МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ» (НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)

Кафедра систем информатики

Маркова Мария Владимировна

АВТОМАТИЧЕСКАЯ СЕМАНТИЧЕСКАЯ РАЗМЕТКА ПРЕДЛОЖЕНИЙ АНГЛИЙСКОГО ЯЗЫКА

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ
по направлению высшего профессионального образования

230100.68 ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

ФАКУЛЬТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Тема диссертации утверждена распоряжением по НГУ № 9 от «11» января 2012 г.
Тема диссертации скорректирована распоряжением по НГУ № 183 от «14» мая 2013 г.

Руководители:

Береснев В.Л.
д.ф.-м.н., профессор

Ващенко В.В.

Новосибирск, 2013г.

МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ» (НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)

Кафедра систем информатики
(название кафедры)

УТВЕРЖДАЮ

Зав. кафедрой Лаврентьев М.М.

……
(подпись, дата)

ЗАДАНИЕ
на магистерскую диссертацию
студент Маркова Мария Владимировна
(фамилия, имя, отчество)

факультета ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Направление подготовки 230100.68 ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

Магистерская программа: «Технология разработки программных систем»

Тема: «Автоматическая семантическая разметка предложений английского языка»

Цели работы: разработать эффективный метод автоматической семантической разметки на основе известных лингвистических теорий, и создать его практическую реализацию путем интеграции крупных лингвистических ресурсов, дополнения их слабых мест и добавления алгоритма означивания.

Руководители:
Береснев В.Л.
д.ф.-м.н., профессор

……..
(подпись, дата)
Ващенко В.В.

……..
(подпись, дата)

Содержание

ВВЕДЕНИЕ

Задачу распознавания смысла текстов на естественном языке машиной научный мир решает уже давно. Заманчивой является идея о том, чтобы компьютерная система могла анализировать тексты, в частности, извлекать из них различные данные, факты, сведения, полезные при решении разных практических задач. Такие задачи могут возникнуть, например, в системах машинного перевода, в системах принятия решений, в вопросно-ответных системах, в задачах информационного поиска, в системах голосового управления и распознавания, список можно продолжать и далее.

Задачу извлечения смысла иногда считают NP-полной, потому как распознавание живого языка требует огромных знаний системы об окружающем мире и возможности с ним взаимодействовать. Несмотря на это, теоретическая и прикладная лингвистика не стоят на месте, за годы их существования созданы мощные лингвистические ресурсы, подкрепленные научными теориями. За разработкой и поддержкой таких ресурсов стоит работа огромного числа специалистов этой области, что обеспечивает их качество.

Цель – на основе известных лингвистических теорий разработать эффективный метод автоматической семантической разметки; путем интеграции ресурсов, охватывающих большой объем языка, дополнения их слабых мест и добавления алгоритма означивания, создать практическую реализацию метода.

В качестве единицы означивания выбрано простое предложение английского языка. Под семантической разметкой подразумевается как определение ситуации, скрытой в предложении с выявлением ее участников и атрибутов, так и определение смыслов слов и словосочетаний, входящих в предложение. Такое определение семантической разметки принципиально отличает данный метод от wsd-алгоритмов[1].

Разработанный метод опирается на существующий лингвистический опыт: он предполагает совместное использование известных лингвистических теорий, а его практическая реализация – интеграцию крупных лингвистических баз данных. Таким образом, через совмещение сильных сторон существующих теорий и дополнение слабых, решается задача распознавания текста на естественном языке на некотором уровне.

Основными задачами, поставленными для достижения цели стали:

- разработка метода автоматической семантической разметки;
- практическая реализация разработанного метода.

Лингвистическими теориями, лежащими в основе предложенного метода, являются: теория семантики фреймов Ч. Филлмора [1] и классификация глаголов Б.Левин [2]. Для

[1] WSD (word sense disambiguation, *рус.* «разрешение лексической неоднозначности») задача выбора значения многозначного слова или словосочетания в зависимости от контекста, в котором оно находится.

практической реализации метода была произведена интеграция наиболее популярных реализаций этих теорий: FrameNet [3, 4], реализующий семантику фреймов, и VerbNet [5], построенный на классификации глаголов Б. Левин. В качестве лексической базы данных был выбран проект Принстонского университета – лексический ресурс WordNet [6]. Также, большое влияние на разработку метода семантической разметки оказали идеи, заложенные в фундаментальной классификации предикатов Д.Ю. Апресяна [7].

Существенным плюсом данного метода можно назвать использование накопленного опыта прикладной лингвистики. Таким образом, в данной работе упор делается не на составление словарных статей, наполнение лексических баз и другую объемную лингвистическую работу, требующую специальной подготовки, а на анализ существующих лингвистических теорий, их совместное использование и, с точки зрения программной реализации, на интеграцию лексических ресурсов.

Глава 1 Обзор предметной области и существующих решений.

Подходы к извлечению смысла из текста строятся на основе формальных моделей, которые способны в той или иной степени выполнять анализ текста на естественном языке. Основные отличия этих подходов заключаются в методах реализации компонента понимания смысла, в используемых средствах анализа, а также в объеме и способах представления знаний. Именно знания, представленные в различной форме, являются базой, от которой зависит процесс распознавания, глубина проникновения в смысл и, соответственно, качество самой модели.

Понимание высказываний включает анализ и интерпретацию. В задачу анализа входит выделение смысла входного текста (под смыслом будем понимать семантику – информацию, которая заложена в исходном тексте) и выражения этого смысла на внутреннем языке системы. Интерпретация заключается в отображении входного текста на знания системы. Проанализируем наиболее проработанные модели естественного языка с этих точек зрения.

В существующих моделях лингвистического анализатора можно выделить следующие способы выделения и представления смысла:

- компонентный анализ;
- сеть концептуализаций;
- идентификация смысла по образцу;
- интегральный подход.

1.1 Компонентный анализ

Одна из первых попыток формализации входного текста принадлежит компонентному анализу [8, 9], который исходит из предположения о том, что посредством конечного набора семантических компонентов можно описать неограниченнее множество лексических единиц. Техника выделения семантических множителей состоит в рассмотрении слов и выделении некоторых признаков, разбивающих слова на разные классы и семантические группы, например, по таким признакам, как *одушевленный*/*неодушевленный*, *мужской*/*женский пол* и т.д. можно выделить и более дифференцированные признаки для классов слов, например, *животные*, *рыбы*, *птицы*, *люди* и т.д. Значение каждого слова, таким образом, представляется как множество таких семантических множителей.

Рассмотрим на конкретном примере этот метод. Возьмем для анализа слово «*журнал*». Сначала, нужно найти слово или словосочетание, обозначающее род вещей, видом которого является журнал. Им будет словосочетание «*периодическое издание*».

Значение общеродового понятия (гипероним) будет первым семантическим компонентом, входящим в определение слова «журнал». Он отображает общие признаки журнала с другими изданиями такого рода, таких признаков два: «*издание*» и «*периодичность*». Эти общие признаки называются *интегральными семантическими признаками*.

Затем ищут все слова, обозначающие другие виды периодических изданий и выявляют те признаки, по которым журналы отличаются от других видов периодики. Такие признаки называют *дифференциальными семантическими признаками*. Помимо журналов, периодическими изданиями являются газеты, бюллетени, каталоги. От газет журналы отличаются тем, что они сброшюрованы. Если печатное издание не сброшюровано, оно не является журналом. От бюллетеня и каталога журнал отличается по другим признакам, относящимся не к оформлению издания, а к его содержанию. Например, каталоги создают для публикации данных о товаре. Таким образом, толкование слова «журнал» включает, кроме интегральных признаков, два дифференциальных. Для журнала это компоненты, характеризующие его со стороны внешнего вида и со стороны содержания.

Метод компонентного анализа активно развивается за рубежом. Существуют различные теории композиционной семантики (ученые Катц и Фодор - родоначальники, Барбара Патти, Анна Вежбицкая и др.).

В основном, для метода компонентного анализа выделяется тезис: «смысл предложений есть сумма смыслов входящих в него слов». Но при кажущейся естественности данный метод связан с существенными трудностями при реализации и не лишен слабостей. Он становится сложным при выражении смысла целого предложения и громоздким при анализе многозначных слов, при этом нет достаточного объяснения слова, что может привести к неправильному его употреблению.

1.2 Семантические падежи

Идея описания входного текста с помощью компонентного анализа нашла свое продолжение в модели Чарльза Филлмора «Семантические падежи», также известной как «Семантические роли» [8, 9]. Он принял гипотезу компонентной структуры значения и идею последовательного разложения смысла слова на более простые компоненты вплоть до семантических примитивов, или атомов смысла. Разделяя общепринятые взгляды на аргументную структуру предиката, он пришел к выводу, что необходимо указывать не только число аргументов данного предиката, но и их роли, т.е. семантическое содержание. Филлмор выделил следующие роли:

1. агент - одушевленный инициатор действия;

2. объект - вещь, являющаяся объектом действия;
3. контрагент - сила, против которой направлено действие;
4. адресат - лицо, для которого совершается действие;
5. пациент - вещь, которая испытывает эффективность действия;
6. результат;
7. инструмент - физическая причина действия/ стимул;
8. источник - исходное состояние объекта до действия.

В модели предложена более детальная концепция смысла высказывания. Общепринятая концепция лексического значения исходит из представлений о многослойности, т.е. включает оттенки значений, их стилистически и эмоционально экспрессивные элементы. Филлмор пошел дальше и расширил значение на две части: *собственно значение* и *пресуппозицию*.

Различия между ними проявляются, например, в различном влиянии на них отрицания. В область действия отрицания попадает только значение, а не пресуппозиция. Например, в высказывании «Вася - не холостяк» не утверждается, что Вася не мужчина. То есть если считать, что значение слова «холостяк» - «взрослый мужчина никогда не состоявший в браке», то отрицательным является только вторая часть после запятой, которая является собственно значением.

Основным результатом исследований Филлмора является пересмотр обычной схемы словарной статьи в толковом словаре. Согласно его теории, словарь считается основным средством задания семантических ролевых структур и правил их перевода в поверхностные структуры, общие с понятиями таких структур в российских исследовательских моделях управления Ю.Д. Апресяна.

1.3 Интегральный подход

Модели, в которых достаточно глубоко продуманы процедуры морфологического, синтаксического и проблемного анализов, можно отнести к моделям, основанным на интегральном подходе описания языка. Одной из них является модель «Смысл ↔ Текст» [8, 10].

Модель «Смысл ↔ Текст», разработанная И.А Мельчуком и А.К Жолковским, представляет собой многоуровневый транслятор текстов в смыслы и наоборот. Эта модель является наиболее детальной моделью русского языка, а также его кибернетической моделью (ориентирована на реализацию на ЭВМ). Для описания семантического сходства лексически различных слов авторами был представлен специальный язык семантических множителей. Существенными особенностями этого языка были следующие:

- «атомное» строение смысла, т.е. семантические примитивы;
- высокая структурность значений, т.е. сильная взаимосвязь слов друг с другом;
- наличие правил преобразования.

В связи с распространением омонимии и синонимии в естественном языке, в модели представлен не одношаговый переход от смысла к тексту, а многошаговый. При этом используется уровневая структура по аналогии с естественным языком. Выделяются четыре основных уровня – фонетический, морфологический, синтаксический и проблемный. Каждый из них, за исключением проблемного, подразделяется на два других уровня – поверхностный и глубинный.

Данная модель может быть применима в системах, где необходимо понимание текста в полном смысле (например, вопросно-ответные системы, системы принятия решений). Но для реализации полной схемы анализа и синтеза модели «Смысл ↔ Текст» придется учесть индивидуальные свойства сотен тысяч словарных, морфологических и лексических единиц и индивидуальные свойства громадного числа пар единиц. Их полное формальное описание представляет собой громадную и объемную теоретическую работу, поставленную в лингвистике в последнее время и еще далекую от решения.

1.4 Сеть концептуализаций

К следующему классу относятся модели, в которых смысл текста представляется в виде сети концептуализаций. Среди моделей данного класса наибольший интерес представляет модель «Концептуальной зависимости» [10, 11].

Данная модель разрабатывалась группой исследователей Йельского университета под руководством Роджера Шенка с 1968-го по 1975-ый год. Результат – новое описание семантики языка. Грамматика Шенка подчеркивает отрицание влияния синтаксиса и морфологии на семантику. Основой семантического представления модели является сеть концептуализаций. Сеть концептуализаций есть квазиграф, подобный размеченному ориентированному графу, в котором, кроме бинарных отношений, есть тернарные и кварнарные, а дуги связывают не только вершины, но и другие дуги. Вершины помечены символами семантических примитивов, на основе которых определяется смысл. Дуги задают отношения между ними.

В рамках теории концептуальной зависимости рассматриваются четыре типа примитивов. К ним относятся:

- ACT — действия (actions);
- PP — объекты (picture producers);
- AA — модификаторы действий (action aiders);

- PA — модификаторы объектов (picture aiders).

Предполагается, что каждое из действий должно приводить к созданию одного или нескольких примитивов ACT. Также существует ряд примитивов, выбранных в качестве компонента действия, например:

- ATRANS — передавать отношение (давать);
- PTRANS — передавать физическое расположение объекта (идти);
- PROPEL — применять физическую силу к объекту (толкать);
- MOVE — перемещать часть тела (владельцем) и другие.

Эти примитивы используются для определения отношений концептуальной зависимости, которые описывают смысловые структуры, такие как падежные отношения или связи объектов и значений. Отношения концептуальной зависимости — это концептуальные синтаксические правила, которые выражают семантические связи в соответствии с грамматикой языка. Эти отношения могут быть использованы для конструирования внутреннего представления английского предложения.

Таким образом, основным в модели концептуальной зависимости является понятие концептуализации, что представляет собой основную единицу семантического уровня, из которого конструируется высказывание и смысл текста. Концептуализация включает в себя действие, множество его концептуальных падежей и участников действия или их состояние. Будучи моделью языка, она не учитывает модели пользователя, что приводит к полному перебору при построении умозаключений. Наличие модели пользователя позволило бы определить его цели (намерения) в диалоге и использовать их для направления процедуры построения умозаключений.

Теория концептуальной зависимости имеет ряд важных преимуществ. Во-первых, за счет построения формальной теории семантик естественного языка уменьшаются проблемы, связанные с двусмысленностью. Во-вторых, само представление в процессе построения канонической формы смысла выражений охватывает много естественно-языковых семантик. Это означает, что все выражения, которые имеют один и тот же смысл, будут внутренне представлены синтаксически идентичными, а не только семантически эквивалентными графами. Это каноническое представление — попытка упростить выводы, требуемые для понимания. Например, мы можем продемонстрировать, что два выражения означают одну и ту же сущность, с помощью простого сопоставления графов концептуальной зависимости. Представление, которое не обеспечивает канонической формы, может потребовать большого количества операций на графах различной структуры.

Исследования показывают, что процесс преобразования выражений в каноническую форму вряд ли можно полностью алгоритмизировать. Более того, нет уверенности, что люди запоминают свои знания в какой-нибудь канонической форме. Критические замечания высказывались также по поводу вычислительной сложности преобразования предложения в набор примитивов низкого уровня. Кроме того, сами примитивы не позволяют охватить многие более тонкие понятия, играющие важную роль в использовании естественного языка.

1.5 «Семантика предпочтения»

Другая модель - «Семантика предпочтения» [8, 10] относится к классу моделей, идентификация смысла в которых осуществляется по образцам. Отличительной чертой таких моделей является то, что в них отсутствуют блоки морфологического и синтаксического анализов, что является принципиальным их недостатком, так как не обеспечивается глубина анализа значений слов, необходимая для точного установления семантической связности текста.

Разработки модели были начаты в Стэндфордском университете США Уилксом с 1964 года, сама модель была представлена в 1972. В этой модели текст характеризуется следующими сущностями: смыслами слов, сообщениями, фрагментами текста и семантической совместимостью. Сообщение рассматривается как теоретический конструкт, посредством которого для каждого слова, входящего во фрагмент текста, может быть выбран один из смыслов слова, за счет чего снимается многозначность. Слову назначается тот из его многих смыслов, который образует «сообщение», согласующееся, в конце концов, с рассматриваемым фрагментом текста. Если слово может подойти к нескольким сообщениям, то выбирается такое, которое согласуется с рассматриваемым текстом.

Анализ фрагмента текста протекает по следующей схеме. С помощью специальных слов-маркеров выполняется фрагментация текста, затем словам приписывают из словаря все их значения. Далее на анализируемый фрагмент текста поочередно накладываются простые шаблоны, известные системе. С помощью специальных правил расширения простой образец преобразуется в полный образец путем добавления слов из текста, которые не вошли в образец. Указанная процедура осложнена тем, что может подойти не один простой образец. Используя процедуры установления семантической близости полученных образцов, формируется окончательное представление обрабатываемого текста. К недостаткам анализа следует отнести то, что анализ текста осуществляется с

помощью словаря шаблонов, которые способны различать только классы событий, а не сами конкретные события.

Другой подход к способу анализа по образцу представлен в моделях, использующих табличный метод. Он основан на анализе ключевых слов, встречающихся в предложениях. Суть табличного метода состоит в идентификации смысла всего предложения на основании нескольких ключевых слов или их групп. После процесса идентификации слова предложения заменяются на их каноническую форму - коды. Замена осуществляется с помощью словаря словоформ. При этом также выделяются некоторые группы слов, несущие тематическую нагрузку. Далее производится распознавание и замена стандартных словосочетаний. Данный метод обладает рядом недостатков. Главным из которых можно назвать ограниченность понимаемых конструкций. Преимуществом является его простота для однозначных естественно-языковых предложений, в которых не требуется полного понимания смысла предложения (например, запросы к базе данных).

1.6 Выводы по обзору

Предлагаемый в работе метод автоматической семантической разметки нельзя в точности отнести ни к одному из описанных подходов. Но общее все же можно отметить. Он сочетает в себе элементы компонентного анализа, во многом обращается к работам Ч.Филлмора, точнее к теории семантики фреймов, выросшей из теории семантических падежей. Также метод имеет общее с интегральным подходом, но в отличие от него при реализации данного метода имеет место обращение к более поверхностной синтаксической структуре предложения. От подходов семантики предпочтений и шаблонных подходов метод принципиально отличается.

Таким образом, предложенный в данной работе метод не повторяет устоявшихся методов семантического означивания, но обращается к их сильным сторонам, что позволяет выдвинуть гипотезу о высокой эффективности предлагаемого метода.

Глава 2 Формальная постановка задачи.

Целью работы является решение задачи означивания английского предложения. Поставленная цель подразумевает разработку эффективного метода автоматической семантической разметки на основе существующих лингвистических теорий и его практическую реализацию. Для практической реализации метода предлагается использовать подкрепленные существующими теориями крупные лингвистические ресурсы, охватывающие большой объем языка, произвести их интеграцию, дополнить слабые места и добавить алгоритм означивания.

Таким образом, возникли задачи изучения существующих лингвистических теорий из области извлечения смыслов из текстов на естественном языке и практическое освоение реализующих их ресурсов.

В рамках программной реализации метода возникла задача интеграции лингвистических ресурсов и, как следствие, изучения средств интеграции. Ими стали отображения между ресурсами. Также в ходе практической реализации возникла задача дополнения слабых мест лексических ресурсов, требующая определенного объема лингвистической работы.

Глава 3 Описание подхода.

В данной главе определяется подход к автоматической семантической разметке предложений естественного языка, и объясняются теоретические основания его появления. Теоретической базой предлагаемого подхода является представление о предикатной структуре языка, или предикатной лексике, и актантной структуре слова [12].

3.1 Предикатная лексика. Актантная структура слова.

Лексику принято делить на предикатные и непредикатные слова в зависимости от того, какого рода объекты и действия они обозначают. Непредикатные слова называют конкретные предметы. Предикатные слова – это слова, обозначающие разного рода ситуации.

Среди предикатных слов выделяют два разряда:

а) слова, обладающие только понятийным содержанием и сами по себе не приспособленные к обозначению предметов действительности, – это прилагательные и глаголы;

б) слова, наделённые полной семантической структурой, способные получать как понятийное, так и денотативное содержание, – имена нарицательные, семантика которых приспособлена и к тому, чтобы называть, и к тому, чтобы обозначать ситуацию, иметь социально закреплённое значение.

Предикатная лексика – это, прежде всего, глаголы и существительные (особенно отглагольные), которые способны обозначать действия, процессы, отношения, т. е. разного рода ситуации.

Под ситуацией понимается такой фрагмент действительности, в котором на фоне тех или иных обстоятельств можно выделить одного или нескольких участников ситуации либо отметить их отсутствие. При этом участником ситуации может быть не только лицо, но и предмет, информация, любое понятие.

Например:

- в ситуации рассвета нет явно выраженного участника ситуации (сравните: *Светает; Рассвело*);
- в ситуации сна – один участник: тот, кто спит (*Ребёнок спит*);
- в ситуации знания – два участника: тот, кто знает, и то, что знает субъект (*Он знает ответ*);
- в ситуации преподавания – три участника: тот, кто преподаёт, тот, кому преподают, и то, что преподают (*Она преподаёт студентам стилистику*);

- в ситуации купли-продажи – четыре участника: продавец, покупатель, товар и деньги (*Бабушка продала нам дачу за один миллион рублей*);
- в ситуации аренды – пять участников: хозяин, арендатор, объект аренды, деньги, период времени (*Она сдала нам квартиру на один месяц за 30 тысяч рублей*) и т. д.

Типичные участники ситуации:

- субъект - одушевленный инициатор действия;
- объект - вещь, являющаяся объектом действия;
- контрагент - сила, против которой направлено действие;
- адресат - лицо, для которого совершается действие;
- пациент - вещь, которая испытывает эффективность действия;
- результат;
- инструмент - физическая причина действия/ стимул;
- источник - исходное состояние объекта до действия.

Типичные участники ситуации получили название **актантов**. Помимо основных участников, описание ситуации нуждается в дополнении: кроме актантов, ситуацию характеризуют **сирконстанты** (от франц. *circonstance* – 'обстоятельство, условие') – имена тех обстоятельств, на фоне которых разворачивается ситуация (место, время, условия, причина и т. д.).

Если актанты обозначают обязательных участников ситуации, то сирконстанты называют её факультативных участников. Так, например, ситуация, описываемая глаголом *говорить,* предполагает трёх участников-актантов: субъект, объект (содержание речи) и адресат – тот, кому говорят (*Он говорил ей о любви*). Без учёта этих трёх актантов семантика глагола не полностью раскрывается и предикатное слово не может быть правильно понято и употреблено. Учёт сирконстантов необязателен при осмыслении и употреблении слова, обозначающего данную ситуацию. Так, например, *говорить о любви* можно *в парке на скамейке, в купе поезда, при свете луны, во время лекции* и т. д. Ясно, что ни одно из этих обстоятельств не раскрывает новых аспектов в значении глагола *говорить* и не влияет на правильное его осмысление.

Актанты предикатных слов образуют определённую структуру: каждый актант занимает своё «место» в соответствии со степенью его обязательности для осмысления данного слова, с невозможностью опустить его при истолковании лексического значения.

Способность предикатного слова «притягивать» к себе определённое количество актантов обычно называют **семантической валентностью** слова. Семантическая

валентность соотносится с синтаксической валентностью, описывающей способность слов вступать в синтаксические связи с другими элементами предложения. Между семантической и синтаксической валентностями существуют достаточно регулярные соотношения. Эти отношения варьируются в зависимости от языка. Так, например, в русском языке каждый актант может выражаться практически всеми падежными формами. Но чаще всего субъект выражается формой именительного падежа, объект – формой винительного падежа, адресат – формой дательного падежа, а инструмент – формой творительного падежа (*Он пишет записку другу карандашом*). В английском соответствие более строгое.

Данное представление структуры языка указывает на тот факт, что синтаксис тесно связан с семантикой, и слово не имеет смысла без своего контекста, отсюда можно сделать вывод о том, что по форме исходного предложения можно попытаться получить его содержание. На этой идее и базируется предложенный в данной работе метод автоматической семантической разметки предложений.

Опишем подробнее теории, составившие базу рассматриваемого метода семантической разметки.

3.2 Семантика фреймов Ч. Филлмора

Представление структуры языка через предикатную лексику и актантную структуру слова описывает общие принципы отношений между словами и группами слов в предложении. Они присуще не только русскому, но и другим языкам. Для английского языка эти принципы описаны Чарльзом Филлмором в теориях падежной грамматики (1968) [9] и семантики фреймов (1976) [1]. Падежная грамматика, или «ролевая грамматика», – метод описания семантики предложения как системы семантических валентностей, в которой значение корневого глагола диктует роли, исполняемые зависимыми от него именными составляющими. Семантика фреймов – это расширение падежной грамматики. Ее базовая идея заключается в невозможности определения смысла изолированного слова. Согласно данной теории нельзя понять смысл слова самого по себе, если дополнительная информация о связанных с ним словах отсутствует. Например, слово *«sell»* не является однозначно осмысленным, если нет дополнительной информации о проведении коммерческой операции, неотъемлемыми участниками которой являются *Продавец* и *Покупатель,* а также объекты *Товары* и *Оплата*, помимо этого присутствуют связи между *Товаром* и *Оплатой*, *Товаром* и *Продавцом*, *Покупателем* и *Оплатой* и так далее. Так поведение слова в рамках той или иной ситуации порождает «семантический фрейм». Фрейм – это структура события, отношения или объекта в виде атрибутов и их

значений. Фрейм описывает характерные особенности образующего его понятия, так называемого ядра, а также особенности взаимодействия со связанными с ним объектами. Помимо общего описания фреймы задают точку зрения, с которой рассматривается ситуация. Например, *«sell»* в предложения *«John sold a car to Mary»* описывает ситуацию с точки зрения продавца, а *«buy»* в предложении *«Mary buy a car from John»* описывают ту же ситуацию с точки зрения покупателя.

Практической реализацией семантики фреймов является проект FrameNet [3, 4], также созданный при участии Ч. Филлмора. Этот ресурс будет описан подробнее в главе 4.

Ключевым элементом падежной грамматики, а, следовательно, и семантики фреймов является, образующий предложение глагол. Глагол – самостоятельная часть речи, которая описывает действие или состояние. Нас интересуют глаголы, выступающие в роли сказуемого в предложении. Традиционно выделяют различные лексико-семантические категории глагола. В различных языках глаголы могут иметь разные показатели, такие как время, залог, наклонение, валентность и другие. В рамках данной работы интересна его синтаксическая валентность. Именно она, и информация о ее пересечении с семантической валентностью, позволяет переходить от поверхностного синтаксического уровня к смыслу предложения.

Обозначим место глагола в предикатной лексике, описав фундаментальную классификацию предикатов Д.Ю.Апресяна [7].

3.3 ФКП Д.Ю. Апресяна и классификация английских глаголов Б. Левин

Как уже было сказано, все лексемы и другие эквивалентные им лексические единицы, в том числе многословные, делятся на два основных типа: предметные и предикатные. Когда-то казалось, что для всех типов лексем, возможны полные толкования. Впоследствии, при составлении толково-комбинаторных словарей, Д.Ю. Апресян пришел к выводу, что предметные и предикатные лексемы требуют двух разных стратегий толкования. Для предметных лексем нужно строить дифференциальные толкования[2], поскольку исчерпывающие невозможны, а для предикатных – исчерпывающие[3], потому что дифференциальные недостаточны. Предикаты должны быть истолкованы исчерпывающим и неизбыточным образом, потому что любой семантический компонент их толкования может оказаться объектом какого-либо правила взаимодействия значений.

[2] Имеются в виду толкования в терминах компонентного анализа.
[3] Имеются в виду толкования в виде разложения смысла на слои (собственно толкование, пресуппозиция), выделение семантических ролей предиката, см. «Семантические падежи» Филлмора.

Двум разрядам лексики, и, соответственно, двум типам толкований соответствуют и две разные семантические классификации языковых единиц. Предметным – таксономическая, а предикатным – фундаментальная. Далее речь пойдет о второй из них.

Фундаментальной называется такая классификация лексики, понятия которой имеют универсальный характер, то есть используются во всех лингвистических правилах – морфологических (категории вида, залога, наклонения), словообразовательных, синтаксических, семантических, сочетаемостных и других. Из них самыми важными являются семантические правила.

Центральные понятия ФКП – это лексема и лексикографический тип. Лексикографические типы – группы лексем с общими свойствами, через которые формулируются те или иные лингвистические правила.

Опишем основные принципы и основания. Классифицируются, в первую очередь, глаголы, потому что любой класс предикатов представлен каким-то количеством глаголов. При этом глаголы являются прототипическими именами действий[4], занятий, воздействий, процессов. Чем дальше от действий, тем чаще прототипическим представителем класса оказывается другая часть речи. Для свойств это прилагательные (красный, низкий, умный, смелый, волевой), а для параметров – существительные (высота, длина, масса; карьера, нация, профессия; запах, форма, цвет).

Сама классификация строится на строго семантических основаниях. Классы определяются через системообразующие смыслы. В их число входят:

- семантические примитивы. Их примеры: *делать, находиться, существовать, мочь, время* и т.п.;
- некоторые более сложные смыслы, например «цель»;
- некоторые семантические кварки, т. е. смыслы, более простые, чем семантические примитивы, и поэтому не имеющие соответствие среди лексем языка, но реально в нем существующие, например, кварк стативности[5].

Кроме них, в ряде определений используется нестрогое, но самоочевидное понятие раунда наблюдения (например, «действие» - один раунд наблюдения, «деятельность» - последовательность действий – несколько раундов наблюдения).

Классификация представляет собой нестрогую многоуровневую иерархию с многочисленными пересечениями классов. Классы упорядочиваются на основе принципа изоморфизма макро- и микромира языка: в самых общих чертах иерархическая структура

[4] Прототипическое действие — это глагол, у которого в вершине ассертивной части толкования (в собственном толковании) на последней ступени семантической редукции обнаруживается семантический примитив ‘делать’, причем, время существования ситуации, названной этим глаголом, укладывается в один раунд наблюдения.

[5] Глаголы со значением *быть* и *иметь* в плане содержания представляют собой кварк стативности.

значений всей глагольной лексики повторяет структуру значений многозначного глагола. В словарях значения сильно многозначных глаголов упорядочиваются в следующем направлении: прямые («свободные») ⇒ производные (переносные) ⇒ лексически или синтаксически связанные ⇒ грамматикализованные (в том числе связочные и чисто служебные). Пример многозначного глагола, на глаголе *входить*:

- *входить в комнату* (действие),
- *входить в транс* (процесс),
- *входить в комиссию* (состояние),
- *входить в ведро* (параметр, а именно — вместимость ведра).

Классы предикатов внутри словаря упорядочиваются так же, как значения внутри многозначного слова.

Было выделено свыше 15 основных классов ФКП, условно говоря, классов верхнего уровня. Например, к ним относятся:

- действия (*писать, идти*),
- деятельности (*торговать, воевать*),
- занятия (*играть, гулять*),
- воздействия (*размывать, прогревать*),
- события (*встретить, найти; происходить, случаться*),

и так далее.

Внутри классов выделяют подклассы. Приведем некоторые из них в качестве примеров. Действия делятся на физические, ментальные и речевые; такие же подклассы обнаруживаются и в классах деятельностей и занятий. Например, *дебатировать* (речевая деятельность), *разглагольствовать* (речевое занятие). Состояния могут быть физическими (*видеть*), материальными (*нуждаться*), волевыми (*хотеть*), ментальными (*знать*) и эмоциональными (*бояться*). Параметры бывают числовые (*весить, вмещать, длиться, достигать, занимать, насчитывать, составлять, стоить*) и качественные (*относиться к кому-л. хорошо <плохо>, питаться хорошо <плохо>*). Внутри событий выделяются происшествия (*случаться, происходить*).

Классы могут пересекаться друг с другом и по горизонтали, и по вертикали, так что в целом границы между ними размываются. Особенно хрупки границы между действиями, деятельностями, занятиями, воздействиями и процессами. Тем не менее, нельзя отказываться от выделения ясных прототипов там, где они есть.

Разным лексикографическим типам, то есть семантическим классам или подклассам предикатов, соответствуют разные наборы семантических, синтаксических и сочетаемостных свойств. Нас в большей степени интересуют синтаксические свойства, а

именно – модели управления. Модель управления в данном контексте – это схема, наглядно отражающая семантические и синтаксические актанты лексемы, а также способы их морфосинтаксического оформления, другим словами, какие части речи данное слово может иметь в качестве своих зависимых и в каких формах они должны стоять. Для примера распишем деятельности, обычно они имеют не более трех актантов:

- *бороться (кто, с кем, за что),*
- *вести переговоры (кто, с кем, о чем),*
- *торговать (чем, с кем)* и т. п.

Как видно из примера, точно так же проявляется актантная структура слова, описанная в семантике фреймов Ч.Филлмора.

Фундаментальная классификация Ю.Д. Апресяна классифицирует глаголы русского языка. Но разумно предположить, что подобные классификации должны существовать и для других языков. В ходе изучения предметной области, была обнаружена классификация глаголов английского языка, разработанная в 1993 году Бет Левин в Стэндфордском университете. Ее исследования показали, что существует корреляция между изменением смысла глагола и списком его возможных синтаксических вариаций. В итоге, было классифицировано свыше 3000 английских глаголов согласно их значению и поведению и выделено более 250 семантических классов глаголов с оригинальными списками синтаксических вариаций.

Эта классификация несколько отличается от ФКП, в первую очередь тем, что в основе разбиения на классы лежит синтаксическое поведение глаголов, а не их семантические признаки. Внутри классы сгруппированы по таким семантическим критериям как, тематические роли (напр., Агент, Объект, Реципиент) и элементы Лексической концептуальной структуры, задающей связи между лексическими элементами предложения через специальные теги, например «HAS_POSSESSION», «CAUSE», «TO» и другие. Общее у этих теорий в том, что глаголы сгруппированы в классы так, что внутри класса они подчиняются общим правилам управления, то есть имеют одинаковые семантические валентности и модели управления. Именно в этой точке классификация глаголов Б.Левин пересекается с семантикой фреймов Ч. Филлмора, что позволяет использовать эти теории совместно для анализа глаголов английского языка.

Дадим некоторые пояснения к сравнению ФКП и классификации Б.Левин. Безусловно, эти классификации построены на разных принципах: ФКП базируется на семантике предикатов, а классификация Левин на их синтаксическом поведении. Но заметим принципиальные отличия английского и русского языков. Русский язык богат на глаголы, их синонимы, различные формулировки предложений, он не имеет таких

жестких правил построения предложений, как английский. Возможно, именно, по этой причине классификация русских глаголов исключительно по синтаксическому поведению становится необъятной задачей. И по этой же причине, именно потому, что правила использования английских глаголов жестко заданы в языке, их можно группировать согласно их синтаксическому поведению, не привязываясь к их смыслу. Но, несмотря на такой способ классификации, семантика не теряется, она задается через тематические роли, свойственные классу. И забегая вперед, скажем, что совместное использование классификации глаголов Левин и теории семантики фреймов, через задание связей между ними, позволяет достаточно однозначно определять смыслы глаголов в классах.

Таким образом, сравнение ФКП и классификации Левин считаем правомочным, несмотря на различия в подходе к классификации, поскольку они были созданы для двух принципиально разных языков.

3.4 Совместное использование теорий

Выше были описаны основные теории, лежащие в основе метода автоматической семантической разметки. Ознакомимся подробней с самим методом.

Как уже было сказано, предлагаемый подход к разметке основан на идее о том, что синтаксис тесно связан с семантикой, а слово не имеет смысла без своего контекста, а значит, имея лишь синтаксическую структуру предложения можно попытаться получить его содержание.

В рамках подхода предлагается определять ситуационный фрейм, описывающий данное предложение. Это сопоставление осуществляется за счет определения принадлежности ключевого глагола предложения к тому или иному классу глаголов. Глаголам из разных классов, в свою очередь, соответствуют разные рамки валентности. Сравнив рамку валентности глагола с информацией о синтаксической структуре предложения, становится возможным приписать тематические роли словам и группам слов, составляющим предложение. Тематические роли в совокупности описывают некоторую стандартную ситуацию (или несколько ситуаций), так мы выходим на ситуационный фрейм, описывающий предложение. В рамках фрейма, уже определяются смыслы отдельных слов и словосочетаний, на основе описания актантов семантического фрейма.

Глава 4 Практическая реализация

В данной главе описан алгоритм создания автоматической семантической разметки и лексические ресурсы, используемые при его реализации.

4.1 Входные данные.

В первую очередь определим входные данные алгоритма. Предметом нашего исследования является предложение на естественном языке, поэтому в первую очередь ознакомимся подробнее с этим понятием.

Предложение — это минимальная единица языка, которая представляет собой грамматически организованное соединение слов (или слово), обладающее смысловой и интонационной законченностью. Предложение не всегда выражает мысль, оно может выражать вопрос, побуждение, волю, эмоции. В соответствии с этим предложения бывают повествовательные, вопросительные побудительные. Предложение считается простым, если содержит в себе одну предикативную единицу, если больше – сложным.

В текущей реализации метода автоматической семантической разметки входными данными является простое утвердительное предложение на естественном языке. В качестве языка исследования был выбран английский язык, ввиду существования крупных лингвистических ресурсов именно для этого языка.

В предложении выделяют поверхностную и глубинную структуры [13, 14] – это два способа представления (абстрактного описания устройства) предложения. Эти структуры противопоставляются друг другу, являясь представлением предложения на уровне поверхностного и глубинного синтаксиса соответственно. Для перехода от одной структуры к другой используются специальные правила перестройки – трансформации, которые сохраняют лексический состав предложения, но могут изменять грамматические значения, например, переставлять слова местами, добавлять или снимать некоторые служебные слова.

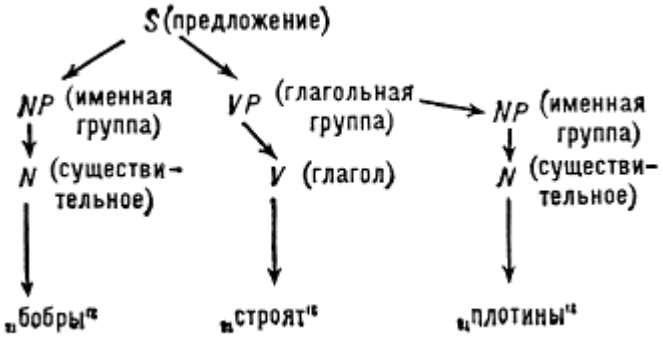
Поверхностная структура лучше отражает лексический состав, синтаксические связи и линейный порядок лексем самого́ конкретного предложения, тогда как глубинная структура приближена к описанию смысла предложения. Поэтому иногда термином «поверхностная структура» называют также и само конкретное предложение, а глубинную структуру именуют «семантической структурой».

Рассмотрим подробнее глубинную структуру. Определение синтаксической структуры предложения является одним из основных этапов в цепочке анализа текста. Зная структуру предложения, мы можем сделать более глубокий анализ и другие

интересные вещи. Например, мы можем создать систему автоматического перевода, или, как в нашем случае систему семантической разметки.

Глубинная структура позволяет отразить смысловую близость ряда предложений, которые содержат одни и те же лексические единицы и отличаются друг от друга только некоторыми грамматическими значениями. Так, например, единая глубинная структура у предложений «Бобры строят плотины» и «Плотины строятся бобрами». Глубинная структура формально изображается в виде размеченной скобочной записи $[[[\text{бобры}]_N]_{NP}[[\text{строят}]_V[[\text{плотины}]_N]_{NP}]_{VP}]_S$ или в виде, так называемого, дерева составляющих (рис. 1). Оба графических средства представляют синтаксическое устройство предложения.

Рисунок 1. Дерево составляющих



Дерево составляющих может быть получено из предложения в результате его синтаксического анализа. Инструментами синтаксического анализа являются синтаксические анализаторы, или парсеры. На вход синтаксическому анализатору подается исходное предложение, а на выходе будет получено его синтаксическое дерево и другая дополнительная информация. Таким образом, синтаксический анализатор обеспечивает переход от поверхностной структуры предложения к его глубинной структуре. В нашем случае, на вход синтаксическому анализатору подается простое утвердительное предложение на английском языке.

4.2 Используемые ресурсы

Важную часть данной работы составляют лингвистические ресурсы, используемые при реализации метода автоматической семантической разметки, а именно: WordNet, FrameNet, VerbNet. Также был задействован синтаксический анализатор Stanford Syntax Parser. Конечная реализация метода автоматической семантической разметки является интеграцией озвученных ресурсов. Использованные ресурсы описаны подробно в данной главе.

4.2.1 Синтаксический анализатор

Обработка входного предложения синтаксическим анализатором – первый этап работы алгоритма семантической разметки предложения. Необходимость использования синтаксического анализатора была обусловлена следующими задачами: определение ключевого предиката и ролей участников, для чего было необходимо ключевое слово предложения и его зависимые слова; помимо этого необходимо было определить части речи и начальные формы слов в предложении для их последующей обработки.

В качестве синтаксического анализатора в текущей реализации был выбран Stanford Syntax Parser [15]. Это многоязычный синтаксический анализатор, относится к анализаторам вероятностного, или статистического типа. Вероятностные парсеры используют знания о языке, полученные из ручной разметки предложений. И анализ нового предложения осуществляется на основе информации об уже известных похожих предложениях. На данном этапе развития автоматического синтаксического анализа статистические парсеры еще совершают ошибки, но в целом с поставленной задачей справляются достаточно хорошо.

Существует большое количество моделей, описывающих человеческие языки, из которых выделяют две наиболее популярных:

1) грамматика составляющих (constituency grammar);
2) грамматика зависимостей (dependency grammar).

В Stanford Syntax Parser совмещены обе эти модели. Согласно первой из них предложение разбивается на более мелкие структуры – группы, а затем каждая группа на более мелкие группы, и так далее, пока не будут достигнуты отдельные слова или словосочетания.

Рисунок 1 отражает разбор предложения именно с точки зрения грамматики составляющих:

- разбиваем предложение на именную и глагольную группы;
- именную группу разбиваем на существительное («бобры»);
- глагольную группу – на глагол («строят») и вторую именную группу;
- вторая именная группа представлена существительным («плотины»).

Такому строению дерева составляющих уделено особое внимание в алгоритме разметки. Сначала анализируется все предложение, затем ключевая группа, затем зависимые от нее группы, зависимые группы зависимых групп и так далее до единичных лексем. Такой анализ производится с целью выявления в зависимых группах самостоятельных ситуаций в рамках общей задачи определения ситуационного фрейма предложения и ролей его участников.

Вторая модель представления предложения – грамматика зависимостей. В этой грамматике порядок слов не важен, т.к. важно определить только то, от какого слова зависит каждое слово в предложении и тип этих связей. На примере это выглядит так (рис. 2):

Рисунок 2. Зависимости между словами в предложении

В терминах используемого синтаксического анализатора Stanford Syntax Parser полученное дерево составляющих и зависимости между словами в предложении выглядят следующим образом (рисунки 3,4 и 5):

Рисунок 3. Синтаксическое дерево

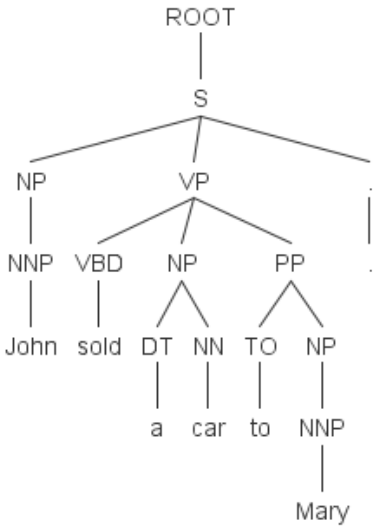


Рисунок 4. Скобочная запись

```
(ROOT
    (S
        (NP (NNP John))
        (VP
            (VBD sold)
            (NP (DT a) (NN car))
            (PP (TO to) (NP (NNP Mary)))
        )
        (. .)
    )
)
```

Рисунок 5. Зависимости между словами в предложении, (2 варианта учета предлогов)

```
[ nsubj(sold-2, John-1),
  root(ROOT-0, sold-2),
  det(car-4, a-3),
  dobj(sold-2, car-4),
  prep_to(sold-2, Mary-6)
]
```

```
[ nsubj(sold-2, John-1),
  root(ROOT-0, sold-2),
  det(car-4, a-3),
  dobj(sold-2, car-4),
  prep(sold-2, to-5),
  pobj(to-5, Mary-6)
]
```

Получение синтаксического дерева предложения и его последующий анализ позволяет решить следующие задачи:

- выявить ключевое слово предложения;
- определить зависимости между словами;

- определить части речи и начальные формы слов в предложении.

Первые два пункта необходимы для дальнейшего определения ситуационного фрейма, описывающего предложение.

Последний пункт необходим для определения смыслов слова, а именно: по начальной форме слова и его части речи, становится возможным получить смыслы данного слова из лексической базы.

4.2.2 Лексическая база. WordNet 3.0.

В качестве лексической базы выбран ресурс WordNet, версии 3.0 [6]. **WordNet** — это электронный тезаурус для английского языка, разработанный в Принстонском университете и выпущенный вместе с сопутствующим программным обеспечением.

WordNet совместим с рядом других лексических ресурсов, например, с используемыми в данном проекте FrameNet и VerbNet. Связи с сущностями ресурсов задаются через отображение на них синсетов WordNet.

Словарь состоит из 4 семантических сетей для основных частей речи: существительного, глагола, прилагательного и наречия. Базовым структурным элементом WordNet является не отдельное слово, а так называемый синонимический ряд (синсет), объединяющий слова со схожим значением и по сути своей являющимися узлами сети. Для удобства использования словаря человеком каждый синсет дополнен определением и примерами употребления слов в контексте. Слово или словосочетание может появляться более чем в одном синсете и иметь более одной категории части речи.

Синсеты связаны между собой семантическими и лексическими отношениями, такими как:

- гиперонимы – более широкие понятия по отношению к данному (*breakfast → meal*);
- гипонимы – понятия, уточняющие данное (*meal → dinner*);
- тропонимы – связь, определенная для глаголов, описывает более узкую манеру совершения некоторого действия (*run → sprint*);
- холонимы – слова, стоящие на месте «целого» в связке «часть-целое»; (*leg → table*), иногда они выражены через отношение типа «member-of» (*pilot → crew*);
- меронимы – слова, стоящие на месте «часть» в связке «часть-целое» (*body → arm*), также могут быть выражены отношением типа «has-member» (*faculty → professor*):

- антонимы – слова с противоположным смыслом, определены для прилагательных (*beautiful → ugly*);
- отношение схожести, определенное для прилагательных (beautiful → pretty)
- причинно-следственные связи – определенные для глаголов отношения, в рамках которых одно действие неизбежно влечет другое действие либо является причиной другого действия;

Среди них особую роль играют отношения гипернимии и гипонимии: они позволяют организовывать синсеты в виде семантических сетей. Для разных частей речи родовидовые отношения могут иметь дополнительные характеристики и различаться областью охвата. Часть алгоритма семантической разметки построена именно на отношениях гипернимии.

Объем базы данных WordNet составляет около 150 000 уникальных слов различных частей речи, а также порядка 250 000 семантических отношений между синсетами. WordNet в настоящее время принимается в качестве стандартного ресурса для систем обработки естественного языка. Простота структуры позволяет сравнительно просто встраивать эту базу знаний в прикладные системы. Интерес исследователей-лингвистов подтверждается и тем, что на данный момент уже созданы аналоги WordNet для других языков, в том числе и для русского, и работы по их улучшению не прекращаются. Кроме того, для WordNet создано большое количество расширений и отображений, связывающих его с другими мощными лексическими ресурсами, что позволяет использовать их совместно для решения поставленной задачи.

В ходе данной работы были использованы отображения WordNet на такие лексические ресурсы как, FrameNet и VerbNet, а также внутренние переходы между версиями WordNet. Часть используемых отображений была получена из внешних источников и использовалась в неизменном виде, часть была подвержена доработке, часть разработана в рамках данного исследования в результате проведения соответствующей лингвистической работы.

В рамках данной работы для обращения к базе данных WordNet 3.0 использовались следующие средства:

- внешняя библиотека JWNL [16], или Java WordNet Library, API[6] для доступа к лексической базе из среды разработки;
- графический пользовательский интерфейс WordNet TreeWalk [17], разработанный в 2007 году Бернардом Боу (Bernard Bou). Иллюстрация данного графического

[6] API (интерфейс программирования приложений) — набор готовых классов, процедур, функций, структур и констант, предоставляемых приложением (библиотекой, сервисом) для использования во внешних программных продуктах. Используется программистами для написания всевозможных приложений.

интерфейса приведена в Приложении А, ее наличие позволит составить более полное представление о ресурсе WordNet.

4.2.3 FrameNet

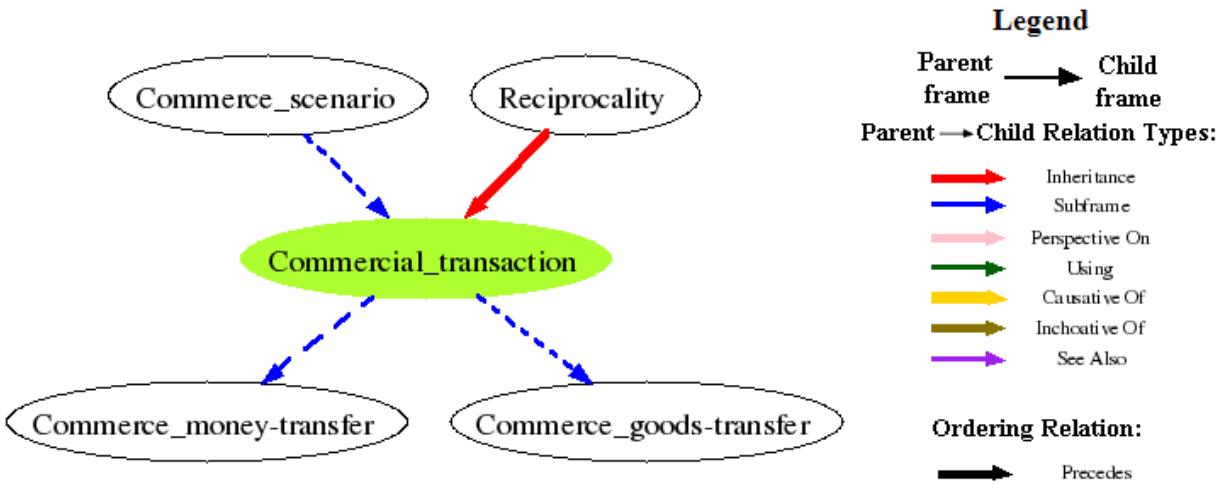
Другим основным ресурсом, использованным при реализации метода автоматической семантической разметки, является FrameNet [3, 4]. Это электронный ресурс, являющийся практической реализацией теории семантики фреймов Ч.Филлмора. Напомним, что семантика фреймов – это расширение падежной грамматики. Ее базовая идея заключается в невозможности определения смысла изолированного слова. Согласно данной теории нельзя понять смысл слова самого по себе, если дополнительная информация о связанных с ним словах отсутствует. Так поведение слова в рамках той или иной ситуации порождает «семантический фрейм». Фрейм – это структура события, отношения или объекта в виде атрибутов и их значений. Фрейм описывает характерные особенности образующего его понятия, так называемого ядра, а также особенности взаимодействия со связанными с ним объектами.

FrameNet представляет собой лексическую базу данных для английского языка[7], основанную на аннотированных примерах предложений из реальных текстов. Структурной единицей FrameNet является **фрейм**. Он содержит название, описание, список участников и атрибутов, каждый из которых также имеет описание и примеры употребления. Участники разделены на корневые и некорневые, что соответствует, в некоторой степени, актантам и сирконстантам ФКП. В FrameNet они носят единое название фрейм-элементов (*с англ.* frame-elements).

Между семантическими фреймами заданы отношения нескольких типов, например, фрейм Commercial_transaction[8] является «подфреймом» фрейма Commerce_scenario[9], его подфреймами являются Commerce_goods-transfer[10] и Commerce_money-transfer[11], и он наследуется от фрейма Reciprocality[12]. Описанные связи между фреймами представлены на рисунке 6.

[7] В данной работе используется база данных именно для английского языка
[8] рус. Коммерческие операции
[9] рус. Сценарий коммерческих операций
[10] рус. Коммерция, операции с товарами
[11] рус. Коммерция, операции с денежными средствами
[12] рус. Взаимодействие

Рисунок 6. Связи между фреймами.



Другой термин, активно используемый в FrameNet – **лексическая единица** (в оригинальном варианте «lexical unit»). Это слово с указанной частью речи и приписанным к нему определенным значением; многозначные слова представлены несколькими лексическими единицами, например слово “run” представлено и как существительное и как глагол. Лексические единицы относятся к тому или иному фрейму. Одна и та же лексическая единица может быть отнесена к нескольким фреймам одновременно.

Следующее важное понятие, определенное в FrameNet и активно используемое в данной работе – понятие **семантического типа**. Оно было добавлено в FrameNet для записи информации, которая не представлена в явном виде иерархии фреймов и фрейм-элементов. Семантические типы задают базовую типизацию для слов-наполнителей фрейм-элементов. Эти базовые типы подразумеваются иерархией фреймов, но не могут быть приписаны произвольно расположенным в иерархии фрейм-элементам, поскольку могут меняться от фрейма к фрейму при совпадении названия. Примеры семантических типов: *Sentient, Duration, Manner*. Так, в фрейме *Building* фрейм-элементу *Agent* сопоставлен семантический тип *Sentient*, а фрейм-элементу *Created_Entity* – тип *Artifact*. Что согласуется с логикой данной ситуации: при построении чего либо (напр., *здание, отношения, муравейник*) его инициатором будет являться *одушевленная сущность*, а объектом строительства – некоторый *неодушевлённый объект*. Семантические типы добавлены в FrameNet преимущественно, чтобы помочь при парсинге фреймов и автоматическом распознавании фрейм-элементов.

Семантические типы согласуются с понятиями, определенными в WordNet, и в данной работе по текстовому описанию семантических типов было разработано отображение между ними и синсетами WordNet. Оно позволяет связывать достаточно абстрактные актанты фрейма с вполне конкретными понятиями из WordNet. Недостатком

использования семантических типов можно назвать неполное покрытие фреймов этими типами. Ввиду этого, использование описанного отображения не дало желаемых результатов, и навело на мысль о необходимости дополнения FrameNet.

Лексическая база FrameNet содержит около 1,200 семантических фреймов, 13,000 лексических единиц (для сравнения, в WordNet определено свыше 150,000 уникальных понятий) и свыше 190,000 примеров предложений.

Лексический ресурс снабжен удобным веб-интерфейсом, его иллюстрация приведена в Приложении А. Нельзя не заметить, что FrameNet – «живой» проект, он регулярно и дополняется новыми фреймами, и обновляются уже существующие.

В текущей реализации была использована версия лексической базы 1.3, доступная на момент начала работы. Также при разработке метода были задействованы ряд отображений между различными сущностями FrameNet (фреймами, актантами и семантическими типами) и синсетами WordNet. Они будут описаны подробнее в последующих главах.

Обобщив вышесказанное, можно отметить, что FrameNet является неотъемлемой частью алгоритма семантической разметки. При создании семантической разметки предложения ему ставится в соответствие описание ситуации или ситуаций. Каждая ситуация, которая может быть приписана предложению – это фрейм из лексического ресурса FrameNet, практической реализации теории семантики фреймов Ч. Филлмора.

В данном исследовании работа велась с англоязычной базой данных FrameNet, но кроме нее существуют базы данных и для других языков: китайского, немецкого, японского, испанского и других. Эти проекты могут быть полезны для семантической разметки предложений из этих языков.

4.2.4 VerbNet

Следующий инструмент, использованный при реализации метода автоматической семантической разметки – лингвистический ресурс VerbNet [5]. Это совместимая с WordNet лексическая база глаголов, явно задающая их синтаксическую и семантическую информацию. VerbNet построен на основе классификации глаголов Б. Левин, но более детализирован, чем исходная классификация. Для VerbNet исходные классы Б.Левин были переработаны, добавлены подклассы для организации семантической и синтаксической связности среди глаголов-участников. В результате, классы могут быть рассмотрены на

разных уровнях детализации, в зависимости от задачи NLP[13], для решения которой они применяются.

Напомним, что фундаментальное предположение, лежащее в основе классификации, заключается в том, что синтаксическое поведение глаголов, как предикатов с актантной структурой, напрямую зависит от их семантики. VerbNet ассоциирует семантику глагола с синтаксическими фреймами, добавляя семантические ограничения к элементам синтаксических фреймов класса. Глаголы, входящие в один и тот же класс, делят общие синтаксические фреймы и, таким образом, они обладают одним и тем же синтаксическим поведением. Это важное свойство, указывающее на то, что VerbNet может быть использован для расширения покрытия лексической базы FrameNet. Кратко говоря, идентифицируя VerbNet-классы, согласованные с фреймами FrameNet, мы получаем возможность анализировать предложения с глаголами, не покрытыми FrameNet. Этого можно добиться, используя транзитивность отношений между VN-классами: глаголы, которые принадлежат одному и тому же классу Левин, скорее всего, относятся и к одному FN-фрейму, и в этом случае их семантические фреймы могут быть проанализированы, несмотря на то, что явно глагол не присутствует в FrameNet.

Каждый класс VerbNet содержит:

- список глаголов – членов класса;
- список синтаксических фреймов, описывающих синтаксическое поведение глаголов – членов класса.
- список тематических ролей, которые встречаются в синтаксических фреймах класса; каждый синтаксический фрейм содержит:
    - синтаксическое описание, дающее представление о порядке слов и их частях речи в предложениях, соответствующих данному фрейму;
    - список ролей, соответствующих элементам синтаксического фрейма;
    - семантические ограничения на синтаксические элементы;
    - пример предложения с описанным синтаксическим фреймом.

VerbNet снабжен веб-интерфесом, который активно использовался в ходе данной работы. Он включает возможность поиска классов по их глаголам-участникам, просмотра классов и составляющих классы элементов: список фреймов с их атрибутами, членов класса, подклассов класса и так далее. Иллюстрация графического интерфейса приведена в Приложении А.

[13] NLP (natural language processing) - общее направление искусственного интеллекта и математической лингвистики, изучающее проблемы компьютерного анализа и синтеза естественных языков.

Так, на основе классификации Б. Левин был развит мощный инструмент для изучения английских глаголов. Кроме того, для него создан ряд отображений в другие широко используемые ресурсы, такие как FrameNet и WordNet.

В результате изучения VerbNet стандартного api для инструмента обнаружено не было, поэтому оно было разработано в рамках данной работы. Были реализованы основные функции, такие как:

- считывание и парсинг файлов с VN-классами;
- получение списка классов по смыслу глагола-участника;
- обход фреймов класса, получение их синтаксических структур, списка ролей и т. д.
- все api для работы с отображениями.

4.3 Выходные данные

На данном этапе описания алгоритма определены основные термины и понятия, теперь можно описать выходные данные, получаемые в результате выполнения алгоритма автоматической семантической разметки. Под семантической разметкой подразумевается как описание предложения в целом, так и описание его отдельных слов и словосочетаний. Таким образом, в разметке можно выделить несколько уровней означивания:

а) семантический фрейм, соответствующий предложению; в том числе, распределение слов и словосочетаний предложения по семантическим ролям фрейма;

б) разрешение смысловой неоднозначности слов и словосочетаний предложения. Данное означивание предполагает сопоставление лексическим единицам предложения смысла (или смыслов) WordNet.

Кроме того, для удобства представления итогового означивания, на выходе алгоритма предоставляется информация о синтаксическом устройстве означенного предложения.

В качестве формата, позволяющего предоставить одновременно несколько уровней означивания, было выбрано представление в виде xml-файла.

Ранее говорилось, что предложению может соответствовать несколько семантических фреймов. От фрейма зависит распределение семантических ролей словам в предложении, что может повлиять на распределение смыслов WordNet. С расчетом на подобную многозначность введено понятие *контекста*.

Выходной XML-файл представляет собой список возможных контекстов входного предложения. Под контекстом подразумевается совокупность семантического фрейма,

синтаксического класса VerbNet и синтаксического фрейма из этого класса (рис. 7), они описаны в соответствующих атрибутах тега CONTEXT.

Рисунок 7. Элемент CONTEXT в файле с означиванием

```
<CONTEXTS>
    <CONTEXT
        fn_frame="Commerce_buy"
        vn_class="get-13.5.1"
        vn_frame="NP V NP PP.asset">
        ...
    </CONTEXT>
</CONTEXTS>
```

В рамках каждого контекста в теге SENTENCE_CORPUS описывается предложение. Предложение разбито на лексические единицы. Теги, описывающие лексические единицы, согласуются с соответствующими им синтаксическими составляющими, полученными при анализе предложения синтаксическим парсером. Сложные лексические единицы, такие как группы слов, не являющиеся устойчивыми словосочетаниями, разбиваются на более простые – слова. Каждой лексической единице, имеющей смыслы в WordNet, приписываются те из них, что были отобраны алгоритмом для данного контекста. В описании смысла дано его текстовое описание и специальная техническая информация, такая как часть речи и offset – уникальный номер в базе данных WN. Рисунок 8 дает представление об описании лексических единиц в выходном файле.

Рисунок 8. Элемент с означиванием предложения

```
<SENTENCE_CORPUS> ...
    <NP order_num="6_7_8" phrase="three_million_dollars">
        <CD phrase="three"> ... </CD>
        <CD phrase="million">
            <SYNSETS>
                <SYNSET
                    description="the number that is represented
                                 as a one followed by 6 zeros"
                    offset="13751533"
                    pos="noun" />
                <SYNSET ... />
            </SYNSETS>
        </CD>
        <NNS phrase="dollar">
            <SYNSETS> ... </SYNSETS>
        </NNS>
    </NP>
</SENTENCE_CORPUS>
```

Формат выходных данных представлен. Как видно из полученной семантической разметки, она содержит переработанную информацию из всех описанных ранее лексических ресурсов. Этот факт подразумевает их интеграцию, о которой и будет сказано далее.

4.4 Интеграция лексических ресурсов.

При попытке совместного использования ресурсов были обнаружены концептуальные и технологические различия между ними. Ресурсы говорят об одном и том же – о языке, но разными словами. Это закономерно, потому что они решают разные задачи:

- WordNet дает описание с лексической точки зрения, он является словарем понятий и задает их иерархию;
- FrameNet дает ситуационного описание языка, то, как понятия языка участвуют в различных ситуациях и какие роли принимают;
- VerbNet ориентирован на описание моделей управления, представляет язык с позиции глагола, образующего предложение.

Интеграция этих ресурсов дает более полное представление о языке, чем каждый из них по отдельности, позволяет представить разрозненные ресурсы как единую лингвистическую базу данных.

Сложность интеграции этих лингвистических ресурсов заключается в разнице организации представления данных в них. В каждом ресурсе знания представлены в собственном формате, для которого разработано собственное api. В нашем случае для интеграции ресурсов необходимо было выделить единые сущности и провести связи между ними, таким образом, создав дополнительный слой перехода от одного формата к другому. Такая работа была проведена и связи между ресурсами были заданы в текстовых или xml-файлах специального вида, которые далее названы отображениями между лингвистическими ресурсами.

Следует отметить, что проблема связи данных ресурсов для их совместного использования была известна и ранее, рассматривалась разными лингвистами задолго до начала данного исследования. Поэтому ряд используемых файлов-связок между ресурсами взят из других работ.

Отображения между лингвистическими ресурсами являются важной частью реализации алгоритма автоматической семантической разметки. Помимо отображений между разными ресурсами, во время работы использовались файлы-связки для перехода между сущностями внутри одного ресурса. Этим ресурсом был WordNet, ввиду

использования разных его версий в разных частях работы, возникла необходимость перехода между версиями ресурса.

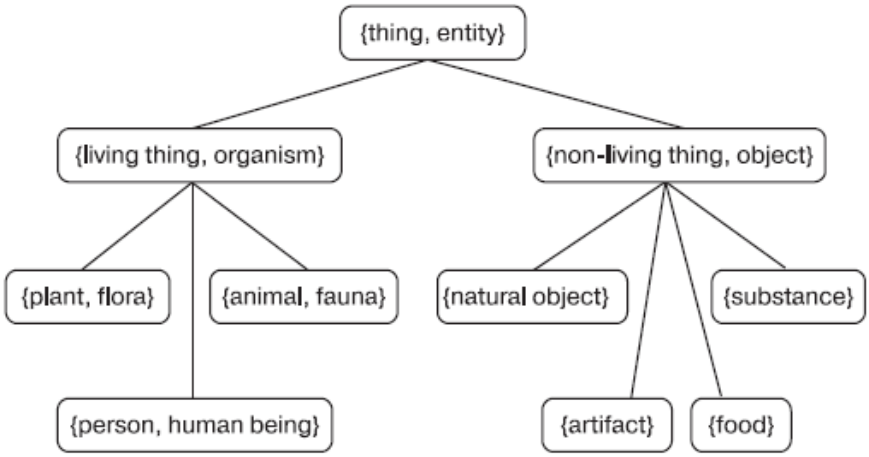
Нельзя не отметить необходимость использования интерфейсов для работы с отображениями, все они были разработаны в рамках данного исследования.

4.4.1 Отображение между FrameNet и WordNet

Большая доля используемых отображений приходится на отображения между синсетами WordNet и разными сущностями ресурса FrameNet.

Отметим, что именно при работе с отображениями для WordNet активно использовались упомянутые ранее отношения гипернимии и гипонимии. Это значит, что при отображении некоторого синсета на фрейм FrameNet или любую другую сущность подразумевается отображение и конкретного понятия и всего дерева под ним. Покажем это на примере самого верхнего понятия WordNet «*Entity*», понятие и дерево иерархии на несколько уровней вниз представлены на рисунке 9:

Рисунок 9. Иерархия понятий WordNet



Согласно рисунку видно, что если некоторый объект будет признан «одушевленным» и отображен на узел «*living thing*», но он может оказаться и животным, и растением и человеком. Это важный момент, который стоит учитывать, при разметке предложений.

В работе были использованы четыре различных отображения между синсетами WordNet и FrameNet: с фреймами, их актантами, с глагольными лексическими единицами и с семантическими типами. Наиболее значимые аспекты каждого из них описаны в данной главе.

MapNet 0.1

Это отображение является сторонней разработкой. Его авторы – Даниэл Пигхин (*Daniele Pighin*) и Сара Тонелли (*Sara Tonelli*), год создания MapNet – 2009 [18].

С его помощью сопоставляются *синсеты* WordNet версии 1.6, *фреймы* и *лексические единицы* FrameNet версии 1.3. Отображение задано в текстовом файле, каждая строчка которого содержит имя семантического фрейма и соответствующие ему лексическую единицу и синсет. Пример такой строки: *«Age old.a a#00917756»*.

Объем – 5162 отображения.

Данное отображение между ресурсами является результатом автоматической генерации, по данным авторов, оценочная точность созданных связей – 79,4 %. Подробно о методологии его составления можно прочесть в статье *«New Features for FrameNet - WordNet Mapping»* [19].

Отображение для глаголов

Данное отображение по своей сути аналогично предыдущему. Оно задает связи между *синсетами* WordNet, описывающими глаголы, и *фреймами* FrameNet.

Формат отображения – текстовый файл объемом 3,094 записи. Каждая запись соответствует одной глагольной лексической единице в FrameNet и содержит следующие поля:

- *фрейм* – имя соответствующего фрейма FrameNet;
- *глагол* – глагольная лексическая единица FrameNet
- идентификатор смысла WordNet, или так называемый *sensekey*. Это тот смысл из WordNet, который соответствует глаголу (второму параметру строки) в рамках ситуации, описанной фреймом (первым параметром).

Пример строки со связью фрейма и глагола: *«choosing select select%2:31:00::»*.

Отображение было построено для версии 2.0 ресурса WordNet. Но, несмотря на использование WordNet версии 3.0, дополнительных преобразований версии при работе с данным отображением не требовалось ввиду особенности параметра «sensekey»: он независим от версии.

Данное отображение, как и предыдущее – сторонняя разработка, его подробное описание дано в статье от его авторов Lei Shi and Rada Mihalcea, *«Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing»* [20].

Отображение между синсетами WordNet и семантическими типами FrameNet

Это отображение отличается от описанных ранее тем, что задает связи между *синсетами* WordNet и *семантическими типами* FrameNet, а не фреймами. Оно было создано полностью в рамках данной работы. Идея отображения почерпнута из самого

ресурса FrameNet. В некотором виде связи между синсетами и семантическими типами были заданы в одном из сопровождающих файлов проекта, в файле «*semtypes.xml*». Он содержит описание семантических типов FrameNet. Каждый тип описан тегом <*semType*>, ключевые элементы в описании типа: *название типа* и его *толкование*. Иерархия между семантическими типами также задана, через уникальные идентификационные номера. Пример описания одного из семантических типов из файла *«semtypes.xml»* представлен на рисунке 10.

Рисунок 10. Описание семантического типа

```
<semType ID="5" name="Sentient" abbrev="Sentient">
    <definition>
        Marks FEs whose fillers are sentient beings;
        dogs, definitely; bacteria, probably not;
        comatose people, planeria, maybe.  This
        is a common-sense type, with fuzzy boundaries,
        not a precise bio-medical category.
    </definition>
    <superTypes>
        <superTypes superTypeName="Animate_being" supId="65"/>
    </superTypes>
</semType>
```

Часть отображения была построена на основе толкований семантических типов. Любое толкование попадает под одну из следующих ситуаций:

а) в толковании типа явно указана ссылка на соответствующий ему синсет. Как, например, для семантического типа Content (рис. 11). В этом случае в соответствие типу ставится синсет из толкования.

Рисунок 11. Элемент толкования семантического типа

```
<definition>
    WN synset: content, cognitive content, mental object
</definition>
```

б) в описании типа нет явной ссылки на синсет, но он смысл семантического типа интуитивно понятен и синсет может быть установлен путем аналитической деятельности: поиска подходящего синсета в иерархии WordNet. Данный тип толкования вызывает большой объем чисто лингвистической работы.

в) ссылки на синсет нет, и он не был установлен по той или иной причине, например, потому что в WordNet нет подходящего смысла. В таком случае в соответствие типу ставится «-1».

Таким образом, было означено 46 семантических типов из 73, стоит отметить, что в отображение были добавлены лишь наиболее очевидные сопоставления, дополнительная

лингвистическая работа позволит создать большее покрытие WordNet семантическими типами, и как результат, задать больше соответствий между типизированным фрейм-элементам и синсетами.

Отображение между синсетами WordNet и актантами фреймов FrameNet

Несмотря на наличие описанных отображений задача покрытия фреймов FrameNet (их фрейм-элементов) синсетами WordNet оказалась нерешенной в достаточной степени. Решить ее за счет за счет покрытия WordNet семантическими типами не удалось, поскольку недостаточное число фрейм-элементов в фреймах FrameNet имели семантические типы. Поэтому было решено создать отображение, напрямую связывающее синсеты WordNet с фрейм-элементами фреймов.

Принцип его построения прост – для выбранного фрейма обходятся все его фрейм-элементы и каждому, согласно его толкованию, ставится в соответствие подходящий синсет или синсеты. Выбранный синсет должен удовлетворять условию: более узкие понятия данного (гипонимы), также должны подходить на роль наполнителя фрейм-элементу.

В рамках данного исследования был вручную создан вариант подобного отображения. Это отразилось на количестве описанных таким образом фреймов, были описаны несколько фреймов из предметной области «Коммерция», выбранной для тестирования. Добавление данного отображения положительно сказалось на результатах тестирования алгоритма на примерах из выбранной предметной области, значительно улучшив их распознавание.

Отмечу, что в рамках данной работы было достаточного небольшого объема отображения, и поэтому задача его более тщательной проработки здесь не решалась. При дальнейшей разработке метода предлагается автоматизация создания данного отображения, либо использование его готовых вариантов. С одним из них можно ознакомиться в источнике [21].

4.4.2 Отображение между VerbNet и WordNet

В качестве отображения между ресурсами VerbNet и WordNet выступают глаголы-участники классов VerbNet, поскольку в функционал ресурса включены ссылки на конкретные смыслы глаголов в WN. Например, класс *«get-13.5.1»*, описывающий процессы получения, связан с первым смыслом глагола «купить» (*buy (WN 1)*) и третьим и

седьмым смыслами глагола «найти» (*find (WN 3,7)*) и данная информация получена напрямую из класса.

4.4.3 Отображение между VerbNet и FrameNet

Это отображение было получено с официального сайта VerbNet, где в качестве его создателей заявлены Andy Dolbey и Russell Lee-Goldman. Отображение представляет собой xml-файл, который содержит в себе описание пар «*фрейм – класс VerbNet*». Для каждой пары задано соответствие семантических ролей, встречаемых в VN-классе, и фрейм-элементов в FN-фреймах. Например, для класса *«get-13.5.1»* это соответствие выглядит следующим образом (рис. 12):

Рисунок 12. Элемент отображения между семантическим фреймом и классом глаголов

```
<vncls class='13.5.1' fnframe='Commerce_buy'>
  <roles>
    <role fnrole='Buyer' vnrole='Agent'/>
    <role fnrole='Goods' vnrole='Theme'/>
    <role fnrole='Money' vnrole='Asset'/>
    <role fnrole='Seller' vnrole='Source'/>
    <role fnrole='Recipient' vnrole='Beneficiary'/>
  </roles>
</vncls>
```

Отметим, что при использовании данного отображения была замечена его неполнота и устранены замеченные пробелы. Именно после этого в указанном на рисунке примере появились пары ролей *«Money-Assert»*, *«Seller - Source»*, *«Recipient - Beneficiary»*, которых ранее не было. В ходе работы были дополнены еще несколько фреймов для тестирования примеров из выбранной предметной области. Но в целом задача дополнения данного отображения является отдельной лингвистической задачей и в данной работе не решалась.

Это отображение выполняет важную роль – позволяет сопоставить конкретному слову (группе слов) в предложении сопоставить фрейм-элемент из фрейма FrameNet. Так отсеиваются неподходящие синтаксические вариации класса, называемые ранее фреймами класса глаголов. Например, в предложении «*Carmen buy dress for $ 10*» поверхностная синтаксическая структура будет: *NP V NP for NP* (где NP – noun phrase, V - verb), ей в классе глаголов соответствуют 2 фрейма: 1) *Agent V Theme {for} Beneficiary* и 2) *Agent V Theme {for} Assert*, каждая из которых отображает сочетание слов «$ 10» либо в фрейм-элемент *Recipient* либо в элемент *Money*. И здесь отображение FrameNet и WordNet, даст понять, что *«$10»* не может быть существом одушевленным, а значит, соответствует фрейм-элементу *Money*, а не *Recepient*.

4.5 Алгоритм разметки

Алгоритм семантической разметки обращается к единой лексической базе, полученной в результате интеграции лексических ресурсов.

Процесс разметки условно можно разделить на два этапа. Первый – подготовительный, на данном этапе с помощью синтаксического парсера выделяется синтаксическая структура предложения, определяется ключевое слово, зависимости между словами, а также начальные формы слов и их части речи. Затем, с использованием полученной информации, всем словам предложения приписываются синсеты WordNet. Также на данном этапе по ключевому слову определяются подходящие семантические фреймы FrameNet и классы глаголов VerbNet. Итог подготовительного этапа – первоначальное означивание, без фильтрации «лишних смыслов».

В качестве примера возьмем предложение ***«Carmen bought a dress for $ 50»****, и далее будем обращаться именно к нему. После подготовительного этапа ему будет приписана синтаксическая структура «****NP V NP PP****», где NP – noun phase, V – verb, PP – preposition phrase, PP = {for} + NP; семантический фрейм – «Commerce_buy» из FrameNet, описывающий ситуацию покупки; класс глаголов «get-13.5.1» из VerbNet.*

Второй этап – собственно означивание, в результате которого отсеиваются неподходящие предложению семантические фреймы и неверные (в контексте предложения) смыслы слов и словосочетаний.

В первую очередь синтаксическая структура предложения сравнивается с моделью управления, описанной в классе глаголов, определенном на прошлом этапе, и определяются подходящие синтаксические вариации класса.

Для нашего примера из класса глаголов «get 13.5.1» будут выделены синтаксические вариации ***NP.Agent V NP.Theme {for} NP.Asset*** *и* ***NP.Agent NP.Theme {for} NP.Beneficiary****. Они синтаксически эквивалентны, но последние элементы несут в себе семантические различия. Отследить эти различия и выбрать «правильную» вариацию можно с помощью отображения между ресурсами FrameNet и VerbNet.*

Используя отображение между синтаксическими вариациями и семантическими фреймами, сопоставляющее их элементы, отбрасывается семантически неподходящая вариация.

Таким образом, в рассматриваемом примере останется только ***NP.Agent V NP.Theme {for} NP.Asset.***

Так от конкретного предложения мы перешли к семантическому фрейму, определив при этом роли слов, входящих в его состав.

В примере «Carmen bought a dress for $ 50»: «Carmen» – NP.Agent – ***Buyer****, «a dress» – NP.Theme –* ***Goods****, «$ 50» - NP.Asset –* ***Money****, где последние элементы последовательности соответствуют актантам семантического фрейма «Commerce_buy».*

Далее, используя отображение между актантами синтаксических фреймов и синсетами WordNet, словам и словосочетаниям предложения сопоставляются конкретные смыслы.

Таким образом, применение описанного алгоритма позволяет добавить смысловую разметку предложению: определить ситуацию в целом и означить лексические единицы, входящие в состав предложения.

Глава 5 Результаты тестирования

Этап тестирования является не менее важным, чем этап проектирования или реализации, поскольку позволяет на практике определить то, насколько верным оказался подход к решению поставленной задачи.

- Для тестирования были определены 4 типа тестовых примеров, позволяющие продемонстрировать корректную работу системы и выявить случаи ее некорректной работы. К случаям, демонстрирующим правильную работу программы, относятся:

- *«позитивно-позитивные»* примеры - смысл лексемы должен определяться и он определяется;

- *«негативно-негативные»* - смысл лексемы не должен определяться и он не определяется.

Некорректная работа программы выявляется в случаях:

- обнаружения примеров, в которых смысл лексемы должен определяться, а он не определяется, это *«позитивно-негативные»* примеры;

- в примерах, в которых смысл лексемы не должен определяться, а он определяется; это *«негативно-позитивные»* примеры.

Следует отметить, что отдельный экземпляр примера может попадать под разные типы одновременно.

«Позитивно-позитивные» примеры:

В качестве примеров возьмем предложения из предметной области «Коммерция, купля-продажа». Первый пример – предложение *«Father sold an apartment»*, его подробное описание приведено в таблице 1:

Таблица 1. Пример работы практической реализации метода

Father sold an apartment.		
Семантический фрейм (FN)	Commerce_sell These are words describing basic commercial transactions involving a buyer and a seller exchanging money and goods, taking the perspective of the seller. The words vary individually in the patterns of frame element realization they allow.	
Класс глагола (VN)	Класс и подклассы[14]	give-13.1 give-13.1-1
	Сентенциальная форма[15]	NP V NP

[14] Класс и подклассы, в которые попадает ключевой глагол предложения (предикат, образующий предложение)

[15] Здесь и далее **сентенциальная форма** – форма из класса глагола, совпадающая с синтаксической структурой предложения.

Толкование (WN)	“father”	Предполагаемый смысл: WN (father 1) 1. “a male parent”
		Приписанные смыслы: WN (father 1, 2, 3, 4, 5, 7, 8) 1. “a male parent” 2. “the founder of family” 3. Syn. Padre 4. “(Christianity) any of about 70 theologians whose writing established and confirmed official church doctrine” 5. “a person who holds an important or distinguished position in some organization” 7. “a person who founds or establishes some institution” 8. “the head of an organized crime family”
		Отброшенные смыслы: WN (father 6) 6. Syn. God
		Вывод: приписанные смыслы с различной вероятностью могут присутствовать в исходном предложении. Система сработала корректно. Отброшенные смыслы отброшены корректно
	“sold”	Приписанный смысл: WN (sell 1) “exchange or deliver for money or its equivalent”, совпадает с предполагаемым смыслом Отброшено 7 смыслов.
	“apartment”	Приписанный смысл совпадает с предполагаемым смыслом: WN (apartment 1) “ a suite of rooms…”, Syn. Flat Отброшено 0 смыслов.

Как видно из примера, означивание предложения создается корректно в пределах допустимой погрешности, то есть предполагаемый смысл всегда попадает во множество приписанных смыслов. А все приписанные смыслы могут иметь место в предложении с разной долей вероятности (это наглядно видно на слове «father»).

Другие примеры правильно размеченных предложений: *«Carmen bought a dress for Mary»* и *«Carmen bought a dress for three thousand rubles»*. Оба предложения попадают в семантический фрейм *«Commerce_buy»*, сказуемое лежит в классе *«get-13.5.1»* и размечается системой как *«obtain by purchase; acquire by means of a financial transaction»*, *WN (buy 1)*, что совпадает с предполагаемым смыслом. Предложения имеют схожую

поверхностную синтаксическую структуру: *NP V NP for NP*, но именно здесь и кроются их принципиальные различия, которые должна обнаружить система.

В предложении *«Carmen bought a dress for Mary»* последняя лексема *«Mary»* подразумевает реципиента, отвечает на вопрос «кому? Для кого?», а в предложении *«Carmen bought a dress for three thousand rubles»* последняя лексема – плата, отвечающая на вопрос «за сколько?». Таким образом, несмотря на совпадающую поверхностную синтаксическую структуру, сентенциальные формы предложений различны. Этот факт отмечается системой, и предложениям приписываются разные сентенциальные формы из глагольного класса, первому – *NP V NP PP.BENEFICIARY*, второму – *NP V NP PP.ASSET*.

К правильно размеченному предложению также можно отнести такой пример: *«Susan sold her old wedding dress»*, в этом предложении интерес представляет дополнение *«her old wedding dress».* Дополнение состоит из трех лексем *«her», «old»* и *«wedding dress»*, каждая из которых успешно распознается системой, и им верно приписываются смыслы, так для последней лексемы – это смысл WN (wedding dress 1): «a gown worn by the bride at a wedding)», а не два смысла каждого из слов, составляющих лексему. Этот тест показывает корректную работу системы на устойчивых словосочетаниях.

В качестве «негативно-негативных» смыслов рассмотрим предложения *«Dog buy a wedding dress for 50 dollars»* и *«Giraffe sold an apartment».* В данных предложениях интерес представляет агенс (подлежащее) – лицо, производящее действие.

1. *«Dog buy a wedding dress for 50 dollars».*

В WordNet есть 7 смыслов для слова *«dog»*, они приведены в таблице 2 вместе с реакцией системы на каждое из них.

Таблица 2. Смыслы слова «dog» и реакция на них практической реализации метода автоматической семантической разметки

Смыслы WN	**Поведение системы**
1. a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times. Syn.: dog, domestic dog, Canis familiaris	Смысл отброшен
2. a dull unattractive unpleasant girl or woman	Смысл принят как верный
3. informal term for a man; Ex.: "you lucky dog"	Смысл принят как верный
4. someone who is morally reprehensible	Смысл принят как верный

5. a smooth-textured sausage of minced beef or pork usually smoked; often served on a bread roll. Syn.: hotdog	Смысл отброшен
6. a hinged catch that fits into a notch of a ratchet to move a wheel forward or prevent it from moving backward	Смысл отброшен
7. metal supports for logs in a fireplace; Ex.: "the andirons were too hot to touch"	Смысл отброшен

Система отбросила все смыслы, которые не попадают под определение «человек или организация», поскольку только эти категории объектов могут совершить процесс покупки, подразумеваемый словом «buy».

2. Аналогично для предложения *«Giraffe sold an apartment»*. Агенс (тот, кто совершает действие) не попадает под категорию «человек или организация», которую требуется предикату для реализации. Это предложение вовсе не будет размечено системой, ввиду несогласованности лексем.

Приведенные пример показывают, как учитываются ограничения, накладываемые сказуемым на свои зависимые слова.

В ходе тестирования были зафиксированы примеры из категорий «позитивно-негативные» и «негативно-позитивные», то есть случаи, когда смысл должен быть приписан лексеме, но этого не происходит и наоборот. Анализ таких ситуаций указывает на точки роста при дополнении ресурсов и совершенствовании программной реализации метода.

Заключение

В ходе исследования была проведена большая работа по изучению существующих лингвистических теорий и практическому освоению лингвистических ресурсов, изучению способов их интеграции. В результате чего был разработан метод автоматической семантической разметки предложений английского языка и создана его программная реализация. Метод базируется на мощных лингвистических теориях, а его практическая реализация представляет собой интеграцию лексических ресурсов, на них построенных.

В заключение отметим, что данная работа имеет как теоретическую, так и прикладную ценность. С одной стороны, на базе метода была построена система автоматической семантической разметки, разработаны недостающие api для работы с лингвистическими ресурсами. С другой стороны, были выявлены слабые места лингвистических ресурсов и отображений между ними, требующие дополнения.

Результаты тестирования показывают, что подход к семантической разметке предложения, опирающийся на семантику фреймов и классификацию глаголов, учитывающий синтаксическую структуру предложения работает, дает положительные результаты, может развиваться и дополняться. Ценности разработанному методу добавляет универсальность относительно используемых ресурсов, а, следовательно, языка, и возможность добавления онтологии, что позволяют создать на его основе мощнейший инструмент распознавания, применимый в различных областях знаний. Тестирование метода, показывает, что подход успешно справляется с озвученной задачей автоматической семантической разметки предложений английского языка.

Изложенному в работе методу автоматической семантической разметки предложений английского языка посвящена одноименная статья [22], представленная для публикации в рецензируемый научно-теоретический и прикладной журнал «Альманах современной науки и образования».

Основные положения работы были представлены на 51-й Международной научной студенческой конференции «Студент и научно-технический прогресс» [23] и были отмечены дипломом второй степени.

Литература

1. Fillmore, Charles J. Frame semantics and the nature of language / Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech. - Berkley, California, 1976. – p. 21 – 23.
2. Levin, B. English Verb Classes and Alternations: A Preliminary Investigation. - Chicago, USA: University of Chicago Press, 1993. - 348 p.
3. Ruppenhofer Josef et al. FrameNet II: Extended Theory and Practice. — Berkeley, California: International Computer Science Institute, 2006. - 166 p.
4. Официальный информационный ресурс проекта FrameNet [Электронный ресурс]. – Режим доступа: https://framenet.icsi.berkeley.edu, свободный.
5. Официальный информационный ресурс проекта VerbNet [Электронный ресурс]. – Режим доступа: http://verbs.colorado.edu/~mpalmer/projects/verbnet.html, свободный.
6. Официальный информационный ресурс проекта WordNet [Электронный ресурс]. – Режим доступа: http://wordnet.princeton.edu, свободный.
7. Аперсян, Ю.Д. Исследования по семантике и лексикографии. Т. I: Парадигматика. – М.: Языки славянских культур, 2009. - С.14 – 57.
8. Евдокимова Е.И. Естественно-языковые системы: курс лекций // Семантический анализ ЕЯ-текстов: Анализ лингвистических моделей. - Улан-Удэ: ВСГТУ, 2006. – С. 89 – 92.
9. Методы семантики // Информационный ресурс по лингвистическим дисциплинам «Языкознание.ру» [Электронный ресурс]. - Режим доступа: http://yazykoznanie.ru/content/view/84/273/, свободный.
10. Модели семантики // Информационный ресурс по лингвистическим дисциплинам «Языкознание.ру» [Электронный ресурс]. – Режим доступа: http://yazykoznanie.ru/content/view/85/274/, свободный.
11. Люгер, Джордж Ф. Искуственный интеллект. Стратегии и методы решения сложных задач. — М.: Вильямс, 2003. — С. 232 - 237.
12. Богданова, Л. И. Стилистика русского языка и культура речи: Лексикология для речевых действий: учеб. пособие. — М. : Флинта: Наука, 2011. – С. 34 – 37.
13. Поверхностная структура предложения // Лингвистический энциклопедический словарь. [Электронный ресурс]. – Режим доступа: http://tapemark.narod.ru/les/379b.html, свободный.

14. Глубинная структура слова // Лингвистический энциклопедический словарь. [Электронный ресурс]. – Режим доступа: http://tapemark.narod.ru/les/110a.html, свободный.
15. Официальный информационный ресурс проекта Stanford Syntax Parser [Электронный ресурс]. – Режим доступа: http://nlp.stanford.edu/software/lex-parser.shtml свободный. .
16. Shiffman, D. Wordnet/ Персональный блог «Programming from A to Z», 2008 [Электронный ресурс]. – Режим доступа: http://www.shiffman.net/teaching/a2z/wordnet/, свободный;
17. Информационный ресурс проекта WordNet TreeWalk [Электронный ресурс]. – Режим доступа: http://wntw.sourceforge.net/, свободный.
18. Pighin Daniel MapNet: a FrameNet to WordNet Mapping [Электронный ресурс]. – Режим доступа: http://danielepighin.net/cms/research/MapNet, свободный.
19. Sara Tonelli and Daniele Pighin New Features for FrameNet - WordNet Mapping // CoNLL'09: Thirteenth Conference on Computational Natural Language Learning. – Boulder, CO, USA, 2009.
20. Lei Shi, Rada Michalcea Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. — Cicling, Mexico, 2005. — p. 100 - 111.
21. FrameNet Extension: repository of senses [Электронный ресурс]. – Режим доступа: https://dkm.fbk.eu/index.php/FrameNet_extension:_repository_of_senses, свободный .
22. *Маркова, М.В. Автоматическая семантическая разметка предложений английского языка* / Альманах современной науки и образования. №6 (73). – Тамбов: Грамота, 2013.
23. *Маркова, М.В. Автоматическая разметка предложений английского языка* / Материалы 51-й Международной научной студенческой конференции «Студент и научно-технический прогресс»: Информационные технологии // Новосиб. гос. ун-т., Новосибирск, 2013. – С. 230.

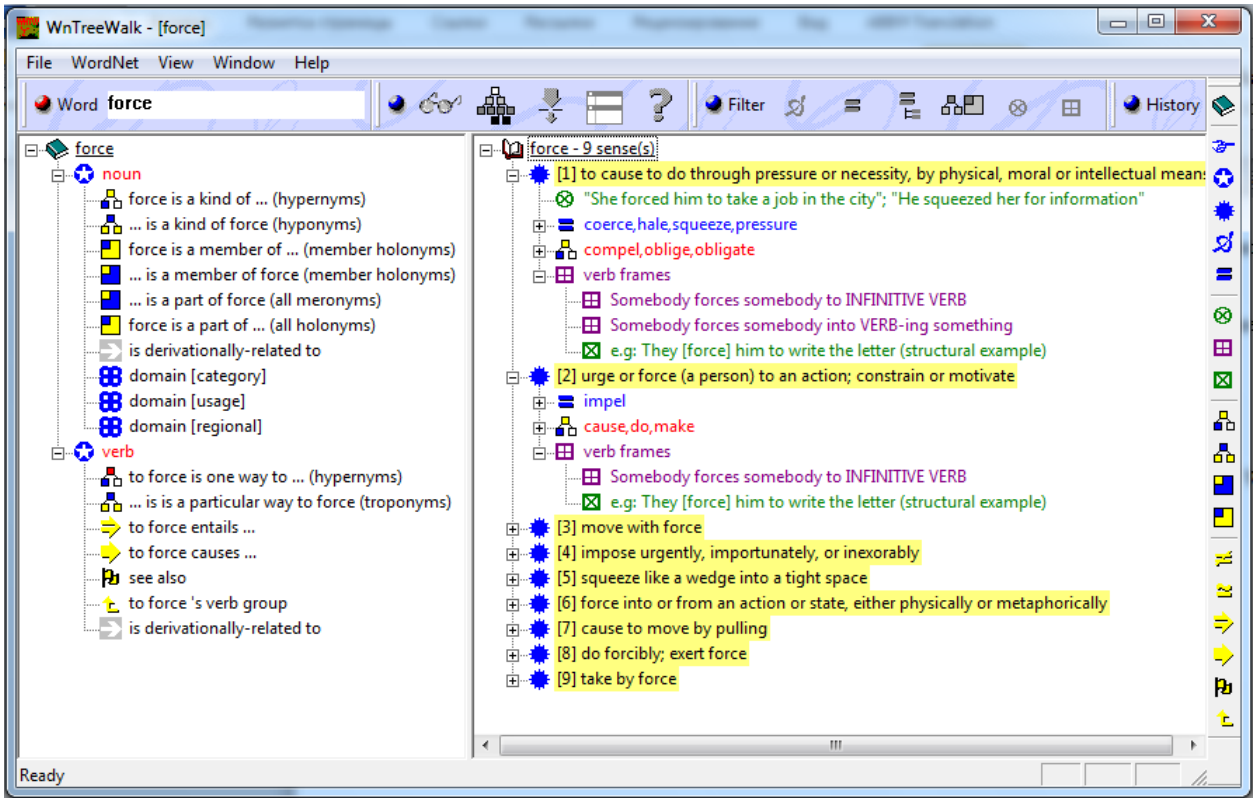
Приложение А

(рекомендуемое)

Графические интерфейсы для работы с лингвистическими ресурсами

Браузер WordNet TreeWalk для работы с WordNet.

Рисунок А.1 WordNet TreeWalk



Функциональность браузера включает:

1) поиск слов в базе данных;

2) просмотр списка понятий слова с отображением краткого описания понятия и примеров употребления, осуществляется сортировка по частям речи;

3) просмотр списка синонимов, антонимов и других семантических отношений, определенных для слова;

4) фильтрация отображения семантических отношений;

5) работа в нескольких окнах;

6) навигация по дереву синсетов, а также по деревьям, построенным на основе других семантических отношений;

7) подсказки на кнопки для удобства использования;

8) сохранение истории поиска.

Веб-интерфейс для работы с FrameNet.

Лингвистический ресурс FrameNet снабжен удобным веб-интерфейсом Окно делится на 2 основных поля: навигация по списку (фреймов или лексических единиц) и окно описания выбранного элемента. На рисунке А.2 представлена навигация по списку фреймов. Также, браузер снабжен такими функциями, как: FrameGrapher (рис. А.3 и А.4) и FrameSQL. Первый позволяет получить визуальное представление отношений между фреймами. Второй – получать данные из базы данных FrameNet с помощью инструмента SQL-запросов.

Рисунок А.2 Окно просмотра фреймов

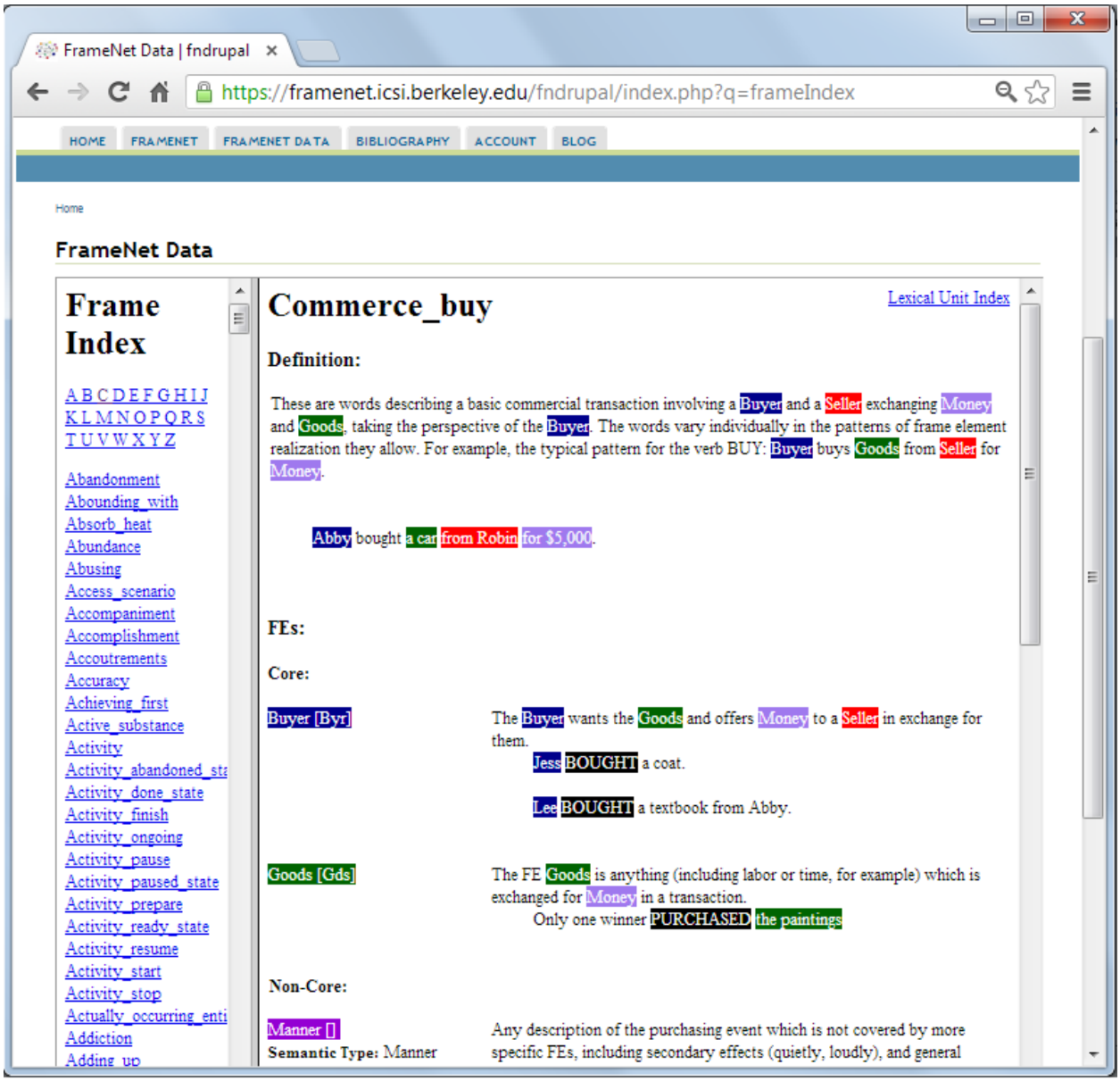


Рисунок А.3 FrameGrapher. Задание условий просмотра отношений между фреймами.

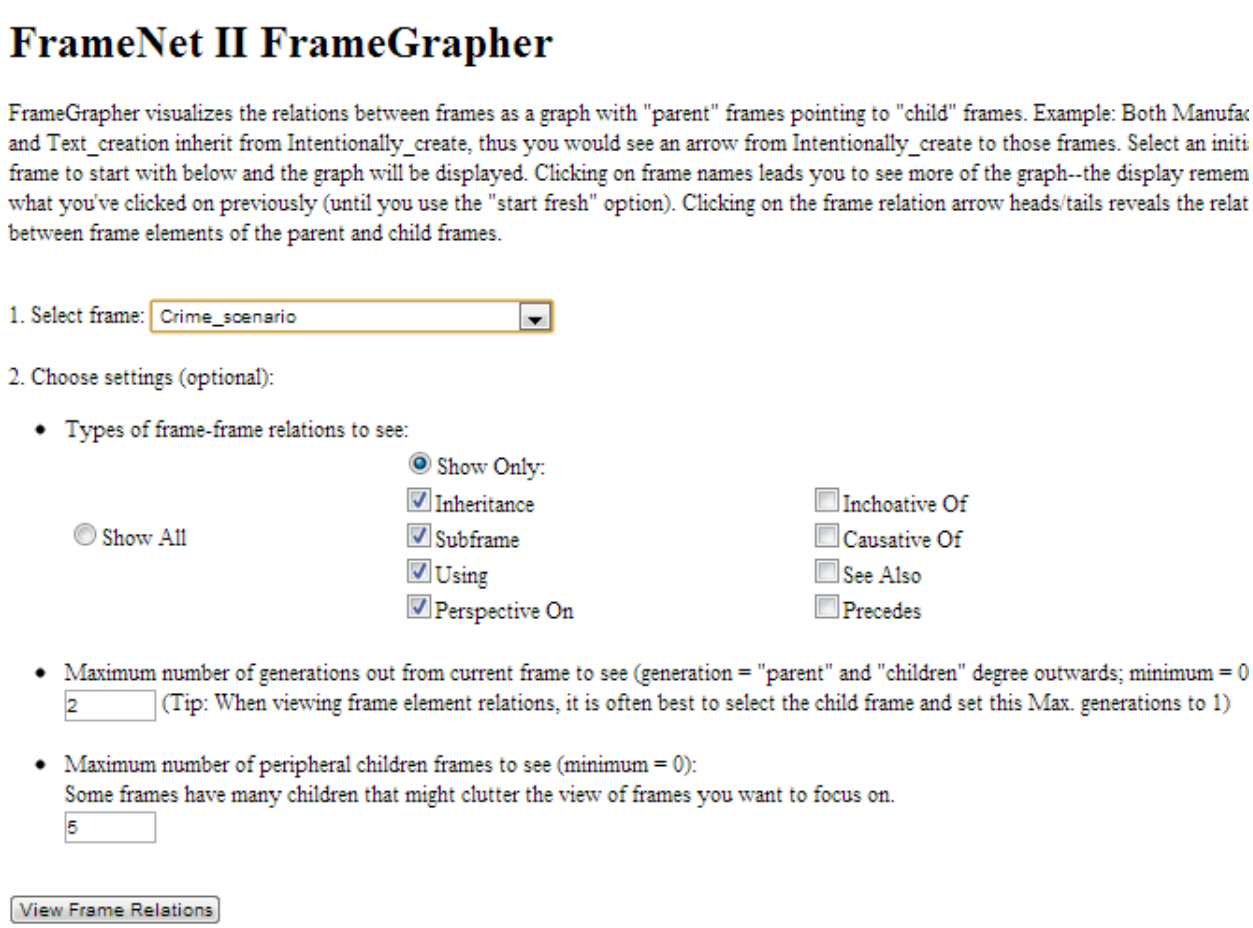
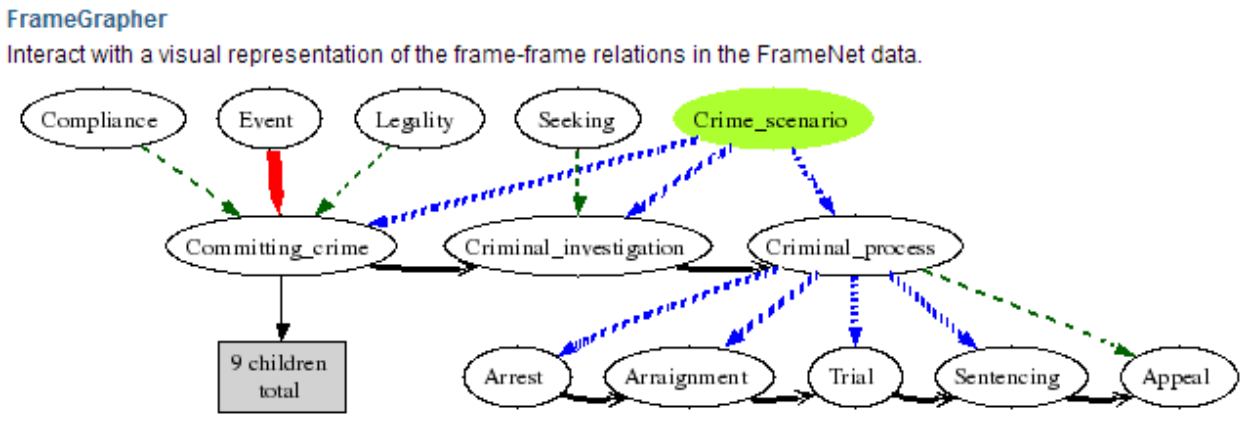


Рисунок А.4 FrameGrapher. Визуальное изображение отношений между фреймами по заданным условиям



Веб-интерфейс для работы с VerbNet

Инструмент для работы с VerbNet является частью объединенного инструмента для работы с несколькими лингвистическими ресурсами, предоставленного университетом Колорадо. На рисунке А.5 показана «стартовая страница», предоставляющая доступ к ресурсами по запросу для некоторого слова. На рисунке А.6 изображена страница с информацией из VerbNet согласно запросу.

Страница, иллюстрирующая класс VerbNet содержит такую информацию как: список членов класса, список ролей участников с возможными логическими ограничениями, список синтаксических вариаций (синтаксических фреймов) с описанием, описание подклассов, если они присутствуют.

Рисунок А.5 Объединенный веб-интерфейс для работы с лингвистическими ресурсами

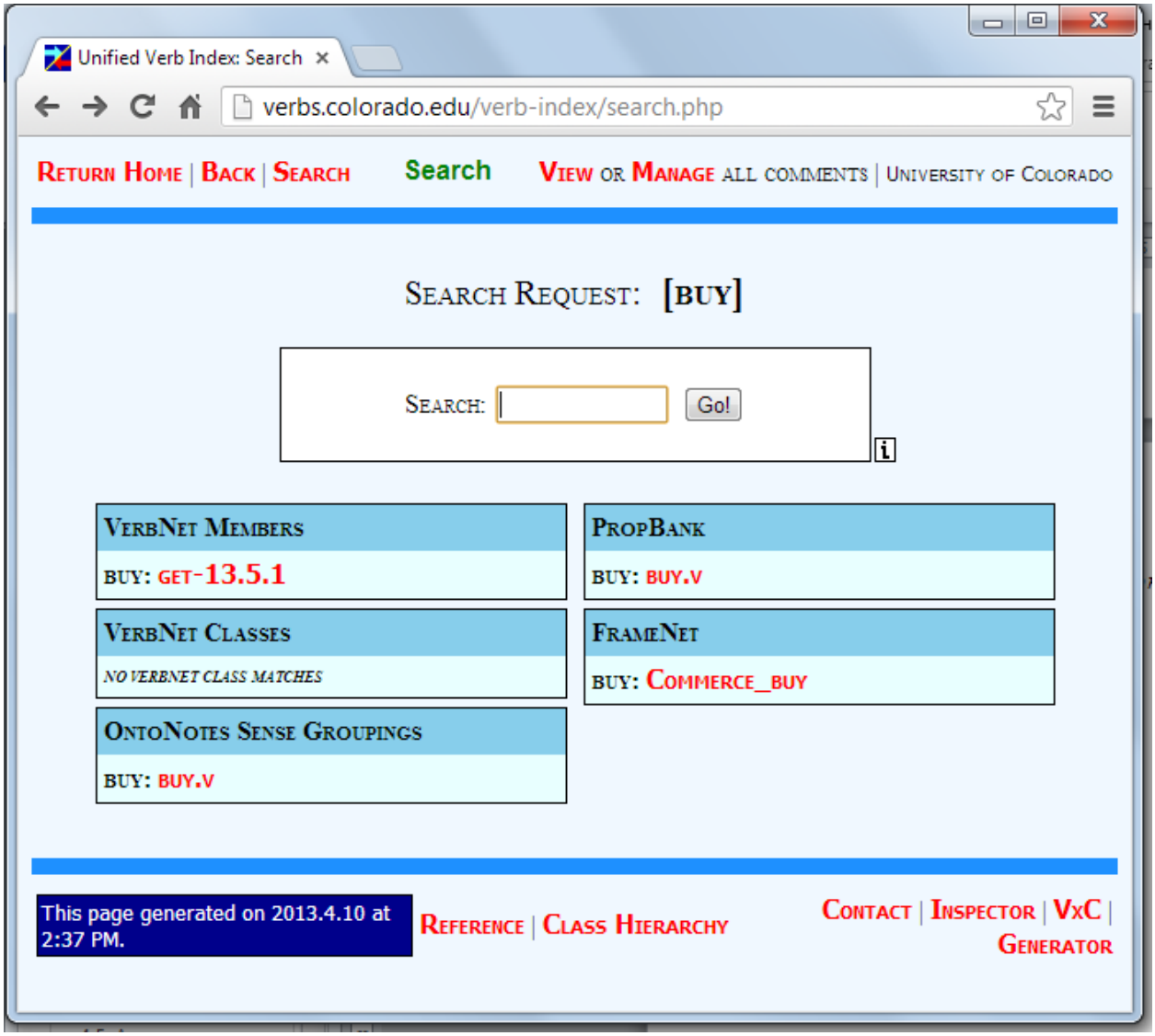


Рисунок А.6 Страница, иллюстрирующая класс VerbNet