеризує числову величину синонімії терму c_a і b_j.

Якщо $\theta(a_a,b_j) = \theta(a_i,b_j)$, то варто видалити кожну з множин $\{a_a,b_j\}$ або $\{a_i,b_j\}$, і використовувати для подальшої обробки показник k, що характеризує числову величину синонімії терму c_i і що остались терму фактичного визначення.

Якщо $\theta(a_a,b_j) \prec \theta(a_i,b_j)$, то видалити множину $\{a_a,b_j\}$ і використовувати для подальшої обробки показник k_a, що характеризує числову величину синонімії терму c_i і b_j.

6. *Деякі різні терми еталонного визначення мають кілька загальних синонімів.*

В цьому випадку для досягнення поставленої мети необхідно:

1. Серед пересічних множин вибрати **множину** $\{a_i,b_j\}$, яка характеризується **найвищим** числовим показником $k_h = \theta\{a_i,b_j\}$;

2. Видалити з подальшого розгляду всі інші множини, у яких присутні обрані елементи a_i і b_j.

3. Серед інших множин продовжувати вибір і видалення множин з максимальним показником $k_a = \theta\{a_b, b_a\}$ по цьому ж правилу доти, поки всі пересічні множини не будуть вичерпані.

У результаті всі пересічні множини виключено. Взаємно однозначну відповідність між значимими термами еталонного і фактичного визначень визначено. Синонімічна відповідність термів враховує сукупність не пересічних множин, що мА ють найбільше значення показника $k = 1$ або $k = 0{,}8$ (множини менше цих значень не враховуються).

Висновки

При визначенні показника k враховується сукупність пересічних множин, які мають найбільше значення із діапазону $0{,}8 \le k \le 1$. За допомогою методики обробки текстових визначень та синонімічного відношення термів еталонного визначення і відповіді можливо обчислити кількісний показник релевантності і застосувати практично в тестувальній системі.

Список літератури

1. *Литвиненко А.Е., Шевченко А.В.* Проблемы моделирования и алгоритмизации в системах сравнительного анализа электронных текстов // I Міжнар. конф. "Математичне та імітаційне моделювання систем МОДС'2006" (Київ, 2006): Тези доп. – К.: ІПММС НАН України, 2006. – С. 107 – 108.

2. *Широков В.А.* Ініформаційна теорія лексикографічних систем, – К.: Довіра, 1998. – 331 с.

3. *Широков В.А.* Семантичні стани мовних одиниць та їх застосування в когнітивній лексикографії // Мовознавство, 2005, № 3-4. – С.1 – 5.

4. *Цаленко М.Ш.* Моделирование семантики в базах данных. – М.: Наука, 1989. – 288 с.

5. *Бадьоріна Л.М.* Метод оцінювання довільних відповідей у комп'ютерних системах тестування знань // Математичні машини і системи. – К.: ІПММС НАН України, 2006. – №4. – С. 138 – 144.