Обучение в контексте для разметки семантических ролей на русском языке

Григорий Казачёнок, Кафедра системных исследований МФТИ
Науч. рук. Смирнов И. В., ИСА РАН

Модели семантики

Семантика — раздел лингвистики, изучающий смысловое значение единиц языка.

Семантика высказываний изучает значение предложений, в отличие, например, от лексической семантики, которая изучает значение слов.

Цель моделирования семантики – создать математическую модель, которая бы описывала значение предложений на уровне структуры.

Модели семантики: мотивация

Построение такой модели поможет решать задачи семантического анализа, такие как ответить на вопрос по тексту; определить, следует ли одно предложение из другого.

Построение явной модели вместо или поверх неявной (LLM) имеет преимущество большей прозрачности и интерпретируемости.

Также, имеется теоретическая польза в построении явной модели языка.

Модели семантики

- Логика первого порядка (Лакофф, Монтегю и др.)
- Семантические сети
- Модель “Смысл-Текст”
- Фреймы
- Вероятностные логики
- Векторные представления
- Процедурная семантика
- …

Модели семантики

Однако, все существующие модели в разной степени неполны или требуют больших человеческих усилий по ручной разметке слов и предложений. Из-за этого их применение на практике очень ограничено.

Последние два десятилетия много усилий было направлено на то, чтобы автоматизировать процесс семантической разметки корпусов.

Логика первого порядка

Постулируется, что каждое высказывание на естественном языке является утверждением о мире, и что ему соответствует утверждение на языке логики первого порядка.

Так, предложение *“**Alexander eats an olive**”* транслируется в формулу:

∃ x, e : Olive(x) ∧ Eat(e, Alexander, x)

(**e** – переменная, обозначающая событие)

Семантические роли

∃ x, e : Olive(x) ∧ Eat(e, Alexander, x)

Здесь “Olive” и “Eat” - предикаты, то есть функции от некоторых логических атомов. “Alexander”, “x” – аргументы функции “Eat”.

Зачастую нас не интересует полная формула, а лишь вопросы “кто?”, “когда?”, “кого?” и т.д. В таком подходе, который называют “поверхностным” представлением семантики, предложение запишется так:

Eat(e, Alexander, olive)

Семантические роли

Однако, нет единой функции, которая бы соответствовала глаголу “eat”: в разных предложениях такая функция может принимать совершенно разные типы аргументов. Эти типы называются *семантическими ролями*. Мы уже видели роль “агенс”, отвечающую на вопрос “кто ест?” (“Alexander”), и роль “тема”, отвечающую на вопрос “что ест?” (“olives”).

Однако могут быть и другие роли: “локатив” (“где?”), инструментатив (“чем?”), и т.д. Возможные роли зависят от конкретного предиката.

Semantic role labeling

Итого:

1. **Предикат** описывает некую ситуацию.
2. **Аргументы** задают “действующих лиц” в этой ситуации.
3. **Семантические роли** – это роли, которые эти лица играют.

Тогда задача разметки ролей состоит из 3-х частей:

1. Нахождение предиката
2. Нахождение аргументов
3. Определение семантических ролей.

(Мы будем считать, что предикат уже дан)

Semantic role labeling

Предлагались различные варианты автоматизации этого процесса. В работе [1] была 1-ая попытка осуществить SRL, используя статистические методы. В [2] был применён механизм self-attention для SRL. В [3] был использован SVM для SRL на русском языке.

[1] Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. Computational linguistics, 28(3):245–288

[2] Emma Strubell, Patrick Verga, Daniel Andor,David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 5027–5038

[3]Ilya Kuznetsov. 2015. Semantic role labeling for Russian language based on Russian framebank. In International Conference on Analysis of Images, Social Networks and Texts, pages 333–338. Springer

Semantic role labeling

В работе [4] авторы используют эмбеддинги, сгенерированные языковыми моделями, для нахождения аргументов и их классификации по ролям. Авторы используют корпус Framebank в качестве тренировочных данных. В [5] предложен метод PromptSRL для разметки, использующей промт к LLM.

[4] Daniil Larionov, Artem Shelmanov, Elena Chistova, and Ivan Smirnov. 2019. Semantic Role Labeling with Pretrained Language Models for Known and Unknown Predicates. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pages 619–628, Varna, Bulgaria. INCOMA Ltd.

[5] Cheng, N. et al. (2024). Potential and Limitations of LLMs in Capturing Structured Semantics: A Case Study on SRL. In: Huang, DS., Zhang, X., Zhang, Q. (eds) Advanced Intelligent Computing Technology and Applications. ICIC 2024. Lecture Notes in Computer Science(), vol 14875. Springer, Singapore.

LLM for Semantic role labeling

Попробуем использовать LLM для этой задачи. На вход будет подаваться промт с инструкцией по разметке, предложение и предикат из него. В предложении нужно найти все аргументы этого предиката и определить их роли.

Также на вход подаются несколько примеров размеченных предложений из корпуса с именно этим предикатом.

Цель – сравнить эффективность с моделью из [4], а заодно проверить, насколько хорошо LLM понимают структуру естественного языка.

FrameBank

В качестве тренировочных данных используется корпус FrameBank.

Этот корпус состоит из предложений на русском с размеченными морфологическими, синтаксическими и семантическими признаками. В частности, в нём есть разметка семантических ролей.

Для каждого предиката найдём все предложения, в которых размечены его аргументы. Выделим из них несколько примеров, так, чтобы для каждой из возможных ролей был хотя бы один пример.

Пример промта

“You are a native Russian linguist specializing in semantic role labeling. You must find all the arguments of a given verb in the sentence and assign each argument a role from the given list. <...> Given a series of few-shot examples, please find all the arguments of the predicate "бесить" and label them with semantic roles from the following list: ["причина","субъект психологического состояния","субъект поведения"]. Here are the few-shot examples:

Example Text:

Пытались через него лазить , но он от этого начинал беситься , сбрасывал с себя людей .

Example Semantic Roles:

он#субъект психологического состояния; этого#причина

Here is the target sentence:

Убежден и готов даже спорить , что именно такое внимание Васильева к слову и смыслу больше всего бесит ту часть критики , которая работает в русле , условно говоря , театра Жолдака .”

LLM for Semantic role labeling

Для тестирования подхода использовались **Gemini 2.5 Flash** и **Mistral Medium 3**.

Для 62% процентов предложений Gemini произвела полностью правильную разметку, включая и поиск аргументов, и разметку ролей.

Проблемы:

- поиск аргумента в сложной фразе ("*Пока мы живём на земле, мы можем себя обмануть, что ещё есть время*")
- LLM может путать схожие по смыслу роли (например, “адресат” и “контрагент”)
- LLM зачастую нарушает данные ей правила, например выдаёт 2 слова вместо одного, или как-то меняет слово

Сравнение моделей

	Argument extraction: precision	Argument extraction: recall	Argument extraction: F1	Role identification: micro F1
Базовая модель [4]	74.5%	85.1%	79.4%	83.4%
Gemini 2.5 Flash	87.8%	85.0%	86.4%	83.1%
Mistral Medium 3	83.6%	85.4%	84.5%	83.9%

Сравнение моделей (роли раздельно)

	Базовая модель (F1)	Gemini 2.5 Flash
агенс	79.5%	82.8%
пациенс	86.9%	85.7%
тема	77.6%	86.9%
субъект псих. сост.	85.2%	90.4%
субъект перемещ.	85.9%	89%
причина	87.4%	88%
место	84.9%	82%
говорящий	75.8%	86.7%

Результаты из [5]

Model	Shot	CoNLL05 WSJ			CoNLL05 Brown			CoNLL12 Test		
		P	R	F1	P	R	F1	P	R	F1
HeSyFu	-	88.86	89.28	89.04	83.52	83.75	83.67	88.09	88.83	88.59
CRF2o	-	89.45	89.63	89.54	83.89	83.39	83.64	88.11	88.53	88.32
MRC-SRL	5-shot	0.04	0.45	0.07	0.02	0.23	0.04	8.65	0.19	0.37
	full-shot	90.34	89.58	89.96	85.47	83.80	84.62	88.52	88.39	88.45
PromptSRL (ChatGPT 3.5)	3-shot	39.19	41.73	40.42	37.59	41.32	39.37	36.57	40.83	38.58

Выводы из работы

- LLM, обучаемые в контексте, могут быть использованы для полу-автоматической семантической разметки предложений
- При этом, такой подход можно применять примерно одинаково для произвольной системы ролей и произвольного корпуса
- LLM обладают неплохим пониманием структурной семантики, и большинство ошибок, допущенных ими, могли бы быть допущены и не-экспертом, знающим язык

Итоги работы

- Первая часть работы была посвящена исследованию существующих моделей и выделению перспективных подходов к моделированию семантики
- Вторая часть была посвящена практической задаче разметки семантических ролей
- Полученный метод работает на уровне SoTA-решений, обучающихся на больших объёмах данных.

Дальнейшие шаги

- Гипотеза: эффективность зависит как от понимания моделью языка, так и от самой системы ролей
- Тогда стоит поэкспериментировать с системой ролей, поставив чёткие и легко применимые правила разметки
- Попробовать применить разметку для предикатов, для которых нету примеров разметки
- Оформить результат в статью