Colors in Context: A Pragmatic Neural Model for Grounded Language Understanding

Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts

Непомнящая М., Родионова Д.

Задача

Задача: создание модели, умеющей по описанию выбирать загаданный цвет из сета цветов. Это задание для слушающего

В большинстве работ с подобной тематикой решали задачу говорящего — порождение описаний цветов

Корпус

- 967 респондентов из Amazon Mechanical Turk
- 1059 игр по 50 раундов
- Референциальная игра на двоих
- Уровни сложности: close, split, far
- RGB
- После фильтрации: 53365 описаний из 469994 раундов 948 играх
- 15519 close, 15693 split, 15782 far

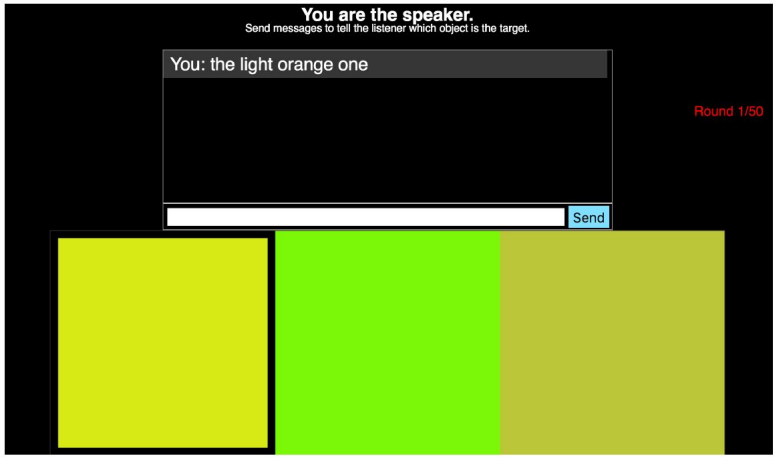


Figure 1: Example trial in corpus collection task, from speaker's perspective. The target color (boxed) was presented among two distractors on a neutral background.

Человеческое accuracy

- Слушающий: 97% far, 90% split, 83% close
- Метрики для говорящего: знаки, слова, сравнительные степени, конкретность, отрицание, превосходные степени
  - Знаки&Слова: для far описания короче, чем для split и close
  - Сравнительные&Превосходные степени: в far меньше, чем в других. В close больше всего превосходных, а в split — сравнительных
  - Отрицание: в far меньше, чем в других
  - Конкретность: в far меньше, чем в других. Иерархия цветов сделана с помощью WordNet

RSA

Listener-based listener:

- $l_0(t \mid u, L) \propto L(u, t)P(t)$
- $s_1(u \mid t, L) \propto e^{\alpha \log(l0(t \mid u, L)) - \kappa(u)}$
- $l_2(t \mid u, L) \propto s_0(u \mid t, L)P(t)$

l_2 напрямую не оценивает результат $L(u, t)$, он оценивает s_1, который оценивает l_0, который уже оценивает $L(u, t)$

Speaker-based listener:

- $s_0(u \mid t,L) \propto L(u, t)e^{-\kappa(u)}$
- $l_1(t \mid u,L) \propto s_0(u \mid t, L)P(t)$

l_1, оценивает s_0, который уже оценивает $L(u, t)$

t — цвет из сета контекстных цветов
u — описание цвета
$L(u, t)$ — функция интерпретации, выдает 1 или 0

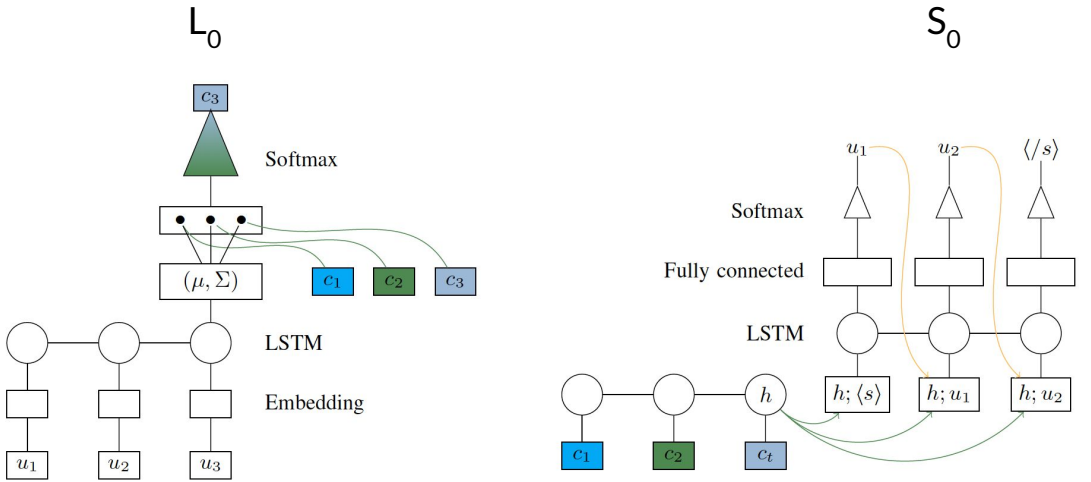
κ — функция потерь
α — параметр, регулирующий «рациональность» модели говорящего

Проблемы

- Ограниченный сет описаний
- Как определить $L(u, t)$?

Решение — RNNs!

RNNs



(a) The L_0 agent processes tokens u_i of a color description u sequentially. The final representation is transformed into a Gaussian distribution in color space, which is used to score the context colors $c_1 \ldots c_3$.

(b) The S_0 agent processes the target color c_t in context and produces tokens u_i of a color description sequentially. Each step in production is conditioned by the context representation h and the previous word produced.

Figure 3: The neural base speaker and listener agents.

Прагматические участники и ансамбль моделей

$$S_1(u \mid t, C; \theta) = \frac{L_0(t \mid u, C; \theta)^\alpha}{\sum_{u'} L_0(t \mid u', C; \theta)^\alpha}$$

$$L_2(t \mid u, C; \theta) = \frac{S_1(u \mid t, C; \theta)}{\sum_{t'} S_1(u \mid t', C; \theta)}$$

$$L_1(t \mid u, C; \phi) = \frac{S_0(u \mid t, C; \phi)}{\sum_{t'} S_0(u \mid t', C; \phi)}$$

$$\mathbf{L}_a \propto \mathbf{L}_0{}^{\beta_a} \cdot \mathbf{L}_1^{1-\beta_a}$$

$$\mathbf{L}_b \propto \mathbf{L}_0{}^{\beta_b} \cdot \mathbf{L}_2^{1-\beta_b}$$

$$\mathbf{L}_e \propto \mathbf{L}_a{}^{\gamma} \cdot \mathbf{L}_b^{1-\gamma}$$

Вместо *L(u, t)* — выученные веса **θ**

Чтобы в знаменатели считать сумму не для всех потенциальных описаний, будем брать 8 сэмплов из $S_0(u \mid i, C; \varphi)$ для каждого таргета

Обучение модели

- Делим датасет на три части: train, dev, test
- Препроцессинг: нижний регистр, токенизация, заменяем слова, которые встретились меньше 2 раз на <unk>, убираем высказывания слушающего
- Adam, ADADELTA и SGD для обучения RNNs
- Гиперпараметры подбирались грид серчем

Результаты

- S_0 по качеству не отличается от S_1. Сравнение с человеком по нашим метрикам: слова, символы, отрицания и конкретность имеют такие же паттерны, как и у человека. Единственное отличие в превосходных и сравнительных степенях: нет такого количества сравнительных степеней для split
- Лучше всего работает ансамбль из моделей

model	accuracy (%)	perplexity
L_0	83.30	1.73
$L_1 = L(S_0)$	80.51	1.59
$L_2 = L(S(L_0))$	83.95	1.51
$L_a = L_0 \cdot L_1$	84.72	1.47
$L_b = L_0 \cdot L_2$	83.98	1.50
$L_e = L_a \cdot L_b$	**84.84**	**1.45**
human	*90.40*	
L_0	85.08	1.62
L_e	**86.98**	**1.39**
human	*91.08*	

Table 3: Accuracy and perplexity of the base and pragmatic listeners and various blends (weighted averages, denoted $A \cdot B$). Top: dev set; bottom: test set.

Анализ результатов

1. L_2 лучше L_0, когда на вход подается сгенерированное S0 описание и L0 не удается по нему идентифицировать загаданный цвет. L_2 в таком случае считает, что этот цвет сложно описать, и поднимает ему вероятность
2. L_2 и L_0 лучше L_1, потому что L_1 не натренирована на выполнение задачи слушающего
3. L_1 помогает L_2 и L_0, потому что ей на вход изначально подается контекст

L_1 — speaker-based, L_2 — listener-based

Вопросы от нас

1. Зачем заменять на <unk>, если можно просто убрать эти слова? И подходит ли нам в качестве критерия фильтрации частота? Важно ли нам что-то, кроме наречий, прилагательных, сравнительных штук и not?
2. Ушла проблема ограниченного кол-ва описаний, но словарный запас все еще ограничен. Возможно ли как-то реализовать генерацию не встретившихся в выборке названий цветов (и нужно ли это)?

Вопросы из зала

1. Вообще цвет — это не универсальное понятие, поэтому цветовая семантика в разных культурах (в том числе в разных языках) отличается. Учитывали ли это авторы статьи, когда искали людей для игры?
2. Как именно listener-based модель помогает speaker-based модели?
3. Модели слушающего вообще не видят контекст, или это касается только L0? Почему так сделано?
4. Как вам кажется, не было ли разумнее ограничить описание цвета в игре каким-то числом слов (1-2-3), как это было сделано в исследовании, которые мы обсуждали на паре?

Вопросы из зала

1. Если между участниками-людьми был возможен диалог, получается, слушающий мог переспросить, если не понял объяснение? Мне кажется, что говорящий в таком случае мог очень сильно изменить исходное описание так, чтобы слушающему стало понятнее, но тогда первое и второе описание вступили бы в конфликт. Кажется, в таких случаях авторы учитывали оба описания ("We also remove listener utterances and concatenate speaker utterances on the same context."). Могло ли это повлиять на «успешность» полученных моделей в худшую сторону?
2. Авторы говорят о том, что они ожидали от говорящих опору на базовые цвета для описания более специфичных. Что делалось в том случае, когда участники не называли никаких цветов? Или, например, сначала могли сказать одно, а потом другое (передумав)? Устная речь в этом плане предоставляет свободу выбора, изменения своего мнения и предоставления новой совершенно другой информации