Statistical evaluation of the syntactic similarity for a collection of texts

Zykova V.I.
NRU HSE, Moscow, Russia
vzykova2001@gmail.com

Abstract

In this article, we present several metrics for evaluating syntactic similarity between texts, as well as analyzing the results of their application to various corpora of texts in Russian. These metrics allow measuring the degree of syntactic complexity of trees in the corpora with varying degree of sensitivity, both in terms of branching in depth and the complexity of the actant structure. Such a measure can be useful in training of text generating algorithms or determining genre attribution or authorship.

Keywords: syntax; syntactic similarity; SynTagRus; Universal Dependencies; statistic analysis; dependency trees

DOI: 10.28995/2075-7182-2021-20-XX-XX

Статистическая оценка сходства синтаксической структуры для коллекции текстов

Зыкова В.И.
НИУ ВШЭ, Москва, Россия
vzykova2001@gmail.com

Аннотация

В этой статье мы представляем несколько метрик для оценки синтаксического сходства между текстами, а также анализируем результаты их применения к различным корпусам текстов на русском языке. Эти метрики позволяют с различным уровнем чувствительности измерять степень синтаксической сложности деревьев в корпусах, как с точки зрения глубины ветвления, так и с точки зрения сложности актантной структуры. Такой способ оценки может быть полезен при обучении алгоритмов генерации текста или определении жанровой атрибуции или авторства.

Ключевые слова: синтаксис; синтаксическое сходство; SynTagRus; Universal Dependencies; статистический анализ; деревья зависимостей

1 Введение

Задача сравнения синтаксической структуры предложений часто оказывается составной частью других задач компьютерной лингвистики. К ним относится, например, задача генерации текста, где требуется определять, насколько порожденные программой предложения удовлетворяют критерию «естественности», которая в данном случае понимается как схожесть по заранее выбранному набору параметров с текстами, написанными человеком.

Существует множество способов решения данной задачи, различающихся в первую очередь способом представления дерева для дальнейшего количественного выражения отличия. Возможно использование сверточных ядер [1], как например в работе [2], где используются поддеревья для репрезентации всего дерева зависимостей. Иным подходом являются эмбеддинги [3], лежащие в основе таких моделей как word2vec [4] или более сложных, основанных на архитектуре Transformers [5]. Принципиально другой взгляд на количественную оценку синтаксиса дает квантитативный поход, полезность которого активно обсуждается в работах [6] и [7]. Например, в работе [8] статистический анализ встречаемости синтаксических паттернов

позволяет сравнивать языки между собой, однако подобный подход возможен не только в случае со сравнением различных языков, но и в случае с корпусами одного языка, при условии наличия метрик, достаточно чувствительных для этого.

Таким образом, целью данной работы является описание нескольких метрик, разработанных для оценки синтаксического сходства текстов, представленных в формате Universal Dependencies [9], и анализ результатов их применения к различным корпусам текстов на русском языке. Для этого в первой части работы будет приведено краткое описание каждой из метрик, а во второй – сформулированы результаты анализа полученных с помощью них данных. В перспективе, описанные метрики могут помочь при определении качества работы порождающих алгоритмов или различении сгенерированных и созданных человеком примеров. В основе всех метрик лежит оценка разницы между исследуемым и референсным корпусами, поэтому если в качестве референсного корпуса взять тексты, написанные человеком (то есть «естественные» тексты), то сравнение позволит увидеть, насколько естественным или неестественным нам покажутся тексты из исследуемой коллекции.

2 Описание метрик

Одной из важных характеристик для синтаксических деревьев является глубина дерева и количество составляющих его вершин. Оценка по данному параметру позволяет в обобщенном виде оценить синтаксическую сложность деревьев. Соотношение между числом вершин и глубиной дерева позволяет выявить преимущественное направление разрастания: в ширину или в глубину. Можно рассчитать совокупность распределений средней по листьям глубины дерева при всех фиксированных значениях числа слов в предложении, используя при этом большую коллекцию синтаксически размеченных предложений.

Для сравнения двух наборов данных (сгенерированного корпуса и корпуса текстов, написанных человеком) в нашем случае будет достаточно визуализации с помощью тепловой карты, которая строится следующим образом: начала требуется получить набор пар <depth, length>, где depth – средняя по листьям глубина дерева, length – количество вершин в нем; после этого строится матрица тепловой карты так, что в строках оказывается глубина дерева, в столбцах - количество вершин, а в ячейках - частота данной комбинации двух параметров, которая и будет отображаться на тепловой карте цветом.

Если требуется получить численное выражение различия, можно использовать любую подходящую меру расстояния, например, корреляцию или дивергенцию Кулльбака-Лейблера. Для этого полученные для обоих корпусов, требующих сравнения, тепловые карты обрезаются так, чтобы исключить из рассмотрения большие «черные» области, заполненные нулями. После этого они нормируются и для них вычисляется выбранная мера расстояния.

Последующие характеристики, описанные в этой работе, специфичны конкретно для синтаксических деревьев зависимостей. Одной из таких характеристик является распределение вероятностей вертикальных (вида корень-…-лист) последовательностей или иначе – вероятностей оказаться в каждом из листьев дерева, рассчитанных из вероятности перехода по каждому из ребер в цепочке. Эта метрика позволяет оценить склонность деревьев к разрастанию в глубину, что с синтаксической точки зрения значит усложнение структуры за счет более разветвленного устройства конкретных зависимых.

Для подсчета данной статистики сначала требуется собрать общую статистику совместной встречаемости конструкций. С целью поиска компромисса между более тонким различением конструкций и излишним разрастанием данных был сделан выбор в пользу статистики, учитывающей в ширину для каждой вершины по одному ближайшему соседу справа и слева, а в глубину – предков с двух уровней выше по дереву. Таким образом, получается, что используемая статистка работает с тройками (см. общую схему на рис. 1).

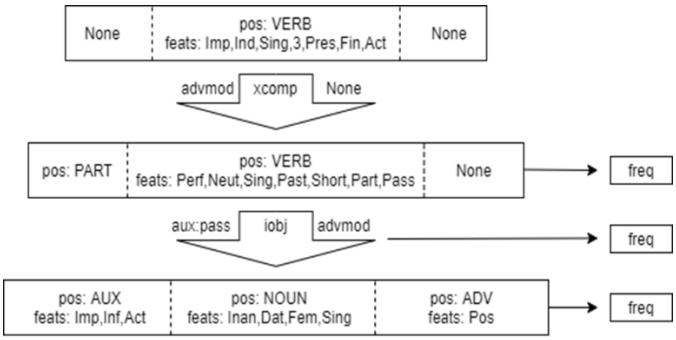

Рисунок 1: Общая схема структуры данных

Алгоритм получения требуемой структуры выглядит следующим образом: для всех уровней, начиная с уровня потомков корня дерева, перебираем дочерние вершины, формируя тройки вида <left, child, right>, где left и right – соответствующие соседи, а child – текущий потомок; после чего для конкретной комбинации из текущей тройки, троек, составленных на предыдущих уровнях тем же способом, и связей между всеми описанными элементами сохраняется частота встречаемости в корпусе.

После того, как была сформирована необходимая структура данных, на ее основе требуется подсчитать для каждого листа в дереве его вероятность. Для этого введем несколько обозначений: пусть $<v_r, v_1, \ldots, v_i, v_l>$ – это путь из корня v_r в лист v_l, состоящий из вершин $v_1, \ldots, v_i$. Тогда можно сказать, что данный путь состоит из ребер $<e_{r,1}, e_{1,2}, \ldots, e_{i-1,i}, e_{i,l}>$, для которых $n_{r,1}, n_{1,2}, \ldots, n_{i-1,i}, n_{i,l}$ – соответствующие частоты, взятые из структуры, описанной ранее, а $N_{r,1}, N_{1,2}, \ldots, N_{i-1,i}, N_{i,l}$ – суммы всех частот на уровнях, соответствующих каждому из ребер. Определив понятия, описанные выше, становится возможным привести итоговую формулу подсчёта вероятности пути $<v_r, v_1, \ldots, v_i, v_l>$ или листа v_l, что в данном случае одно и то же:

$$P_{r,l} = \frac{n_{r,1}}{N_{r,1}} \times \frac{n_{1,2}}{N_{1,2}} \times \cdots \times \frac{n_{i-1,i}}{N_{i-1,i}} \times \frac{n_{i,l}}{N_{i,l}}$$

Произведя описанные подсчеты для всех листьев каждого дерева в корпусе, можно сравнить распределение частот полученных вероятностей. В дальнейшем можно сравнивать подобные распределения для различных корпусов с помощью графиков (преимущественно гистограмм) или описанного ранее алгоритма подсчета расстояний.

Следующим способом оценки является матрица совместной встречаемости потомков при одной вершине. В данном случае используются только типы зависимостей (например, nsubj, obl, parataxis) без характеристик конкретных вершин, чтобы учитывать возможный состав аргументов и адъюнктов и общую склонность деревьев в корпусе к усложнению аргументной структуры.

Статистика собирается следующим образом: для каждой вершины в дереве ее представлением в статистике выбирается связь с родителем (т.е. тип потомка, которым является текущая вершина). После этого в цепочке дочерних вершин смотрим, какие типы зависимостей следуют друг за другом, и вносим полученную информацию в матрицу, в которой на данным момент содержаться частоты. Кроме выделяемых в корпусе типов в матрице также содержаться маркеры «BOS» и «EOS» - начало и конец последовательности соответственно. После того, как были обработаны все деревья в выбранном корпусе, матрица нормируется за счет деления каждого значения в строке или столбце на значение в соответствующих ячейках «BOS» или «EOS». Вид итоговой матрицы для одного из типов зависимостей представлен на рисунке 2 (где a_1, a_2, a_3 – частоты в конкретных ячейках, n – частота для «BOS» или «EOS» и в то же время – частота данного потомка). Всего таких матриц должно получиться столько же, сколько имеется различных связей в корпусе плюс один, так как добавляется элемент «BOS» (в случае SynTagRus – 40 таблиц).

	nsubj	…	case	EOS
nsubj	a_1/n	…	a_2/n	n/n
…	…	…	…	…
case	a_3/n	…		
BOS	n/n	…		

Рисунок 2: Матрица совместной встречаемости потомков

Использовать подобную матрицу можно двумя способами. Первый заключается в том, чтобы, повторив расчеты для второго корпуса, сравнить полученное распределение с референсным, использовав, например, алгоритм, схожий с тем, что был описан для распределения глубины и размера деревьев. Второй способ представляет собой подсчет вероятностей на основе полученной матрицы, что будет описано в следующем разделе данной работы.

Получив матрицу совместной встречаемости потомков, становится возможным оценить вероятность каждой цепочки дочерних вершин при одном родителе. В данном случае под вероятностью последовательности понимается вероятность получить последнего потомка при всех предыдущих. Для подсчета необходимого значения требуется для цепочки потомков текущей вершины взять для работы соответствующую последовательность типов зависимостей, имеющую вид <BOS, d_1, …, d_i, EOS>, где BOS, EOS – маркеры начала и конца последовательности, а d_1, …, d_i – связи текущей вершины с потомками. Следующим шагом среди описанных в предыдущем разделе следует выбрать матрицу M, соответствующую типу связи, которым текущая вершина связана со своим родителем. Дальше в матрице выбирается столбец M[d_i], заданный элементом d_i – последним в цепочке потомков, – и в этом столбце перемножаются ячейки, соответствующие всем потомкам цепочки, кроме последнего. Таким образом, получается следующая формула:

$$P_{d_i} = \prod_{j=1}^{i-1} M[d_i][d_j]$$

Выполнив описанные выше расчеты для каждой цепочки во всех деревьях корпуса и проверяемых данных, можно сравнить два распределения вероятностей цепочек с помощью визуальной оценки гистограмм или подсчета расстояний, несколько вариантов выполнения которого было описано выше.

3 Результаты для различных корпусов

В ходе работы описанные выше метрики были применены к различным синтаксически размеченным корпусам, среди них: четыре корпуса новостей (региональные, литературные, научные и мира моды), каждый содержит более миллиона словоупотреблений; два корпуса с автоматической синтаксической разметкой SynTagRus с помощью библиотек DeepPavlov и UD-Pipe и наборы текстов, сгенерированных с помощью GPT2[1] и GPT3[2], где в качестве затравок для генерации использовались первые три слова из предложений корпуса SynTagRus с целью дать модели возможность подстроиться под стиль корпуса и минимизировать стилистические отличия.

Для более полного понимания рисунков, приведенных в работе, стоит уточнить несколько деталей. Во-первых, gorky, science, vogue, news мы будем называть корпуса текстов, собранные из Gorky.media, Naked Science, Vogue и нескольких источников локальных новостей соответственно; gpt3 и gpt2 – тексты, сгенерированные с помощью предобученных моделей GPT3 и GPT2; str – корпус SynTagRus [10], а str_ud и str_dp – тот же корпус, но размеченный с помощью парсеров от UD-Pipe и DeepPavlov. Во-вторых, при построении тепловых карт каждый из

[1] https://huggingface.co/sberbank-ai/rugpt2large
[2] https://huggingface.co/sberbank-ai/rugpt3large_based_on_gpt2

перечисленных корпусов по очереди выбирался в качестве референсного, с которым потом сравнивались все остальные. Таким образом, на рисунках в строках расположены референсы, в столбцах – объекты сравнения, а на пересечении – дивергенция Кулльбака-Лейблера, полученная в ходе применения описанных метрик.

В результате было обнаружено, что они, с одной стороны, находят различия между корпусами, а с другой – позволяют объединять их в соответствии с описанными группами. Это можно увидеть на рисунке 3, где цветовая шкала тепловой карты показывает среднюю по всем метрикам дивергенцию Кулльбака-Лейблера для упомянутых корпусов (0, минимальная степень отличия, показана темно-синим цветом, а максимальная – белым).

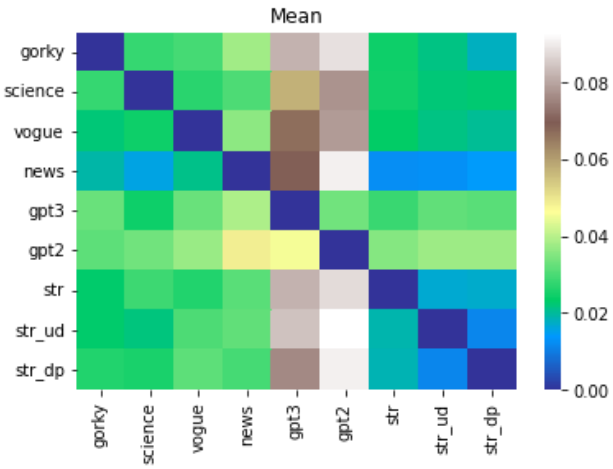

Рисунок 3: Среднее по метрикам расстояние между корпусами (gorky, science, vogue, news – новости; gpt2, gpt3 – автоматическая генерация; str, str_ud, str_dp – автоматическая разметка).

На рисунке можно отчетливо видеть, как новости, сгенерированные тексты и SynTagRus образуют три отдельные группы, причем относительно этого объединения можно отметить, что внутри групп степень взаимного отличия разнится. Наименьшее отличие демонстрируют сравнения различных синтаксических парсеров для SynTagRus, хотя стоит обратить внимание на тот факт, что друг от друга результаты автоматического разбора отличаются меньше, чем каждый из них от ручной разметки. Следующей по степени взаимного отличия является группа новостных корпусов, внутри которой среднее значение дивергенции значительно выше, чем в случае с SynTagRus, однако ниже, чем отличие от результатов автоматической генерации, которые образуют на тепловой карте, кроме собственно отдельной группы, явную «полосу отличия».

Говоря более подробно о результатах автоматической генерации с помощью GPT2 и GPT3, стоит обратить внимание на отдельные метрики. Как можно заметить на рисунке 4, три метрики из четырех (вероятность цепочек в ширину – “Width”, вероятность цепочек в глубину – “Depth” и соотношение глубины и количества вершин – “Size”) явно выделяют GPT в отдельную категорию, демонстрирующую значительные отличия в том числе и внутри самой себя. Четвертая метрика (длина горизонтальных цепочек – “Length”) также регистрирует для результатов автоматической генерации большее отличие от остальных групп, однако похожая ситуация оказывается у корпуса региональный новостей, который по остальным метрикам объединяется с другими новостными корпусами. Однако надо заметить, что порядок отличия для распределения длин цепочек меньше, чем для остальных трех метрик, что делает ее менее говорящей.

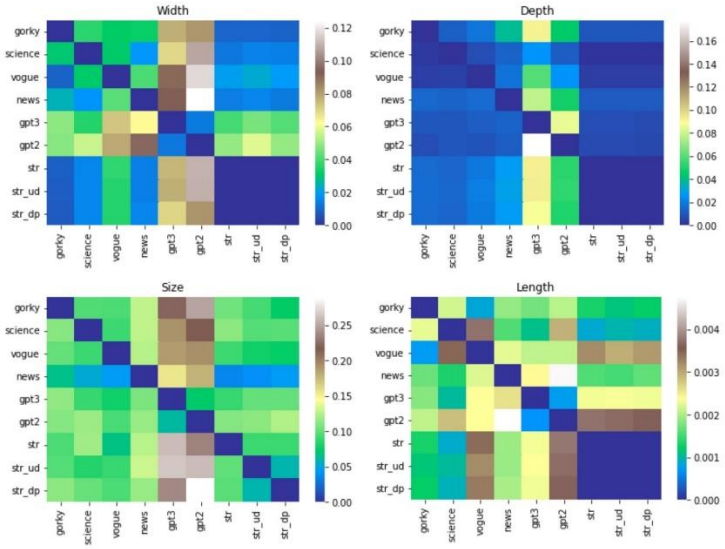

Рисунок 4: Сравнение по четырем метрикам (gorky, science, vogue, news – новости; gpt2, gpt3 – автоматическая генерация; str, str_ud, str_dp – автоматическая разметка).

4 Заключение

Описанные выше способы оценки соответствия деревьев референсным распределениям в корпусе могут быть полезны при решении различных задач. Для более успешного применения в зависимости от задачи требуется выбирать разные референсные данные так, чтобы в одном случае можно было говорить о «естественности» порождаемых алгоритмом предложений, а в другом – о различиях в синтаксической структуре, свойственной конкретным жанрам, авторам или, в общем случае, корпусам.

Как можно видеть из описания результатов применения метрик и рисунков, их иллюстрирующих, описанные метрики позволяют находить и оценивать различия между корпусами с разной степенью схожести, поэтому при правильном выборе референса становится возможным использование описанных метрик при обучении порождающих моделей в качестве части используемой функции потерь или при оценке в качестве одной из метрик. Также возможно использование в качестве одного из признаков текста при решении задачи различения стилей.

References

[1] Haussler D. Convolution kernels on discrete structures // Technical report, Department of Computer Science, University of California at Santa Cruz. — 1999. — Vol. 646.

[2] Smola A., Vishwanathan S. V. N. Fast kernels for string and tree matching // Advances in neural information processing systems. — 2002. — Vol. 15.

[3] Almeida F., Xexéo G. Word embeddings: A survey // Computing Research Repository. — 2019. — Vol. arXiv:1901.09069. — Access mode: https://arxiv.org/abs/1901.09069.

[4] Mikolov T. et al. Efficient estimation of word representations in vector space // Computing Research Repository. — 2013. — Vol. arXiv:1301.3781. — Access mode: https://arxiv.org/abs/1301.3781.

[5] Wolf T. et al. Transformers: State-of-the-art natural language processing // Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. — 2020. — P. 38–45.

[6] Gibson E., Fedorenko E. The need for quantitative methods in syntax and semantics research // Language and Cognitive Processes. — 2013. — Vol. 28. — №. 1-2. — P. 88–124.

[7] Sprouse J., Almeida D. The empirical status of data in syntax: A reply to Gibson and Fedorenko // Language and Cognitive Processes. — 2013. — Vol. 28. — №. 3. — P. 222–228.

[8] Klyshinsky E. S., Karpik O. V. Quantitative Evaluation of Syntax Similarity // Mathematica Montisnigri. — 2019. — Vol. 46. — P. 123–132.

[9] Nivre J. et al. Universal Dependencies v2: An evergrowing multilingual treebank collection // Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020). — Marseille, 2020. — P. 4034–4043.

[10] Droganova K., Lyashevskaya O., Zeman D. Data conversion and consistency of monolingual corpora: Russian UD treebanks // Proceedings of the 17th international workshop on treebanks and linguistic theories (tlt 2018). — Oslo, Norway, 2018. — Vol. 155. — P. 53–66.