Перенос стиля текста

Text Style Transfer

Описание задачи TST

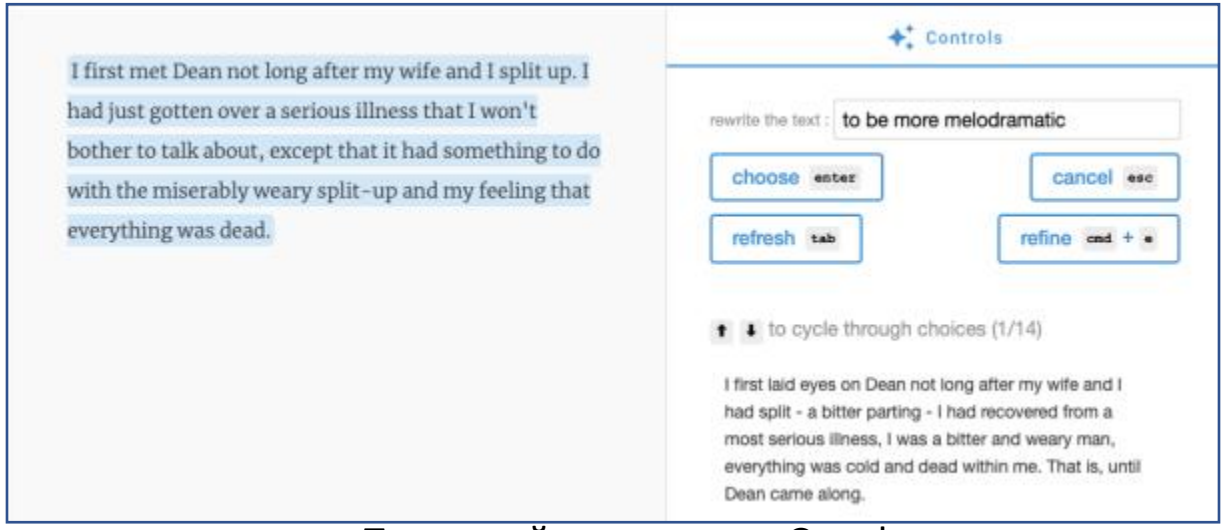
- Вход:
  - Исходный текст , имеющий стиль
  - Желаемый стиль
- Выход:
  - Текст , имеющий стиль

Примеры переноса стиля

	Вход	Выход
Тональность	Отличный телефон по доступной цене! *(позитивная тональность)*	Плохой телефон по завышенной цене! *(негативная тональность)*
Токсичность	Обиделся? Не плачь, нытик вонючий. *(токсичный текст)*	Вы обиделись? Не плачьте, пожалуйста, я же любя и по-дружески. *(нейтральный текст)*
Авторский стиль	Пустите, пожалуйста, я ничего! умолял слабый мужской голос. *(Л. Н. Толстой)*	Пощадите, пожалуйста, я ничего! прервал сильный женский крик. *(Ф. М. Достоевский)*

Применение

- Текстовые редакторы с интеллектуальными помощниками
- Генерация привлекающих новостных заголовков
- Персонализированные диалоговые системы
- Анонимизация авторства



Текстовый редактор от Google

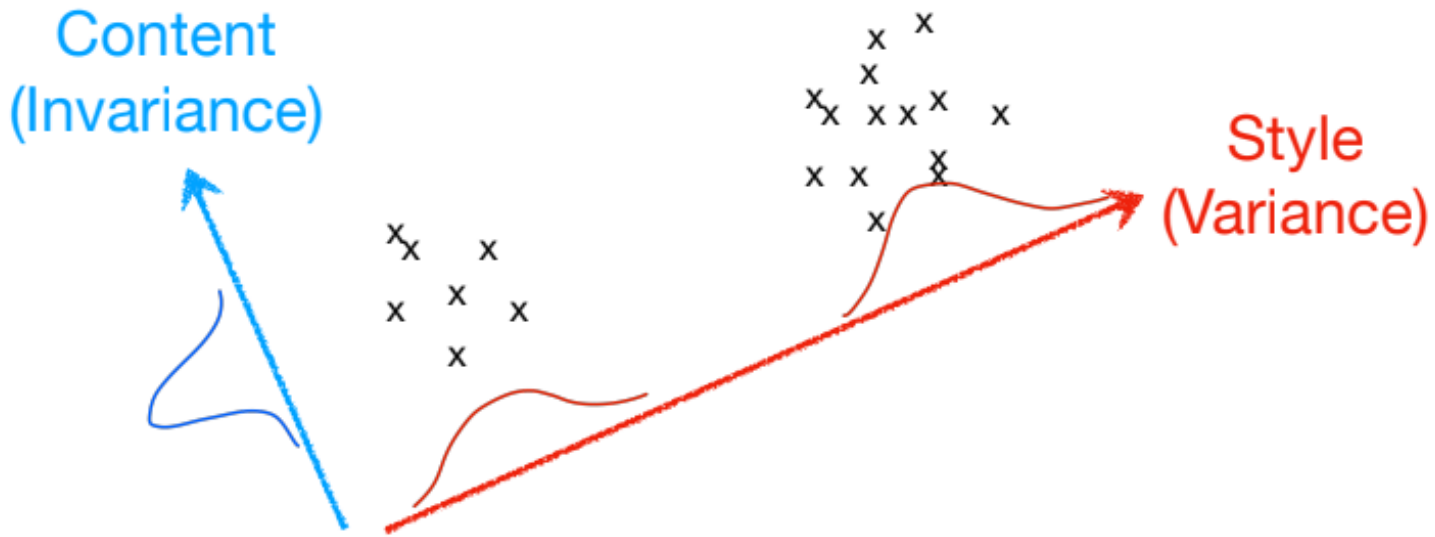
Стиль и содержание: лингвистический подход

- Стиль определяет как автор использует язык:
  - Выбор слов
  - Структура предложений
  - Образные выражения

  При помощи стиля автор может передать дополнительную информацию, интерпретируемую читателем.
- Содержание – суть текста, которую автор хочет донести до читателя.

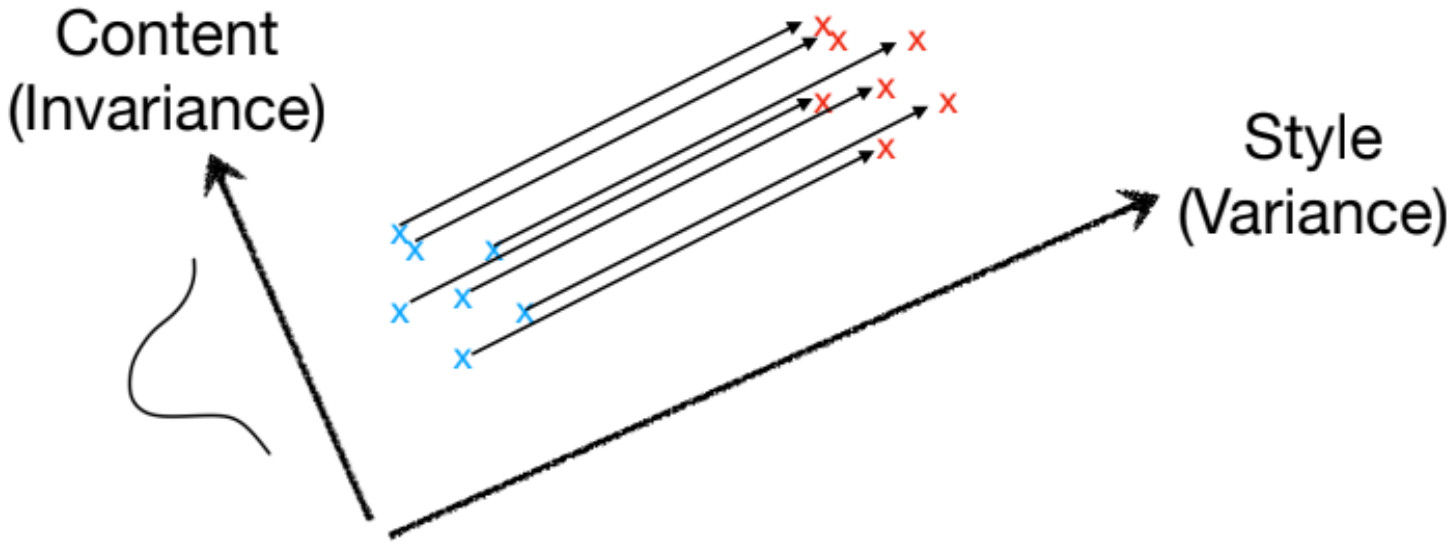
Стиль и содержание: эмпирический подход (data-driven)

- Стиль – множество атрибутов текстовых данных (метки)
- Может быть определен специальной функцией, например классификатором



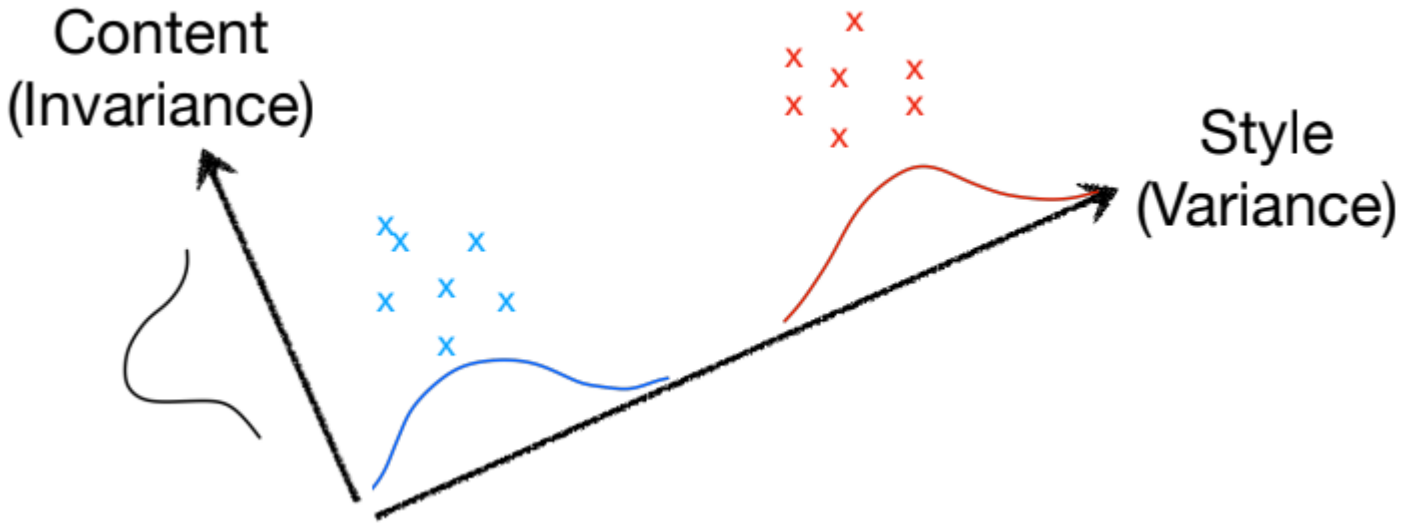
Параллельный корпус

- Пары текстов разных стилей с одинаковым содержанием



Не параллельный корпус

- Каждый текст имеет метку стиля, без сопоставления этого текста с другим стилем.



Критерии оценивания переноса стиля текста

1. Сохранение независимого от стиля содержания исходного текста
2. Соответствие выходного текста целевому стилю
3. Естественность языка

Вклад исследования

- Перенос авторского стиля
  - Составлен корпус русской литературы XIX века
  - Разработан модифицированный метод tf-idf взвешивания слов
  - Разработана модель переноса авторского стиля: tf-idf + word2vec

- Перенос тональности
  - Составлен параллельный корпус отзывов на мобильные телефоны
  - Протестированы модели ruGPT3 (small, medium, large, XL)

Перенос авторского стиля

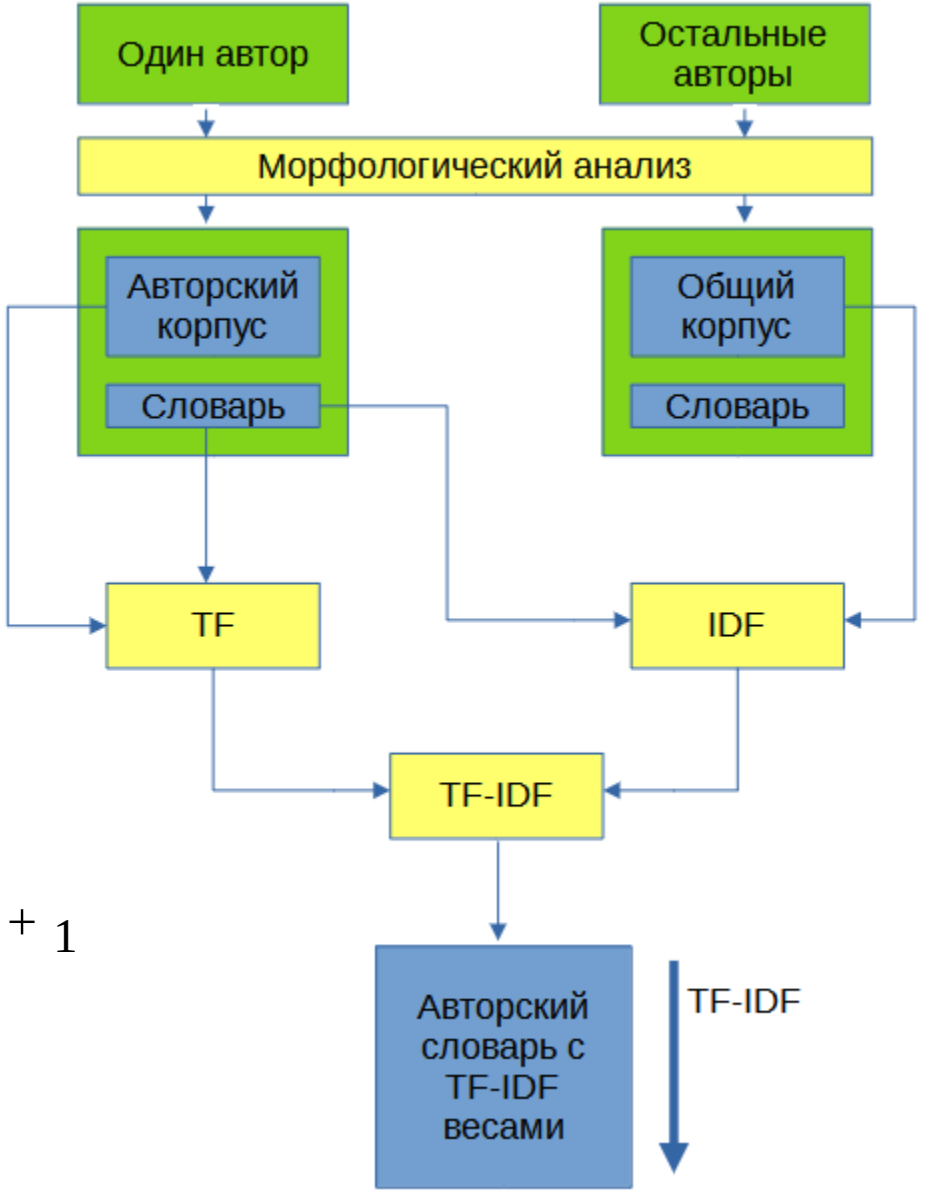
Корпус русской литературы XIX века

Автор	Количество предложений	Доля от общего количества, %
А. П. Чехов	15 310	5,9
Ф. М. Достоевский	54 177	20,8
Н. В. Гоголь	13 262	5,1
И. А. Гончаров	37 878	14,5
Н. С. Лесков	37 575	14,4
А. Н. Островский	19 202	7,4
А. С. Пушкин	6 836	2,6
М. Е. Салтыков-Щедрин	24 382	9,4
Л. Н. Толстой	41 469	15,9
И. С. Тургенев	16 666	6,4
Всего	260 757	

Tf-idf взвешивание

$$tf(t^Y, D^Y) = \log(N(t^Y, D^Y)) + 1$$

$$idf(t^Y, D^X) = \log \frac{|D^X| + 1}{|\{d_i^X \in D^X \mid t^Y \in d_i^X\}| + 1} + 1$$



Н. В. Гоголь	А. С. Пушкин	Л. Н. Толстой	И. А. Гончаров
чичиков	пугачёв	левин	обломов
козак	дубровский	нехлюдов	марфинька
запорожец	оренбург	вронский	райский
акакий	германн	кить	тарантьев
ноздрево	мятежник	пьер	адуев
ковалёв	ибрагим	кутузов	фрегат
акакиевич	яицкий	денисов	японец
манилов	михельсон	долли	марк
андрий	швабрин	ростов	леонтий
собакевич	кузмич	анатоль	штольц
бульб	кирилович	маслов	якут
остап	оренбургский	долох	викентьев
селифан	бибиков	ростовый	посьета
костанжочь	кирил	болконский	адмирал
козацкий	савельич	элен	фаддеев
кошев	сильвио	аркадьич	едо
оксана	комендант	дутлов	кичиб
парубок	кирджать	свияжский	колония
козаков	самозванец	катюша	савич
платонов	троекур	балашева	бен

Замена слов

- Word2vec (наиболее близкие)
- Фильтры:
  - Слово {авторский словарь}
  - Та же часть речи
  - tf-idf исходный tf-idf
  - Разные нормальные формы
- Сортировка по tf-idf

Ф. М. Достоевский → Н. В. Гоголь

1	**Д**	ведь это вы тогда вошли, а? **мать** ее просто **смешная** светская старушонка
	Г	ведь это вы тогда вошли, а? **старуха** ее просто **странная** светская старушонка
2	**Д**	но **девок** всего пришло только три, да и **марьи** еще не было
	Г	но **баб** всего пришло только три, да и **анны** еще не было
3	**Д**	еще от дворника узнал он, что **петрушка** и не **думал** являться
	Г	еще от дворника узнал он, что **селифан** и не **подумал** являться

Ф. М. Достоевский → А. С. Пушкин

4	**Д**	ведь это вы тогда вошли, а? **мать** ее просто **смешная** светская старушонка
	П	ведь это вы тогда вошли, а? **дочь** ее просто **славная** светская старушонка
5	**Д**	но девок всего **пришло** только три, да и марьи еще не было
	П	но девок всего **приходило** только три, да и марьи еще не было
6	**Д**	еще от **дворника** узнал он, что **петрушка** и не **думал** являться
	П	еще от **кабака** узнал он, что **ямщик** и не **чувствовал** являться

Accuracy (SVM + tf-idf) = 0,58

Проблемы

- Согласование форм слов
- Обучение word2vec
- Уни-грамм может быть недостаточно

Спасибо за внимание