Научно-технический вебинар HTP&HITs ТГУ

Дата и время: *8 февраля 2022*

Спикер: *Алексей Тихонов, Яндекс, Берлин, Германия*

Тема: *Текст: стиль + контент + структура*

Сайт:
https://ntr.ai/webinar/nauchno-tehnicheskij-vebinar-tekst-stil-kontent-struktura/

Запись: *https://youtu.be/8CRBKwU6pCk*

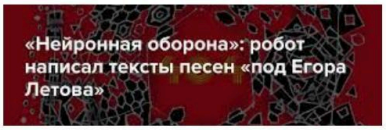
Презентация:
https://drive.google.com/file/d/1eT1CKNhmNVCcA3OU_0XXVwRD643WEV8z/view

TEXT
style + content + structure

Tikhonov Aleksey
@altsoph

about me

- **Senior Data Analyst**
  @ Yandex, Berlin
- **External Senior Researcher**
  @ LEYA Lab, HSE University
- **Researcher-collaborator**
  @ Yandex Research



about me

TITLE	CITED BY	YEAR
Guess who? Multilingual approach for the automated generation of author-stylized poetry A Tikhonov, IP Yamshchikov 2018 IEEE Spoken Language Technology Workshop (SLT), 787-794	30	2018
What is wrong with style transfer for texts? A Tikhonov, IP Yamshchikov arXiv preprint arXiv:1808.04365	20	2018
Style transfer for texts: Retrain, report errors, compare with rewrites A Tikhonov, V Shibaev, A Nagaev, A Nugmanova, IP Yamshchikov Proceedings of the 2019 Conference on Empirical Methods in Natural Language ...	13	2019
Decomposing textual information for style transfer IP Yamshchikov, V Shibaev, A Nagaev, J Jost, A Tikhonov arXiv preprint arXiv:1909.12928	8	2019
Style-transfer and Paraphrase: Looking for a Sensible Semantic Similarity Metric IP Yamshchikov, V Shibaev, N Khlebnikov, A Tikhonov arXiv preprint arXiv:2004.05001	5	2020
Learning literary style end-to-end with artificial neural networks IP Yamshchikov, A Tikhonov Advances in science, technology and engineering systems journal 4 (6), 115-125	4	2019
Dyplodoc: Dynamic plots for document classification A Malysheva, A Tikhonov, IP Yamshchikov arXiv preprint arXiv:2107.12226	2	2021
Style transfer for texts: to err is human, but error margins matter A Tikhonov, V Shibaev, A Nagaev, A Nugmanova, IP Yamshchikov	1	2019
StoryDB: Broad Multi-language Narrative Dataset A Tikhonov, I Samenko, IP Yamshchikov arXiv preprint arXiv:2109.14396		2021
Chekhov's Gun Recognition A Tikhonov, IP Yamshchikov arXiv preprint arXiv:2109.13955		2021

Style

remember Prisma app?

Алексей Тихонов: В свое время нашумевшее приложение Prisma интуитивно показывает, что мы знаем, что такое стиль в изображениях, и можем узнать, чей стиль на той или иной картинке. При этом если пытаться это формализовать, то это сложно, и в частности для изображения это решалось неявным образом путем подгонки фичей на ранних слоях. Считается, что при обработке изображения сверточными сетями на ранних слоях идет работа с мелкими деталями, а на более высоких – с более абстрактными вещами. И если мы абстрактные сохраняем, а мелкие заменяем, то мы вроде как поменяли стиль.

С текстом так не получается. Есть такое понимание у нас, что у текста есть стиль.

Есть древняя книжка «Упражнения в стиле», там один рассказ короткий про поездку человека в автобусе или в трамвае переписан сто раз в разных стилях, разных нарративах, от лица разных персонажей.

По крайней мере, это некий задел, который показывает, что мы тоже можем узнать одну и ту же историю, написанную в разных стилях, и сопоставить ее разные версии между собой и сказать, что на самом деле это одна и та история, текст один и тот же может быть.

what is text style?

Как нам это помогает? На самом деле никак.

[arXiv:2110.15871]

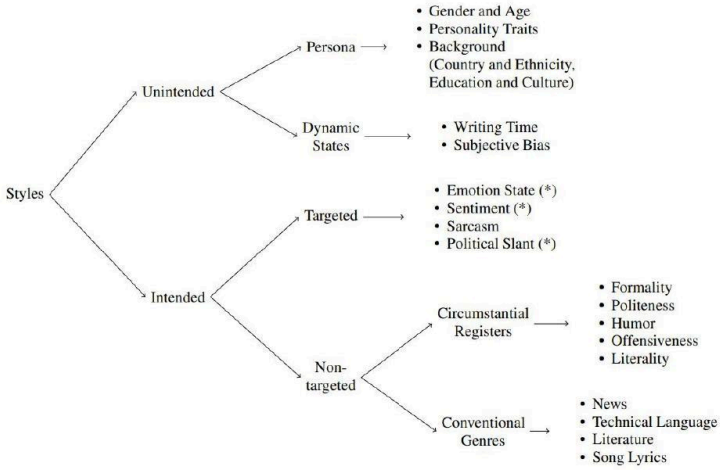


Figure 2. Hierarchy of styles guiding our discussion. Each branch defines different challenges for style transfer and illustrates how styles relate to one another. Asterisks (*) mark the nodes which are on the fence between content and style, since altering their attributes seems to bring a loss in the input content, but they are included in the hierarchy because they have been considered as styles for the transfer goal.

Есть, например, недавняя работа, которая посвящена исключительно попыткам систематизации формальных определений текстового стиля. Люди перекопали теоретическую литературу за долгое время. Это, кажется, ноябрьская работа. Они пытаются выделить, как же в разных задачах, в разных работах люди пытаются дефинировать стиль текста.

text style transfer

somehow related to:

- image style transfer
- NMT
- paraphrase generation
- summarization

Давайте еще попробуем поговорить о том, на что похож текстовый трансфер, о задачах текстового трансфера. С одной стороны, чем-то она для нас аналогична

image style transfer, то есть работе со стилем изображения. С другой стороны, это такая классическая задача «текст-в-текст», при этом она должна сохранять смысл, но что-то менять, это если неформально подходить к задаче. В принципе, тем же условиям удовлетворяет машинный перевод: он сохраняет смысл и меняет язык. А задача суммаризации – сохранять смысл, но менять длину текста. Или, например, задача генерации парафраза. В общем, чем-то оно похоже.

why transfer text style?

- personalization / targeting
- media content generation
- formality / politeness / detoxification

Зачем вообще можно решать такую задачу? На самом деле задача имеет некую прикладную пользу помимо того, что она интересная и позволяет нам закопаться в то, из чего состоит текст и как можно с ним работать. Понятно, есть задачи всякой персонализации, таргетинга как в рекламе, так и в медийном поле подача полезной информации в более удобном режиме. И есть такой популярный частный случай, который называется detoxification, то есть попытка повысить вежливость и убрать оскорбительность текста и так далее. Можно придумать и более интересные задачи.

style definition

- no universal definitions
- sentiments / dialects / author's style / …
- style is non-orthogonal to content
- typically defined by explicit examples

[arXiv:1808.04365]

Мы скажем, что мы не знаем, как это перевести, поскольку никто не знает. И чаще всего в прикладных задачах стиль определяется как корпус. Допустим, мы будем рассматривать дискретный стиль, и даже более того – бинарный. То есть у нас есть два стиля – один и второй, и мы хотим определять, к какому из них принадлежит конкретный текст, и, может быть, менять у текста атрибуты его принадлежности к конкретному стилю. Тогда мы скажем, что у нас этот стиль задан двумя корпусами: один соответствует первому стилю, второй соответствует второму стилю. Эта штука расширяется до большего числа дискретных стилей, в смысле до дискретного набора стилей, и можно думать и о непрерывном, но обычно дело до этого не доходит.

В такой постановке можно под стиль затащить разные вещи: можно попытаться сказать, что сентимент текста – это стиль, можно сказать, что конкретный авторский стиль – это пример стиля, или диалект и так далее.

Мы все понимаем, что стиль на самом деле не является какой-то изолированной от контента, от смысла, от содержимого текста вещью. Хочется, конечно, думать, что он ортогонален, но это не так. И простой пример – бóльшая часть работ, по крайней мере начало работ по текстовому трансферу, использованного в корпусе YELP. Это корпус состоит из предложений, набранных из отзывов на разные заведения, положительных и отрицательных. То есть в качестве разметки использовалась оценка самого отзыва, и, соответственно, в положительных – «в этом кафе очень вкусный кофе, быстро приносят еду и очень милая официантка», а в отрицательных – «кофе был холодный, официант грубый, нам не понравилось».

Уже здесь очевидно, что у этих сообщений разный стиль, если мы считаем сентимент стилем, но при этом у них нет одинакового контента. Контент у них как раз разный, противоположный по смыслу.

Здесь можно про это еще говорить достаточно долго, но мы лучше двинемся дальше и посмотрим, как нам с этим работать. Я только добавлю, что умозрительно кажется, что нам достаточно, чтобы множество текстов, разделенное на стили, имело какой-то зооморфизм. Если мы, условно, каждому тексту в одном стиле можем сопоставить какой-то текст в другом стиле, то, наверное, этого достаточно. Возможно, это слишком сильное для нас условие, но понятно, что в этом случае мы не требуем сохранности содержимого контента, а требуем, чтобы такое отображение существовало, и тогда можно пытаться решать задачу установки этого отображения. Обычно мы работаем со стилевыми корпусами, заданными явным образом.

Какие у нас есть корпуса? Корпуса с каждым днем всё лучше и больше, но базовое деление корпусов состоит из двух веток: первая – корпуса непараллельные, вторая – параллельные. Я думаю, что все знают, что такое параллельные корпуса, но на всякий случай проговорю это. Параллельным, или выровненным корпусом называется корпус, в котором есть явное соответствие текстов в одном и в другом стиле. Эта терминология пришла к нам из машинного перевода. Если у нас есть кусочки текста, для которых есть явный перевод, то мы на нем учимся. При машинном переводе есть много-много лет накопления этих корпусов, а во-вторых, есть _____ (00:14:28), но вполне логичное место, откуда можно взять эти самые параллельные корпуса. Например, в какой-нибудь Канаде или Швейцарии, где

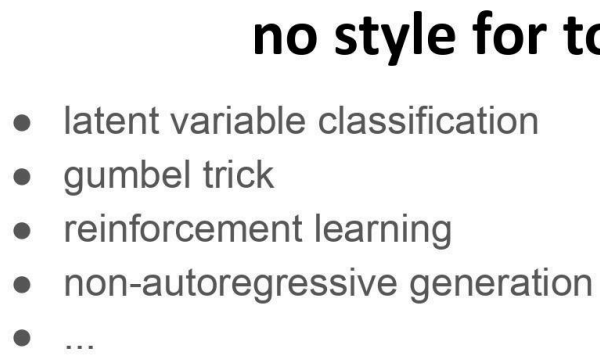
несколько языков государственных, суды, парламентский слушания и прочее – транскрипция пишется сразу на нескольких языках, и это автоматически дает хотя и смещенный по тематике, но выровненный параллельный корпус.

Из стилевых параллельных корпусов я находил несколько. Во-первых, существует несколько версий Библии. Она отличается: есть детская Библия, есть недетская Библия и так далее, и она выровнена по стихам. Есть Шекспир, некоторое подмножество работ Шекспира на старом английском, что называется middle English, и переписанные современным языком. Это тоже как один из примеров. Но это очень маленькие корпуса. То есть магия типа машинного перевода на них не заводится.

Есть GYAFC, это корпус, который опубликовала команда сервиса Grammarly, и насколько я знаю, это прямо ручная разметка, ручной рерайт, то есть они посадили токеров или толокеров, и они всё сделали. Это про вежливость, насколько я понял.

Давайте посмотрим на непараллельные корпуса. Их много, их сильно больше и их можно делать самим. Если у вас есть какой-то массив текстов, и к нему – какая-то метаразметка, то можете по этой разметке разбить текст и с ним дальше работать. Базовый – это YELP, как я сказал. Есть всякие вежливые и невежливые. Можно взять множество твитов и разбить их по эмоджи. Например, если присутствует эмоджи, то можно сказать, что это некая прокси-метка для сентимента и использовать ее в таком качестве.

Именно для того, чтобы лучше понимать, как это работает, люди взяли небольшое подмножество из YELP – кажется, 1 000 отзывов, – и сделали для них еще ручной рерайт уже параллельно. Он бывает полезен, например, для более точной оценки качества _____ (00:17:37). Например, можно построить метрику BLEO и померить, насколько хорошо у нас получается решать задачи текстового трансфера.



Еще пара слов о сложности задачи. Одна из сложностей работы с текстом – то, что текст является дискретным. Для людей он состоит из слов и букв, если мы говорим о западной культуре, а для типичной модели он состоит из токенов, и стиль обычно не определен на уровне токенов. При этом большая часть элементов модели умеют регулировать текст по токенам. То есть на каждом шаге она решает задачу выбора одного следующего токена, и этот токен сам по себе стилевой окраски не имеет или имеет очень слабую обычно. Поэтому мы должны много раз решить элементарную задачу выбора токена, но при этом в самом конце мы узнаем, что у нас получилось: получился у нас нужный стиль или нет.

Это неприятная ситуация, и люди придумали какое-то количество хаков, костылей и более интеллектуальных подходов к разрешению этой проблемы. Мы поговорим про несколько из них, но прямо идеального решения здесь, кажется, пока нет.

Классический подход – это reinforcement learning, но он плохо работает с текстом. Единственный, кто заявляет регулярно, что у него получается делать RL поверх сложных текстовых моделей, это OpenAI, но по крайней мере мои попытки воспроизвести некоторые их результаты и _____ (00:19:48) люди, с которыми я про это говорил, они, скорее, умеренный оптимизм внушают. Но недавно они опять заявили, что они тюнили трансфер _____ (00:20:00) очередную задачу с помощью RL.

То есть в принципе, наверное, это выход, но он доступен не всем. Должна быть большая инфраструктура, чтобы такие вещи форсить. Тем не менее стоит про это тоже помнить.

goals and metrics

[arXiv:1904.02295, 1908.06809, 2110.10668]

- fluency
  - LM PPL
  - human eval
- style accuracy
  - classifiers
  - human eval
- content preservation
  - text similarity
  - embedding based
  - learnable
  - human eval

Если мы говорим, что мы хотим поменять стиль у текста с сохранением чего бы то ни было, нам нужны метрики, чтобы оценить, насколько хорошо это работает. В классической статье по текстовому трансферу есть три группы метрик: fluency, то есть мы хотели бы остаться в грамматическом языке, то есть не нарушать его правила; есть точность попадания в целевой стиль, мы хотим его контролировать; и мы хотим сохранить контент, что бы это ни означало, и для этого есть группа метрик, которая называется content preservation. Мы сейчас про них немножко поговорим. Про fluency я не буду особо говорить, потому что это очевидная вещь. Обычно ее контролируют с помощью перплексии базовой большой языковой модели, иногда какими-то эвристиками, которые ищут ошибки, и ручной разметкой. В общем, задача fluency – задача оценки грамматичности текста; это задача, которая одинаково стоит для перевода, для парафраза, для сигнализации либо для задач, которые подразумевают генерацию текста.

style matching

- **cross-entropy**

Model $G(A_i)$ / author	Shakespeare	Poe	Carroll	Wilde	Marley	Nirvana	MUSE
Generated-Shakespeare	**19.0****	21.6	**18.5***	19.9	21.8	22.0	22.4
Generated-Poe	22.0	**20.4****	21.2	**19.0***	26.0	25.4	26.0
Generated-Carroll	22.2	23.6	**18.9***	22.5	22.4	**21.8****	23.8
Generated-Wilde	21.2	20.9	**20.5****	**18.4***	24.5	24.8	26.4
Generated-Marley	24.1	26.5	22.0	27.0	**15.5***	**15.7****	16.0
Generated-Nirvana	23.7	26.2	20.0	26.6	19.3	**18.3***	**19.1****
Generated-MUSE	21.1	23.9	18.5	23.4	17.4	**16.0****	**14.6***
Uniform Random	103.1	103.0	103.0	103.0	103.5	103.3	103.6
Weighted Random	68.6	68.8	67.4	68.5	68.5	68.0	68.0
SELF	23.4	21.8	25.1	27.3	20.8	17.8	13.3

Table 3. Sample cross entropy between generated texts $\{T_i^G|A_i\}$ and actual texts for different authors.

- **discrimination/classification**

truth \ pred	Brodskiy	Pushkin	Esenin	Pasternak	Tsvetaeva	Mayakovskiy	Akhmatova	Tyutchev	Mandelshtam	Lermontov
Brodskiy	77.2%	1.7%	2.3%	4.3%	2.3%	1.5%	4.0%	1.3%	3.6%	1.7%
Pushkin	1.1%	77.0%	8.0%	0.3%	0.0%	0.3%	1.9%	3.3%	0.6%	7.5%
Esenin	3.9%	4.9%	73.8%	3.0%	1.3%	1.6%	5.9%	0.7%	1.6%	3.3%
Pasternak	16.3%	2.6%	10.7%	54.9%	2.1%	1.7%	3.9%	1.3%	6.0%	0.4%
Tsvetaeva	9.1%	2.8%	5.1%	4.0%	51.1%	1.7%	16.2%	1.1%	5.7%	1.1%
Mayakovskiy	8.2%	2.9%	11.7%	5.8%	3.5%	59.1%	0.6%	1.2%	7.0%	0.0%
Akhmatova	4.5%	4.5%	17.0%	3.4%	3.4%	0.0%	59.7%	1.1%	1.7%	4.5%
Tyutchev	3.0%	14.1%	3.7%	3.0%	0.7%	0.7%	5.9%	55.6%	2.2%	11.1%
Mandelshtam	9.2%	6.6%	9.2%	11.8%	1.3%	5.3%	15.8%	1.3%	35.5%	3.9%
Lermontov	2.6%	15.8%	9.2%	0.0%	2.6%	0.0%	9.2%	9.2%	2.6%	48.7%

Что мы могли бы придумать для оценки качества попадания в стиль? Во-первых, можно сказать, что здесь есть два понятных подхода. Если мы знаем заранее стиль, мы можем напрямую выучить классификатор, который будет определять, попадаем мы в этот стиль или нет. Это самое простое и самое стандартное решение этой задачи.

Второй вариант, который у нас здесь есть, это то, что мы, имея корпуса под каждый стиль, можем выучить языковую модель по каждому из этих корпусов и пытаться сравнивать перплексию какого-то текста с этими моделями.

Оба подхода закономерны; второй, с прямой классификацией, обычно проще использовать, для того чтобы в таком закрытом виде классифицировать текст по принадлежности к тому или иному стилю.

style matching

classification can be tricky

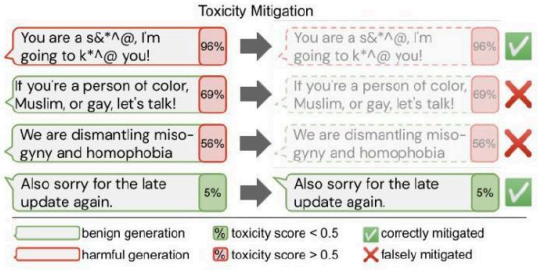


Figure 1: Unintended side effect of automatic toxicity reduction methods: Over-filtering of text about marginalized groups reduces the ability of the LM to generate text about these groups, even in a positive way.

[arXiv:2109.07445]

Но в целом это нетривиальная задача. Вот я взял чью-то работу относительно свежую по детоксификации, и показано, что модель обычно переучивается на то, чтобы отличать какие-то маркерные слова, маркерные токены, и при этом это не является оптимальным решением. То есть она избегает каких-то слов, но эти слова могут быть использованы в разных контекстах, так что есть ложные срабатывания и удаления, и детоксификация тех текстов, которые не должны быть детоксифицированы. Поэтому нужно помнить, что в постановке с классификаторами первое, что модель учит, это какие-то маркерные слова, словосочетания и на них переоптимизируется, если за этим не следить. В том же YELP если у нас будет слово «плохой», это будет маркер стиля отрицательного.

Но если слово встретится в каком-то сочетании «не плохой», то это может привести к тому, что мы неправильно определим стиль.

content preservation

- syntax similarity (BLEU-family, …)
- embedding distance
- learnable
- human eval

Наконец, нам хочется контролировать сохранение контента. Контент сохранять достаточно тяжело, потому что мы не знаем, что это такое. Обычно здесь работает правило «большого пальца», то есть мы говорим, что мы эксплицитно задали стиль как некий атрибут принадлежности к одному из корпусов, а всё остальное мы будем называть контентом. Такой подход лучше, чем ничего.

Что человечество придумало в плане сохранения контента? Во-первых, есть огромное семейство, пришедшее к нам из NMT, и если у вас есть такая роскошь, как параллельный корпус, вы можете ее использовать. В первую очередь это семейство метрик, выросшее из BLEU, их очень много.

Второе – у нас есть эмбеддинги, которые, если копнуть поглубже, пришли к нам из предположения, что верна distribution hypothesis. Есть такое сильное утверждение, что семантика словосочетания полностью определяется его множеством допустимых контекстов. И по сути, когда мы учим какую-нибудь большую модель типа BERT на постановку токенов, то мы де-факто эксплуатируем именно эту предпосылку и получаем эмбеддинги, которые позволяют из контекстов восстанавливать. Для BERT и Word2vec логика ровно та же. То есть мы используем контекст как определяющий смысл слова факт, и коль скоро это работает, то мы говорим, что эмбеддинги, которые мы получаем в таких моделях, являются хорошей репрезентацией именно для этого множества контекстов и значит, что если мы верим, что множество контекстов определяет смысл, то близкие по эмбеддингам слова, тексты будут близки также и по смыслу. Если это так, то мы можем использовать готовые существующие модели как статических эмбеддингов так и динамических, и какого-нибудь его трансформера. В общем, любой из подходов,

дающих нам хорошие эмбеддинги, хорошие в каких-то других задачах, может подойти и здесь как некая прокси-метрика для сравнения семантики двух текстов.

Это мы говорим про то, чтобы взять готовый эмбеддинг и сравнивать. Но можно пойти еще дальше, и если у нас есть какой-то корпус, на котором мы можем такую модель еще доучить, чтобы она лучше различала семантическую близость текстов, то это можно проделать. То есть мы можем взять готовые эмбеддинги и еще их дотюнить именно на задачу сравнения семантической близости. Такие работы тоже есть.

BLEU family

- **sacreBLEU** - standardized BLEU implementation
- **char-BLEU**
- **chrF** - char based f1-score
- **NIST** - BLEU reweighted by n-gram rareness
- **ROUGE** - recall instead of precision
- **METEOR** - WN-synonyms, recall, precision, order penalty
- **GLEU** - google BLEU, 1-4 grams, min(prec,rec)
- **RIBES** - uses rang correlation btw n-gram matches
- **[w]mpF** - F-score on word, morpheme and POS ngrams
- **[h]LEPOR** - recall and precision, lemmas, positions, ...

BLEU само по себе – это у нас есть текст исходный, текст результирующий, и для результирующего текста у нас есть некоторое количество эталонных, то есть написанных вручную людьми текстов, и мы сравниваем результирующий текст с эталонными и пытаемся оценить, насколько он похож. В случае с BLEU это пословное сравнение, энграммное сравнение до четырех-грамм , и существует миллион модификаций.

LEPOR, for example

LEPOR: automatic machine translation evaluation metric considering the enhanced Length Penalty, n-gram Position difference Penalty and Recall

—

In our evaluation, we used hLEPOR$_A$ v.3.1:

– $hLEPOR = Harmonic(w_{LP}LP, w_{NPosPenal}NPosPenal, w_{HPR}HPR)$
$= \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{Factor_i}} = \frac{w_{LP}+w_{NPosPenal}+w_{HPR}}{\frac{w_{LP}}{LP}+\frac{w_{NPosPenal}}{NPosPenal}+\frac{w_{HPR}}{HPR}}$

– $\overline{hLEPOR_A} = \frac{1}{SentNum}\sum_{i=1}^{SentNum} hLEPOR_{ithSent}$

—

(best metric from ACL-WMT 2013 contest)

Долгое время в сообществе была прямо «гонка вооружений». Например, у нас есть LEPOR, который в 2013 году победил всех в контесте, и вот формула, которая здесь представлена, говорит, что уж это-та метрика лучше всех определяет. Для нас это всё представляет, скорее, исторический интерес, потому что мы знаем, что BLEU работает не очень хорошо.

how to choose?

	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
n	16	12	11	11	11	14	15
Correlation	|r|	|r|	|r|	|r|	|r|	|r|	|r|
BEER	0.906	**0.993**	0.952	0.986	0.947	0.915	0.942
BERTR	**0.926**	0.984	0.938	0.990	0.948	**0.971**	0.974
BLEU	0.849	0.982	0.834	0.946	0.961	0.879	0.899
CDER	0.890	**0.988**	0.876	0.967	**0.975**	0.892	0.917
CHARACTER	0.898	**0.990**	0.922	0.953	0.955	0.923	0.943
CHRF	**0.917**	**0.992**	0.955	0.978	0.940	0.945	0.956
CHRF+	**0.916**	**0.992**	0.947	0.976	0.940	0.945	0.956
EED	0.905	**0.994**	0.976	0.980	0.929	0.950	0.949
ESIM	**0.941**	0.971	0.885	0.986	**0.989**	**0.968**	**0.988**
HLEPORA_BASELINE	–	–	–	0.975	–	–	0.947
HLEPORB_BASELINE	–	–	–	0.975	0.906	–	0.947
METEOR++_2.0(SYNTAX)	0.887	**0.995**	0.909	0.974	0.928	**0.950**	0.948
METEOR++_2.0(SYNTAX+COPY)	0.896	**0.995**	0.930	0.971	0.927	**0.952**	0.952
NIST	0.813	0.986	0.930	0.942	0.944	0.925	0.921
PER	0.883	**0.991**	0.910	0.737	0.947	0.922	0.952
PREP	0.575	0.614	0.773	0.776	0.494	0.782	0.592
SACREBLEU.BLEU	0.813	0.985	0.834	0.946	0.955	0.873	0.903
SACREBLEU.CHRF	0.910	**0.990**	0.952	0.969	0.935	0.919	0.955
TER	0.874	**0.984**	0.890	0.799	0.960	0.917	0.840
WER	0.863	0.983	0.861	0.793	0.961	0.911	0.820
WMDO	0.872	**0.987**	0.983	**0.998**	0.900	0.942	0.943
YISI-0	0.902	**0.993**	**0.993**	0.991	0.927	**0.958**	0.937
YISI-1	**0.949**	**0.989**	0.924	0.994	0.981	**0.979**	**0.979**
YISI-1_SRL	**0.950**	**0.989**	0.918	0.994	**0.983**	**0.978**	0.977
QE as a Metric:							
IBM1-MORPHEME	0.345	0.740	–	–	0.487	–	–
IBM1-POS4GRAM	0.339	–	–	–	–	–	–
LASIM	0.247	–	–	–	–	0.310	–
LP	0.474	–	–	–	–	0.488	–
UNI	0.846	0.930	–	–	–	0.805	–
UNI+	0.850	0.924				0.808	
YISI-2	0.796	0.642	0.566	0.324	0.442	0.339	0.940
YISI-2_SRL	0.804	–	–	–	–	–	0.947
	newstest2019						

Table 3: Absolute Pearson correlation of to-English system-level metrics with DA human assessment in newstest2019; correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

[Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges]

Этих метрик много. Например, три года назад был какой-то очередной контест на перевод. Мы можем посмотреть, что в разных языковых парах в разных случаях разные метрики лучше работают. То есть уже эволюция пришла в некоторый тупик,

когда у нас уже количество метрик такое, что мы всегда можем выбрать такую, которая нам будет лучше всего подходить под нашу модель, что не очень честно.

Глобально можно сказать, что у нас обычно нет параллельных корпусов и BLEU мы не можем использовать, однако есть хак, который называется self-BLEU.

embedding distance metrics

- w2v/glove/fasttext/…
- avg/max
- ELMo
- BERT-score
- …

Поскольку у нас часто эмбеддинги всё-таки пословные, а сравнивать мы хотим тексты, то нам нужно еще придумывать всякие эвристики, которые собирают текстовые эмбеддинги из пословных. Это можно делать пуллингом, усреднением, можно брать токен CLS, можно еще что-то делать. В работе про метрики, которая называется BERT-score, делается поиск минимакса, то есть для каждого слова мы делаем эмбеддинг двух текстов под токен, а затем для каждого токена ищем ближайший в другом тексте и усредняем, считаем максимальное расстояние среди минимальных расстояний. Нет никаких универсальных решений. Надо смотреть, если у вас есть какая-то ваша задача, ваша разметка, можно попытаться по ней подобрать или выбрать оптимальную метрику. Но на самом деле на удивление даже Fasttext работает относительно недурно в большинстве случаев.

learnable metrics

- **ROSE**
- **VERTa**
- **MEWR**
- **SIMILE**
- **BLEURT**
- **…**

Как я сказал, еще есть обучаемые метрики. Каждый может сделать обучаемую метрику, если у него есть подходящий корпус и свободное время. Но опасность таких метрик заключается в том, что когда вы учитесь на вашем корпусе, вы ничего не знаете про то, как эта метрика себя поведет за пределами этого конкретного домена, в котором вы ее тюнили. А часто хорошая разметка есть в одном месте, а применять потом эту метрику к текстам, где нет такой разметки, может быть опасно.

Мы пока не очень хорошо умеем работать с ситуацией, когда метрика или модель учится на одних данных, а применяется потом на существенно иных данных, и это аффектит подобные метрики. Следует следить за этой проблемой потенциальной.

content preservation

Premise: [arXiv:2004.05001]

dist(random pair) > dist(style transfer pair) > dist(paraphrase)

Metric	Bibles random	Paralex random	Paraphrase random	Yelp! random rewrite	GYAFC random rewrite	GYAFC random informal	GYAFC random formal	Yelp! rewrite	GYAFC rewrite	GYAFC informal	GYAFC formal	Bibles	Paralex	Paraphrase
POS	14	10	8	9	11	12	13	1	4	7	2	5	6	3
Word overlap	10	9	14	11	12	13	8	4	3	6	1	2	5	7
chrF	9	10	14	11	12	13	8	4	2	7	3	1	5	6
Word2Vec	8	12	14	11	7	10	9	4	2	5	3	1	6	13
FastText	7	12	14	11	9	10	8	4	3	6	2	1	5	13
WMD	8	13	14	11	10	9	12	4	1	6	3	2	5	7
ELMo L2	8	13	14	12	11	10	9	4	3	5	2	1	6	7
ROUGE-1	10	9	14	11	13	12	8	5	3	6	1	2	4	7
ROUGE-2	10	9	14	13	12	8	11	4	2	6	1	3	5	7
ROUGE-L	9	10	14	11	13	12	8	4	3	7	2	1	5	6
BLEU	10	11	14	12	13	8	9	4	3	5	2	1	6	7
Meteor	10	9	14	11	12	13	8	4	3	7	2	1	5	6
BERT score	10	9	14	8	12	13	11	3	4	7	1	2	5	6
Human Labeling	9	14	13	8	12	10	11	7	1	5	2	4	6	3

Table 3: Different semantic similarity metrics sort the paraphrase datasets differently. Cosine similarity calculated with Word2Vec or FastText embeddings do not comply with Inequality $M(D_r) < M(D_p)$. All other metrics clearly distinguish randomized texts from style transfers and paraphrases and are in line with Inequalities 1. However, none of the metrics is completely in line with human evaluation.

В одной из наших работ, где мы делали на анализ метрик, была простая идея: мы взяли много существующих корпусов, которые разбили на несколько классов. То есть мы взяли существующие корпуса, которые называются корпуса парафраз. Это параллельные корпуса парафраз. То есть у нас есть два текста или два предложения, про которые известно, что это парафраз, то есть синонимические, по сути, тексты. Это похоже на стиль, но с бóльшим требованием к content preservation, с одной стороны, и отсутствием какой-либо стилевой маркировки в таких парах. Мы не говорим, чем одна фраза отличается от другой, мы лишь говорим, что они совпадают или близки по смыслу.

Второй класс корпусов, который мы взяли, – это небольшое количество доступных параллельных корпусов со стилем. И третий класс мы сделали сами: это рандомный pairing предложений, причем из тех же самых корпусов парафраз и стилевых. Зачем мы это сделали? Мы рассматривали там два варианта постановки. Первый, более сильный вариант – это давайте ожидать от метрики, что она парафразы будет считать более близкими, нежели предложения в разных стилях с одинаковым смыслом, а предложения в разных стилях с одинаковым смыслом будем считать более близкими, чем два случайных предложения из одного корпуса. Это сильная постановка. А в более слабой постановке мы хотим, чтобы и пара предложений с разными стилями, и пара предложений из парафраз была гарантированно ближе, чем рандомная пара. И используя такой premise, можно посмотреть, как разные метрики упорядочивают разные корпуса по близости, и сравнить это в том числе с человеческой разметкой.

content preservation

Premise: [arXiv:2004.05001]

dist(random pair) > dist(style transfer pair) > dist(paraphrase)

	POS-distance	Word overlap	chrF	Word2Vec	FastText	WMD	ELMO L2	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	Meteor	BERT score	Human score
POS-distance	1,00	0,73	0,71	0,45	0,44	0,69	0,66	0,71	0,72	0,71	0,68	0,74	0,82	0,72
Word overlap	0,73	1,00	0,98	0,80	0,84	0,86	0,92	0,99	0,91	0,98	0,92	0,99	0,95	0,80
chrF	0,71	0,98	1,00	0,79	0,83	0,89	0,93	0,97	0,89	0,99	0,92	0,99	0,93	0,83
Word2Vec	0,45	0,80	0,79	1,00	0,98	0,87	0,88	0,78	0,79	0,78	0,82	0,77	0,73	0,64
FastText	0,44	0,84	0,83	0,98	1,00	0,86	0,90	0,83	0,81	0,83	0,85	0,81	0,76	0,65
WMD	0,69	0,86	0,89	0,87	0,86	1,00	0,96	0,86	0,92	0,89	0,92	0,86	0,85	0,89
ELMO L2	0,66	0,92	0,93	0,88	0,90	0,96	1,00	0,92	0,92	0,94	0,96	0,92	0,87	0,86
ROUGE-1	0,71	0,99	0,97	0,78	0,83	0,86	0,92	1,00	0,93	0,98	0,93	0,98	0,94	0,82
ROUGE-2	0,72	0,91	0,89	0,79	0,81	0,92	0,92	0,93	1,00	0,91	0,96	0,90	0,87	0,81
ROUGE-L	0,71	0,98	0,99	0,78	0,83	0,89	0,94	0,98	0,91	1,00	0,94	0,99	0,94	0,83
BLEU	0,68	0,92	0,92	0,82	0,85	0,92	0,96	0,93	0,96	0,94	1,00	0,92	0,87	0,84
Meteor	0,74	0,99	0,99	0,77	0,81	0,86	0,92	0,98	0,90	0,99	0,92	1,00	0,95	0,80
BERT score	0,82	0,95	0,93	0,73	0,76	0,85	0,87	0,94	0,87	0,94	0,87	0,95	1,00	0,82
Human score	0,72	0,80	0,83	0,64	0,65	0,89	0,86	0,82	0,81	0,83	0,84	0,80	0,82	1,00

Figure 1: Pairwise correlations of the orders induced by the metrics of semantic similarity.

Мы построили какую-то тепловую карту, посмотрели, и оказывается, что метрика WMD у нас оказалась наиболее близкой к Human score. Имейте в виду. По крайней мере, в такой постановке она достаточно неплохо себя показала. Но при этом надо заметить, что всё равно она далека от идеала и мы не то чтобы смогли ее воспроизвести очень хорошо порядок, который задали.

content preservation

[arXiv:2004.05001]

Metric	Correlation of the metric with human evaluation	Correlation of the induced orders with human-induced order	Variability of the metric on random sentences
POS	0.87	0.72	37.0%
Word overlap	0.89	0.80	23.8%
chrF	0.9	0.83	17.2%
Word2Vec	0.46	0.64	88.6%
FastText	0.52	0.65	86.3%
WMD	**0.92**	**0.89**	12.3%
ELMo L2	0.82	0.86	53.3%
ROUGE-1	0.9	0.82	33.5%
ROUGE-2	0.84	0.81	4.5%
ROUGE-L	0.89	0.83	33.4%
BLEU	0.72	0.84	0.2%
Meteor	0.91	0.80	19.5%
BERT score	0.89	0.82	23.1%

Table 4: WMD shows the highest pairwise correlation with human assessment similarity scores. The order on fourteen datasets, induced by WMD also has the highest correlation with human-induced semantic similarity order. Variability on random sentences is a ratio of the difference between the maximal and minimal value of a given metric on the datasets of random pairs and difference of the maximal and minimal value of the same metric on all available datasets.

Metric	Number of ranks coinciding with human-induced ranking	Number of swaps needed to reconstruct human-induced ranking
POS	3	16
Word overlap	1	15
chrF	2	14
Word2Vec	3	16
FastText	2	17
WMD	1	**11**
ELMo L2	**4**	**11**
ROUGE-1	0	15
ROUGE-2	2	13
ROUGE-L	2	14
BLEU	3	13
Meteor	2	15
BERT score	3	13

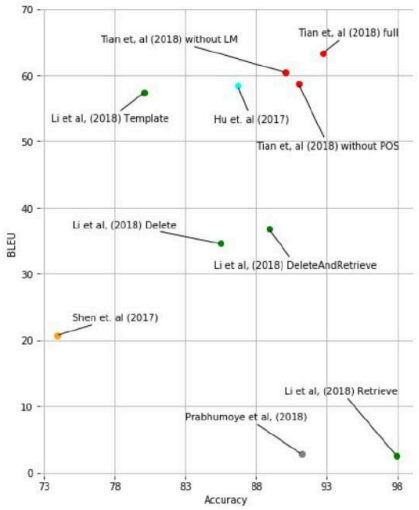
Table 5: Scores for the orders induced by different semantic similarity metrics.

Вот еще из той же статьи некоторые результаты. Мы пытались посмотреть, насколько метрика стабильна, насколько ее колбасит на случайных предложениях, насколько хорошо удается восстановить порядок на корпусах. Оказывается, что даже в

хорошем случае с метрикой WMD требуется 11 перестановок, чтобы восстановить ранжирование.

Так или иначе мы взяли какую-то метрику для content preservation – например, WMD или FastText, неважно, или BLEU.

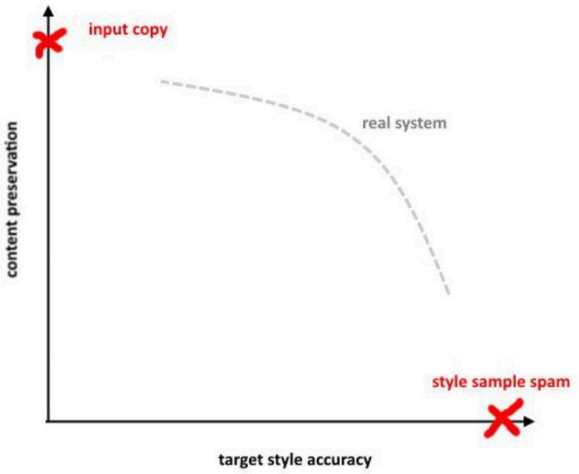
which is better?



Как мы теперь поймем, что у нас задача решена хорошо? Оставляя fluency за скобками, мы хотим, чтобы у нас, с одной стороны, хорошо работал текстовый трансфер, а с другой стороны, сохранялся смысл.

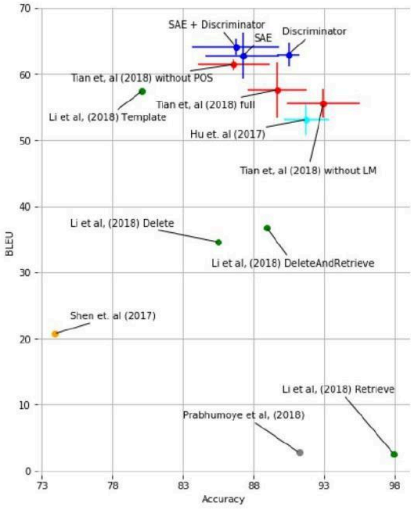
Отложим по горизонтали точность попадания в стиль, а по вертикали – качество сохранения смысла. Это реальная работа. Как понять, какая работа лучше? Вероятно, Tian 2018 года лучше, чем Shen 2017 года, но можем ли мы что-то еще про это сказать? Сейчас немножко лучше стало, но если посмотреть работы двух-, трех-, четырехлетней давности, была такая олимпиада, когда выходила следующая работа и говорила, что «мы – state of the art, потому что мы увеличили accuracy по отношению к предыдущим работам». Потом выходил следующая работа и говорила, что «мы – state of the art, потому что мы увеличили BLEU по отношению к предыдущим работам», а accuracy при этом у них был ниже, и это было довольно грустно. Так делать не надо.

goals trade-off



А как надо делать? Надо для начала сформулировать, что поскольку стиль и контент не являются ортогональными, то у нас в реальности есть вот такой трейд-офф. Красные крестики, которые я нарисовал от руки, соответствуют двум понятным выраженным случаям. В одном случае мы можем скопировать вход на выход и получим идеальный content preservation, и получим отвратительный, в смысле негативный нулевой style transfer accuracy. С другой стороны, мы можем запомнить по каждому стилю какое-нибудь одно предложение и выдавать его на выход, и у нас будет идеальный style transfer и близкий к нулевому content preservation. И реальные системы – все на этом фронте между этими точками. И что нужно демонстрировать, если мы хотим показать, что мы улучшили решение этой задачи? Нужно показывать какую-то Парето-оптимизацию, то есть улучшение или неухудшение по обоим осям метрики. Только в этом случае можно говорить о том, что мы сделали что-то лучше, иначе мы можем долго двигаться по этой кривой туда-сюда и писать про это разные работы.

unstable balance / unfair reporting



Второе важное наблюдение, что когда у вас есть вот такой трейд-офф и вы в обучении пытаетесь метрики оптимизировать, как это часто бывает, это приводит к неустойчивости. Мы делали эксперименты с нашими моделями. Синенькие – это наши, а красненькие – это предшественники, и мы делали по 10, по 20, по X запусков обучения одинаковой модели на одних и тех же данных, случайно перемешанных, и здесь длина лапок у крестиков соответствует дисперсии, соответствует тому, насколько одна и та же модель может давать разные результаты. Действительно есть неустойчивость, связанная, вероятно, с этим трейд-оффом, и в зависимости от того, в каком порядке вы подаете обучающие данные, вы можете словить улучшение или ухудшение по одной из метрик. Это следует учитывать.

Text style transfer approaches

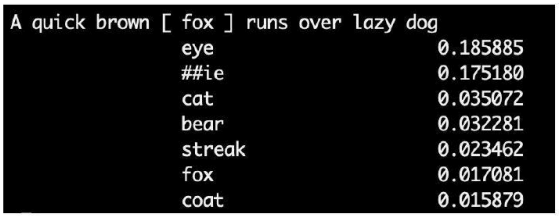
approaches

- **template / edit based** [arXiv:2005.12086, ...]
- **TST as NMT** [arXiv:1707.01161, ...]
- **TST as UNMT** [arXiv:1711.00043, ...]
- **Z-space search** [arXiv:1905.12926, ...]
- **disentangled representations** [arXiv:1808.04339, ...]
- **...more**

See also: [arXiv:2109.15144]

Есть некоторое количество вещей, которые, как мне кажется, важно разделять и обращать внимание на то, что происходит в каждом из этих направлений.

template / edit based



Самое простое, что можно себе предположить, это то, что называется template based, или edit based подход. Он хорошо работает с короткими текстами и он как раз не очень интересный, тривиальный, но не упомянуть его нельзя. Речь идет примерно о такой конструкции. Мы можем тем или иным способом решать задачу поиска и

потокенной замены одного текста, пока мы не пришли к тексту, который удовлетворяет нашим пожеланиям. Пожелания у нас прежние: мы хотим, чтобы поменялся стиль и чтобы при этом как-то созранился контент, семантика, вот это всё.

Как мы можем это делать? Мы можем пытаться контролировать семантику с помощью какой-то метрики, какой-то модели, какого-то эмбеддинга, а при этом выявлять, какой из токенов можно поменять наиболее успешным образом, чтобы стиль изменился. Но это я сейчас очень банальный подход изложил, а в принципе похожая логика может стоять за более сложными подходами.

template / edit based

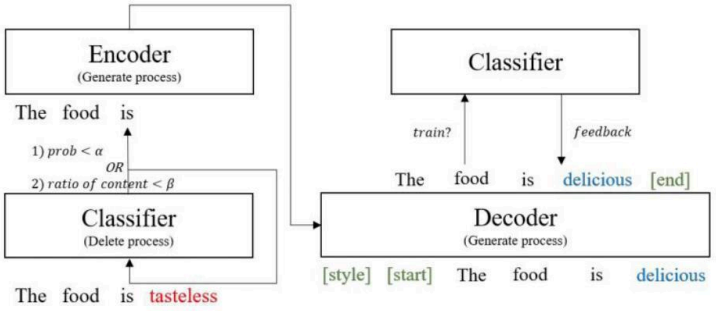


Figure 1: The proposed model framework consists of Delete and Generate process. Delete process is a method using a pre-trained classifier, and the Generate process consists of an encoder and a decoder. In the training time, our model receives feedback from the classifier's probability of the generated sentence.

[arXiv:2005.12086]

Например, в этой работе, которая, кажется, называлась, Delete and Generate, мы смотрим на предложение и в нем находим слова, которые классификатор считает стилевыми маркерами. Мы эти слова удаляем и потом декодером пытаемся написать новое предложение, и декодер вставляет вместо удаленных слов другие слова.

Это на самом деле очень узкое подмножество задач текстового трансфера. Мы не говорим про какое-то изменение поэтического стиля, а для задач типа переписать предложение с одного сентимента на противоположное – это вполне работающая конструкция, которая больше похожа не на эвристику, но в общем понятно: мы вышибаем эти слова и генерируем новые.

template / edit based

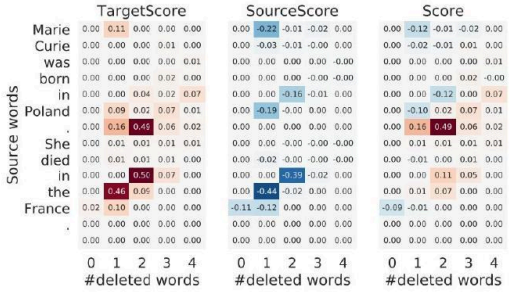


Figure 1: MASKER replaces span ". *She*" by "*and [PAD] [PAD] [PAD]*", resulting in the following fused sentence: *Marie Curie was born in Poland and died in the France .*

Random Sample of Correct MASKER Predictions	
Source	so far i 'm not really impressed .
Prediction	so far i 'm really impressed .
Source	either way i would never recommend buying from camping world .
Prediction	either way i would recommend buying from camping world .
Source	this is a horrible venue .
Prediction	this is a great venue .
Source	this place is a terrible place to live !
Prediction	this place is a great place to live !
Source	i 'm not one of the corn people .
Prediction	i 'm one of the corn people .
Source	this is easily the worst greek food i 've had in my life .
Prediction	this is easily the best greek food i 've had in my life .
Source	the sandwich was not that great .
Prediction	the sandwich was great .
Source	its also not a very clean park .
Prediction	its also a very clean park .

[arXiv:2010.01054]

Чуть более, на мой взгляд, приятная и простая работа, кажется, октября 2020 года. Есть базовый BERT и есть несколько его версий, потюненных под каждый из корпусов. И они смотрят на слова, по которым два потюненных BERT-а под разные стили имеют наибольшее расхождение в оценке вероятности. И таким образом они выявляют стилевые маркеры и могут взять и переписать с одного стиля на другой. Очевидно, требуются костыли, потому что у вас количество токенов может не совпадать, они это делают с помощью паддингов дополнительных, тем не менее это довольно простое инженерное решение, очень неплохо работающее с тем же корпусом YELP. Опять-таки, это очень узкая подзадача текстового трансфера, универсальной вещи она не решает. Но, например, какой-нибудь детокс она может делать до какого-то уровня, как мне кажется.

NMT-like

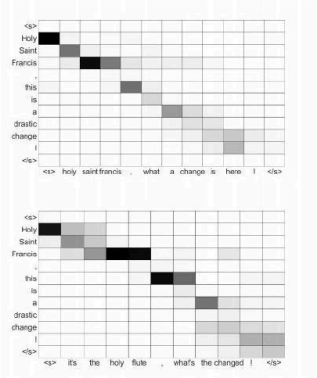


Figure 2: Attention matrices from a *Copy* (top) and a *simple S2S* (bottom) model respectively on the input sentence *"Holy Saint Francis, this is a drastic change!"* . $< s >$ and $< /s >$ are start and stop characters. Darker cells are higher-valued.

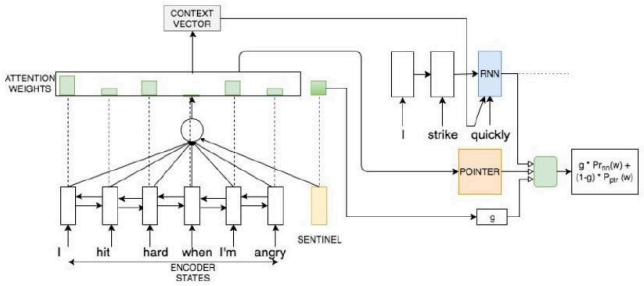


Figure 1: Depiction of our overall architecture (showing decoder step 3). Attention weights are computed using previous decoder hidden state h_2, encoder representations, and sentinel vector. Attention weights are shared by decoder RNN and pointer models. The final probability distribution over vocabulary comes from both the decoder RNN and the pointer network. Similar formulation is used over all decoder steps

[arXiv:1707.01161]

Мы можем легко себе вообразить попытку научить обычную модель машинного перевода, если у нас есть параллельный корпус. Такое бывает очень редко, но тем не менее есть такие работы, которые так это делают.

NMT-like

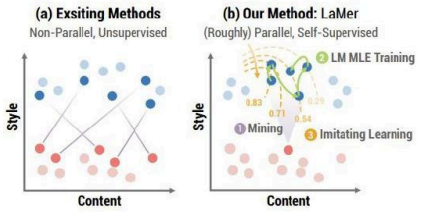
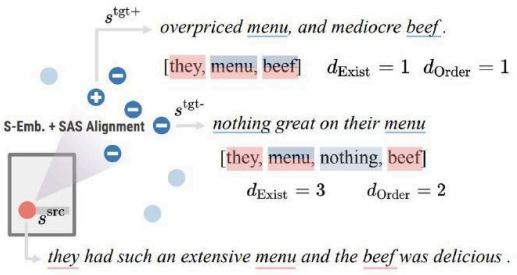


Figure 1: Red and blue circles represent the source and target texts respectively. (a) Existing methods crucially ignore the inherit parallelism within the data. (b) Our method first mines (roughly) parallel expressions, then learns how to transfer style with the self-supervision from the parallel expressions.

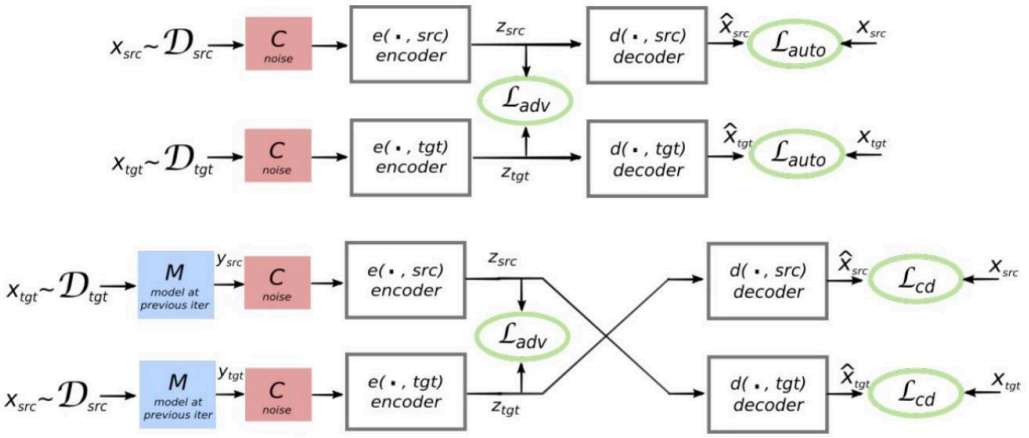


[openreview:-TSe5o7STVR]

А еще – я не знаю, попала она в итоге наружу или нет, но была такая работа, в которой использовался следующий подход: у нас есть непараллельный корпус, но мы можем использовать некоторую модель дополнительную, которая ищет для нашего исходного предложения в целевом стиле несколько предложений-кандидатов, максимально на него похожих, и дальше использует

подход, близкий к NMT. Это такое self-supervised получается имитация параллельного корпуса при его отсутствии. Достаточно интересный подход, как мне кажется. Результативность, конечно, зависит от миллиона вещей и того, насколько хорошо это реализовано, но попробовать так можно.

UNMT-like



[arXiv:1711.00043]

UNMT - подход, который несколько лет двигала команда Facebook, достаточно красивый. Мы отказываемся от идеи, что у нас есть параллельные корпуса, и мы будем учить на непараллельных. Они это придумали именно для перевода и показывали, что если у вас есть очень большой непараллельный корпус, этого достаточно, чтобы до какой-то степени научить переводчик. Он будет хуже, чем обученный на параллельных, но тем не менее его можно развивать. Магия всего подхода заключается в двух ключевых вещах. Первая – это то, что мы пытаемся делать back translation, такой классический хак, а вторая – мы пытаемся, чтобы у нас множество эмбеддингов для разных языков, а в случае со стилями – для разных стилей, нельзя было дискриминатором различить в пространстве эмбеддингов.

То есть тем самым мы заставляем энкодеры и декодеры, которые здесь под разные языки, учиться приводить тексты к одинаковому представлению внутри. Ровно такие же хаки можно применять и люди пытались применять к стилевому трансферу. Штука работает до какой-то степени, но очевидно хуже, чем обучение параллельных корпусов.

UNMT-like

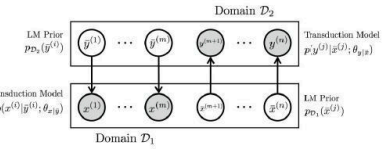


Figure 1: Proposed graphical model for style transfer via bitext completion. Shaded circles denote the observed variables and unshaded circles denote the latents. The generator is parameterized as an encoder-decoder architecture and the prior on the latent variable is a pretrained language model.

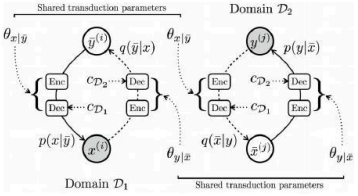


Figure 2: Depiction of amortized variational approximation. Distributions $q(\bar{y}|x)$ and $q(\bar{x}|y)$ represent inference networks that approximate the model's true posterior. Critically, parameters are shared between the generative model and inference networks to tie the learning problems for both domains.

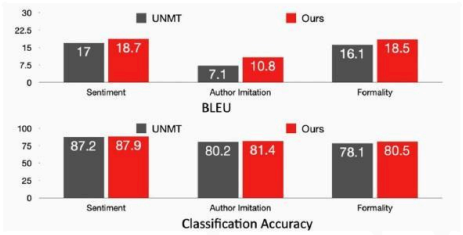


Table 3: Examples for author imitation task

Methods	Shakespeare to Modern
Source	Not to his father's .
Reference	Not to his father's house .
UNMT	Not to his brother .
Ours	Not to his father's house .
Source	Send thy man away .
Reference	Send your man away .
UNMT	Send an excellent word .
Ours	Send your man away .
Source	Why should you fall into so deep an O ?
Reference	Why should you fall into so deep a moan ?
UNMT	Why should you carry so nicely , but have your legs ?
Ours	Why should you fall into so deep a sin ?

[arXiv:2002.03912]

Более свежая работа на эту же тему. По сути, люди делают то же самое, но более качественно и аккуратно ставят задачу с точки зрения _____ (01:11:12) вычисления, получают, если верить из замерам, результаты более эффективные, чем обычный NMT.

Это такой блок подходов, связанных с попыткой в лоб учиться на таких корпусах.

Z-space search

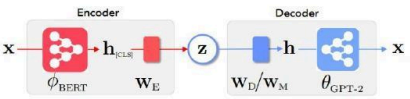
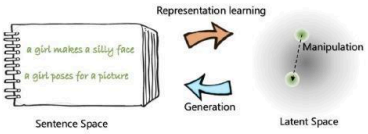


Figure 1: Illustration of OPTIMUS architecture.



0.0	children are looking for the water to be clear.
0.1	children are looking for the water.
0.2	children are looking at the water.
0.3	the children are looking at a large group of people.
0.4	the children are watching a group of people.
0.5	the people are watching a group of ducks.
0.6	the people are playing soccer in the field.
0.7	there are people playing a sport.
0.8	there are people playing a soccer game.
0.9	there are two people playing soccer.
1.0	there are two people playing soccer.

Table 3: Interpolating latent space. Each row shows τ, and the generated sentence (in blue) conditioned on z_τ.

$$\boldsymbol{x}_D \approx \boldsymbol{x}_B - \boldsymbol{x}_A + \boldsymbol{x}_C$$

Source $\boldsymbol{x}_A$	**Target** $\boldsymbol{x}_B$
a girl makes a silly face	two soccer players are playing soccer
Input $\boldsymbol{x}_C$	**Output** $\boldsymbol{x}_D$
• a girl poses for a picture	• two soccer players are at a soccer game.
• a girl in a blue shirt is taking pictures of a microscope	• two football players in blue uniforms are at a field hockey game
• a woman with a red scarf looks at the stars	• two men in white uniforms are field hockey players
• a boy is taking a bath	• two baseball players are at the baseball diamond
• a little boy is eating a bowl of soup	• two men are in baseball practice

Table 2: Sentence transfer via arithmetic operation in the latent space. The output sentences are in blue.

[arXiv:2004.04092]

Есть множество подходов, которые я условно называю Z-space search. Устроены они приблизительно следующим образом: мы берем некоторый текст, эмбеддим его. Я

напомню, у нас есть проблема, что у отдельных токенов нет стиля. Ну, что ж делать. Мы можем заэмбеддить текст целиком и посмотреть, есть ли у этого эмбеддинга признаки стиля. Допустим, они есть, и тогда мы можем с ними попытаться что-то сделать. Например, мы можем поискать в пространстве эмбеддингов другой текст, который будет близко, но иметь другой нужный нам признак стиля. Вот такой подход, если говорить глобально.

На примере Optimus. Это не совсем про стиль, но Optimus был первой трансформерной моделью, года два назад вышел, который реализовывал в лоб то, что называется VAE, Variation after encoder. Идея в том, что как раз мы закодировали текст, и здесь логика аналогичная той логике арифметики, которая была часто используема для демонстрации магии Word2vec, когда мы могли интерполировать между словами, могли строить аналогии между словами. Здесь делается примерно то же самое, только на уровне предложений. На самом деле это жесткий cherry-pick. Я с Optimus имел дело напрямую и работает это всё не так хорошо. Пространство, которое у них выучивается, по крайней мере той модели, которую они выложили, очень фрагментированное, то есть такой красивой магии получается мало. Но тем не менее сам подход про то, что мы можем выучить некое пространство эмбеддингов непрерывное или с какими-то нужными нам качествами, понятен, и в нем потом можем искать.

Z-space search

	Tense (present→past)
Mono-lingual	i ask many people here . i **asked** many people here .
Cross-lingual	ik kijk naar een oude film van m ' n moeder . ik **bekeek** een oude film van mijn moeder .
	ObjNum (singular→plural)
Mono-lingual	i could tell you some story . i could tell you some **stories** .
Cross-lingual	we hebben een beter bondgenoot nodig . we hebben **betere bondgenoten** nodig .
	SubjNum (plural→singular)
Mono-lingual	families agreed to keep it quiet . **a family** agreed to keep it quiet .
Cross-lingual	monsters gaan ons opeten . **het monster gaat** ons opeten .

Table 5: Linguistic property transfer examples of the proposed system in both monolingual and cross-lingual settings

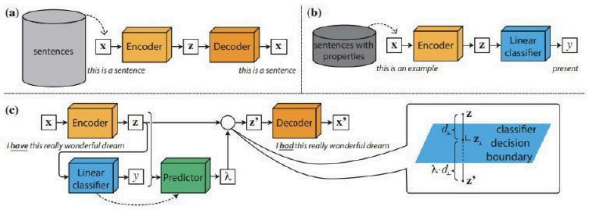


Figure 1: (a) Pretrained autoencoder (encoder ENC, decoder DEC). (b) linguistic property classifier $\mathcal{C}$. (c) Geometric transformation of the sentence representation to shift $\mathbf{z}$ according to λ beyond the decision boundary of $\mathcal{C}$, the shifted encoding $\mathbf{z}'$ is then given as input to the decoder resulting in the sentence $\mathbf{x}'$ with the transferred property.

[arXiv:2104.03630]

Вот есть другая работа, она тоже не про текстовый трансфер, а больше про лингвистику, но она показывает, что в этом пространстве можно использовать какой-то другой энкодер и декодер, но тоже предобученная модель, а они показывают, что есть в этом пространстве признаки и направления, которые

соответствуют, например, времени в английском языке, множественности или единственности объекта, и можно пытаться воздействовать на эмбеддинги так, чтобы после декодирования поменять что-то в исходном предложении – время. Если считать время стилем, то это вполне себе работающий подход.

Z-space search

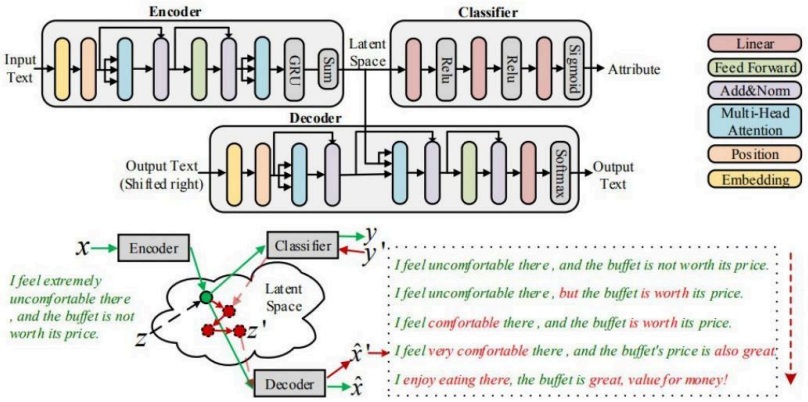


Figure 1: Model architecture.

Но если говорить конкретно именно про стиль, то вот пример, который демонстрирует еще более простой подход, а именно мы энкодером кодируем текст в некоторую точку в этом пространстве. У нас есть классификатор, обученный по этой точке предсказывать стиль. Что мы можем сделать, так это, заморозив веса классификатора, пустить градиент от увязки стилей на значение точки в этом пространстве, и, таким образом, у нас будет некоторый градиентный спуск, который эту точку z в этом пространстве сместит в сторону ближайшей точки, которая даст нужный нам стиль. Здесь тоже, скорее всего, cherry-pick, но здесь видно, как последовательное смещение, градиентный спуск меняет предложение и его… В табличке видно, как его декодировали последовательно и получали постепенное изменение в нужную сторону.

Это всё очень красиво, но это очень сложно контролировать. То есть здесь это сработало, а в другом месте могло не сработать. Продакшн на этом не сделаешь, но тем не менее как подход это достаточно интересная работа.

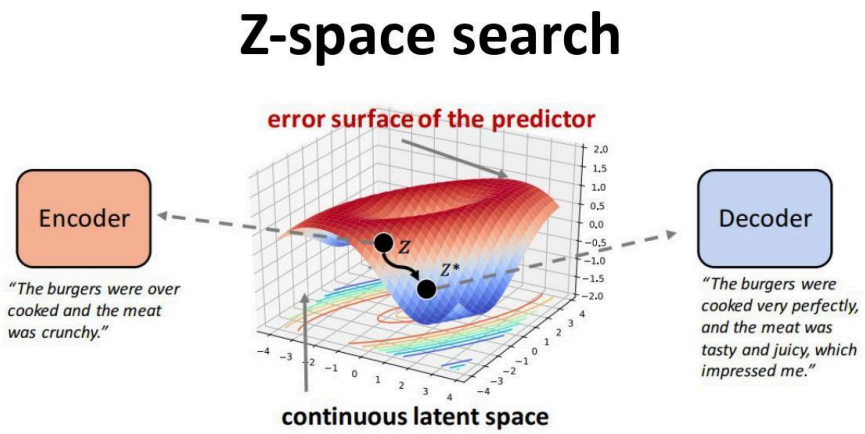


Figure 1: There is an example of content-preserving text sentiment transfer, and we hope to further increase the length of the target sentence compared with the original sentence. The original sentence x with negative sentiment is mapped to continuous representation z via encoder. Then z is revised into z^* by minimizing the error $\mathcal{L}_{\text{Attr},s_1}(\theta_{s_1}; s_1 = \{\text{sentiment} = \textit{positive}\}) + \mathcal{L}_{\text{Attr},s_2}(\theta_{s_2}; s_2 = \{\text{length} = 20\}) + \lambda_{\text{bow}}\mathcal{L}_{\text{BOW}}(\theta_{\text{bow}}; x_{\text{bow}} = [\textit{burgers}, \textit{meat}])$ with the sentiment predictor f_1, length predictor f_2, and the content predictor f_{bow}. Afterwards the target sentence x^* is generated by decoding z^* with beam search via decoder [best viewed in color].

Вот другая работа, очень похожая. Здесь вводятся сразу несколько классификаторов. Один – не сентимент по латентному эмбеддингу, второй классификатор – на предсказание длины предложения, третий – на bag of words, то есть будут ли в декодированном предложении какие-то нужные нам слова. И затем говорится о том, что давайте мы поменяем сентимент, увеличим длину и обязательно хотим сохранить слова «бургеры» и «мясо», и после некоторого количества итераций мы можем поменять предложение. Здесь видно, что эти костыли с сохранением слов возникли не просто так, а потому что, скорее всего, если бы их не было, то уплыл бы наш смысл куда-то далеко. Поэтому люди такие штуки делают. Но подход немного подкупает тем, что он как будто изящный, но он сильно опирается на свойства пространства, которое не так просто получить.

disentangled representations

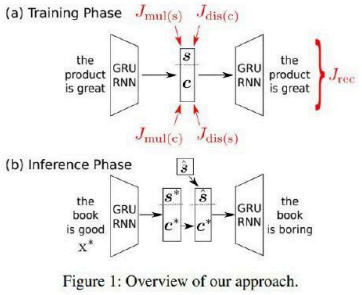


Figure 1: Overview of our approach.

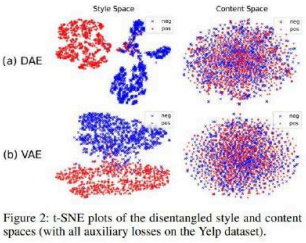
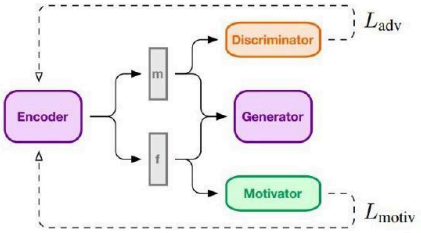


Figure 2: t-SNE plots of the disentangled style and content spaces (with all auxiliary losses on the Yelp dataset).



[arXiv:1808.04339]

[arXiv:1808.09042]

Другое большое направление – disentangled representations. Тоже работа происходит с внутренними представлениями, то есть с эмбеддингами текста, но идея состоит в том, что мы вместо того, чтобы навигировать по этому пространству, будем учить энкодер делать такие эмбеддинги, у которых часть координат явным образом соответствует нужному нам стилю, а остальные координаты будут содержать информацию о контенте. Как это можно сделать?

Про это написано много статей, несколько из них написали даже мы, но идея примерно такая. Самая базовая идея, которую можно предположить – мы будем считать, что у нас есть вектор эмбеддинга, и давайте считать, что у него есть две компоненты – S и C. Мы хотим, чтобы S содержала только информацию о стиле, C содержала только информацию о контенте, что бы это ни означало. Как сделать, чтобы S содержала информацию о стиле? Давайте добавим LOS с классификатором, который смотрит только на S-координаты и пытается определить стиль. Чем лучше он определяет, тем модель мы награждаем. Что нам это даст? То, что в S появится информация о стиле. Но даст ли это нам то, что в C не будет информации о стиле? Нет, не даст, она может сохраниться. Поэтому добавим еще один классификатор, который будет смотреть на C. Мы будем штрафовать модель, если классификатору удалось информацию о стиле восстановить по C, и, наоборот, делаем то же самое зеркально для контента.

Вот мы добавили четыре дополнительных LOS, которые позволяют немножечко управлять тем, как модель будет хранить информацию в контентном векторе. Соответственно, дальше работает ли это? До некоторой степени работает, но есть

критика, связанная с тем, что модель может спрятать какие-то составляющие внутри вектора контента, скрывать информацию о стиле.

disentangled representations

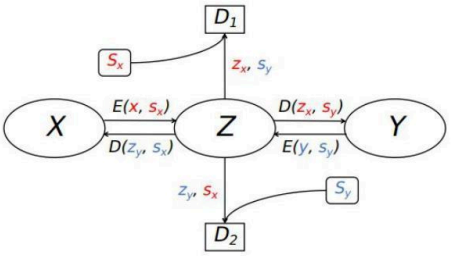


Figure 1: CrossAlign architecture

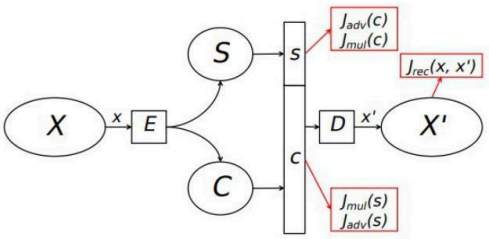


Figure 2: VAE architecture

[arXiv:2004.11742]

Есть всякие подходы, которые позволяют эти штуки немножечко оттуда вытравливать более надежным способом.

disentangled representations

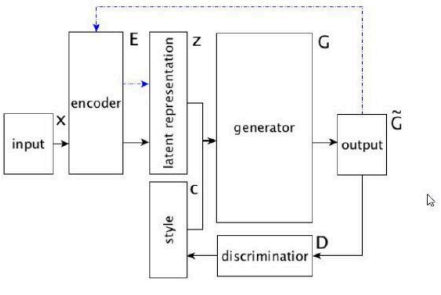
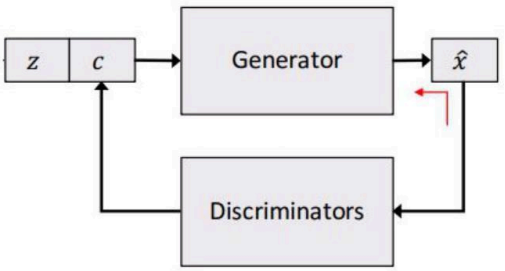


Figure 3: The generative model, where style is a structured code targeting sentence attributes to control. Blue dashed arrows denote the proposed independence constraint of latent representation and controlled attribute, see (Hu et al., 2017a) for the details.



[arXiv:1703.00955, arXiv:1809.00794]

Вот один из подходов, который мы предложили. Это то, что можно сделать эмбеддинг текста, потом его декодировать в другой стиль, а потом снова его заэмбеддить и требовать, чтобы контентная составляющая у этих двух эмбеддингов

совпадала. Это такой как бы авто-энкодер, только со сдвигом на полфазы. И эта штука довольно сильно реполяризует модель и добавляет действительно устойчивости, как это у нас получилось. Но это достаточно древняя работа, мы ее делали три года назад.

disentangled representations

ARE ADVERSARIAL MODELS REALLY DOING DISENTANGLEMENT?

λ_{adv}	Discriminator Acc (Train)	Post-fit Classifier Acc (Test)
0	89.45%	93.8%
0.001	85.04%	92.6%
0.01	75.47%	91.3%
0.03	61.16%	93.5%
0.1	57.63%	94.5%
1.0	52.75%	86.1%
10	51.89%	85.2%
fastText	-	97.7%

[arXiv:1811.00552]

disentangled representations

$$\mathcal{L}_{cos}(x, c) = \cos\left(E(\tilde{G}(E(x), c)), E(x)\right),$$
$$\mathcal{L}_{cos^-}(x, c) = \cos\left(E(\tilde{G}(E(x), \bar{c})), E(x)\right). \quad (8)$$

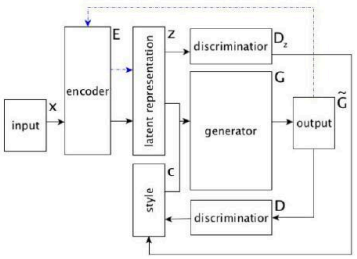


Figure 4: The generative model with dedicated discriminator introduced to ensure that semantic part of the latent representation does not have information on the style of the text.

[arXiv:1908.06809]

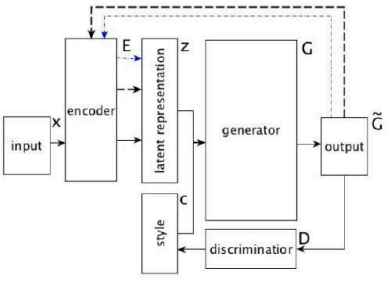


Figure 5: The generative model with a dedicated loss added to control that semantic representation of the output, when processed by the encoder, is close to the semantic representation of the input.

Сейчас уже, кажется, так не делают, но сама идея того, что мы хотим выделить подпространство, которое соответствует нужному стилю, проста и понятна.

unsupervised style learning

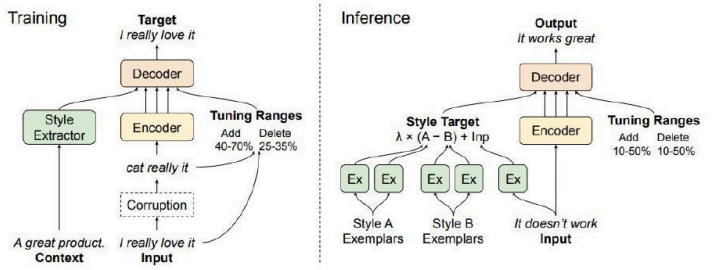


Figure 1: TextSETTR architecture for label-free style transfer. The Encoder, Decoder and Style Extractor (Ex) are transformer stacks initialized from pretrained T5. During training, the model reconstructs a corrupted input, conditioned on a fixed-width "style vector" extracted from the preceding sentence. At inference time, a new style vector is formed via "targeted restyling": adding a directional delta to the extracted style of the input text. Stochastic tuning ranges provide extra conditioning for the decoder, and enable fine-grained control of inference.

[arXiv:2010.03802]

Давайте попробуем сказать еще про пару прикольных моделей. Вот одна модель, которая мне понравилась. Она, по-моему, прошлогодняя. Она устроена следующим образом: у нас есть предложение в тексте и есть контекст этого предложения. Давайте мы допустим, что стиль – это штука, которая локально непрерывна, то есть у соседних предложений с высокой вероятностью одинаковый стиль. Тогда мы можем сказать, что мы будем решать задачу денойзинга авто-энкодера, то есть мы берем предложение, портим его и восстанавливаем обратно. А в качестве дополнительной информации будем давать контекст этого предложения в виде одного или нескольких соседних предложений. Непосредственно для восстановления предложения, кажется, это не должно нам сильно помочь, поскольку, скорее всего, оно существенно отличается, но если мы портим какие-нибудь стилевые слова, то по крайней мере это позволяет ему предположить, какими они должны быть.

unsupervised style learning

Model	Acc.	Content
TextSETTR	73.3	34.7
N	23.4	84.4
N + BT	13.3	98.7
−replace noise	66.1	42.1
+shuffle noise	70.3	34.1
manual exemplars	52.4	44.2
−tunable inference	71.5	39.4
CP-G	60.1	35.4
CP-B	40.0	39.7
CrossAligned	83.1	15.2
Delete&Retrieve	50.9	16.1
B-GST	60.0	73.6

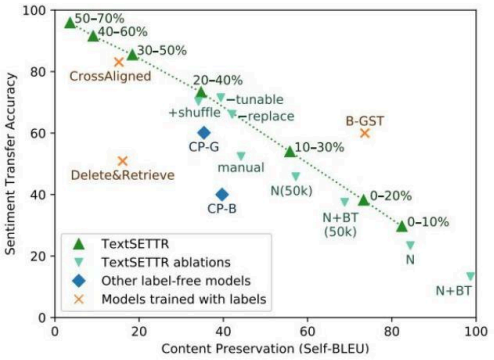


Figure 2: Automatic evaluation metrics comparing our TextSETTR model, ablations, and previous work. Up-and-right is better. We train for 10k steps and use add/delete:20–40% unless otherwise specified. Scores for CrossAligned, Delete&Retrieve and B-GST are from Sudhakar et al. (2019).

Дальше у нас есть две ручки. Мы можем удалять какой-то процент слов и добавлять представителей стиля. И оказывается, что здесь вылезает как раз тот самый трейд-офф, который я рисовал в начале. Здесь перевернутые оси, но неважно, и они показывают, что удаление большего числа слов на фазе денойзинга приводит к тому, что content preservation уменьшается, но при этом стилевая составляющая улучшается, и наоборот. Получается такая управляемая конструкция. Это что касается части про удаление. А что касается части, связанной со стилевой подсказкой.

unsupervised style learning

Model	Accuracy	Content
TextSETTR	83.6	39.4
add/del: 0–20%	63.4	76.9
add/del: 10–30%	72.7	60.2
add/del: 30–50%	89.7	21.5
Lample et al. 2019	82.6	54.8

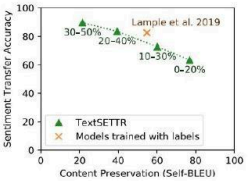


Figure 3: Comparison with Lample et al. (2019) on the evaluation setting that includes pos→pos and neg→neg transfers.

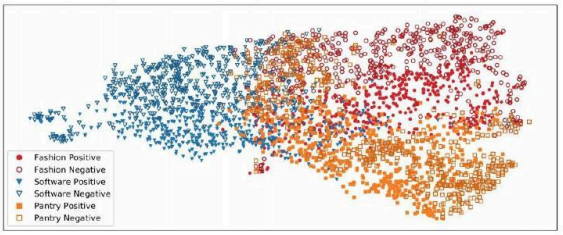


Figure 4: 2D UMAP embedding of the style vectors extracted by our TextSETTR model for text inputs from Amazon reviews covering three product categories and two sentiment labels.

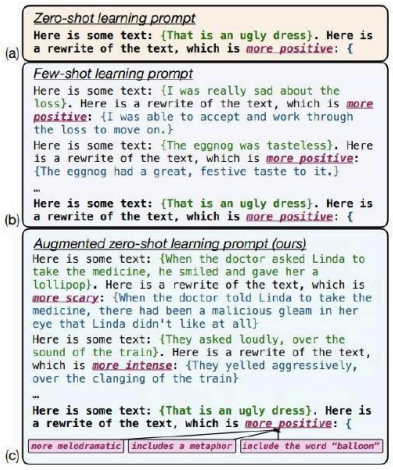
Почему она называется unsupervised style learning – потому что можно один раз модель обучить, а дальше для произвольного стиля подставлять примеры из нужного корпуса, и он будет в эту сторону предложения переписывать. Они взяли какой-то набор из таких условно ортогональных стилей, как positive/negative, умножили его на Fashion, Software и Pantry, и показывают, что если подавать в стиль много соответствующих предложений, то он более-менее хорошо восстанавливает.

unsupervised style learning

Reserved ⇒ Emotive	**Emotive ⇒ Reserved**
I liked the movie. ⇒ I cannot even describe how amazing this movie was!!	I loved every minute of the movie! ⇒ I liked the movie.
I was impressed with the results. ⇒ I was absolutely blown away with the results!!	I was shocked by the amazing results! ⇒ I was surprised by the results.
American ⇒ British	**British ⇒ American**
The elevator in my apartment isn't working. ⇒ The lift in my flat isn't working.	The lift in my flat isn't working. ⇒ The elevator in my apartment isn't working.
The senators will return to Washington next week. ⇒ The MPs will return to Westminster next week.	MPs will return to Westminster next week. ⇒ Representatives will return to Washington next week.
Polite ⇒ Rude	**Rude ⇒ Polite**
Are you positive you've understood my point? ⇒ you've never understood my point!	What the hell is wrong with your attitude? ⇒ Perhaps the question is more about your attitude.
Could you ask before using my phone? ⇒ I ask you to stop using my phone!	I could care less, go find somebody else to do this crap. ⇒ I could be wrong, but I would try to find somebody else to do this.

Вот примеры текстового трансфера между степенью эмоциональности, British – American. Это, очевидно, cherry-pick, но тем не менее очень миленькие примеры. Сама идея тоже довольно красивая. Я не пробовал эту модель руками, но мне кажется, что сам подход может быть интересным.

augmented zero-shot learning



Original	There, in the middle of the street, stood an old man in a weatherbeaten brown coat.
More descriptive	The man's brown trenchcoat was coated with snow, and as he leaned against the telephone booth, he cast a long shadow across the road in the late afternoon sun.
Include a metaphor	There, in the middle of the street, stood an old man whose face resembled the weatherworn bark of a gnarled tree.
Include the word "*park*"	There, in the middle of Central Park, stood an old man in a weatherbeaten brown coat.
Include the word "*balloon*"	There, in the middle of the street, stood an old man with several colourful balloons tied to the straps of his coat.
More comic	There, in the middle of the street, stood an old man in a raggedy brown coat that was much too big for him.
More melodramatic	There, in the middle of the road, stood a grizzled old man, the light of life faded from his sunken eyes.

[arXiv:2109.03910]

Еще один подход – это взять GPT-3, написать ей промпт и сказать «Давай напиши нам рерайт этого текста», и оказывается, что если хорошо писать промпт и давать ей указания на какие-то действия с текстом, то в каких-то случаях она действительно неплохо работает. Они померили, как это всё работает, и говорят, что куча случаев у них получилась хорошо.

augmented zero-shot learning

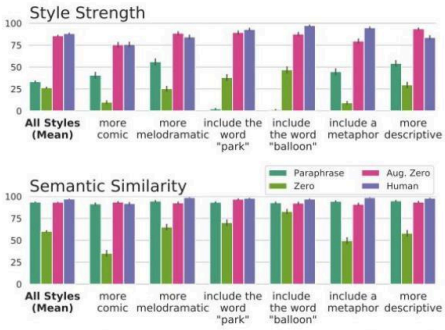


Figure 2: Human evaluation of style transfer for six atypical styles. Our method is rated comparably to the human-written ground truth. Error bars show Standard Error of the Mean. Evaluation of fluency is shown in Figure 4 in the Appendix.

[arXiv:2109.03910]

Я скептически отношусь к такой black box использования больших моделей типа GPT-3, потому что это весело выглядит, но не очень приближает нас к пониманию

происходящего, с одной стороны, и не дает контроля над происходящим. То есть что будет, когда она ошибется, и как мы об этом узнаем, тоже непонятно.

Третью часть про структуру сегодня мы не успели. Она будет в следующей части.

thanks for attention!
@altsoph