ОБЩЕЕ ЯЗЫКОЗНАНИЕ

Кондюрин Иван Андреевич

Санкт-Петербургский государственный университет (СПбГУ),
Санкт-Петербург, Россия
ivan.kondyurin@gmail.com
Научный руководитель — А. В. Добров, канд. филол. наук

РАЗРЕШЕНИЕ НЕОДНОЗНАЧНОСТИ В ОДНОРОДНЫХ ИМЕННЫХ ГРУППАХ СРЕДСТВАМИ ОНТОЛОГИЧЕСКОЙ СЕМАНТИКИ

Ключевые слова: синтаксическая неоднозначность, онтологическая семантика, семантические связи, однородные именные группы, язык СМИ.

Статья посвящена снятию специфических типов синтаксической неоднозначности, порождаемых однородными именными группами, которые неверно интерпретируются существующими алгоритмами. Предложены способы создания правил на основе онтологической семантики, проанализирована проблема неразрешимой неоднозначности и проведена модификация существующей онтологии для увеличения точности разбора.

Kondyurin Ivan

Saint Petersburg State University (SPbSU),
St. Petersburg, Russia

SYNTACTIC DISAMBIGUATION IN HOMOGENEOUS NOUN PHRASES USING ONTOLOGICAL SEMANTICS

Keywords: syntactic disambiguation, ontological semantics, semantic links, homogeneous noun phrases, mass media language.

The article is concerned with the disambiguation of specific consistently ill-interpreted ambiguous constructions in homogeneous noun phrases by the means of ontological semantics. These constructions have been analyzed; certain ontology-based disambiguation rules have been proposed; the ontology has also been modified in order to improve the precision of the syntactic analysis. The paper also reviews the problem of insoluble ambiguity and suggests probabilistic criteria.

Введение

При анализе синтаксических структур одной из главных проблем является снятие неоднозначности. Поскольку число вариантов синтаксического разбора каждого неоднозначного фрагмента при анализе умножается на общее число вариантов разбора предложе-

ния, многократная омонимия в рамках одного предложения нередко приводит к комбинаторным взрывам. Такая проблема особенно актуальна для предложений с однородными рядами, где в ряде случаев, как будет показано, число способов разбора умножается по крайней мере на количество членов в ряде. Ярким примером такого неоднозначного предложения может служить пример из собранной в ходе исследования коллекции:

Традиционная лекция выдающегося лингвиста, академика, ученого с мировым именем Андрея Анатольевича Зализняка (…) *прошла 1 октября 2015 года.*

Здесь однородные именные группы создают более 10 возможных трактовок, по одной из которых лекцию прочли лингвист, академик, учёный и *мировое имя Зализняка.* Для правильного понимания этого предложения недостаточно обычных статистических методов или анализа контекста. Выбор версии разбора данного предложения, по-видимому, можно осуществить только на основании знаний о мире — в частности, о том, что «имя» не может читать лекцию (то есть, в данном случае, не может быть связано соответствующим семантическим отношением), а А. А. Зализняк является и лингвистом, и академиком, и учёным.

Снятие синтаксической неоднозначности: семантический подход

В целом методы снятия синтаксической неоднозначности можно разделить на несколько групп: связанные с семантическими ограничениями, синтаксические и статистико-вероятностные методы [Митренина, 2005, с. 86]. На практике наибольшую эффективность показали решения, основанные на сочетании нескольких методов и включающие семантику.

Семантический подход позволяет с высокой точностью разграничить синтаксические функции, основываясь на семантических валентностях лексических единиц. Его главным недостатком на данном этапе является низкая масштабируемость. Решением данной проблемы может стать применение единого источника, содержащего формализованное и структурированное описание значений основных лексических единиц языка. Такими источниками являются электронные тезаурусы и компьютерные онтологии.

Неоднократно предпринимались попытки использования данной семантической сети для снятия неоднозначности. В ряде случа-

ев они оказались успешными, но в целом исследования показали недостаточный прирост точности при добавлении этого метода в существующие системы анализа: отметим работу К. Фрагоса (Fragos) по дополнению алгоритма Леска отношениями из WordNet, в которой использование отношений меронимии только понизило эффективность по сравнению со стандартным алгоритмом на 6% [Fragos et al., 2003, с. 642]. Недостатком WordNet является слишком малое количество конкретных отношений, применимых для АОТ. Они не позволяют отразить все возможные интерпретации семантики: для установки семантических ограничений неприменимы такие абстрактные связи, как «относящийся к чему-л», «связанный с чем-л». Такие отношения позволяют лишь обозначить наличие связи, но не детализировать её.

Онтологии вместо электронных тезаурусов использовались для снятия неоднозначности в основном на небольших предметных областях — как, например, в [Tine et al., 2006, с. 73]. Наиболее общее определение даётся Т. Грубером: «An ontology is a specification of a conceptualization» [Gruber, 1993, с. 200]. Для задач данного исследования онтологию можно определить как формализацию упрощённого представления о некоторой предметной области, включающую концепты (формальные модели понятий) этой области и логические выражения для описания отношений между ними. Концепты обладают атрибутами (свойствами) и связаны между собой отношениями, причём участие в отношении может являться атрибутом, и наоборот, а отношение само по себе является концептом [Nirenburg, Raskin, 2004, с. 28].

Лингвистические онтологии учитывают большее, чем в компьютерных тезаурусах, число отношений для разных значений и оттенков значения слов, но содержат и некоторую лингвистическую информацию (например, сведения о части речи). Поскольку лингвистическая онтология по определению является упрощённой формализованной моделью языковой картины мира, та неоднозначность, которая не может быть устранена в этой модели, в том или ином виде будет присутствовать в действительности.

Неоднозначность однородных именных рядов

Для выполнения исследования использовались программные наработки системы AIIRE, включающей лингвопроцессор, онтологию, корпус-менеджер для загрузки текстов и разметки корпуса. Система AIIRE обладают некоторыми ценными особенностями: в ней

присутствует непосредственная связь с лингвопроцессором, интерфейс коллективного редактирования, средства для ускоренного редактирования глагольных концептов (Ontohelper), и большой набор отношений между объектами, а также детализированы концепты верхнего уровня. Подробный алгоритм построения синтксических структур описан в работе А. В. Доброва [Добров, 2014, с. 156]. В процессе разбора предложения проблемой является не только неоднозначность, но и разрывы — маркеры отсутствия связывания. Их причины — необработанные в онтологии единицы, дефекты описания и дефекты формальной грамматики. Устранение разрывов представляет собой трудоёмкий процесс, требующий модификации не только онтологии, но и лингвопроцессора, поэтому представляется целесообразным сформулировать методы разрешения неоднозначности на материале отдельных примеров из корпуса, для которых был предварительно обеспечен разбор без разрывов.

Исследование синтаксической неоднозначности было проведено на материале модифицированных предложений из Национального корпуса русского языка (НКРЯ). В использованный подкорпус вошли 500 наиболее показательных предложений с возможной неоднозначностью однородных ИГ из текстов, относящихся к категории «публицистические тексты информационного содержания». Выбор однородных рядов связан с тем, что такие конструкции наиболее существенны при анализе насыщенных фактической информацией новостных сообщений.

Однородные члены предложения, связанные сочинительной связью, образуют группу, аналогичную по синтаксическим функциям любому её элементу. В них включаются только лексически сопоставимые члены предложения, не связанные отношением гиперонимии или меронимии (напр. рыба, овощи, морковь) и относиться к разным абстрактным классам (напр. результат, красота, здоровье).

Анализ материала

В собранной коллекции предложения были сгруппированы по типам неоднозначности именных групп (ИГ):

Пояснительные конструкции. Согласно «Русской грамматике» [Шведова, 1980, с. 173], они могут входить в однородный ряд и связываются с поясняемым элементом сочинительной связью. В любом открытом однородном ряду при координации числа присутствует неоднозначность, связанная с невозможностью определить, является ли конструкция сочинительной или поясняющей:

В экспозицию (…) *включены этюды, графика и акварель.*

Возможный способ снятия неоднозначности с помощью семантики — проверка отношений между ИГ. Исходя из определения, члены сочинительного ряда не могут быть связаны отношением гипонимии и синонимии. Если последующее — гипоним или мероним первого, то речь идёт о включении или уточнении. Если же в ряду присутствуют пространственные или временные понятия одного уровня («леса, поля», «Швеция, Финляндия»), они не могут являться уточнениями.

Отношения, выраженные генитивом

Если однородный ряд стоит в родительном падеже, некоторые ИГ могут как являться членами этого ряда, так и состоять с его элементами в генитивных отношениях. В силу многозначности данного падежа, количество возможных отношений высоко. При этом требуется не только определить структуру зависимостей, но и установить правильный тип отношения. Например:

По словам руководителя центра по борьбе со СПИДом академика Вадима Покровского

В данном случае для слов в родительном падеже потенциально возможны следующие отношения: *центр* (принадлежность) *академика Вадима, СПИД* (принадлежность собственнику или материальная принадлежность) *академика Вадима, академик* (принадлежность или работа под начальством) *Вадима и т. д.* Однородный ряд может быть приложением имени собственного, и тогда неясно, относится к нему весь ряд или только последний его член:

140 лет со дня рождения художника и графика Леонида Пастернака.

Разрешимыми являются только те случаи, в которых о называемом человеке известно, может ли он обладать указанными характеристиками (то есть связывается ли ИГ в функции приложения как его профессия, занятие и т. д.). К примеру, Леонид Пастернак — российский живописец и график — может обладать обеими характеристиками.

Неоднозначность зачастую остаётся неразрешимой на данном этапе, но число версий можно уменьшить несколькими ограничениями. Во-первых, ИГ, обозначающая имя, может иметь в качестве приложения только одушевлённую ИГ. Во-вторых, для частотных в официальных текстах генитивных конструкций производится

идиоматизация: «*Министерство природных ресурсов и экологии*» — отдельное понятия, и для него неоднозначность возникать не будет. Отдельный сложный вопрос — проблема кавычек. Названия в кавычках могут не быть существительными и в общем случае ничего не сообщают о называемом ими объекте. Для частотных, общеизвестных наименований самое очевидное решение — учёт всех именованных сущностей в онтологии, причём значение названия не должно оформляться как одно из значений исходного слова (значение «*Москва*»-*гостиница* не должно находиться в том же концепте, что и *Москва-город*).

Зависимые слова

Проблемы однородных рядов почти всегда осложняются зависимыми словами. Они могут относиться как только к первому или только к последнему, так и к каждому слову из однородного ряда. Это фактически означает эллипсис повторяющихся зависимых слов: «смена паспорта и имени» — «смена паспорта и смена имени»:

Большой опыт и прозрачность сделок обеспечивают доверие клиентов и партнеров.

В таких конструкциях чаще всего возникает неразрешимая неоднозначность. Возможность отнесения зависимого слова ко всему ряду можно формально определить: требуется, чтобы оно могло семантически связываться со всеми элементами ряда одним и тем же отношением. Если такое связывание невозможно (*жителей страны и туристов*), зависимое слово однозначно будет относиться только к одному элементу. В иных случаях, однако, нельзя однозначно выбрать одну из версий.

Заключение

Исследование возможностей использования средств онтологической семантики для снятия синтаксической неоднозначности при автоматическом анализе новостных текстов новостных текстов позволило сделать следующие выводы:

1. Существуют случаи, при которых применение онтологии действительно позволяет уменьшить количество версий синтаксического разбора составляющих с однородными ИГ — когда отбрасываемые версии противоречат существующим в онтологии ограничениям.

2. Отсутствие противоречия для какой-либо версии не означает, что именно она является правильной. Если нарушения ограничений нет, остаются возможными все существующие версии.

В дальнейшем планируется задать правила ранжирования вероятности версий на основе сведений из онтологии и исследовать их эффективность выходят за пределы данного исследования.

ЛИТЕРАТУРА

Добров, 2014 — *Добров А. В.* Автоматическая рубрикация новостных сообщений средствами синтаксической семантики: дис. … канд. филол. наук: 10.02.21. СПб., 2014.

Митренина, 2005 — *Митренина О. В.* Проблемы неоднозначности синтаксического анализа: дис. … канд. филол. наук: 10.02.21. СПб., 2005.

Шведова, 1980 — Русская грамматика. Т. 2: Синтаксис / Н. Ю. Шведова (гл. ред.). М.: Наука, 1980.

Fragos et al., 2003 — *Fragos K., Maistros Y., Skourlas C.* Word sense disambiguation using WordNet relations // First Balkan Conference in Informatics. Thessaloniki, 2003.

Gruber, 1993 — *Gruber. T. R.* A translation approach to portable ontologies // Knowledge Acquisition, 1993, 5(2): 199–220.

Nirenburg, Raskin, 2004 — *Nirenburg S., Raskin V.* Ontological Semantics — Cambridge, Mass: MIT Press, 2004.

Tine et al., 2006 — *Tine L., Terney, Vestskov T.* An Ontology-Based Approach to Disambiguation of Semantic Relations // Association for Computational Linguistics, 2006.