Сравнение стратегий подбора обучающих примеров для few-shot оценки языковой компетенции LLM
(на примере датасета BLiMP и модели Mistral)

Проектные задачи компьютерной лингвистики
Элеонора Измайлова
24.02.2025

1. Датасет

- **BLiMP:** The Benchmark of Linguistic Minimal Pairs for English [Warstadt et al., TACL 2020]
- создавался для выявления сильных и слабых сторон моделей (*n-gram, LSTM, Transformer-XL, GPT-2*) по сравнению с человеческими оценками
- состоит из **67** подразделов (‘парадигм’), каждый из которых содержит **1 000** минимальных пар предложений на одно грамматическое явление
- был сгенерирован **автоматически** с сохранением структурной аналогии до ключевой позиции контраста и использованием словаря с многоуровневой разметкой размером **3000** слов

Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
ARG. STRUCTURE	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
BINDING	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
CONTROL/RAISING	5	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*
DET.-NOUN AGR.	8	*Rachelle had bought that chair.*	*Rachelle had bought that chairs.*
ELLIPSIS	2	*Anne's doctor cleans one important book and Stacey cleans a few.*	*Anne's doctor cleans one book and Stacey cleans a few important.*
FILLER-GAP	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
IRREGULAR FORMS	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
ISLAND EFFECTS	8	*Whose hat should Tonya wear?*	*Whose should Tonya wear hat?*
NPI LICENSING	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
QUANTIFIERS	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
SUBJECT-VERB AGR.	6	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*

Table 2: Minimal pairs from each of the twelve linguistic phenomenon categories covered by BLiMP. Differences are underlined. *N* is the number of 1,000-example minimal pair paradigms within each broad category.

2. ‘Предыдущие’ исследования

Тестирование в [Warstadt et al., TACL 2020]: «forced-choice» (модель сравнивает вероятности двух вариантов)

- *минимальные пары создавались таким образом, чтобы различия в вероятностях объяснялись исключительно грамматическим контрастом, а не контекстом или длиной предложений*

Model	Overall	Ana. agr	Arg. str	Binding	Ctrl. rais.	D-n agr	Ellipsis	Filler. gap	Irregular	Island	NPI	Quantifiers	S-v agr
5-gram	60.5	47.9	71.9	64.4	68.5	70.0	36.9	58.1	79.5	53.7	45.5	53.5	60.3
LSTM	68.9	91.7	73.2	73.5	67.0	85.4	67.6	72.5	89.1	42.9	51.7	64.5	80.1
TXL	68.7	94.1	69.5	74.7	71.5	83.0	77.2	64.9	78.2	45.8	55.2	69.3	76.0
GPT-2	80.1	99.6	78.3	80.1	80.5	93.3	86.6	79.0	84.1	63.1	78.9	71.3	89.0
Human	88.6	97.5	90.0	87.3	83.9	92.2	85.0	86.9	97.0	84.9	88.1	86.6	90.9

Table 3: Percentage accuracy of four baseline models and raw human performance on BLiMP using a forced-choice task. A random guessing baseline would achieve an accuracy of 50%.

3. Данное (квази-)исследование

- ➔ тестирование путем **промптинга**
  - **zero-shot** [Acc. **83.58**]
  - **zero-shot CoT** [Acc. **87.01**]
- ➔ размер тестовой выборки: **670** предложений (**10** предложений на каждое из **67** грамматических явлений)
- ➔ **Цель:** определить лучшую стратегию подбора обучающих примером для ***few-shot*** оценки

3. Данное (квази-)исследование

Эксперименты:

1. **Без использования NNS по эмбеддингам**
   a. случайный выбор из ***всего*** датасета
   b. случайный выбор из ***подраздела*** датасета на то же грамматическое явление
2. **С использованием NNS *(по эмбеддингам предложений)***
   a. поиск по близким к ***грамматичному*** предложению в тестовой паре
   b. поиск по близким к ***неграмматичному*** предложению в тестовой паре
3. **С использованием NNS *(по эмбеддингам пар предложений)***
   a. поиск по близким к ***разности*** эмбеддингов тестовой пары
   b. поиск по близким к ***покомпонентому умножению*** эмбеддингов тестовой пары
   c. поиск по близким к ***усредненному*** эмбеддингу тестовой пары
   d. поиск по близким к ***конкатенированному*** эмбеддингу тестовой пары
4. **С использованием NNS *(комбинированный подход)***
   a. поиск по близким к конкатенации разности и ***усреднения***
   b. поиск по близким к конкатенации разности и ***покомпонентного умножения***
   c. поиск по близким к конкатенации разности и ***конкатенации***

3. Данное (квази-)исследование

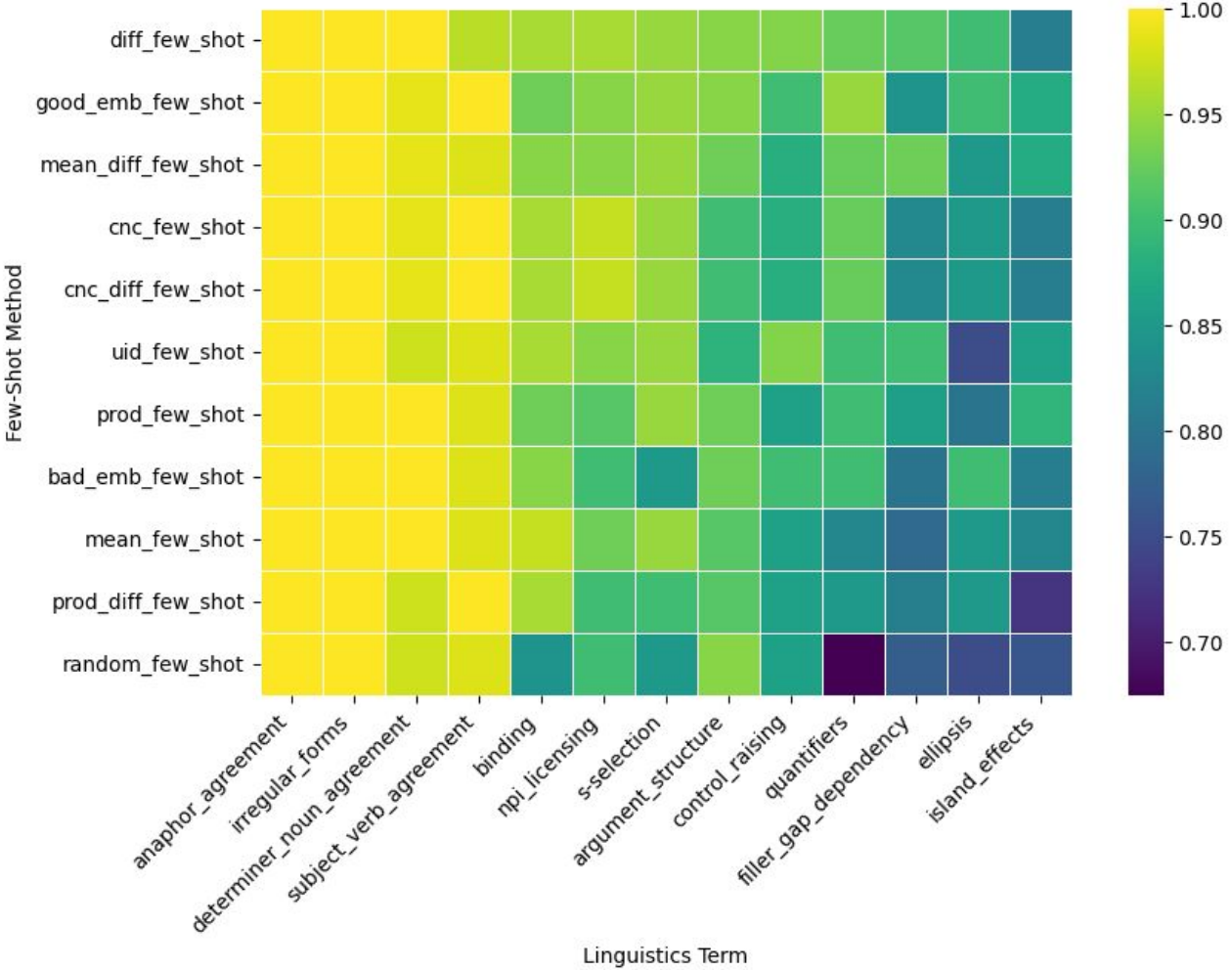
Эксперименты:

1. **Без использования NNS по эмбеддингам**
   a. случайный выбор из ***всего*** датасета [Acc. **86.87**]
   b. случайный выбор из ***подраздела*** датасета на то же грамматическое явление [Acc. **92.69**]

2. **С использованием NNS *(по эмбеддингам предложений)***
   a. поиск по близким к ***грамматичному*** предложению в тестовой паре [Acc. **93.28**]
   b. поиск по близким к ***неграмматичному*** предложению в тестовой паре [Acc. **91.04**]

3. **С использованием NNS *(по эмбеддингам пар предложений)***
   a. поиск по близким к ***разности*** эмбеддингов тестовой пары [Acc. **93.73**]
   b. поиск по близким к ***покомпонентому умножению*** эмбеддингов тестовой пары [Acc. **92.24**]
   c. поиск по близким к ***усредненному*** эмбеддингу тестовой пары [Acc. **90.09**]
   d. поиск по близким к ***конкатенированному*** эмбеддингу тестовой пары [Acc. **92.09**]

4. **С использованием NNS *(комбинированный подход)***
   a. поиск по близким к конкатенации разности и ***усреднения*** [Acc. **93.58**]
   b. поиск по близким к конкатенации разности и ***покомпонентного умножения*** [Acc. **89.94**]
   c. поиск по близким к конкатенации разности и ***конкатенации*** [Acc. **91.94**]


diff_few_shot
good_emb_few_shot
mean_diff_few_shot
cnc_few_shot
cnc_diff_few_shot
uid_few_shot
prod_few_shot
bad_emb_few_shot
mean_few_shot
prod_diff_few_shot
random_few_shot
Few-Shot Method
anaphor_agreement
irregular_forms
determiner_noun_agreement
subject_verb_agreement
binding
npi_licensing
s-selection
argument_structure
control_raising
quantifiers
filler_gap_dependency
ellipsis
island_effects
Linguistics Term
1.00
0.95
0.90
0.85
0.80
0.75
0.70

4. Дальнейшие планы

- ➔ сравнение с NNS по нормализованным эмбеддингам
- ➔ дообучение матрицы проекций эмбеддингов + NNS
- ➔ *TBD...*

5. Использованная литература

★ Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.