Микола Сажок

Проблеми побудови корпусів для лінгвістичної моделі мовлення

Жукин’2010

Основна ідея лінгвістичної моделі

Оцінювання імовірності речення із l слів $W_{0,l} = (w_1, w_2, \ldots w_l)$ за умови спостережень мовленнєвого сигналу $O = (o_1, o_2, \ldots o_m)$:

$$P(W_{0,l}/O) = \frac{P(O/W_{0,l})\,P(W_{0,l})}{P(O)},$$

Основна ідея лінгвістичної моделі

Імовірність речення із l слів $W_{0,l} = (w_1, \quad w_2, \ldots w_l)$ розкладається на добуток умовних імовірностей:

$$P(w_1, w_2, \ldots, w_l) = \prod_{i=1}^{l} P(w_i / w_1, w_2, \ldots, w_{i-1}) = \prod_{i=1}^{l} P(w_i / W_{0,i-1}),$$

Основна ідея лінгвістичної моделі

Обмежуємо контекст:

$$P(w_1, w_2, \ldots, w_l) \cong \prod_{i=1}^{l} P(w_i / w_{i-n+1}, \ldots, w_{i-1}) = \prod_{i=1}^{l} P(w_i / W_{i-n,i-1}),$$

де n – порядок моделі.

При $n = 1$ модель вироджується до відсутності елементарних контекстів, тобто розглядаються **уні-грами**.

При $n = 2$ отримуємо **бі-грамну** модель, коли в елементарний контекст входить граматична функція одного попереднього слова.

Збільшуючи далі параметр n, ми розширюємо елементарний контекст, підвищуючи точність моделі.

Основна ідея лінгвістичної моделі

Враховуючи:

$$\hat{P}\left(W_{0,l}\right) = \prod_{i=1}^{l} P\left(w_i / W_{i-n,i-1}\right),$$

Оцінюємо ймовірність послідовності слів при надходженні наступного слова w_{i+1}

$$\hat{P}\left(W_{0,i+1}\right) = \hat{P}\left(W_{0,i}\right) P\left(w_{i+1} / W_{i-n,i}\right)$$

Основна ідея лінгвістичної моделі

Оцінка параметрів лінгвістичної моделі

$$P\left(w_i / W_{i-n,i-1}\right) = \begin{cases} \alpha\left(W_{i-n,i-1}\right) & : \quad c\left(W_{i-n,i}\right) = 0 \\ d_{c\left(W_{i-n,i}\right)} \dfrac{c\left(W_{i-n,i}\right)}{c\left(W_{i-n,i-1}\right)} & : \quad 1 \le c\left(W_{i-n,i}\right) \le k \\ \dfrac{c\left(W_{i-n,i}\right)}{c\left(W_{i-n,i-1}\right)} & : \quad c\left(W_{i-n,i}\right) > k \end{cases}$$

Основна ідея лінгвістичної моделі

Розбиття слів на класи $g = G(w)$

Основні вимоги до корпусу

- Фільтрування вхідних текстів
- Розшифровування чисел, символів і скорочень
- Узгодження типів словоформ розшифрованого
- Узагальнення власних назв, чисел
- Зняття омографії
- Виявлення орфографічних та пунктуаційних помилок, та їх виправлення
- Виявлення новотворів

Фільтрування вхідних текстів

- Проводиться токенізація сегментів тексту
- Формуються множини токенів:

  A. слова, що належать до базового словника

  B. слова, що не належать до базового словника

  C. числа або символи

- Обчислюються співвідношення потужностей A, B і C і приймається рішення щодо “рідності”
- Включаємо слова з “рідних” сегментів до базового словника та закручуємо ітерацію

Розшифровування чисельників і символів

Вхід	Вихід
У нас в квартирі було прописано 11 чоловік:батько, мати, я з 6 дітей та мій брат з сином.	У нас в квартирі було прописано **одинадцять** чоловік:батько, мати, я з **шість** дітей та мій брат з сином.
Постанова Кабінету Міністрів Україн від 18 листопада 1994 р. № 784 “Про мінімальні ставки авторської винагороди за Використання творів літератури і мистецтва”	Постанова Кабінету Міністрів Україн**и** від **вісімнадцять** листопада **тисяча дев'ятсот дев'яносто чотири рік номер сімсот вісімдесят чотири** “Про мінімальні ставки авторської винагороди за Використання творів літератури і мистецтва”
Заробіна плата зросла на 22%	Заробіна плата зросла на **двадцять два відсоток**

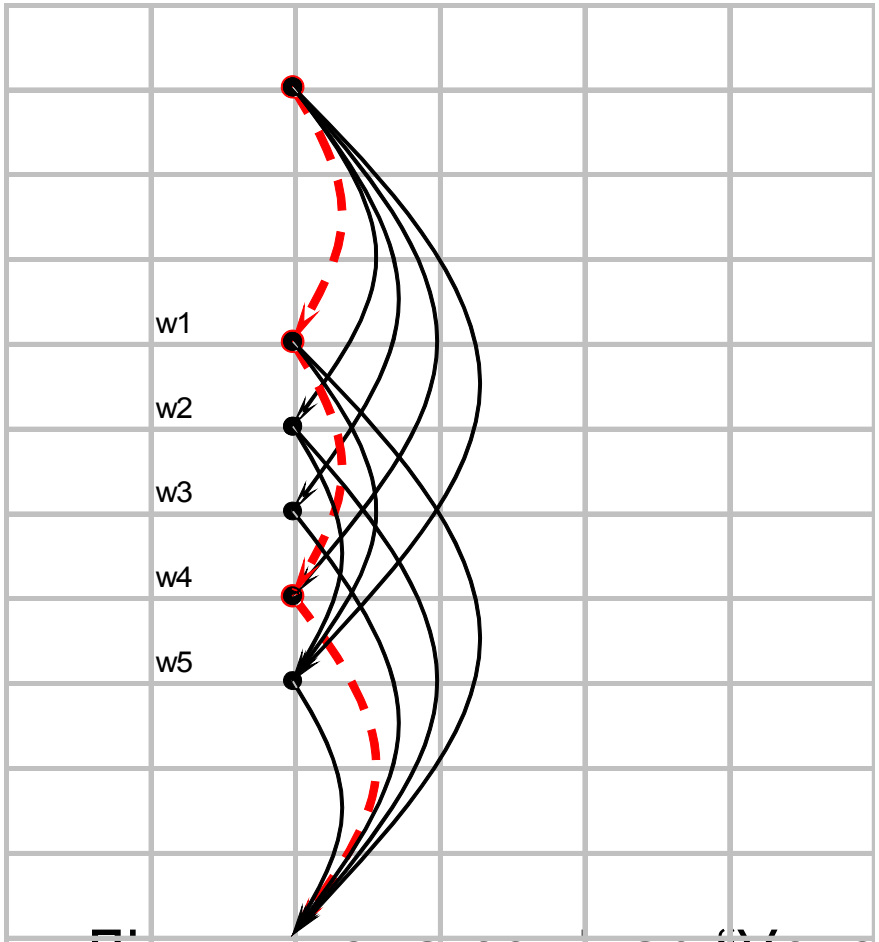
Узгодження чисельників і символів

Вхід	Вихід
Постанова Кабінету Міністрів Україн від 18 листопада 1994 р. № 784 “Про мінімальні ставки авторської винагороди за Використання творів літератури і мистецтва”	Постанова Кабінету Міністрів Україн**и** від вісімнадцят**ого** листопада тисяча дев'ятсот дев'яносто **четвертого року** номер сімсот вісімдесят чотири “Про мінімальні ставки авторської винагороди за Використання творів літератури і мистецтва”
Заробіна плата зросла на 22%	Заробіна плата зросла на двадцять два відсот**ки**

Основні вимоги до корпусу

- Фільтрування вхідних текстів
- Розшифровування чисел, символів і скорочень
- Узгодження типів словоформ розшифрованого
- Узагальнення власних назв, чисел
- Зняття омографії
- Виявлення орфографічних та пунктуаційних помилок, та їх виправлення
- Виявлення новотворів

Відновлення пунктуації



Моделі розставляння пунктуації

Характеристика корпусу

	News	Publ	Lit	Just	Hum
MB	204,1	241,9	171,4	106,3	21,4
%	27,4	32,5	23,0	14,3	2,9
Млн. слів	28,7	34,1	24,1	15,0	3,0

Кількість реалізацій слів: 105 млн.

Обсяг словника: 227 тис.

ukrcenter
kazka.in.ua
vesna.org.ua
libr.org.ua
jokes2
anegdot.com.ua
radio_svoboda
unian
liga_net
vartainfo.com.ua
911.kiev.ua
justinian.com.ua
ovu.com.ua
yurist-online.com
primacus

Плани на майбутнє

- **Автоматичне визначення власних назв, слів іншомовного походження, хибно написаних слів.**
- **Удосконалення модуля верифікації мови тексту.**
- **Аналіз омографів.**
- **Розшифровування скорочень та абревіатур.**

Дякую за увагу!

mykola@uasoiro.org.ua