

ПРОБЛЕМИ ЗНЯТТЯ МОРФОЛОГІЧНОЇ ОМОНІМІЇ НА ОСНОВІ МЕТОДУ ПРИХОВАНОГО МАРКІВСЬКОГО ЛАНЦЮГА

Жерновнікова А.В., Кочуєва З.А.

*Національний технічний університет
«Харківський політехнічний інститут»,
м. Харків*

Проблеми зняття омонімії актуальні на даний час. В автоматичній обробці природної мови часто використовують корпуси морфологічно-розмічених текстів. Розмітка яких виконується завдяки заздалегідь сконструйованим морфологічним аналізаторам. Основною проблемою роботи таких систем є неоднозначність. Через це виникає необхідність у виборі єдиного варіанта аналізу, правильного для даного контексту [1].

Існують наступні методи зняття морфологічної омонімії: використання нейронної мережі; методи, засновані на правилах; використання алгоритму прихованої Марківської моделі [2].

Для вирішення завдання зняття морфологічної омонімії обрано широко відомий імовірнісний підхід, заснований на використанні прихованої Марківської моделі (Hidden Markov Model (HMM) tagging).

Для реалізації алгоритму необхідно виконати морфологічну розмітку таким чином, щоб максимізувати функцію:

$$P(\text{word}/\text{tag}) * P(\text{tag}/\text{previous } n \text{ tags}),$$

де $P(\text{tag} / \text{previous } n \text{ tags})$ - умовна ймовірність (обчислена за розміченим корпусом), появи даного тега tag , за умови, що попередні n тегів вже визначені. $P(\text{word} / \text{tag})$ - умовна ймовірність (також обчислена за корпусом) появи в даному місці слова word , за умови, що це слово має даний граматичний клас tag [3].

Алгоритм прихованої Марківської моделі має досить високу точність для англійської мови, а саме 96%. При застосуванні даної моделі для української мови виникають труднощі, оскільки потребуються розмічені корпуси дуже великого обсягу, враховуючи багатство українського словотворення і словозміни порівняно з англійською мовою.

Література:

1. Jurafsky, Martin. Speech and Language Processing // – Upper SaddleRiver, NJ, USA: Prentice-Hall, Inc., 2009.
2. Б.П. Кобрицов. Методы снятия семантической неоднозначности. НТИ, Сер.2, Вып. 3, 2004
3. Daniel Jurafsky, James H. Martin. Speech and Language Processing, 2000