И.В. Азарова, С.В. Бичинева, Д.Т. Вахитова

АВТОМАТИЧЕСКОЕ РАЗРЕШЕНИЕ ЛЕКСИЧЕСКОЙ НЕОДНОЗНАЧНОСТИ ЧАСТОТНЫХ СУЩЕСТВИТЕЛЬНЫХ (в терминах структурных единиц RussNet)

Разрешение лексической неоднозначности является необходимым модулем систем автоматического анализа текста, в основном, тех, в которых создается семантическое представление текста. Несмотря на обширную литературу и разнообразие методов снятия неоднозначности, данная проблема имеет ряд аспектов, существенным образом влияющих на ее схему и эффективность.

В рамках предлагаемого доклада рассматривается снятие неоднозначности для частотных существительных в терминах структурных единиц RussNet – семантических деревьев и синсетов (синонимических рядов). Существительные, особенно конкретные, как показывают данные WordNet 2.x[1] по большей части моносемантичны (79% существительных vs. 5% глаголов), при этом многозначные существительные имеют довольно высокий индекс неоднозначности 2,79. Этот показатель мог бы быть существенно выше при его вычислении по данным традиционных толковых словарей, что связано как с общим набором зафиксированных значений, так и их градуированием (выделением оттенков значений и проч.). Поскольку в RussNet включаются только те значения, которые встречаются в корпусе современных текстов объемом 21 млн словоупотреблений[2], при анализе структуры

[1] URL: **http://wordnet.princeton.edu/**

[2] *Азарова И.В., Синопальникова А.А.* Использование статистико-комбинаторных свойств корпуса современных текстов для формирования структуры компьютерного тезауруса RussNet // Труды международной конференции «Корпусная лингвистика – 2004». 11–14 октября 2004 г. СПб., 2004. С. 5–15.

семантических значений производится разметка случайной выборочной совокупности контекстов корпуса, при этом Словарь русского языка в 4-х томах (под ред. А.П. Евгеньевой, М., 1981 – МАС) используется в качестве основы предварительной схемы значений. Были выявлены случаи, когда невозможно разграничить реализацию пар словарных значений в контекстах, поскольку контексты в равной степени могут быть отнесены к обоим значениям, или различие в словарных значениях не сформулировано четко и не проиллюстрировано соответствующими примерами.

Очевидно, что задача снятия неоднозначности в значительной степени связана с разграничением значений в словаре RussNet. Таким образом, схема градуирования значений при составлении RussNet должна быть гармонизирована с дальнейшим использованием этих значений в процедурах снятия неоднозначности. Поэтому при разграничении значений явным образом фиксируется список контекстных маркеров, включающий морфосинтаксические, семантические и статистические особенности контекстов, реализующие те или иные значения. Эта информация оформляется в виде активных и пассивных рамок валентностей, которые составляют часть структуры RussNet и хранятся при соответствующих синсетах[1].

Однако, определить параметры рамок валентностей можно только для частотных значений слов. Поскольку большую долю составляют умеренно частотные и низкочастотные значения, то требуется выявить способы экстраполирования этой информации

[1] *Azarova I.V., Sinopalnikova A.A., Ovchinnikova E.A., Ivanov V.L.* RussNet as a Semantic Component of the Text Analyser for Russian // Proceedings of the Third International WordNet Conference. Brno, 2005. P. 19–28.

на базе частотных слов, например, наследования свойств активных рамок валентностей[1]. В упомянутом исследовании было показано, что сочетание свойств рамок валентностей имеет более сложный характер, чем характерный вариант наследования меронимических свойств в гипонимических деревьях.

Для существительных неоднозначная ситуация с наследованием свойств рамок валентностей усложняется еще и тем, что многие значения довольно сложным образом соединяют активные и пассивные рамки.

Для поэтапного разрешения поставленных выше проблем были независимо протестированы две автоматические процедуры разрешения неоднозначности.

Первая процедура использует однозначную морфологическую разметку обучающей совокупности контекстов для выявления кластеров групп значений существительных. Применение этой процедуры для автоматической классификации глагольных контекстов дало удовлетворительные результаты[2]. Однако перенос методики на группировку значений существительных, возможно, потребует ее корректировки, на что указывали авторы этого метода К. Ликок, М. Ходоров и Дж. Миллер[3].

---

[1] *Азарова И.В., Иванов В.Л., Овчинникова Е.А*. Использование схемы наследования рамок валентностей в тезаурусе RussNet для автоматического анализа текста // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2006». М., 2006. С. 18–25.

[2] *Azarova I.V., Marina A.S., Sinopalnikova A.A.* Verification of Valency Frame Structures by means of Automatic Context Clustering in RussNet // Proceedings of the Fourth Global WordNet Conference, Szeged, Hungary, January 22–25, 2008. P. 35–43.

[3] *Leacock C., Chodorow M., Miller G.* Using Corpus Statistics and WordNet Relations for Sense Identification // Computational Linguistics. 1998. Vol. 24, № 1. 147–165.

Вторая процедура была построена на базе выявления маркеров в тысяче контекстов корпуса для ряда многозначных существительных (*мир, образ, пол, пора, путь*). Контекстные маркеры были организованы в последовательно применяемые наборы правил, обусловленные синтаксически, морфологически и фразеологически. Чтобы оценить вклад каждой группы правил на первом этапе, использовалась информация о значении морфологических категорий и лемматизации текстовых форм.

В докладе сопоставляются результаты работы двух процедур автоматического снятия неоднозначности на базе значений морфологических категорий для группы частотных существительных, имеющих несколько вхождений в синсеты RussNet. Оценивается эффективность процедур для значений из разных семантических деревьев.