ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет

«Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа
«Фундаментальная и компьютерная лингвистика»

КУРСОВАЯ РАБОТА

На тему «Разработка стандарта по разметке семантических ролей и фреймов в формате PropBank и FrameNet для русского языка.» *Название темы на английском* «Towards the Gold Standard for Semantic Role Labeling and Frames Recognition in Russian: The PropBank and FrameNet Formats.»

Студентка 3 курса
группы № 141

Сонина Полина Алексеевна

Научный руководитель:
Ляшевская Ольга Николаевна
Кандидат филологических наук,
профессор Школы Лингвистики

Москва, 2017

Оглавление

1. Введение

На данный момент в лингвистике существует потребность в текстовом материале с добавленным уровнем разметки семантических ролей и фреймов.

Для английского языка собрано значительное количество подобного материала в рамках таких ресурсов, как FrameNet, PropBank и других. Создаются ресурсы и для других языков, в основном, ориентируясь на структуру англоязычных. Для русского языка идёт разработка FrameNet-ориентированного ресурса – система FrameBank.

Цель данной работы – сделать шаг к созданию полноценного стандарта разметки семантических ролей и фреймов, который может в дальнейшем пополняться, а в будущем быть использован для решения исследовательских задач, иллюстрации семантических явлений, в качестве материала для машинного обучения.

Задачи данного исследования состояли в изучении особенностей семантических ролей и фреймов в русском языке; в анализе и сравнении существующих ресурсов и стандартов разметки; в разработке стандарта разметки предложений на основе изученной информации.

2. Семантические роли и фреймы

«Семантическая роль членов предложения - роль, определяемая коммуникативной значимостью информации, которую содержит тот или иной член предложения в семантике всего предложения.»[1]

«Фреймовая семантика - общее название для разных типов формализованного описания деятельности человека в контексте ситуации; направление, соотносящее значение слов, словосочетаний, предложений, текстов со сценами в рамках общей теории семантического знания.»[1]

Фрейм – схематическое представление о ситуации, представленной в тексте. Он задает участников ситуации и отношения между ними.

3. Существующие ресурсы

3.1 FrameNet

FrameNet[2] - лексикографическая система, созданная на базе теории фреймовой семантики. Цель проекта – описать и снабдить примерами все возможные валентности, в которых может выступать слово в каждом его значении. Результатом проекта является обширная лексическая база данных, включающая в себя более 13000 лексических единиц (7000 – полностью проаннотированы), которые входят в состав более 1000 связанных между собой фреймов, а также более 200 тысяч проаннотированных предложений.

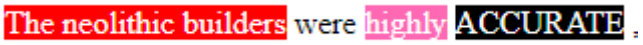
Все данные находятся в открытом доступе либо в интерактивной форме на сайте проекта, либо в виде файлов, которые предоставляются по запросу. Файлы в формате xml открываются в браузере. Разметка данных для проекта происходит в специальной программе, где разметчик может добавлять уровни семантических конструкций к текстовому материалу.

В основном режиме просмотра различные элементы фрейма выделяются определёнными цветами. Также можно переключать режим просмотра разметки между цветами и выделением элементов с помощью квадратных скобок и текстовых примечаний.

---

[1] Словарь лингвистических терминов: Изд. 5-е, испр-е и дополн. — Назрань: Изд-во "Пилигрим". Т.В. Жеребило. 2010.

[2] Проект доступен по адресу: https://framenet.icsi.berkeley.edu/fndrupal/



(1) [$_{Agent}$The neolithic builders] were [$_{Degree}$highly] ACCURATE^{Target}

3.1.1 Тип данных: фреймы

Файлы описывающие фреймы содержат:

- описание фрейма
- примеры предложений
- списки центральных элементов фрейма (ролей) и дополнительных
- список связей фрейма с другими фреймами, если они существуют в системе
- список с лексических единиц, представляющих данный фрейм, и таблицу со ссылками на соответствующие файлы, на примеры употребления.

3.1.2 Тип данных: лексические единицы

Файлы содержат:

- название лексической единицы
- фрейм, к которому она принадлежит
- определение
- таблицу с элементами фрейма и их синтаксическими реализациями для данной лексической единицы
- таблицу с синтаксическими структурами, в которых могут находиться элементы фрейма

3.1.3 Тип данных: тексты

Помимо данных описывающих фреймы и лексические единицы с отдельными примерами, в рамках проекта ведётся также полная аннотация текстов. Коллекция текстов не велика, так как проект направлен в первую очередь на лексикографическое описание фреймов.

3.2 PropBank

В противоположность FrameNet ресурс PropBank[3] направлен на создание материала для обучения статистических программ. Его задача заключается в создании аннотация для каждой конструкции в Penn Treebank.

Материал PropBank изначально покрывал только глаголы, в последующих версиях были добавлены категории существительных и прилагательных в качестве предикатов. Теперь же в последней версии PropBank было принято решение не разделять предикаты на категории, а объединять в один предикат слова вне зависимости от части речи. Например, *create* и *creation* теперь будут объединяться.

3.2.1 Тип данных: предикаты

Первый тип данных в PropBank – файлы-«фреймсеты», содержащие набор предикатов, связанных с определённой леммой, а также с фразами, содержащими лемму. Например, файл для *keep* будет содержать информацию и для *keep from*. Файлы представлены в формате xml[4].

Каждый предикат содержит набор ролей (roleset). Наборы ролей описывают структуру аргументов и маркеры для аннотации конструкций. Разным значениям предиката могут соответствовать разные наборы ролей. Комбинации предикатов и наборов ролей снабжаются примерами. В примерах сначала даётся полное предложение, а затем - соответствие маркеров ролей частям предложения.

Обязательные составляющие файла:

- Открывается тегом *<!DOCTYPE frameset SYSTEM "frameset.dtd">*
- Основная часть файла заключена между тегами *<frameset>…</frameset>*
- Каждый вариант предиката заключён в теги *<predicate lemma="ЛЕММА">…</predicate>*
- Набор ролей между тегами *<roleset id=" ЛЕММА.01" name="ОПИСАНИЕ КОНСТРУКЦИИ">…</roleset>*

[3] Материалы и документация доступны по адресу: http://propbank.github.io/
[4] Подробная документация формата: https://github.com/propbank/propbank-documentation/blob/master/data-format/frameset.dtd

- Между тегами *<aliases>…</aliases>* находятся названия соответствующих предикатов/фреймов в FrameNet и VerbNet, если они существуют.
- Список ролей заключён между *<roles>…</roles>*
- Информация о каждой роли: *<role descr="ОПИСАНИЕ" f="ФУНКЦИЯ" n="НОМЕР">…</role>* (внутри тега могут быть описания роли из FrameNet и VerbNet в соответствующих тегах *<vnrole/>, <fnrol/>*)
- Примеры: *<example name="НАЗВАНИЕ" src="ИСТОЧНИК" type="ТИП">…</example>*
- Форма леммы в примере: *<inflection МОРФОЛОГИЧЕСКИЕ ПАРАМЕТРЫ />*
- Текст примера: *<text>ТЕКСТ</text>*
- Части предложения занимающие роли: *<arg f="ТЕГ ФУНКЦИИ" n="НОМЕР">ЧАСТЬ ТЕКСТА</arg>*

Кроме обязательных элементов, в файле могут присутствовать пояснения, уточнения, дополнительная информация, заключённые в теги *<note>…</note>*.

3.2.2 Тип данных: тексты

Текстовые данные из Ontonotes и English Web Treebank с аннотацией предикатов и ролей представлены в PropBank формате. Данный формат содержит экземпляры (PropBank instances), по одному на каждой строке.

Формат разметки[5]:

<tree_path> <tree_id> <predicate_id> <annotator_id> <frameset>
<lemma>.<roleset_id> <aspects>(<argument>)+

Где:

<argument> ::= <terminal_id>:<height>-<label>

<tree_path> ::= путь к файлу в Treebank

<tree_id> ::= индекс дерева, содержащего предикат

<predicate_id> ::= индекс предиката (число)

<annotator_id> ::= ID разметчика

[5]Подробная документация формата: https://github.com/propbank/propbank-documentation/blob/master/data-format/EPB-data-format.txt

<*framefile*> ::= название файла, содержащего данный предикат

<*roleset_id*> ::= ID набора ролей

<*terminal_id*> ::= ID первого терминального узла в составляющей

<*height*> ::= уровень составляющей относительно её первого терм. узла

<*label*> ::= PropBank тег

3.3 FrameBank

FrameBank[6] — это русскоязычный FrameNet-ориентированный ресурс, развивающийся в направлении формата - «корпусного словаря конструкций». Ресурс создаётся с учётом традиций русской лексической семантики и специфики русского языка. Присутствует ориентация сбора данных на отобранные примеры, а не на полную разметку текста.

4. Стандарт для представления русских данных

Созданы файлы для 10 предикатов:

- 4 глагола (*удивить*, *превратиться*, *осуществить*, *описать*)
- 3 имени существительных (*подарок, отзыв*, *конец*)
- 3 прилагательных/наречия (*точный, быстрый*, *важный*)

Примеры предложений взяты из материалов FrameBank и из НКРЯ.

См. приложение.

[6] Проект доступен по адресу: http://www.framebank.ru/.

6. Заключение

В ходе выполнения данной работы были изучены форматы представления семантических данных в двух популярных англоязычных проектах (FrameNet, PropBank), а также рассмотрен развивающийся русскоязычный ресурс – FrameBank. Был создан вариант представления русских фреймов в международных форматах.

Дальнейшая работа возможна в направлении уточнения и дополнения стандарта, а также создания скрипта для автоматического перевода русских данных в существующие международные форматы.

Литература

Апресян Ю. Д. *Избранные труды, том I. Лексическая семантика.* М., 1995. 1-е изд.: М., 1974.

Жеребило Т.В. *Словарь лингвистических терминов: Изд. 5-е, испр-е и дополн*. — Назрань: Изд-во "Пилигрим", 2010.

Ляшевская О. Н., Кашкин Е. В. Типы информации о лексических конструкциях в системе ФреймБанк // *Труды института русского языка им. В.В. Виноградова.* 2015. № 6. С. 464-555.

Ляшевская О.Н., Кузнецова Ю.Л. Русский фреймнет: к задаче создания корпусного словаря конструкций // *Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27-31 мая 2009 г.).* Вып. 8 (15). М.: РГГУ, 2009. С. 306-312.

Падучева Е.В. *Динамические модели в семантике лексики.* М., 2004.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L Petruck, Christopher R. Johnson, Collin F. Baker, Jan Scheffczyk: *FrameNet II: Extended Theory and Practice* (Revised November 1, 2016.)

Martha Palmer, Dan Gildea, Paul Kingsbury, The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics Journal*, 31:1, 2005.

Приложение

Материалы в репозитории на Github:

https://github.com/SoDipole/roles-and-frames