Формализация семантики высказываний на естественных языках

Основные подходы, идеи

Введение

Семантика — раздел лингвистики, изучающий смысловое значение единиц языка.

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

Хотим создать математическую модель, которая бы описывала значение предложений и связывала бы их с объектами реального мира, о которых в них говорится.

Мотивация

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

Но такие модели уже существуют! Нейросети умеют это и многое другое.

Однако нейросети непрозрачны: из-за огромного количества параметров, LLM практически невозможно анализировать и никто не скажет, почему они дают один ответ, а не другой.

Мотивация

Другая проблема: хорошие LLM дорого обучать и использовать, а также требуется огромный объём данных.

Людям требуется на порядки меньше данных, что подсказывает, что можно построить модель лучше!

	Training Set (Words)	Training Set (Tokens)	Relative size (Llama 3 = 1)
Recent LLMs			
Llama 3	11 trillion	15T	1
GTP-4	5 trillion	6.5T	0.5
Humans			
Human, age 5	30 million	40 million	10^{-6}
Human, age 20	150 million	200 million	10^{-5}

source: https://www.educatingsilicon.com/2024/05/09/how-much-llm-training-data-is-there-in-the-limit/

Мотивация

Наше понимание языка включает в себя и другие задачи: например, прочитать меню в ресторане и на основе этой информации сделать заказ, или прочитать инструкцию и научиться по ней что-то делать.

Хорошая модель должна уметь не только связывать между собой предложения, но и связывать их с объектами реального мира.

Проблемы

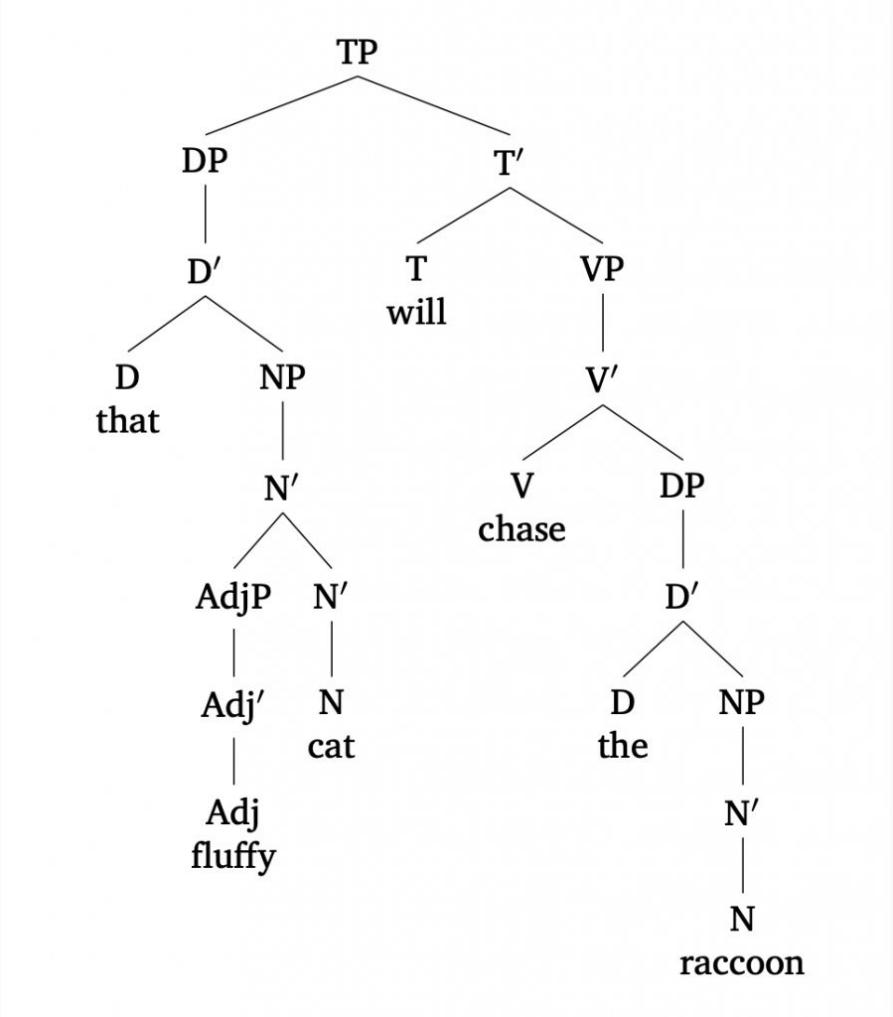
Было бы хорошо, если бы было так просто. Однако семантика естественных языков содержит огромное количество преград для её описателя:

- Неоднозначность: “если бы у Ивана был осёл, он бы его бил”
- Многое просто подразумевается: всем понятно, о ком говорит словосочетание “отец Гаусса”, однако у Карла Гаусса также был сын, которого тоже звали Гаусс
- …

Проблемы

Зависимость от контекста: согласно принципу Фреге, смысл составного выражения является функцией от его составляющих и способа, которым оно составлено.

Однако, смысл в естественных языках очевидно зависит и от слов вне выражения.



Модельный подход

Смысл определяется как выражение на формальном языке (обычно логике 1-го порядка, хотя подходит даже лямбда-исчисление, как у Монтегю).

Субъекты и объекты моделируются как переменные. Действия и описания моделируются с помощью предикатов и отношений.

Все переменные, в свою очередь, соответствуют объектам реального мира или предметной области.

Модельный подход

Пример:

I have a car.

Переводится в:

$$\exists e, y \; Having(e) \wedge Haver(e, Speaker) \wedge HadThing(e, y) \wedge Car(y)$$

Модельный подход

Изначальные положения добавляются в качестве аксиом. Когда мы хотим проверить, следует ли какое-то другое утверждение из данных, или вывести новое, запускаем процедуру вывода по формальным правилам.

Помимо тонкостей перевода предложений в логические формулы, есть проблемы, связынные с самой моделью. Например, как добавить аксиомы так, чтобы выводились все верные положения? Как добавить переменные для всех возможных объектов? Как быть с неточными высказываниями, которые нельзя выразить обычной логикой?

Вариации: вероятностные логики

В *Semantic Parsing using Distributional Semantics and Probabilistic Logic (Beltagy et al., 2014)*, авторы предлагают вместо обычной логики использовать Марковские логические сети (каждой формуле присваивается некая вероятность) и мягкую вероятностную логику.

Используя статистику, авторы составляют список объектов и “мягких” аксиом, например, правило $\forall x\ man(x) \Leftrightarrow guy(x) \mid p$, добавляется на основе того, что “man” и “guy” встречаются в похожих контекстах, со степенью уверенности p.

Векторные представления

Подойти можно и с другой стороны: представлять смысл не как выражение на формальном языке, а как набор чисел.

Word2vec и другие модели позволяют представлять слова как вектора.

В *Composition in distributional models of semantics (Mitchell and Lapata, 2010)*, авторы предлагают по некоторым правилам совмещать векторы слов в вектор словосочетания такой же размерности.

Процедурная семантика

Предлагает, по аналогии с (императивными) языками программирования, смотреть на семантику не как на математическую структуру, а как на последовательность команд. Это также ближе к тому, как человек воспринимает язык.

Такой подход применялся в робототехнике для понимания роботами команд на естественном языке (*Roy et al. 2005*, *Regier 1996*).

Хорошей модели всё ещё не придумали.

Спасибо!

Подготовил Григорий Казачёнок.