

Пробинг: локализация грамматики в нейронных сетях

Корнилов Альберт, Степанова Ангелина, Сухарева Мария, Шумакова Лада

Немного контекста



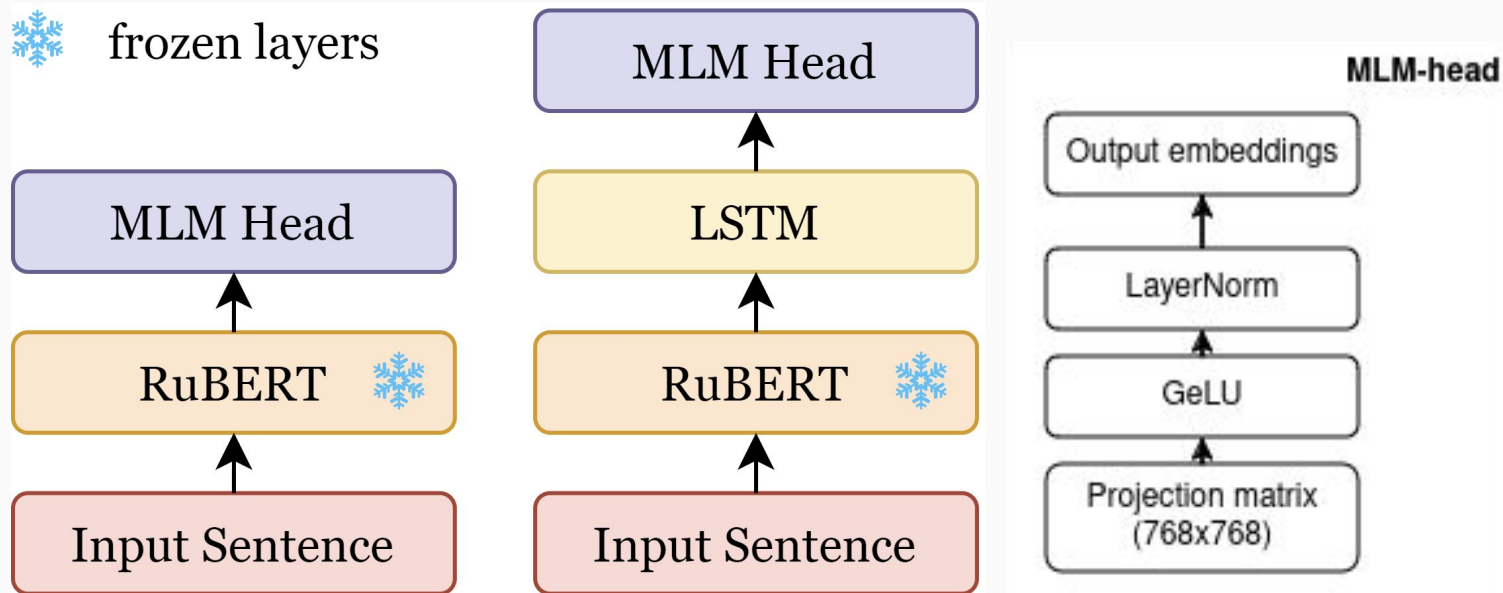
Исследование предыдущей команды [Kudriashov et al. 2024]

- данные на «поломанном» русском языке:
введение полиперсональности

Она делала кашу

- цель – локализовать полиперсональность в модели
- «замораживание» модели

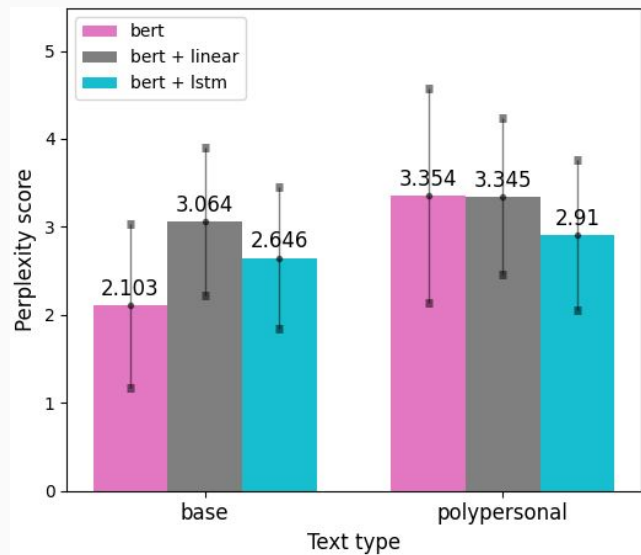
Метод локализации знаний в модели



Дообучение RuBERT на стандартных текстах корпуса было проведено Сергеем Кудряшовым

Оценка перплексии

Показатели перплексии обученных без флагов моделях





Пробинг на стандартном RuBERT

На выходах слоёв обучили классификаторы, которые должны были по вектору распознать, есть ли показатель полиперсональности в предложении или нет – **диагностический пробинг.**

Пробинг показал, что на выходах моделей, не обученных на полиперсональных текстах, примеры разделяются на три класса с высокой ассурасу (0,94), в то время как показатель ассурасу на случайных значениях равен 0,5.



Пробинг на стандартном RuBERT

К диагностическому пробингу есть вопросы:

- насколько корректно оценивать знания модели по результатам работы классификатора?
- действительно ли вся заложенная в векторах информация, на которую опираются линейные классификаторы, используется в моделях?

Итог: нужна дальнейшая интерпретация моделей



Дизайн эксперимента

- Так как при полиперсональности в форме глагола содержится информация об объекте (число и род), гипотеза:
- Будет ли модель, обученная на полиперсональных предложениях, лучше предсказывать, в какой форме должен быть объект?



Дизайн эксперимента

- Удалить из тестовой выборки прямой объект (всю составляющую, а не только вершину)

Исходный пример:

- Не на все планеты потащишь **ет** с собой **большую академическую** машину.

После удаления объекта:

указание на число и
род объекта

- Не на все планеты потащишь **ет** с собой [MASK].



Дизайн эксперимента

- Смотрели на заполнение только первой маски, чтобы снизить влияние отдельных примеров на общий скор
- *Старик читает только [MASK] и [MASK].*



только эта маска

- Смотрели только на первый токен, так как там, скорее всего, уже будет объект (или прилагательное, которое с ним уже согласовано по числу)



Выделение прямого объекта

Задача: подготовить датасет таким образом, чтобы все прямые объекты были заменены на MASK → проверка на полиперсональной модели.

Вопрос: что может пойти не так?

Ответ: все.

0. Мелочи жизни

- Парсинг притяжательных местоимений 3 лица (морфологическая неоднозначность)
- Маскирование согласуемых зависимых (и степень их согласуемости)
- Spasy не совместима с Python 3.13.2





1. Alignment предложений

- Предложения с полиперсональным глаголом и без могли парситься значительно по-разному → либо не выделялось ничего, либо выделялось что-то лишнее
- Решение:
 - сначала парсится норма
 - потом с нормальной частью выравнивается полиперсональная часть и на основе этого выравнивания ищется объект



2. Дети объекта

- Вассал моего вассала не мой вассал: метод `children` для токенов `zrasu` возвращают только зависимые на расстоянии 1 (в отличие от многих библиотек для графов)
- При этом работа `zrasu` в рамках грамматики зависимостей
- Решение: рекурсивная функция
- Теперь надо всего лишь понять, какие из согласуемых аргументов относятся конкретно к объекту



2. Дети объекта

А третьего дня лейтенант Франк принес ему **хлебец** с таинственным узором , **обнаруженный** в третьей роте .

А третьего дня лейтенант Франк принес ему MASK .



3. Синтаксическая неоднозначность

- *Топот услышал Ходсон , топот ног , бегущих в разные стороны , шум толпы , крики ...*
- Где объект?
- Правильно: Ходсон!



4. Кресло-качалка и другие сложные слова

- Сел в плетеное кресло-качалку и , оттолкнувшись ногой , слегка раскачал кресло .
- Сел в плетеное **кресло** - **MASK** и , оттолкнувшись ногой , слегка раскачал **MASK** .
- Решение: Ы (именно капслоком), re.sub в начале и в конце



5. Удаление генетива

- *Когда лотос попадает в руки последнего сипая , тот так же безмолвно исчезает и уносит цветок на соседнюю станцию .*
- *Когда лотос попадает в руки MASK , тот так же безмолвно исчезает и уносит MASK на соседнюю станцию .*



6. Спорные примеры

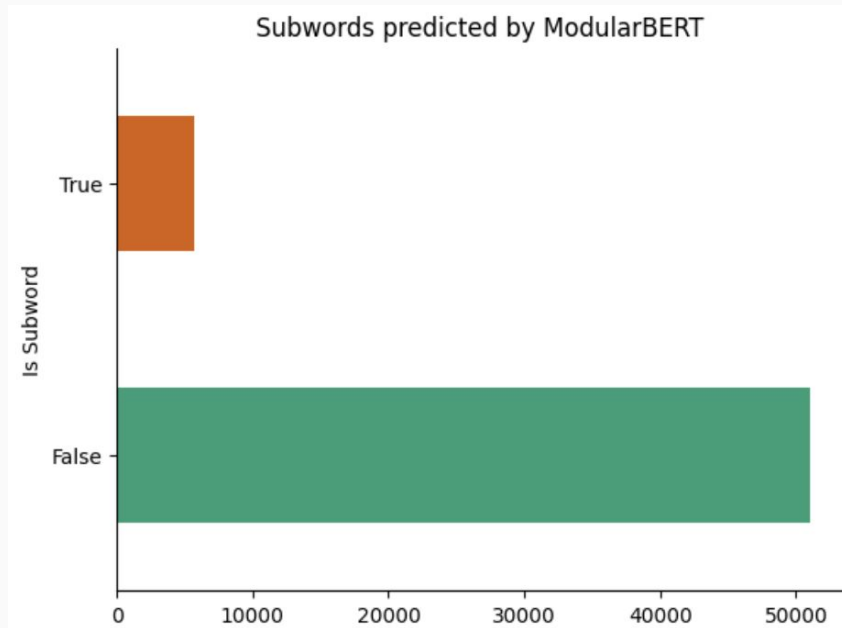
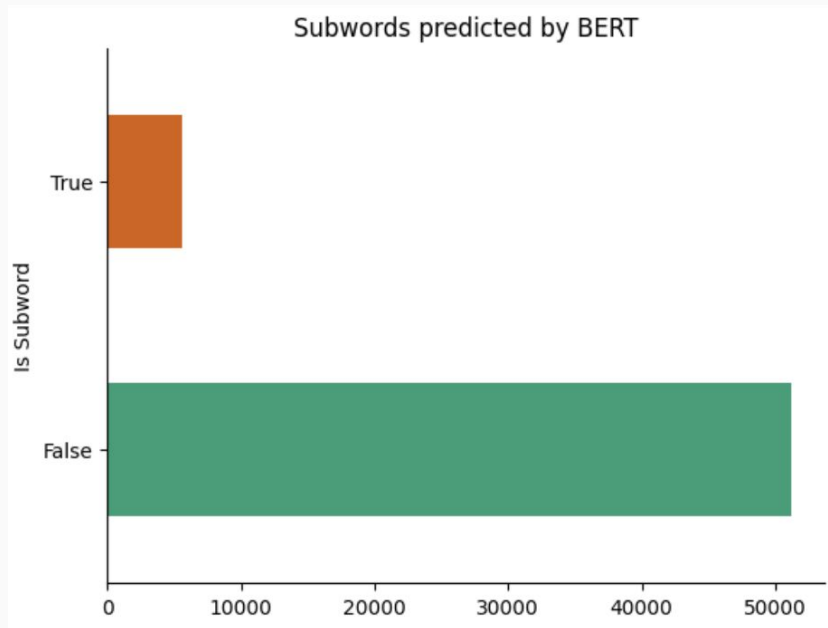
- Что делать?
 - Она показала Леле **свои руки : почерневшие суставы , обнаженное до костей гниющее мясо .**
 - Она показала Леле MASK MASK .
- Удаление субъекта и объекта



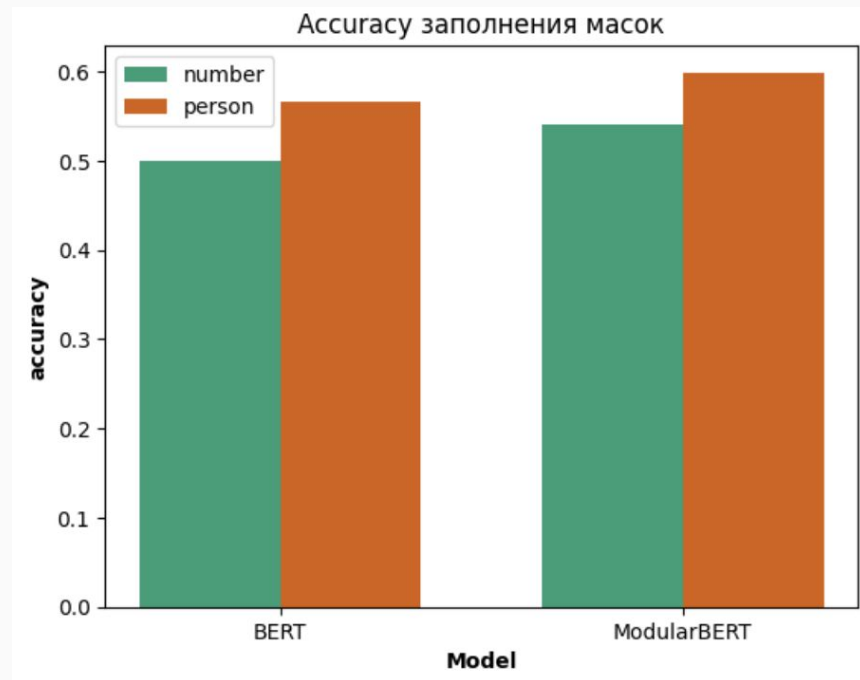
7. Время работы кода

- Процессинг 3 часа
- Нужна оптимизация
- Исправили баг и перешли к 14 минутам на 200,000 предложений
- Оптимизировали процессинг spasy и перешли к 3-4 минутам

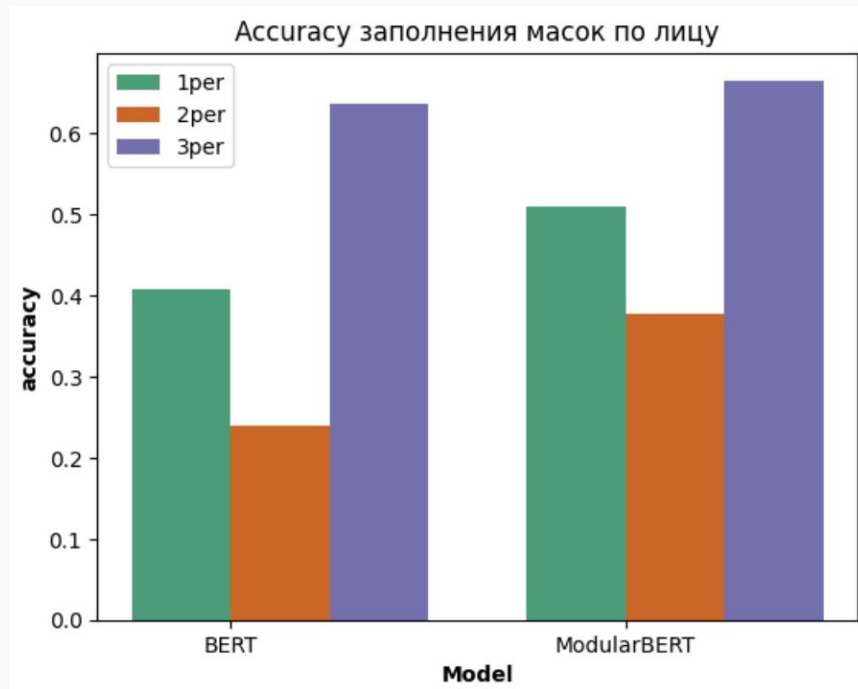
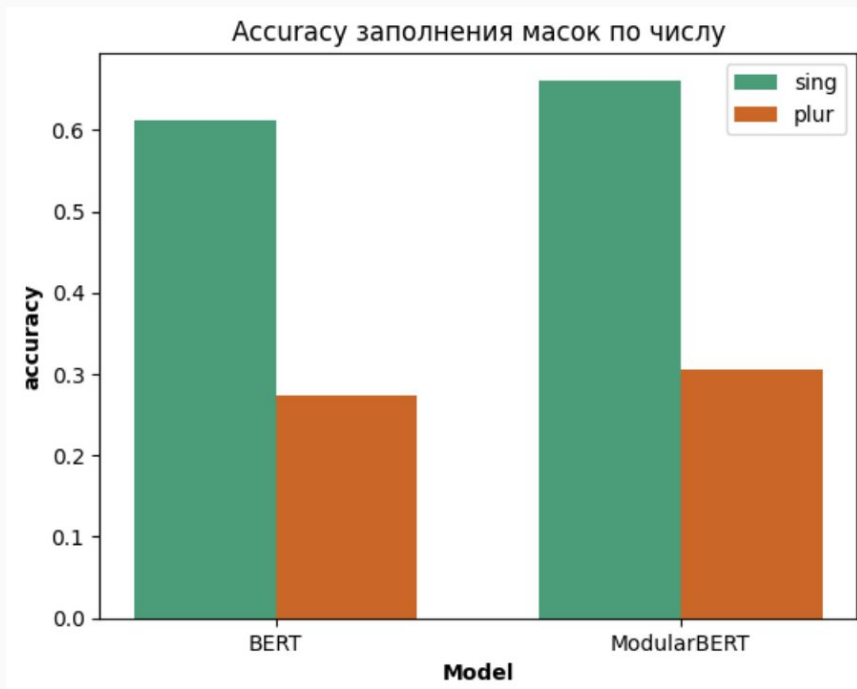
Результаты



Результаты



Результаты






Выводы

- Возможно, модель с полиперсональным модулем действительно выучила полиперсональность, так как чаще на месте объектов генерирует объекты с верным лицом и числом, когда информацию о лице и числе объекта в предложении передаёт только полиперсональный суффикс
- Но стоит в этом убедиться надежнее и порефлексировать над дизайном эксперимента



Дальнейшие планы

- Сделать такой же эксперимент на модели прошлого года с другой архитектурой
- Дообучить БЕРТ на полиперсональных примерах и повторить эксперимент
- Сравнить результаты
- Посмотреть на матрицы внимания моделей
- Если по результатам видно, что успешно локализовать полиперсональность не получается, то 



Дальнейшие планы (другое направление)

- Feature transfer
 - проверить все параметры в UDPipe
 - придумать новые признаки, которые можно добавлять как полиперсональность в разные языки (с автоматизацией)
 - обучить БЕРТ тому, что он умеет делать плохо
- Conditional discriminator
 - сгенерировать тексты с противоречивой разметкой
 - модифицировать RussianSuperGLUE
 - conditional embeddings
- Выходные веса от нетяжелой БЯМ → обучаем БЕРТ

Место для ваших вопросов





**Спасибо за
внимание!**