

СНЯТИЕ ОМОНИМИИ КАК ОДИН ИЗ ЭТАПОВ СИНТАКСИЧЕСКОЙ СЕГМЕНТАЦИИ

Гаевская Я.К., Канищева О.В.

Научный руководитель – к.т.н. Канищева О.В.

Национальный технический университет

«Харьковский политехнический институт»

(61002, Харьков, ул. Фрунзе, 21,

каф. Интеллектуальных компьютерных систем, тел. (057) 707-63-60)

e-mail: yanna.rey@gmail.com, olya-kanisheva@rambler.ru

The given work is devoted to the problem of homonymy resolution in Ukrainian language. A review of methods allowing relieve some syntactic homonymy for the Ukrainian language. The authors propose an approach based on contextual rules allows implementation of effective and fast standalone module for disambiguation during syntactic segmentation.

Как известно, на данный момент ни одна система автоматического анализа или перевода текста не является совершенной или хотя бы близкой к таковой. Одной из основных причин неудач является высокий уровень неоднозначности естественного языка.

Еще до проведения синтаксического анализа имеется возможность выдвинуть некоторые предположения о структуре разбираемого предложения, выделить его фрагменты (сегменты), которые можно разбирать независимым образом. Дальнейший синтаксический анализ будет опираться на эти предположения и получит возможность сразу отбросить часть вариантов. Для использования этих возможностей вводится этап синтаксической сегментации. Одним из этапов синтаксической сегментации является уменьшение количества омонимов, соответствующих каждой словоформе [1].

Актуальность проблемы определяется еще тем, что практически все существующие алгоритмы снятия омонимии включаются в состав синтаксического анализа, что создает трудноразрешимое противоречие, когда для успешного снятия омонимии необходимы точные результаты синтаксического анализа, для получения которых, в свою очередь, нужно предварительно снять омонимию. Кроме того, значительный объем исходного числа связей существенно замедляет обработку, приводя к так называемому «комбинаторному взрыву».

Омонимия – тип семантических отношений, который устанавливается между словами, значения которых абсолютно не связаны друг с другом, но формально эти значения выражаются сходными звуковыми оболочками. *Омонимы* (греч. homo's – одинаковый и o'пута – имя) – слова, имеющие одинаковое звучание, но разные значения.

Для автоматизированного разрешения неоднозначностью используется несколько основных подходов: а) детерминированные

правила, работающие на основе лексических и грамматических данных; б) базы знаний об окружающем мире и онтологии, дающие возможность учитывать экстралингвистические данные; в) вероятностные анализаторы, учитывающие статистические данные языка, как правило, обучающиеся в процессе работы. К сожалению, все эти подходы имеют свои ограничения по эффективности и не дают результата желаемого уровня.

Существует четыре основных метода разрешения многозначности:

- методы, основанные на знаниях (dictionary- и knowledge-based methods): эти методы преимущественно используют словари, тезаурусы, лексикографические базы данных.
- методы обучения с учителем (supervised methods): эти методы используют размеченные корпуса текстов для тренировки классификатора.
- методы частичного обучения с учителем (minimally-supervised methods): эти методы используют вторичные знания, такие как определения терминов в толкованиях слов или выровненный двуязычный корпус.
- методы обучения без учителя (unsupervised methods): большинство этих методов не предполагает использование каких-либо внешних данных и используют только raw unannotated corpora; также, они известны под термином кластеризации и «word sense discrimination».

Также существуют другие методы, основанные на совершенно отличающихся от вышеперечисленных принципах: определение доминантности значения слова (Determining Word Sense Dominance); разрешение, основанное на темах корпуса (Domain-Driven Disambiguation); WSD, использующее кросс-языковые данные (Cross-Lingual Evidence) [2].

В работе рассмотрена проблема разрешения омонимии в украинском языке. Проведен обзор методов позволяющих частично снять синтаксическую омонимию для украинского языка. Авторами предложен подход, основанный на контекстных правилах, позволяющий реализовать эффективный и быстродействующий автономный модуль для снятия омонимии на этапе синтаксической сегментации.

Список источников информации:

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е. И., Клышинский Э. С., Ландэ Д. В., Носков А. А., Пескова О. В., Ягунова Е. В. – М.: МИЭМ, 2011. – 272 с.

2. Интерактивное разрешение лексической и синтаксической неоднозначности в системах автоматической обработки естественного языка / Бердичевский А. С., Крейдлин Л. Г., Лазурский А. В., Митюшин Л. Г., Сизов В. Г. // Интернет-математика 2005. – М.: Яндекс, 2005. – С. 44-66.