НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Компьютерная лингвистика и информационные технологии

Неделя 4: Аугментация данных, линейная регрессия

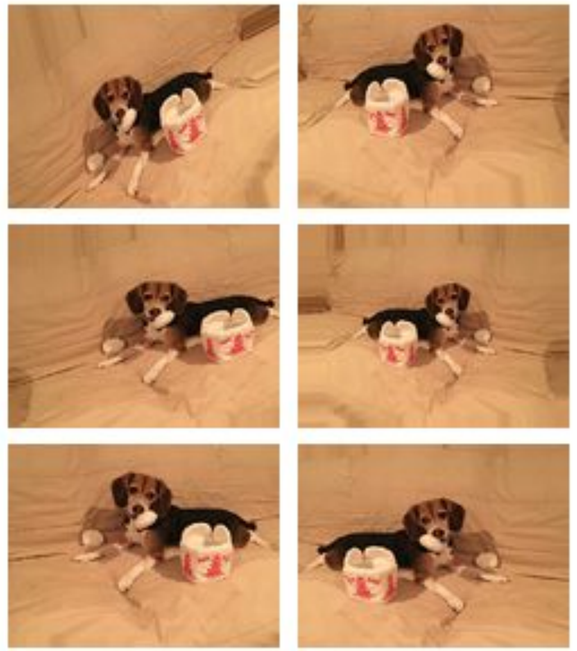
Data Augmentation

- Back-translation
- Word replacement
- Search engines
- Character-level augmentation
- Templates
- Syntactic tree augmentation
- Automatic annotation

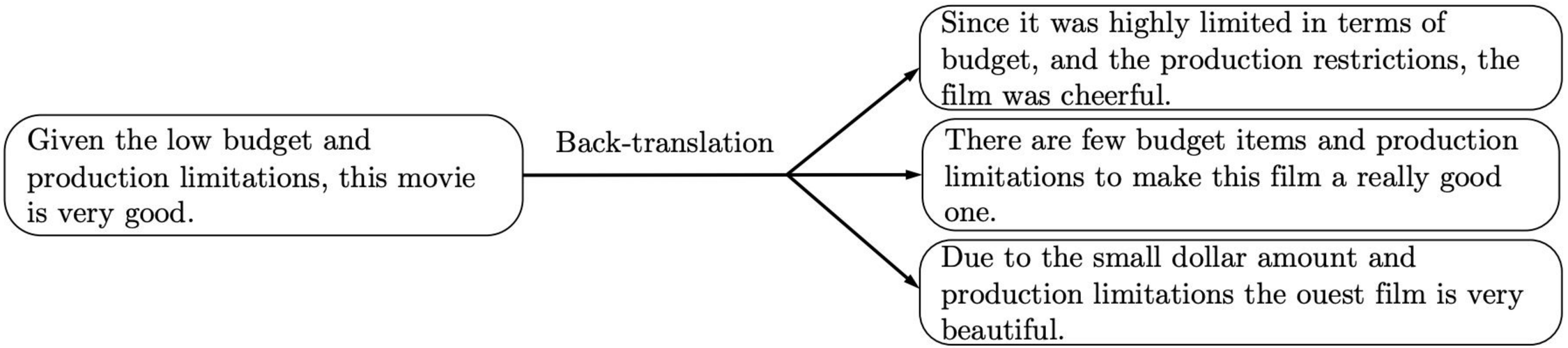
Input Image

Keras

Augmented Images



Back-translation



Word replacement

- Embeddings

> *Я так и знал! - бормотал он в смущении, - я так и думал!*

> *Я так и думал! - шептал он в замешательстве, - я так и подумал!*

- TF-IDF

> Замена слов с низким скором по коллекции текстов

https://arxiv.org/pdf/1904.12848.pdf

Search Engines

Пушкин Александр Сергеевич 1799 - 1837: биог…
histrf.ru › Личности › Биографии › p/pushkin-alieksandr… ▾
Александр **Пушкин родился** 26 мая (6 июня) 1799 года в небогатой дворянской семье. Начальное образование, как это было принято у дворян, маленький **Пушкин** получил дома, его обучением занимались учителя и гувернеры... Читать ещё ›

Александр **Пушкин** - биография, жизнь и смерт…
biographe.ru › znamenitosti/aleksandr-pushkin/ ▾
Родился Александр **Пушкин** 6 июня(26 мая по старому стилю) 1799 года в Москве. Его отец Сергей **Пушкин** был поэтом-любителем, светским острословом. Мама Надежда Ганнибал, приходилась внучкой Абраму Ганнибалу. Дед по... Читать ещё ›

Пушкин, Александр Сергеевич — Википедия
ru.wikipedia.org › Пушкин, Александр Серге… ▾
Алекса́ндр Серге́евич **Пу́шкин** — русский поэт, драматург и прозаик, заложивший основы русского реалистического направления, критик и теоретик литературы, историк, публицист...

Пушкин Александр Сергеевич — биография по…
culture.ru › persons/8195/aleksandr-pushkin ▾
Александр **Пушкин родился** в обедневшей дворянской семье 6 июня 1799 года. В раннем детстве он был молчаливым и малоподвижным ребенком... Читать ещё ›

Биография александра сергеевича **пушкина**...
zen.yandex.ru › Яндекс.Дзен › …-pushkina-1799-1837… ▾
Александр Сергеевич **Пушкин родился** 6 июня 1799 года (по старому стилю 26 мая) в Москве в дворянской помещичьей семье (отец его был майор в отставке) в день праздника Вознесения. В тот же день у императора Павла **родилась**... Читать ещё ›

Character-level Augmentation

- Sequence-to-sequence

> мамп мвла рам

> мфма мылд аму

- Keyboard distance
- Metaphone

ВИТАФСКИЙ → Витавский, Витовский.

ВИТИНБИРК → Витенберг, Виттенберг.

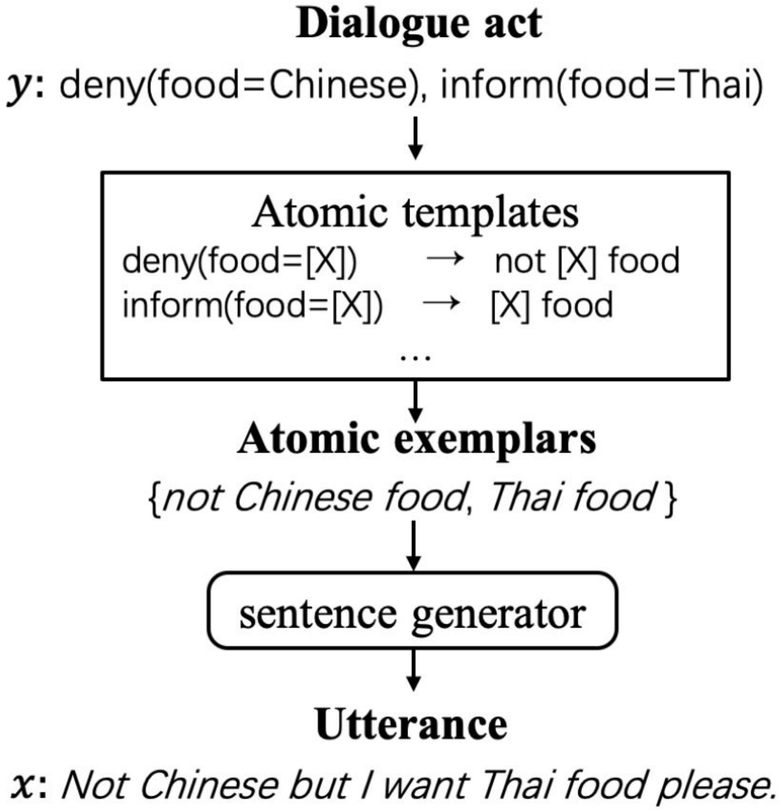
НАСАНАФ → Насанов, Насонов, Нассонов, Носонов.

ПИРМАКАФ → Пермаков, Пермяков, Перьмяков.

Templates

- мама мыла [slot]
  > мама мыла пол
  > мама мыла яблоко
  > мама мыла кошку

- погода (в) [city]
  > погода в санкт петербурге
  > погода нск



Syntax Tree Augmentation

https://arxiv.org/pdf/1903.09460.pdf

“Приемная была обставлена просто, но по-деловому”

- *Rotate*

> Просто, но по-деловому была обставлена приемная

> Просто, но по-деловому была приемная обставлена

- *Crop*

> Приемная была обставлена просто

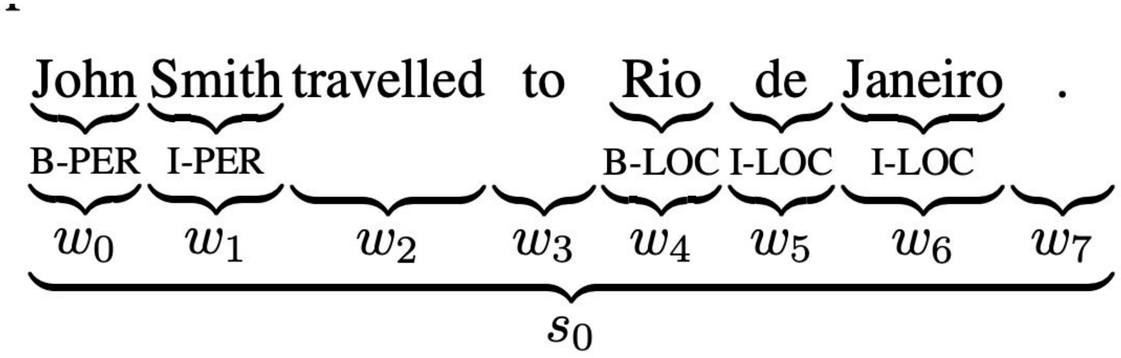
> Приемная была обставлена

Automatic Annotation

- Pre-trained model to annotate raw texts
  > e.g. general-domain NER
  > high probability scores

- Language models
  > LAMBADA (Anaby-Tavor et al., 2019)

- Algorithm-based automatic annotation
  > dictionaries
  > knowledge graphs



Notes

- Аугментированные данные используются только для обучения
- Аугментация данных не предполагает, что все будет хорошо