Разметка семантических ролей

Григорий Казачёнок

Напоминание: модели семантики

Цель моделирования семантики - создать математическую модель, которая бы описывала смысл высказываний на естественном языке.

Такая модель должна также позволять практические задачи обработки текстов.

В прошлый раз мы рассмотрели различные подходы к моделированию семантики, применявшиеся в науке.

Напоминание: 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

Задача заключается в нахождении аргументов и определении их ролей. Делать это вручную – очень трудоемкий процесс. 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.

[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.

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.

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

FrameBank

В качестве тренировочных данных используется корпус FrameBank.

Этот корпус состоит из предложений на русском с размеченными синтаксическими и семантическими ролями. Каждый аргумент представлен одним словом, которому приписана соответствующая роль.

Для каждого предиката найдём все предложения, которые его содержат. 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 использовалась Gemini 2.5 Flash.

Для 62% процентов предложений она 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”)

Сравнение с базовой моделью [4]

	Argument extraction: precision	Argument extraction: recall	Argument extraction: F1	Role identification: micro F1
Базовая модель	74%	85%	79%	83%
Gemini 2.5 Flash	88%	85%	86%	83%

Итоги

- LLM справляется с задачей SRL на уровне нейросети, специально обученной под SRL
- Большинство ошибок вызваны не непониманием языка, а сложностью системы ролей, и их мог бы допустить и человек
- При должной предобработке, LLM можно использовать для автоматической или полуавтоматической разметки корпусов

Дальнейшие шаги

- Поэкспериментировать с разными моделями и промтами
- Попробовать эту технику на более простой системе ролей, предназначенной для разметки непрофессионалами (например, QA-SRL)