курс «Глубокое обучение»

Аугментация для текстов

Александр Дьяконов

22 июля 2020 года

Аугментация для текстов

В отличие от изображений при аугментации текстов

- **больше шансов поменять смысл**
- **менее «автоматические» преобразования**

не простое растяжение, а подбор синонима

Недостижимая мечта аугментации – перефразирование:

«Петя купил машину»

↓

«А автомобиль был приобретён Петром»

Аугментация для текстов

замена	перестановки	зашумление
• замена синонимом • сокращение или распаковка сокращения • используя представления слов • замена несущественного • замена существенного • замена на спецтокен или удаление • анализ окружения	• слов • предложений • кроссовер	• ошибки в буквах • изменение регистра • изменение пунктуации • типичные ошибки

замена + вставка	генерация
• случайная вставка синонима • используя контекстные представления / маскирование	• Text MixUp • трансформация синтаксического дерева • генеративные модели

Аугментация для текстов – замена синонимом

Thesaurus – аугментация для text CNN с помощью словаря

Zhang et al. introduced synonyms Character-level Convolutional Networks for Text Classification // https://arxiv.org/pdf/1509.01626.pdf

Synonym Replacement (SR) – выбор нескольких случайных не стоп-слов, замена синонимом

Wei et al. «EDA: Easy Data Augmentation Techniques for Boosting Performance onText Classification Tasks» // https://arxiv.org/abs/1901.11196

«полуувядших лилий аромат мои мечтанья лёгкие туманит»

«полуувядших лилий запах мои мечтанья грудь туманит»

аромат	лёгкие
запах	грудь (1)
дух	дыхалка (1)
благоуханье	невесомые (2)

Аугментация для текстов – сокращения

contractions – есть списки сокращений

https://en.wikipedia.org/wiki/Wikipedia%3aList_of_English_contractions

Contraction ⬍	Meaning
'aight	alright
ain't	am not / is not / are not / has not / have not / did not (colloquial)[1]
amn't	am not[2]
aren't	are not[3]
can't	cannot
'cause	because
could've	could have
couldn't	could not
couldn't've	could not have
daren't	dare not / dared not
daresn't	dare not
dasn't	dare not

не все сокращения допускают однозначную «распаковку»:

«He's» → «He is» / «He has»

библиотека https://github.com/kootenpv/contractions

Аугментация для текстов – использование представлений слов
Word Embeddings

выбираем соседа в пространстве представлений

Wang and Yang «That's So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors using #petpeeve Tweets» https://aclweb.org/anthology/D15-1306

«полуувядших лилий аромат мои мечтанья лёгкие туманит»
«полуувядших лилий запах мои мечтанья лёгкие туманит»

аромат
запах 0.75
благоуханье 0.67
пряный 0.65
сладковатый 0.63

```
# pip install gensim
import gensim.downloader as api

model = api.load('glove-twitter-25')
model.most_similar('cat', topn=5)
```

Аугментация для текстов – Контекстные представления

Contextualized Word Embeddings

двунаправленная LM для генерации подходящего, но редкого слова

Marzieh Fadaee et al. «Data Augmentation for Low-Resource Neural Machine Translation»

https://arxiv.org/pdf/1705.00440.pdf

«полуувядших лилий аромат мои мечтанья лёгкие туманит»

«полуувядших роз аромат мои мечтанья лёгкие туманит»

«полуувядших листьев аромат мои мечтанья лёгкие туманит»

«полуувядших листьев аромат мои мечтанья лёгкие туманит»

Сейчас можно использовать Masked Language Model

```
from transformers import pipeline
nlp = pipeline('fill-mask')
nlp('This is <mask> cool')
```

Аугментация для текстов – контекстные представления (продолжение)

bi-directional LSTM-RNN LM, сэмплирование с температурой

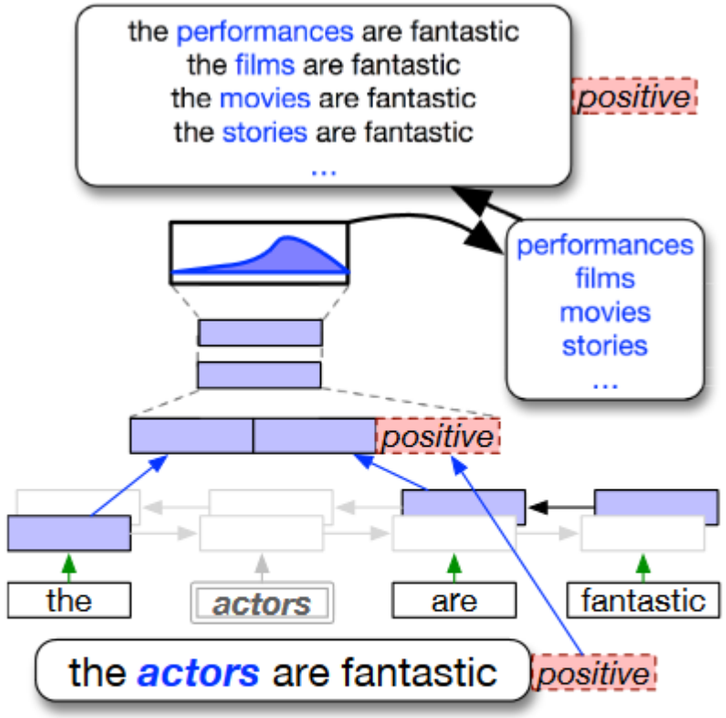


Figure 1: Contextual augmentation with a bi-directional RNN language model, when a sentence *"the actors are fantastic"* is augmented by replacing only *actors* with words predicted based on the context.

Kobayashi «Augmentation: Data Augmentation by Words with Paradigmatic Relation»

https://arxiv.org/pdf/1805.06201.pdf

Аугментация для текстов – маскирование

Проблемы в маскировании – могут изменить смысл

Изменение предложения с помощью маскирования BERT (замена, вставка слева, вставка справа) – поиск состязательных примеров

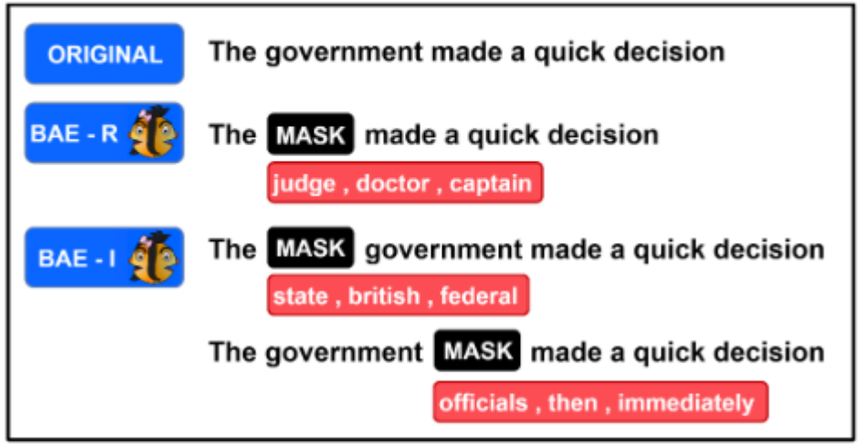


Figure 1: We use BERT-MLM to predict masked tokens in the text for generating adversarial examples. The MASK token replaces a word (BAE−R attack) or is inserted to the left/right of the word (BAE−I).

https://github.com/QData/TextAttack/

Garg. et al. «BAE: BERT-based Adversarial Examples for Text Classification»

https://arxiv.org/abs/2004.01970

Аугментация для текстов – маскирование (продолжение)

Original [Positive Sentiment]: This film offers many delights and surprises.
TextFooler: This flick citations disparate revel and surprises.
BAE-R: This movie offers enough delights and surprises
BAE-I: This lovely film platform offers many pleasant delights and surprises
BAE-R/I: This lovely film serves several pleasure and surprises .
BAE-R+I: This beautiful movie offers many pleasant delights and surprises .
Original [Positive Sentiment]: Our server was great and we had perfect service.
TextFooler: Our server was tremendous and we assumed faultless services.
BAE-R: Our server was decent and we had outstanding service.
BAE-I: Our server was great enough and we had perfect service but.
BAE-R/I: Our server was great enough and we needed perfect service but.
BAE-R+I: Our server was decent company and we had adequate service.

Table 3: Qualitative examples of each attack on the BERT classifier (Replacements: Red, Inserts: Blue)

TextFooler – чуть хуже читаемость

Di Jin et al. «Is bert really robust? natural language attack on text classification and entailment» // arXiv:1907.11932

Аугментация для текстов – замена несущественного

Заменяем слова с маленьким TF-IDF

«полуувядших лилий аромат мои мечтанья лёгкие туманит»
«полуувядших лилий аромат наши мечтанья лёгкие туманит»

Xie et al. «Unsupervised Data Augmentation» https://arxiv.org/abs/1904.12848

Аугментация для текстов – замена существенного (продолжение)

не просто удаляем слова, а удаляем «самые сентиментные» (по словарю)

в задаче определения сентимента

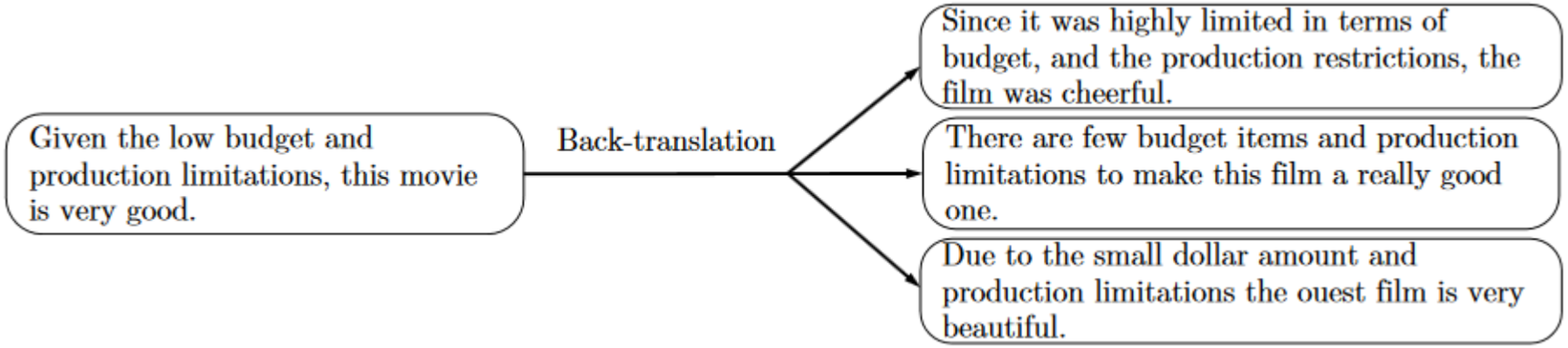
добавляем контрпримеры Adversarial Examples

Original text	the only problem is that, by the end, no one in the audience or the film seems to really care
DA-EK	the only that , by the end, one in the audience or the film seems to care
Adversarial example	the only difficulty is that, by the end, no one in the audience or the movie seems to really caring
DA-ADV	the only is that, by the end, no one in the audience or the seems to really
Original text	michel piccoli's moving performance is this films reason for being
DA-EK	michel piccoli's this films reason for being
Adversarial example	michel piccoli's moving play is this movie reason for being
DA-ADV	michel piccoli's moving is this reason for being

Table 2: Some examples of the augmented data created by DA-EK and DA-ADV.

Hanjie Chen «Improving the Explainability of Neural Sentiment Classifiers via Data Augmentation» https://arxiv.org/abs/1909.04225

Аугментация для текстов – обратный перевод
Back Translation



Xie et al. «Unsupervised Data Augmentation» https://arxiv.org/abs/1904.12848

«полуувядших листьев аромат мои мечтанья лёгкие туманит»

↓

«Half-withered leaves the scent of my dreams fogs my lungs»

↓

«Полусохшие листья, запах моих снов затуманивает мои легкие»

изначально перевод целевой язык → исходный, и пополнение выборки // Sennrich et al., 2016

Аугментация для текстов – обратный перевод (продолжение)

можно использовать несколько языков

средства для Back Translation

TextBlob https://textblob.readthedocs.io/en/dev/

https://amitness.com/2020/02/back-translation-in-google-sheets/

https://amitness.com/back-translation/

Kaggle 1top «Toxic Comment Classification Challenge» https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/discussion/52557

Аугментация для текстов – зашумление

Random Noise Injection

ошибки в буквах

+ изменение регистра, пунктуации

«полуувядших лилий аромат мои мечтанья лёгкие туманит»

«полуувяйших лилий аройат мои мечт7нья лёгкие туманит»

типичные ошибки

«полуувядших лилий аромат мои мечтанья лёгкие туманит»

«полуувчдших лидий аромат мои мечтанья дёгкие туманит»

Аугментация для текстов – зашумление (продолжение)

Замена спец-токеном / Blank Noising
или удаление слов

«полуувядших лилий аромат мои мечтанья лёгкие туманит»
«полуувядших лилий __ мои мечтанья лёгкие туманит»

Ziang Xie et al. «Data noising assmoothing in neural network language models»
https://arxiv.org/abs/1703.02573

перестановки слов / Random Swap

«полуувядших лилий аромат мои мечтанья лёгкие туманит»
«полуувядших аромат лилий мои мечтанья лёгкие туманит»

перестановки предложений / Sentence Shuffling

Аугментация для текстов – зашумление (продолжение)

Random Insertion (RI)
берём произвольное слово, находим синоним, вставляем в случайную позицию предложения

«полуувядших лилий аромат мои мечтанья запах лёгкие туманит»

Wei et al. «EDA: Easy Data Augmentation Techniques for Boosting Performance onText Classification Tasks» // https://arxiv.org/abs/1901.11196

Аугментация для текстов – кроссовер

берём представителей одного класса

Полуувядших лилий аромат Мои мечтанья легкие туманит. Мне лилии о смерти говорят, О времени, когда меня не станет.	Мир — успокоенной душе моей. Ничто ее не радует, не ранит. Не забывай моих последних дней, Пойми меня, когда меня не станет.
↓	
Полуувядших лилий аромат Мои мечтанья легкие туманит. Не забывай моих последних дней, Пойми меня, когда меня не станет	Мир — успокоенной душе моей. Ничто ее не радует, не ранит. Мне лилии о смерти говорят, О времени, когда меня не станет.

увеличивает F1-меру, в задаче классификации твитов:

Franco M. Luque «Atalaya at TASS 2019: Data Augmentation and Robust Embeddings for Sentiment Analysis» // https://arxiv.org/abs/1909.11241

Аугментация для текстов – MixUp for Text

$$\widetilde{x}^{ij} = \lambda x^i + (1-\lambda)x^j$$
$$\widetilde{y}^{ij} = \lambda y^i + (1-\lambda)y^j$$

Смешивать можно представления

1) на уровне слов

+ можно выбирать ближайшее слово к л/к в пространстве представлений

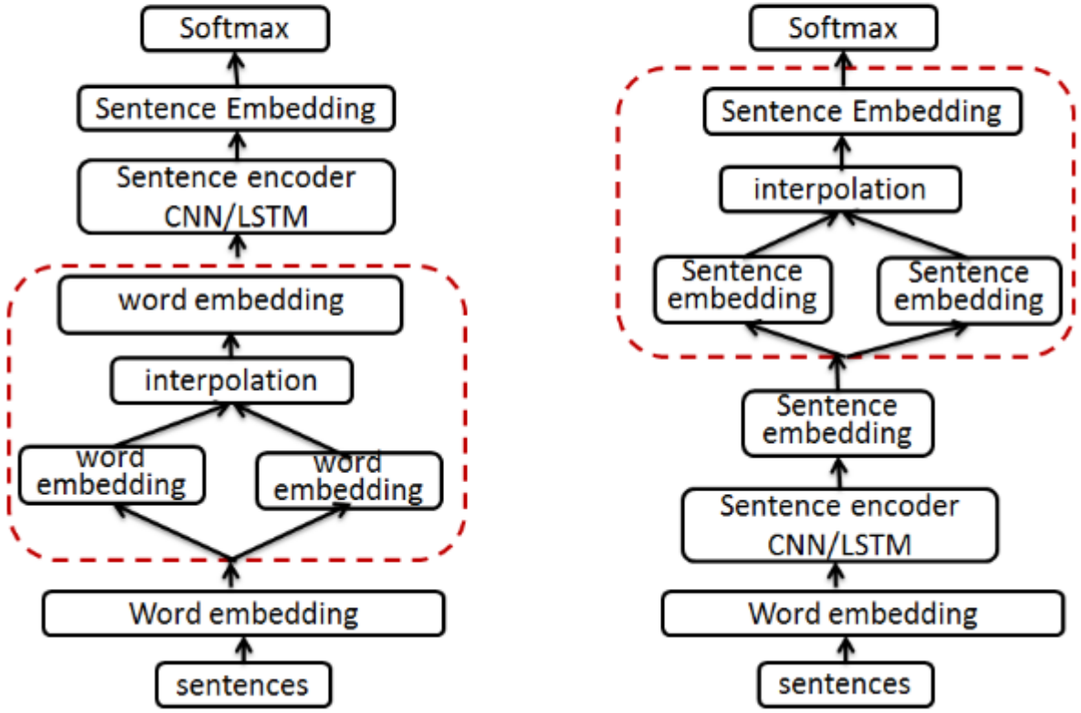
2) на уровне предложений

Можно выбирать слова с вер-ми (λ, 1–λ)

Предварительно выравниваем длины спец-токенами

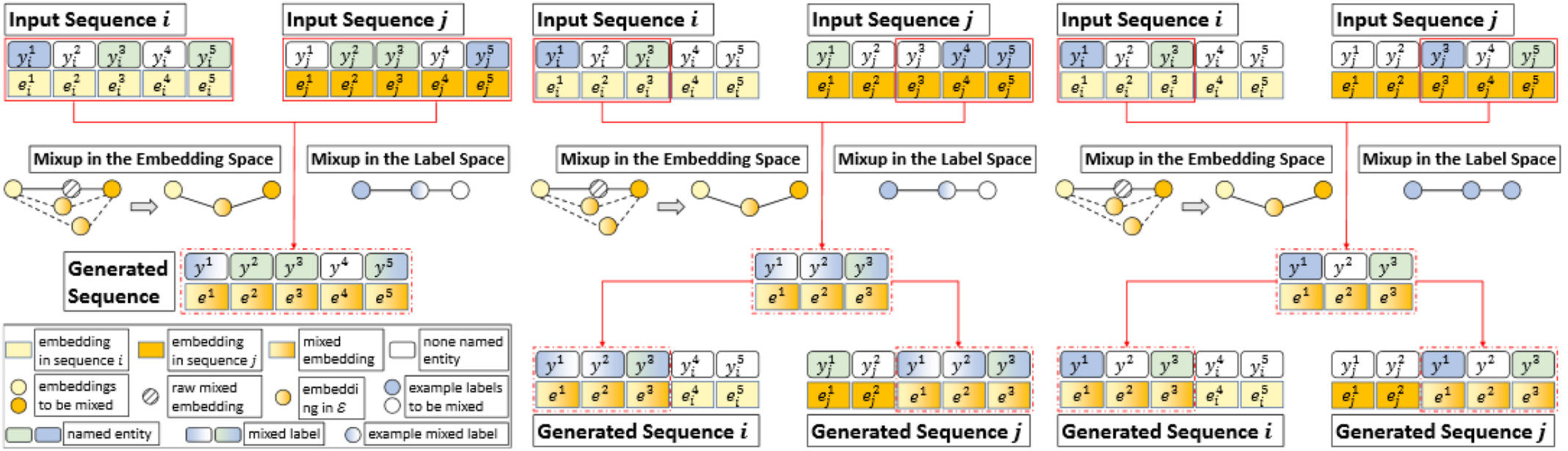
в любом случае, смешиваем представления

Figure 1: Illustration of wordMixup (left) and sen-Mixup (right), where the added part to the standard sentence classification model is in red rectangle.



Hongyu Guo«Augmenting Data with Mixup for Sentence Classification: An Empirical Study» // https://arxiv.org/abs/1905.08941

Аугментация для текстов – MixUp for Text



(a) Whole sequence mixup (b) Sub-sequence mixup (c) Label-constrained sub-sequence mixup

Figure 1: Illustration of the three variants of SeqMix. We use $s = 5, \eta_0 = \frac{3}{5}$ for whole-sequence mixup and $s = 3, \eta_0 = \frac{2}{3}$ for sub-sequence mixup and label-constrained sub-sequence mixup. The solid red frames indicate paired sequences or sub-sequences, and the red dotted frames indicate generated sequence or sub-sequence. In the original sequences, the parts not included in the solid red frames will be unchanged in the generated sequences. For the mixup in the embedding space, we take the embedding in $\mathcal{E}$ which is closest to the raw mixed embedding as the generated embedding. For the mixup in the label space, the mixed label can be used as the pseudo label.

R Zhang «Seqmix: Augmenting active sequence labeling via sequence mixup»

https://arxiv.org/abs/2010.02322

Аугментация для текстов – синтаксическим деревом

syntax trees transformations

«Полуувядших лилий аромат мои мечтанья легкие туманит»

↓

«Мои мечтанья туманятся полуувядших лилий ароматом»

Coulombe «Text Data Augmentation Made Simple by Leveraging NLP Cloud APIs»
https://arxiv.org/abs/1812.04718

Аугментация для текстов – синтаксическим деревом (продолжение)

Example of dependency tree produced by SyntaxNet

A man eats an apple in the kitchen.

```
root eats/eat/verb/proper_unknown/present/third/gender_unknown/singular
    nsubj man/man/noun/proper_unknown/tense_unknown/person_unknown/gender_unknown/singular
        det A/A/det/proper_unknown/tense_unknown/person_unknown/gender_unknown/number_unknown
    dobj apple/apple/noun/proper_unknown/tense_unknown/person_unknown/gender_unknown/singular
        det an/an/det/proper_unknown/tense_unknown/person_unknown/gender_unknown/number_unknown
    prep in/in/adp/proper_unknown/tense_unknown/person_unknown/gender_unknown/number_unknown
        pobj kitchen/kitchen/noun/proper_unknown/tense_unknown/person_unknown/gender_unknown/singular
            det the/the/det/proper_unknown/tense_unknown/person_unknown/gender_unknown/number_unknown
    p ./punct/proper_unknown/tense_unknown/person_unknown/gender_unknown/number_unknown
```

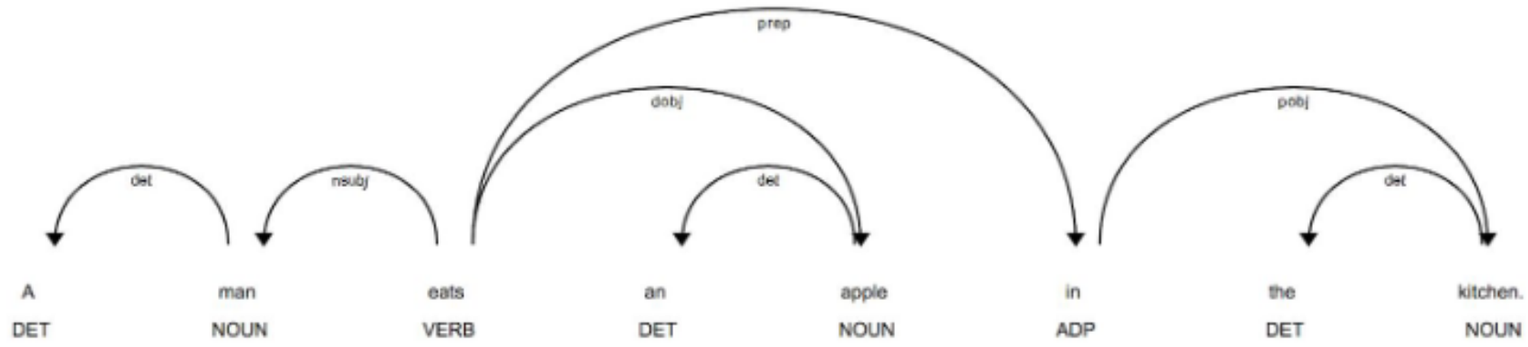


Diagram drawn with the help of spaCy [Honnibal & Montani, 2017]

Аугментация для текстов – анализ окружений

GECA = good-enough compositional augmentation

The cat sang.
The wug sang. →
The cat daxed.

The wug daxed.
~~*The sang daxed.*~~

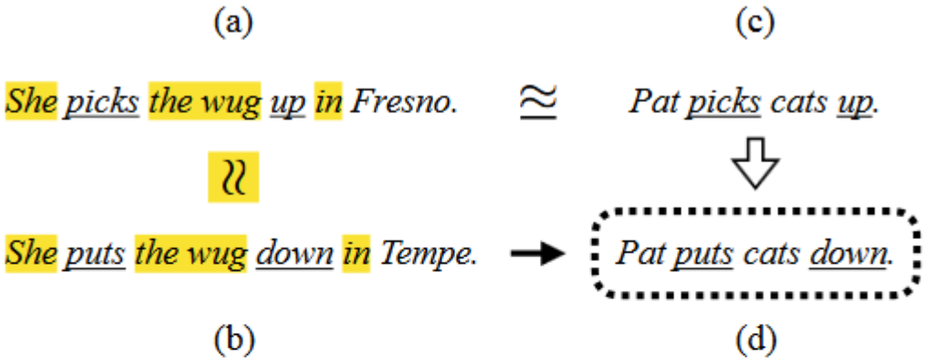


Figure 1: Visualization of the proposed approach: two discontinuous sentence fragments (a–b, underlined) which appear in similar environments (a–b, highlighted) are identified. Additional sentences in which the first fragment appears (c) are used to synthesize new examples (d) by substituting in the second fragment.

если какой-то фрагменты A и B появляются в одном лексическом окружении, то они взаимозаменяемые

Jacob Andreas «Good-Enough Compositional Data Augmentation» // https://arxiv.org/pdf/1904.09545.pdf

Аугментация для текстов – генеративная

Генерация текста (Text Generation) языковой моделью

Kafle et al. «Data Augmentation for Visual Question Answering»

https://aclweb.org/anthology/W17-3529

можно

- **предобучать на специальных корпусах**
- **обуславливать**

например, генерировать объекты нужного класса

тогда вход метка класса + сепаратор

«<Positive><SEP> ...»

- **менять закон сэмплирования**

больше вариативность

Аугментация для текстов – генеративная (продолжение)

Algorithm 1: Data Augmentation approach

Input : Training Dataset D_{train}

Pretrained model $G \in \{AE, AR, Seq2Seq\}$

Fine-tune G using D_{train} to obtain G_{tuned}

$D_{synthetic} \leftarrow \{\}$

foreach $\{x_i, y_i\} \in D_{train}$ **do**

  Synthesize s examples $\{\hat{x}_i, \hat{y}_i\}_p^1$ using G_{tuned}

  $D_{synthetic} \leftarrow D_{synthetic} \cup \{\hat{x}_i, \hat{y}_i\}_p^1$

end

Kumar et al. «Data Augmentation using Pre-trained Transformer Models»

https://arxiv.org/abs/2003.02245

Аугментация для текстов – генеративная (продолжение)

LAMBADA (Language Model Based Data Augmentation)

Algorithm 1: LAMBADA

Input: Training dataset D_{train}
Classification algorithm $\mathcal{A}$
Language model $\mathcal{G}$
Number to synthesize per class $N_1, \ldots, N_q$

1 Train a baseline classifier h from D_{train} using $\mathcal{A}$
2 Fine-tune $\mathcal{G}$ using D_{train} to obtain $\mathcal{G}_{tuned}$
3 Synthesize a set of labeled sentences D^* using $\mathcal{G}_{tuned}$
4 Filter D^* using classifier h to obtain $D_{synthesized}$
5 **return** $D_{synthesized}$

Anaby-Tavor et al. «Not Enough Data? Deep Learning to the Rescue!»
https://arxiv.org/abs/1911.03118

Аугментация для текстов – генеративная (продолжение)

В последнее время много применений трансформеров в медицине

Медицинская задача

1) кодировщик-декодировщик

2) GPT-2 с контекстным токеном

Exploring Transformer Text Generation for Medical Dataset Augmentation

https://www.aclweb.org/anthology/2020.lrec-1.578/

Что дальше? – RL для выбора аугментации и весов объектов

Algorithm 1 Joint Learning of Model and Data Manipulation

Input: The target model $p_\theta(y|x)$
The data manipulation function $R_\phi(x, y|\mathcal{D})$
Training set $\mathcal{D}$, validation set $\mathcal{D}^v$

1: Initialize model parameter θ and manipulation parameter ϕ
2: **repeat**
3: Optimize θ on $\mathcal{D}$ enriched with data manipulation through Eq.(7)
4: Optimize ϕ by maximizing data log-likelihood on $\mathcal{D}^v$ through Eq.(8)
5: **until** convergence

Output: Learned model $p_{\theta^*}(y|x)$ and manipulation $R_{\phi^*}(y, x|\mathcal{D})$

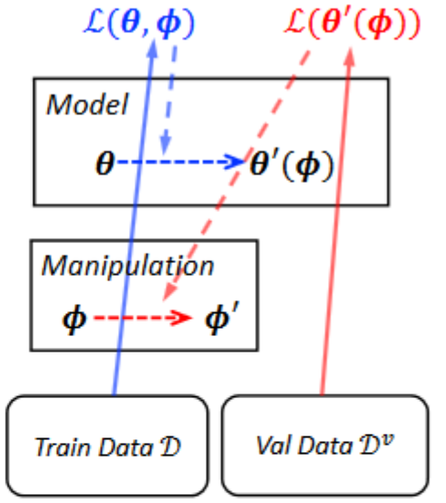


Figure 1: Algorithm Computation. Blue arrows denote learning model θ. Red arrows denote learning manipulation ϕ. Solid arrows denote forward pass. Dashed arrows denote backward pass and parameter updates.

Параллельно вывод: LM data augmentation плоха для дисбаланса.

Zhiting Hu et al «Learning Data Manipulation for Augmentation and Weighting»
http://papers.nips.cc/paper/9706-learning-data-manipulation-for-augmentation-and-weighting

Аналогичные обзоры

Data augmentation in NLP

https://towardsdatascience.com/data-augmentation-in-nlp-2801a34dfc28

Data augmentation for NLP

https://amitness.com/2020/05/data-augmentation-for-nlp/

SOTA

https://paperswithcode.com/task/text-augmentation

Библиотеки

https://github.com/makcedward/nlpaug

там и для звуков…

и библиотека https://github.com/QData/TextAttack