Автоматическое извлечение семантических ролей, вводимых формами с показателем дательного падежа

Екатерина Волошина, НИУ ВШЭ
Научный руководитель: Рыжова Дарья Александровна, к.ф.н., доцент

26 мая 2021

Код и данные находятся по ссылке `https://github.com/EkaterinaVoloshina/classification_of_dative_semantic_roles`

1 Введение

Семантические роли - модель, объединяющая похожих участников различных ситуаций на основе сходства их значения и морфосинтаксического поведения.

Цель: проверить, могут ли модели классификации различать семантические роли, имеющие похожее морфосинтаксическое выражение на примере семантических ролей участников, маркированных дательным падежом

Гипотеза: предсказания моделей будут основываться во основном на семантических признаках, при этом ошибки классификатора будут отражать структуру категории дательного падежа: более прототипические и более периферийные семантические роли, выражаемые дательным падежом

2 Семантические роли, выражаемые дательным падежом в русском языке

Описание семантических ролей основывается на работах (Haspelmath 2003) и (Janda и Clancy 2002). Для целей нашего исследования мы используем классификацию из 6 семантических ролей:

- Реципиент: *Отец дал игрушку* ***своему сыну****.*
- Бенефактив: *Налейте* ***мне****, пожалуйста, еще чашечку чаю.*
- Экспериенцер: ***Мне*** *не нравится это платье. Брату минуло двадцать лет.* ***Вашим спутникам*** *придется здесь ночевать.*
- Поднятый Посессор: *Ветки били* ***охотнику*** *в лицо.*
- Направление: *Режиссёр, поддерживая балерину под локоть, вывел её к* ***кулисе****. Люблю уезжать надолго и болтаться по* ***улицам****.*
- Контрагент: *Иван подражал* ***отцу*** *в манере говорить. Дважды два равняется* ***четырем****.*

3 Данные

Данные взяты из оффлайн-версии FrameBank. В FrameBank используется 91 семантическая роль (Lyashevskaya и Kashkin 2015). Мы вручную объединили и переразметили роли, следуя описанной выше классификации. В нашей выборке составило 1703 примера.

Семантическая роль	**Количество примеров**
Бенефактив	225
Контрагент	275
Направление	396
Поднятый Посессор	125
Реципиент	400
Экспериенцер	282

Таблица 1: Распределение данных по классам

Кроме этого, FrameBank содержит много примеров, неразмеченных по семантическим ролям, но имеющим синтаксическую и морфологическую разметку. Из них мы отобрали 94864 примеров с существительным или местоимением в дативе.

Все примеры были автоматически размечены по следующим признакам:

- *синтаксические*: наличие или отсутствие прямого объекта, наличие или отсутствие предлога, управляющего местоимением в дативе;
- *морфологические*: часть речи, одушевленность, число для объекта в дативе и субъекта, часть речи для предиката и, если это глагол, наклонение и вид глагола;
- *семантические*: леммы предиката, непрямого объекта и субъекта, а также их семантические характеристики, взятые из Национального корпуса русского языка: для непрямого объекта и субъекта - разряд существительного (предметное, непредметное и имя собственное), таксономический класс существительного, топологический класс, мереологический класс, коннотация и словообразовательная структура (например, диминутивы, аугментивы), для предиката - семантический класс глагола, каузативный / некаузативный глагол, служебный ли глагол, словообразовательная структура.

4 Эксперименты

В исследовании использованы два типа методов: обучение с учителем и обучение с частичным привлечением учителя. Основное отличие между двумя подходами заключается в том, что для обучения с учителем используются только размеченные данные, в то время как методы обучения с частичным привлечением учителя используют небольшую выборку размеченных данных, и в процессе обучения модели размечают неаннотированные данные и в последующие итерации обучаются на них. Результаты моделей обучения с учителем представлены в таблице:

Модель	**F1-Score (weighted)**
Random Forest	*0.71*
Gradient Boosting	*0.72*
Логистическая регрессия	*0.73*
Полносвязная нейронная сеть	***0.75***

Таблица 2: Сравнение результатов моделей обучения с учителем

В качестве метода обучения с частичным привлечением учителя в этой работе использовался метод Self-training. Для экспериментов с методами обучения с частичным привлечением учителя мы использовали датасет без признаков лемм непрямого объекта, глагола и субъекта, так как размеченных данных гораздо больше, чем неразмеченных, и эти признаки могут только испортить качество моделей: большая часть предикатов из неразмеченных данных не будет содержаться в размеченном датасете.

Алгоритм	**С леммами**	**Без лемм**
Логистическая регрессия	*0.73*	*0.6*
XGBoost	*0.72*	*0.6*

Таблица 3: Сравнение результатов моделей обучения с частичным привлечением учителя

Обучение с частичным привлечением учителя не дает улучшения качества, так как модели не выучивают ничего нового из данных, аннотированных на основании уже имеющейся разметки.

5 Обсуждение

- *Качество классификации*: лучшая модель в нашем исследовании показала качество 0.744 (micro F1-Score) и 0.742 (macro F1-Score). В исследовании (Larionov и др. 2019), где был разработан первый пайплайн для русского языка, модель, основанная на признаках, показала качество 0.769 (Micro F1) и 0.736 (macro F1-Score).

- *Важность признаков*: для логистической регрессии важными оказываются леммы глаголов, а также семантический класс глагола и леммы объекта в дательном падеже, для модели XGBoost - семантический класс предиката, словообразовательная структура субъекта и семантический класс непрямого объекта.

- *Качество классификации по отдельным ролям*: Самое низкое качество классификаторы показывали на примерах с ролью Бенефактива. В отличие от других семантических ролей, предикаты, которые вводят роли Бенефактива, не составляют естественного класса и имеют тенденцию к коэрции:

  (1) *Среди ночи она жарила **ему** яичницу на электроплитке.*

  Роли Реципиента и Экспериенцера вводятся глаголами из нескольких семантических классов (см. (Janda и Clancy 2002): обе роли имеют три подкатегории), что влияет на качество модели.

  Сочетаемость с большим количеством предикатов, не ограниченным одним семантическим классом, - одна из характерных черт прототипа радиальной категории. Таким образом, Реципиент, Экспериенцер и Бенефактив являются одними из центральных значений категории датива.

- *Ошибки классификаторов*: модели часто присваивают метку **Реципиента**, что соответствует представлению о Реципиенте как о прототипе дательного падежа, который служит источником для других ролей. Семантический переход от одной роли к другой основывается на общих чертах двух ролей, поэтому роль Реципиента имеет больше всего общих черт с другими ролями.

  Модели совершают ошибки при различении Реципиента и **Бенефактива**. Как показывают данные диахронии (Haspelmath 2003), роль Бенефактива часто служит источником для развития других семантических ролей, в том числе и для Реципиента. Примеры с Бенефактивом могут определяться как Экспериенцер: эти роли оказываются близкими в русском языке (Janda и Clancy 2002).

Семантическая роль	XBoost	RF	LogReg	FNN	Average
Бенефактив	*0.6*	*0.68*	*0.63*	*0.65*	0.64
Контрагент	*0.76*	*0.78*	*0.81*	*0.84*	0.8
Направление	***0.82***	***0.82***	***0.85***	***0.87***	**0.84**
Поднятый посессор	*0.69*	*0.54*	*0.65*	*0.71*	0.65
Реципиент	*0.71*	*0.69*	*0.67*	*0.65*	0.68
Экспериенцер	*0.7*	*0.65*	*0.71*	*0.73*	0.7

Таблица 4: Сравнение F1-Score разных классификаторов по семантическим ролям

Некоторые примеры из классов Реципиента и Экспериенцера получают метку **Направления**. Направление в русском языке является периферийным значением для дательного падежа, так как он не может выражаться дательным падежом без предлога, однако Направление является единственной семантической ролью, имеющей пространственное значение, что является одним из признаков прототипического значения. Типологически Направление является источником для диахронического перехода к другим семантическим ролям (Haspelmath 2003).

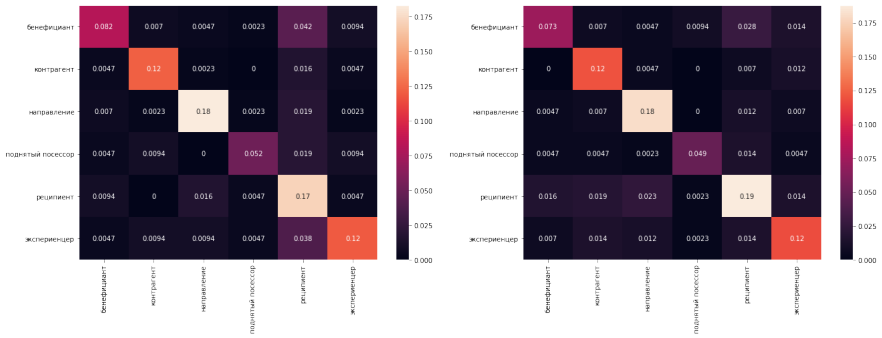

Рис. 1: Матрица ошибок логистической регрессии и XGBoost

Литература

Janda, Laura A and Steven J Clancy (2002). *The case book for Russian*. English. Vol. 1. Slavica Pub.

Haspelmath, Martin (2003). "The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison". English. In: *The new psychology of language* 2.1976, pp. 1–30.

Lyashevskaya, Olga and Egor Kashkin (2015). "FrameBank: a database of Russian lexical constructions". English. In: *International Conference on Analysis of Images, Social Networks and Texts*. Springer, pp. 350–360.

Larionov, Daniil et al. (2019). "Semantic role labeling with pretrained language models for known and unknown predicates". English. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pp. 619–628.