Experiments on Adaptation of End-to-end Coreference Resolution Models For Russian

Gutnik Gleb Konstantinovich
Moscow Lomonosov State University,
Moscow, Russia / Leninskie Gory, 1
glebgytnik@gmail.com

Abstract

In the present study, we conduct a series of experiments to improve the accuracy of the integrated ("end-to-end") model of coreference for the Russian language based on BERT. We introduce metadata on the speaker into a Russian-language coreference model for the first time, which results in an improvement of 2% as per F1-score as compared to the baseline model. We also propose an algorithm for automatic translation of the CONLL-2012 Shared Task corpus into Russian and present a part of this corpus for potential augmenting of data for coreference resolution.
Keywords: coreference resolution, end-to-end modelling, BERT, mention detection.

Опыт адаптации интегральных моделей разрешения кореферентности для русского языка

Гутник Глеб Константинович
Московский государственный
университет им. М.В. Ломоносова,
Москва, Россия / Ленинские горы, 1
glebgytnik@gmail.com

Аннотация

В настоящей статье мы представляем ряд экспериментов для улучшения точности интегральной ("end-to-end") модели кореферентности для русского языка на основе BERT. Мы размечаем метаданные о субъекте речи на двух русскоязычных корпусах для разрешения кореферентности и отмечаем, что это приводит к улучшению F-меры на 2% по сравнению с базовой моделью. Мы также предлагаем метод автоматического перевода корпуса CONLL-2012 на русский язык и представляем часть этого корпуса для возможного аугментирования обучающих данных.
Ключевые слова: разрешение кореферентности, end-to-end модель, BERT, определение упоминаний.

1 Введение

Разрешение кореферентности является одной из фундаментальных проблем (Wu, 2020) в обработке естественных языков. На этапе предобработки для таких задач как аспектный анализ тональности, извлечение информации, машинный перевод и т. д. автоматическое построение кореферентных цепочек может значительно повышать качество последующих задач.

Разрешение кореферентности – это выделение всех упоминаний в тексте, относящихся к одной и той же сущности во внеязыковой реальности. Например, в тексте ниже «Google», «компании», «её» и «корпорация Google» образуют кореферентную цепочку:

[#1] M1 Google хочет привлечь аудиторию M2 социальными играми . В M1 компании уверены , что именно M2 игры заставляют пользователей проводить больше времени в социальных сетях , чем на сайтах M1 её многочисленных сервисов . M1 Корпорация Google , M1 которой принадлежит популярнейший одноимённый поисковик , решила осваивать новый сегмент - M2 социальные игры .

Несмотря на относительную простоту и однозначность (хотя и во многом мнимую: см., например, (Nedoluzhko, 2013)) формулировки проблемы, моделирование кореферентности было и остается трудноразрешимой задачей даже в отношении английского языка, для которого кореферентные цепочки размечены на значительных объемах данных.

За последние несколько лет пальма первенства в освоении этой комплексной задачи перешла от систем на основе правил (например, (Raghunathan, 2010)) к моделям машинного обучения и глубокого обучения. При этом такие системы, как правило, представляли собой две последовательно соединенные модели: mention detection и mention clustering, каждая из которых выполняла свою отдельную задачу (выделение упоминаний в тексте и соединение этих упоминаний в цепочки, соответственно). Однако отделить «mentions» («упоминания», то есть такие именные фразы, которые участвуют в кореферентных цепочках) от «singletons» (одиночные упоминания сущностей, которые больше не представлены в тексте) без учета самих кореферентных связей оказалось трудно, в связи с чем впоследствии была предложена архитектура end-to-end (далее «интегральная модель», см. (Lee, 2017)), различные варианты и усовершенствованные формы которой остаются наиболее совершенными по настоящее время.

В рамках настоящего исследования мы адаптировали для русского языка одну из передовых моделей разрешения кореферентности и провели ряд экспериментов для выяснения факторов, которые оказывают влияние на точность разрешения кореферентности:

1) Различные способы учета информации о субъекте коммуникации
2) Архитектура контекстного слоя нейронов
3) Увеличение объема обучающих данных[1]

2 Базовая модель

Для оценки влияния новых показателей и/или архитектурных решений мы адаптировали для русского языка модель, представленную в (Lee, 2018) на основе эмбеддингов BERT (Joshi, 2019), (Joshi, 2020). При этом алгоритм coarse-to-fine pruning, отличающий модель (Lee 2018) от (Lee 2017), не применялся. Выбор базовой модели обусловлен тем, что такая конфигурация (BERT/SpanBERT + Coarse-to-fine Coreference) практически не уступает модели (Joshi, 2020) по показателю F-меры (79,2%[2] против 79,6% соответственно). Наиболее передовым по точности на корпусе CONLL-2012 подходом считается Coref-QA (Wu, 2020), в котором задача кластеризации упоминаний формулируется как предсказание промежутков текста на основе запроса («query-based span prediction»). Несмотря на это, мы приняли решение отложить эксперименты с данной моделью для будущих исследований, так как ее обучение связано с рядом комплексных задач при переносе на русский язык (так, авторы используют для предобучения набор данных SQUAD, примерно в 3 раза превосходящий соответствующий русскоязычный корпус Sber-SQUAD, а также применяют трехэтапное обучение модели на TPU).

Эмбеддинги ELMO из оригинальной модели (Lee 2018) заменяются на эмбеддинги BERT. С этой целью этого на вход в трансформер подаются id токенов и маска внимания, на выходе

[1] Данные и код для воспроизведения экспериментов настоящей статьи можно доступны по ссылке: C2F-RuBERT (github.com).

[2] allennlp-models (github.com) [Электронный ресурс] // URL: https://github.com/allenai/allennlp-models/pull/339/commits/3c63d84a1d19b67640854397236879eb17566591 (дата обращения: 13.06.2022).

фиксируются все веса скрытых состояний. Для извлечения репрезентаций промежутков мы инициализируем веса модели RuBert от Deeppavlov[3] (в базовом подходе SpanBERT+C2F применяется SpanBERT, однако, насколько нам известно, кодировщик этой архитектуры еще не обучен для русского языка).

Цель обучения состоит в том, чтобы предсказать распределение $P(y_i)$ по антецедентам для каждого промежутка i:

$$P(yi) = \frac{e^{s(i,y_i)}}{\sum_{y' \in Y(i)} e^{s(i,y')}}$$

где $s(i,j)$ — попарная оценка кореферентной связи между промежутком i и промежутком j. Модель включает три фактора для этой парной кореферентной оценки: (1) $s_m(i)$, является ли диапазон i упоминанием, (2) $s_m(j)$, является ли диапазон j упоминанием, и (3) $s_a(i, j)$ является ли j антецедентом i:

$s(i, j) = s_m(i) + s_m(j) + s_a(i, j)$

В частном случае фиктивного (нулевого) антецедента показатель $s(i,ϱ)$ вместо этого принимается равным 0. Общим компонентом, используемым во всей модели, являются векторные представления g_i для каждого возможного промежутка i. Они вычисляются с помощью двунаправленных LSTM (Hochreiter, 1997), которые отражают контекстно-зависимые представления границ промежутков. Оценочные функции s_m и s_a принимают в качестве входных данных следующие представления диапазона:

$sm(i, j) = w_a^{\top} FFNN_a([g_i, g_j, g_i \circ gj])$

где ◦ обозначает поэлементное умножение, *FFNN* обозначает нейронную сеть с прямой связью, а антецедентная оценочная функция $s_a(i, j)$ включает явное поэлементное сходство каждого промежутка $g_i \circ g_j$ и расстояние между двумя промежутками.

Из всех промежутков используется k наибольших элементов в соответствии с оценкой $sm(i)$, которая считается для каждого промежутка. Число k рассчитывается как $k = 0{,}4 * L$, где L – длина текста.

Модель обучается путем максимизации предельного логарифмического правдоподобия возможных антецедентов. Эта маргинализация необходима, поскольку наилучший антецедент для каждого промежутка является скрытой переменной.

2.1 Детали обучения

Максимальное число потенциальных антецедентов в наших экспериментах ограничивалась 150 или 200 (конкретные показатели каждого эксперимента указаны в таблицах результатов). Максимальное число эпох было задано равным 150, с ранней остановкой, если F-мера валидации не улучшалась за последние 10 итераций. Все эксперименты проводились при помощи графического ускорителя Nvidia Ampere A100. Время обучения на совмещенном корпусе RuCor (S. Toldova, 2017) + AnCor (E. Budnikov, 2019) составляет 30-60 минут. В качестве метрик оценки качества использовались MUC, B^3 и CEAFE с усреднением показателей точности, полноты и F-меры по трем метрикам. Другие гиперпараметры экспериментов представлены в файле конфигурации в репозитории проекта (см. стр. выше).

2.2 Данные обучения и тестирования

Для обучения модели (обучающая выборка и выборка валидации) мы использовали комбинированный датасет корпусов RuCor (S. Toldova, 2017) и AnCor (E. Budnikov, 2019), для тестирования использовалась только тестовая выборка AnCor (это было сделано с той целью, чтобы иметь возможность сопоставить результаты соревнования Dialogue-Evaluation-2019 по разрешению кореферентности со значениями, полученными в рамках экспериментов.

[3] DeepPavlov/rubert-base-cased · Hugging Face [Электронный ресурс] // URL: https://huggingface.co/DeepPavlov/rubert-base-cased (дата обращения: 13.06.2022).

2.3 Учет информации о говорящем

В ходе экспериментов мы заметили очевидное расхождение между существующими моделями для русского языка и общепринятыми векторами признаков для англоязычной базовой модели, включающих метаданные о говорящем (или пишущем) как фактор попарного скоринга потенциальных антецедентов: вектор признаков *φ(i,j)*, кодирующий информацию о говорящем и жанре из метаданных. Ввод такой информации интуитивно понятен: например, такой показатель необходим для отделения «я» говорящего в прямой речи от «я» автора. Выражение одной и той же сущности может быть разным у разных субъектов речи. Для того, чтобы учесть эту информацию, мы конвертировали наборы данных RuCor и AnCor в формат CONLL-2012, выполнили автоматическую детекцию прямой речи на основе правил и вручную разметили метаданные о субъектах речи с указанием их порядкового номера.

Для корпуса RuCor, в котором положение и длина упоминаний указаны в соответствующей колонке, мы перенесли тэг, начинающий упоминание (открывающая скобка и номер кластера), в соответствии с разметкой CONLL-2012, и закрывающий тэг (номер кластера и закрывающая скобка). Для корпуса AnCor предоставленные текстовые файлы были разбиты на токены и собраны в поле формата CONLL-2012. Далее по индексам, предоставленным в файлах для обучения, были расставлены открывающие и закрывающие тэги. Иная метаинформация (об именованных сущностях, грамматических показателях и т.д.) была опущена. Для разметки говорящих были взяты все случаи открывающих и закрывающих кавычек, для токенов между ними расставлены тэги (типа «speaker#1»), затем такая разметка была вручную дополнена точной классификацией говорящих («speaker#2», «speaker#3» и т.д.).

Результаты экспериментов показали, что инкорпорирование информации о говорящих (модель Coref-SP в таблице ниже) повышает F-меру на тестовой выборке на 1,5-2%.

Модель	Выборка	Precision	Recall	F1
e2e-Coref-LSTM (150 ант.)	Валидация	63,52%	52,75%	57,63%
	Тест	**63,42%**	**48,73%**	**55,09%**

e2e-Coref-LSTM-SP (150 ант.)	Валидация (max)	63,31%	54,19%	57,93%
	Тест	**64,91%**	**50,56%**	**<u>56,82%</u>**

Мы также применили метод передачи информации о говорящих “as input” (передача имени говорящего как эмбеддинга BERT, а не как бинарного вектора-признака), предложенную (Wu, 2020), однако такой способ подачи метаданных оказал отрицательное влияние на предсказательную точность модели.

2.4 Модификации архитектуры контекстуального слоя

Важное значение для кластеризации упоминаний имеет отражение данных о контексте в репрезентации промежутков: действительно, если анафор представляет собой местоимение или синоним/гипероним по отношению к антецеденту, репрезентация конкретного промежутка имеет важное значение детекции упоминаний, но не их кластеризации.

В end-to-end-модели кореферентности, представленной в (Lee, 2017), а также в C2F (Lee, 2018) для ввода контекстуальной информации использовалась двунаправленная LSTM. Мы заменяем единицы LSTM на единицы GRU, что при равных гиперпараметрах несколько улучшает F-меру

базовой модели на тестовой выборке и дает лучшую максимальную F-меру при валидации среди всех экспериментов, представленных в настоящей работе.

Модель	Выборка	Precision	Recall	F1
e2e-Coref-LSTM (200 ант.)	Валидация (max)	64,18%	54,24%	58,77%
	Тест	**63,45%**	**50,13%**	**55,97%**
e2e-Coref-GRU (200 ант.)	Валидация (max)	63,10%	55,40%	59,00%
	Тест	**64,51%**	**50,59%**	**56,68%**[4]

2.5 Увеличение выборки обучающих данных

Для повышения эффективности модели мы предприняли попытку увеличить набор данных для обучения. Известно, что корпуса RuCor и AnCor в совокупности примерно в 36 раз меньше обучающей выборки CONLL-2012 Shared Task (Le T. A., 2019). С этой целью мы выполнили автоматический перевод части корпуса CONLL-2012 следующим образом:

1) Конвертация корпуса CONLL-2012 в формат SACR (Oberle, 2018).
2) Перевод полученных SACR-документов на русский язык при помощи Google Translate API.
3) Конвертация переведенного корпуса SACR в формат CONLL-2012.

Формат SACR – это метод аннотации кореферентности, в котором промежутки помещаются непосредственно в текст документа в фигурных скобках с указанием их кластера, например:

The chief prosecutor at {C4 the International War Crimes Tribunal} has {C5 demanded} {C8 the new Yugoslav President} hand {C1 Slobodan Milosevic} over to face {C2 trial for war crimes} , but {C0 the U.S. , which has spent enormous amounts of time and energy fighting {C1 Milosevic}} , seems more willing to wait {C9 tonight} . Here 's {C6 our national security correspondent , John M Wethy} .

Главный обвинитель {C4 Международного трибунала по военным преступлениям} {C5 потребовал} от {C8 нового президента Югославии} передать {C1 Слободана Милошевича}, чтобы он предстал перед {C2 судом за военные преступления} , но {C0 США , которые потратили огромное количество времени и энергии на борьбу с {C1 Милошевичем}}, {C9 сегодня вечером}, кажется, больше готовы ждать . А вот и {C6 наш корреспондент по национальной безопасности, Джон М. Уэти}.

В отличие от формата CONLL-2012, где данные представлены в вертикальных полях, такой формат приемлем для автоматических переводчиков, при этом фигурные скобки сохраняются и в тексте перевода в начале и конце каждого промежутка. Далее полученный перевод может быть конвертирован обратно в формат CONLL-2012 (мы воспользовались для этого уже существующими инструментами)[5].

[4] Результаты в таблице отличаются от результатов, представленных на следующей странице, так как F-мера рассчитана не как среднее от трех метрик, а на основании средних precision и recall по метрикам.
[5] boberle/corefconversion: Conversion scripts for coreference (github.com). [Электронный ресурс] // URL: https://huggingface.co/DeepPavlov/rubert-base-cased (дата обращения: 13.06.2022).

«Слабым звеном» такого метода является именно этап перевода: было замечено, что из-за тэгов SACR разметки кореферентных цепей машинный перевод документа ухудшается. По этой причине мы отфильтровали полученные документы CONLL по выявленным наиболее частотным ошибкам (таким, как участки непереведенного текста в случае, если он обрамлен тэгами) и получили 474 лучших документа (Ontonotes-RU), которые позволили увеличить объем обучающих данных примерно на 25%.

Несмотря на улучшение точности, полнота при добавлении большего объема данных, вопреки ожиданиям, понижается. Вероятно, это объясняется различиями в принципах и правилах аннотации для английского и русского языка. Так, в русскоязычных наборах данных не применяются вложенные упоминания («nested mentions»), тогда как в CONLL-2012 они повсеместны. Следует отметить и ухудшение качества синтаксиса предложений, так как тэги SACR нередко «привязывают» содержимое к позиции в предложении, которая не является естественной для русского языка. Для лучшего применения новых данных необходимо редактирование разметки в соответствии с правилами и инструкциями, по которым аннотировались корпуса RuCor и AnCor, а также повышение их лингвистической приемлемости с точки зрения русского языка.

	Выборка	Precision	Recall	F1
e2e-Coref-LSTM +Onto-notes-RU	Валидация (max)	65,13%	51,44%	56,69%
	Тест	**65,59%**	47,45%	54,05%

В сопоставлении с результатами соревнования RuEVAL-2019 мы констатируем незначительное опережение лучшего результата дорожки. Колебание в размере F-меры может объясняться тем, что авторы (Petrov 2019) инициализировали отдельные слои весов BERT, а не каждый из них, а также особенностями тестирования (в своих экспериментах мы не применяем десятикратную кросс-валидацию).

Модель	MUC	BCUBE	CEAFE	Avg.F1
SCRb (Petrov 2019)	60.00%	48.89%	50.39%	53.61%
Baseline+ RuBERT(1–6–12) + RuCor (Petrov 2019)	66.74%	54.88%	51.72%	57.78%
e2e-Coref-LSTM-SP	**69,14%**	53,99%	51,11%	**58,08%**

2.6 Анализ распространенных ошибок модели

Тестирование предсказаний модели на текстах, не относящихся к корпусу RuCor + AnCor, выявило ряд распространенных случаев, когда модель ошибается:

1) Невыявление кореферирующих упоминаний, например:

Тактические группы Балтийского флота ({M1: БФ }) приступили к учениям в Балтийском море и на полигонах боевой подготовки в Калининградской области под руководством командующего БФ вице - адмирала Виктора Лиины . В тренировочных мероприятиях задействованы около 60 надводных боевых кораблей , катеров и судов обеспечения , свыше 40 самолетов и вертолетов , а также до 2 тыс . единиц вооружения , военной и специальной техники {M1: Балтийского флота } .

Вероятно, избежать подобных ошибок может разметка негативных примеров на обучающей выборке для повышения точности определителя упоминаний (согласно отдельным исследованиям, «negative sampling» может быть полезен для определения упоминаний: (Yu, 2020)). Трудность для автоматической разметки искусственных синглтонов может представлять только отсутствие «groundtruth»-разметки синтаксических зависимостей в корпусе AnCor.

2) Объединение в одной кореферентной цепи близких по значению слов разных частей речи (например, «Британия» - «британский» - «британцы»). Предположительно, это происходит из-за наличия подобных примеров в обучающей выборке, но они не являются кореферентными в строгом смысле слова и, как представляется, не появляются в тестовой выборке. Пересмотр и редактирование разметки выборки обучения/валидации могут устранить эту погрешность.

3 Заключение

Таким образом, в рамках настоящей работы мы представили комбинированный корпус RuCor + AnCor, размеченный данными о субъекте речи и тем самым добились прироста значения F-меры относительно базовой модели. Сравнение с лучшим результатом соревнования RuEval-2019 показывает конкурентность двух моделей. На основе анализа ошибок мы предлагаем пути улучшения точности существующих подходов. Кроме того, мы представили дополнительный аугментированный корпус для разрешения кореферентности, переведенный с английского языка, и предложили метод автоматического перевода данных, размеченных для разрешения анафоры и кореферентности, на русский язык.

Поскольку кореферентность – комплексное явление, являющееся ключевым для многих задач автоматической обработки текста, для обучения соответствующих моделей необходимы большие объемы размеченных корпусов. По этой причине представляется перспективным использовать современные возможности краудсорсинговых платформ для ручного редактирования автоматически переведенных данных CONLL-2012 Shared Task.

Мы также планируем продолжать исследования внешних показателей упоминаний (таких как субъект речи, дискурсивный статус предложения по классификации RST (Khosla, 2021) и т.д.) и проектировать модели автоматической разметки таких показателей (включая аттрибуцию прямой речи). Кроме того, планируется осуществить оценку обученных моделей на размеченных упоминаниях («gold mentions»).

Список литературы

1. **E. Budnikov Toldova S., Zvereva D., Maksimova D., Ionov M.** Ru-eval-2019: разрешение анафоры и кореферентности на русском языке. [Конференция] // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 29 мая — 1 июня 2019 г.). - Москва : [б.н.], 2019.
2. **Joshi, Mandar, Omer Levy, Daniel S. Weld and Luke Zettlemoyer.** BERT for Coreference Resolution: Baselines and Analysis. [Конференция] // EMNLP, 2019.
3. **Joshi, Mandar, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer and Omer Levy.** SpanBERT: Improving Pre-training by Representing and Predicting Spans. [Конференция] // Transactions of the Association for Computational Linguistics 8, 2020.
4. **Hochreiter S., Schmidhuber, J. Long.** Short-Term Memory [Статья] // Neural Computation. - 1997 г.. - 9 : Т. 8.
5. **Le T. A. Petrov M. A., Kuratov Y. M., Burtsev M. S.** Sentence Level Representation and Language Models in The Task of Coreference Resolution for Russian [Конференция] // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». - Москва : РГГУ, 2020., 2019.
6. **Lee Kenton, Luheng He, Luke Zettlemoyer.** Higher-Order Coreference Resolution with Coarse-to-Fine Inference [Конференция] // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

Technologies. - New Orleans, Louisiana : Association for Computational Linguistics, 2018. - T. Volume 2 (Short Papers).
7. **Lee Kenton, Luheng He, Mike Lewis, Luke Zettlemoyer.** End-to-end Neural Coreference Resolution. [Конференция] // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. - Copenhagen, Denmark : Association for Computational Linguistics, 2017.
8. **Nedoluzhko, Anna.** Generic noun phrases and annotation of coreference and bridging relations in the Prague Dependency Treebank. [Конференция] // Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. - Sofia, Bulgaria : Association for Computational Linguistics, 2013.
9. **Oberle, Bruno**. SACR: A Drag-and-Drop Based Tool for Coreference Annotation [Конференция] // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). - Miyazaki, Japan : European Language Resources Association (ELRA), 2018.
10. **Raghunathan Karthik, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, Christopher Manning.** A Multi-Pass Sieve for Coreference Resolution [Конференция] // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. - Cambridge, MA : Association for Computational Linguistics, 2010.
11. **S. Toldova M. Ionov**. Coreference resolution for russian: The impact of semantic features. [Конференция] // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». - Москва : Изд-во РГУУ, 2017. - Т. 1.
12. **Wu Wei, Fei Wang, Arianna Yuan, Fei Wu and Jiwei Li.** CorefQA: Coreference Resolution as Query-based Span Prediction. [Конференция] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. - Online : Association for Computational Linguistics, 2020.
13. **Yu, Juntao, Bernd Bohnet and Massimo Poesio.** Neural Mention Detection. [Конференция] // LREC, 2020.