Trabalho 2:
Métodos Computacionais

Matéria: Tópicos Avançados em Inteligência Computacional

Autor: Matheus Muniz Damasco
Professores: Heder Soares e Alex Borges Vieira

Data Augmentation

Em NLP existem várias técnicas que nos permitem pegar um pequeno conjunto de dados e usar alguns truques para criar mais dados. Esses truques também são chamados de aumento de dados (data augmentation) e buscam explorar propriedades da linguagem para criar textos que sejam sintaticamente semelhantes aos dados de origem.

Técnicas de Data Augmentation

Synonym replacement:

Selecione aleatoriamente "k" palavras em uma sentença que não sejam stop words e substituímos elas pelos seus sinônimos. Para obter sinônimos, podemos utilizar os Synsets do WordNet [1,2].

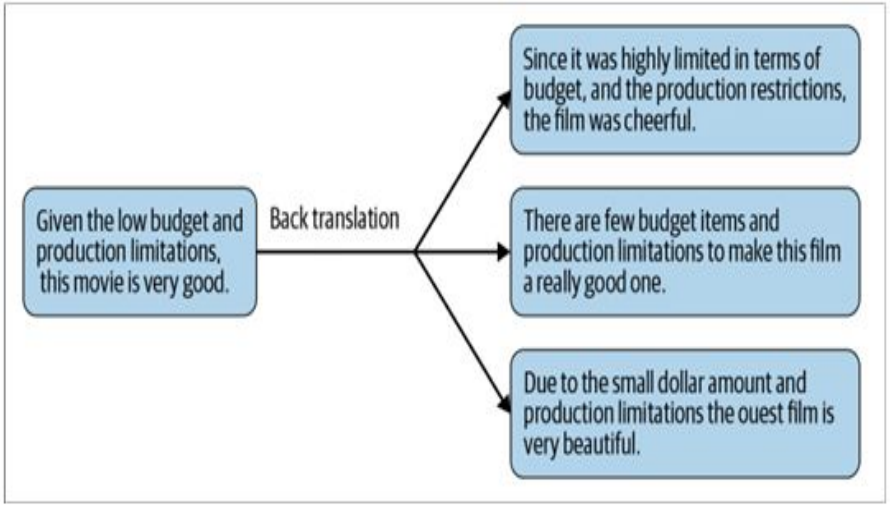
Bigram flipping:

Divida a sentença em bigrams. Selecione aleatoriamente um bigrams e inverta sua ordem. Por exemplo: "Eu estou indo ao supermercado." Se escolhermos o bigram "indo ao", ao invertê-lo, a sentença resultante seria: "Eu estou ao indo supermercado."

Técnicas de Data Augmentation

Back-Translation:

Considere uma sentença S1 em inglês. Utilizamos uma biblioteca de tradução automática, como o Google Tradutor, para traduzi-la para outro idioma, por exemplo, alemão, resultando na sentença S2. Em seguida, traduzimos S2 de volta para o inglês, obtendo a sentença S3. Observamos que S1 e S3 possuem significados muito semelhantes, mas apresentam variações sutis. Podemos então adicionar S3 ao nosso conjunto de dados. Este método é especialmente eficaz para classificação de texto.



Técnicas de Data Augmentation

Replacing entities:

Substitua entidades como nomes de pessoas, locais, organizações, etc... por outras entidades da mesma categoria. Por exemplo, em "Eu moro na Califórnia", substituímos "Califórnia" por "Londres".

Substituição de Palavras Baseada em TF-IDF:

A back-translation pode omitir certas palavras cruciais na sentença. Para lidar com isso, os autores do artigo 'Unsupervised Data Augmentation for Consistency Training' em utilizam o TF-IDF [3].

Técnicas de Data Augmentation

Adding noise to data:

Em muitas aplicações de NLP, os dados recebidos contêm erros ortográficos, principalmente devido às características da plataforma onde são gerados (por exemplo, Twitter). Nesses casos, podemos adicionar um pouco de ruído aos dados para treinar modelos mais robustos. Por exemplo, escolhemos aleatoriamente uma palavra em uma sentença e substituímos por outra palavra com grafia semelhante. Outro tipo de ruído é o problema de "dedo gordo" em teclados móveis. Podemos simular erros de teclado QWERTY substituindo alguns caracteres por seus vizinhos no teclado.

Técnicas de Data Augmentation

Snorkel [4,5]:

Este é um sistema para construir dados de treinamento automaticamente, sem rotulagem manual. Usando o Snorkel, um grande conjunto de dados de treinamento pode ser “criado” — sem rotulagem manual — usando heurística e criando dados sintéticos ao transformar dados existentes e criar novas amostras de dados. Essa abordagem demonstrou funcionar bem no Google no passado [6].

Easy Data Augmentation (EDA) [7,8] e NLPAug [9]:

Estas duas bibliotecas são usadas para criar amostras sintéticas para NLP. Elas fornecem a implementação de várias técnicas de aumento de dados, incluindo algumas das que discutimos anteriormente.

Técnicas de Data Augmentation

Active learning [10]:

Um paradigma especializado de ML onde o algoritmo de aprendizado pode consultar interativamente um ponto de dados e obter seu rótulo. É utilizado em cenários onde há abundância de dados não rotulados, mas a rotulagem manual é cara. Nesses casos, a questão torna-se: para quais pontos de dados devemos solicitar rótulos para maximizar o aprendizado enquanto mantemos baixo o custo de rotulagem?

Data Augmentation Using GANs [11]:

Este estudo propõe o uso de GANs para gerar dados artificiais de treinamento, especialmente útil em situações de conjuntos de dados desbalanceados ou com informações sensíveis.

Técnicas de Data Augmentation

Paraphrasing Revisited with Neural Machine Translation [12]:

Este estudo revisita a técnica de tradução automática onde um texto é traduzido para uma língua intermediária antes de ser traduzido para a língua alvo, apresentando um modelo que reescreve frases ou textos usando palavras diferentes, mas mantendo o mesmo significado, baseado em redes neurais. O modelo representa as frases ou textos reescritos em um espaço contínuo, estima o grau de similaridade semântica entre segmentos de texto de qualquer comprimento e gera frases ou textos candidatos para qualquer entrada.

Técnicas de Data Augmentation

Text Data Augmentation Using Generative Adversarial Networks [13]:

Este estudo aborda as Redes Adversárias Generativas, compostas por duas redes neurais baseadas em GANs: o Gerador e o Discriminador. O Gerador cria dados sintéticos que imitam os dados reais de treinamento, enquanto o Discriminador avalia esses dados, distinguindo entre os reais e os gerados. Durante o treinamento, ambas as redes competem entre si, aprimorando suas habilidades e resultando em dados sintéticos cada vez melhores.

Referências

- [1] Miller, George A. “WordNet: A Lexical Database for English.” Communications of the ACM 38.11 (1995): 39–41.
- [2] NTLTK documentation. “WordNet Interface”. Last accessed June 15, 2020.
- [3] Xie, Qizhe, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. “Unsupervised Data Augmentation for Consistency Training”. (2019).
- TF-IDF - 3 Basic Approaches in Bag of Words which are better than Word Embeddings
- [4] Snorkel. “Programmatically Building and Managing Training Data”. Last accessed June 15, 2020.
- [5] Ratner, Alexander, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. “Snorkel: Rapid Training Data Creation with Weak Supervision.” The VLDB Journal 29 (2019): 1–22.

Referências

- [6] Bach, Stephen H., Daniel Rodriguez, Yintao Liu, Chong Luo, Haidong Shao, Cas sandra Xia, Souvik Sen et al. “Snorkel DryBell: A Case Study in Deploying Weak Supervision at Industrial Scale”. (2018).
- [7] Wei, Jason W., and Kai Zou. “Eda: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks”, (2019).
- [8] GitHub repository for [10]. Last accessed June 15, 2020.
- [9] Ma, Edward. nplaug: Data augmentation for NLP, (GitHub repo). Last accessed June 15, 2020.
- [10] Shioulin and Nisha. “A Guide to Learning with Limited Labeled Data”. April 2, 2019.

Referências

- [11] Data Augmentation Using GANs
- [12] Paraphrasing Revisited with Neural Machine Translation
- [13] Text Data Augmentation Using Generative Adversarial Networks – A Systematic Review