МЕТОД ОПРЕДЕЛЕНИЯ КОРЕФЕРЕНТНЫХ СВЯЗЕЙ В МИНИМАЛЬНОЙ ЕДИНИЦЕ ДИСКУРСА

Терещенко В. И.
Национальный технический университет
«Харьковский политехнический институт»
г. Харьков, ул. Фрунзе, 21, тел. (057) 700-15-64
e-mail: omsroot@kpi.kharkov.ua

Одной из основных задач обработки естественного языка (Natural Language Processing) является установление связей между объектами в одном и том же тексте. Такие связи между элементами фраз в лингвистике называются кореферентными. На сегодняшний день существует достаточно много алгоритмов и моделей определения кореферентных связей в тексте. Однако большинство из данных разработок являются неприменимыми для флективных языков с хорошо развитой морфологией [1]. Хотя решение данной задачи является насущно-необходимым в приложениях Машинного перевода, Opinion Mining и автоматического реферирования, задача определения кореферентных связей в минимальной единице дискурса еще не была разрешена.

Кореферентность – отношение между местоимением (анафором) и его антецедентом, при котором и местоимение и его антецедент соотносятся с одним и тем же предметом объективной действительности [2]. Пример предложения с кореферентной связью:

*«Чистая прибыль Visa[**antecedent**] составила 1,27 млрд дол, говорится в сообщении компании. По итогам II квартала 2012 года она[**anaphor**] сообщила о чистой прибыли в 1,29 млрд дол.»*

Можно выделить следующие этапы разрешения данной задачи:

1) определение минимальной единицы дискурса;
2) выделение типов кореферентных отношений;
3) анализ текстов в которых встречаются кореферентные отношения;
4) разработка алгоритма определения кореферентных связей в минимальных единицах дискурса;
5) создание информационно-лингвистического обеспечения задачи определения кореферентных отношений в текстах заданного языка;
6) создание программной реализации разработанного алгоритма.

Дискурс (франц. discours – язык) – текст в некотором событийном аспекте; речь, которая рассматривается как целеустремленное социальное явление или компонент который берет участие во взаимодействии людей и механизмах их мышления (когнитивных процессах).

Выделяют несколько уровней письменного дискурса [3]:

1) синтаксический (фраза, предложение, высказывание);
2) лексический (токен, слово);
3) морфологический (морфема);
4) фонологический (фонема).

Следующим этапом исследования является выделение типов кореферентных связей. Текстам флективных языков наиболее присущи такие типы кореферентности [4]:

1) анафора;
2) катафора;
3) кореферентность именных групп;
4) сведение;
5) расширение.

В процессе исследования были определены особенности каждого из вышеперечисленных типов и для более глубокого исследования был выбран анафорический тип кореферентных связей, поскольку он встречается в более чем 80% проанализированных текстов. Этот тип кореферентности имеет структуру при которой антецедент всегда предшествует анафору в тексте. При этом анафор как правило является местоимением а антецедент существительным. Однако, исследования показывают, что не каждое местоимение является анафором и далеко не каждое существительное в тексте относится к анализируемому местоимению [5]. Взяв во внимание эти и другие установленные в ходе исследования особенности организации текстов на русском языке, было разработано множество грамматических правил для определения кореферентных связей в минимальных единицах дискурса. В результате чего к настоящему моменту было разработано 15 грамматических правил, включающих как синтаксические, так и морфологические закономерности анафорических связей в минимальной единице дискурса текстов русского языка. Был проанализирован массив текстов включающий более 200 предложений находящихся в свободном доступе на сайте GoogleNews.

Результатом исследования является программная реализация разработанная на основе алгоритма разрешения задачи определения кореферентных связей в минимальной единице дискурса учитывающего особенности русского языка. Программная реализация была выполнена средствами языка программирования Python 2.7. Работа приложения была протестирована на базе из 100 русскоязычных текстов общим объемом в 5,214Мб. Средняя точность автоматического определения анафоры приблизительно равна 69%.

Список литературы

1. *Clark J.H., Gonzalez-Brenes J.P*. “Coreference Resolution: Current Trends and Future Directions”, 2008. – 11-16p.

2. *Hobbs J. R.* “Pronoun resolution” – California, 1976. – 10-18p.

3. *Кашкин В.Б.* Сопоставительные исследования дискурса «Концептуальное пространство языка». Тамбов: ТГУ, 2005. С. 337-353.

4. *Скатов Д.* “Разрешение кореференции: обзорная экскурсия” – Н. Новгород, ДИКТУМ, 2012.

5. *Grosz B.J.* “Readings in natural language processing” – California, Morgan Kaufmann Publishers, Inc., 1986. - p. 339-352.