**Summary of references**

Stanza：

<https://stanfordnlp.github.io/stanza/depparse.html>

Pytextrank：

https://derwen.ai/docs/ptr/overview/

|  |  |  |  |
| --- | --- | --- | --- |
| Style transfer |  |  |  |
| Title | Dataset | Method | Evaluation matrix |
| 《Delete, Retrieve, Generate:  A Simple Approach to Sentiment and Style Transfer》 | 1.YELP  2.Reviews on Amazon  3.Image captions | Baseline models：  RETRIEVEONLY：Returns, word by word, sentences retrieved that are similar to the original sentence with the deleted attribute. Guaranteed to produce a grammatically correct sentence with the target attributes, but its content may differ from the original sentence.  TEMPLATEBASED：Replace the deleted attribute in the original sentence with the target attribute  Neural modes：  DELETEONLY：An RNN is used to embed the sentences with the attributes removed into a vector. It then connects the final hidden state to the learned embedding of the target attribute and feeds it to the RNN decoder to generate the final sentence. The decoder attempts to generate words that represent the source content and target attributes while maintaining fluency.  DELETEANDRETRIEVE：Similar to DELETEONLY, but uses the attributes of the retrieved sentence rather than the target attributes. As with DELETEONLY, it encodes the sentence with the original deleted attribute using an RNN. The RNN decoder uses the splicing and content embedding of this vector to generate the final sentence. |  |
| Text Style Transfer: A Review and Experimental Evaluation | 1. Yelp a corpus of restaurant reviews 2. Amazon product review datasets   3.IMDb movie reviews datasets | Creating a taxonomy to organize the TST models, and provide a comprehensive summary of the state of the art. We review the existing evaluation methodologies for TST tasks, and conduct a large-scale reproducibility study where we experimentally benchmark 19 state-of-the-art TST algorithms on two publicly available datasets. Finally, we expand on current trends and provide new perspectives on the new and exciting developments in the TST field. |  |
| Style Transfer Through Back-Translation | Gender  1. Reddy and Knight’s (2016) dataset of reviews from Yelp annotated for two genders corresponding to markers of sex.  Political slant  2.Top-level comments on Facebook posts from all  412 current members of the United States Senate and House who have public Facebook pages.  Sentiment.   1. Yelp a corpus of restaurant reviews. | Focus on transferring author attributes:  gender and political slant, and on sentiment modification.  The second task is novel:  given a sentence by an author with a particular political leaning, rephrase the sentence to preserve  its meaning but to confound classifiers of political slant. | Style Transfer Accuracy    Preservation of Meaning    Fluency |
| Dear Sir or Madam, May I Introduce the GY AFC Dataset:  Corpus, Benchmarks and Metrics for Formality Style Transfer | Yahoo Answers,a question answering forum,contains a large number of informal sentences and allows redistribution of data. Hence, we use the  Yahoo Answers L6 corpus5 to create our GYAFC  dataset of informal and formal sentence pairs. | 1.Rule-based Approach  Develop a set of rules to automatically make an informal sentence more formal where we  capitalize first word and proper nouns, remove repeated punctuations, handcraft a list of expansion for contractions etc.  2.Phrase-based Machine Translation  Use a combination of training regimes to develop our model. Train on the output of the rule based approach when applied to GYAFC. This is meant to force the PBMT model to learn generalizations outside the rules.  3.Neural Machine Translation  Experiment with three NMT models  NMT baseline: Our baseline model is a bidirectional LSTM encoder-decoder model with attention.  NMT Copy: Jhamtani et al., (2017) introduce a copy-enriched NMT model for style transfer to better handle stretches of text which  should not be changed. We incorporate this mechanism into our NMT Baseline.  NMT Combined:We augment the  data used to train NMT Copy via two techniques:  1) we run the PBMT model on additional source data, and 2) we use back-translation (Sennrich et al., 2016c) of the PBMT model to translate the large number of in-domain target style sentences  from GYAFC. | Human-based Evaluation：  Formality,fluency,meaning preservation,overall Ranking  Automatic Metrics:  Formality,fluency,meaning preservation,overall Ranking |
| MULTIPLE-ATTRIBUTE TEXT STYLE TRANSFER | **Yelp Reviews**  This dataset consists of restaurant and business reviews provided by the Y elp Dataset Challenge  **Amazon Reviews**  The amazon product review dataset (He & McAuley, 2016) is comprised of  reviews written by consumers of Amazon products.  **Public social media content**  3 independent pieces of available information about that content: 1) gender  (male or female) 2) age group (18-24 or 65+), and 3) writer-annotated feeling (relaxed or annoyed). |  | 1) produce sentences that conform to the set of pre-specified  attribute(s), 2) preserve the structure and content of the input, and 3) generate fluent language. We  therefore evaluate samples from different models along three different dimensions:  **Attribute control**: We measure the extent to which attributes are controlled using fastText  classifiers, trained on our datasets, to predict different attributes.  **Fluency**: Fluency is measured by the perplexity assigned to generated text sequences by a  pre-trained Kneser–Ney smooth 5-gram language model using KenLM (Heafield, 2011).  **Content preservation**:We measure the extent to which a model preserves the content  present of a given input using n-gram statistics, by measuring the BLEU score between  generated text and the input itself, which we refer to as self-BLEU. |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Machine translation |  |  |  |
| Title | Dataset | Method | Evaluation matrix |
|  |  |  |  |

There is a delicate balance between retaining the original content and removing the original attributes, and existing adversarial models tend to sacrifice one or the other.

It introduces a simple approach to text attribute transfer whose main advantage comes from induction bias, i.e. attributes are usually represented as locally recognisable phrases, which allows it to outperform previous models based on adversarial training. It also reflects some problems in that content and attributes cannot be so cleanly separated by word boundaries.

A fruitful direction would be to develop a concept of attributes that is more general than N-grams, more general than N-grams, but with more inductive bias than arbitrary potential vectors.

|  |  |
| --- | --- |
| **Evaluation of machine translation** |  |
| **Evaluation indicators** | **Meaning** |
| BLEU | BLEU was one of the first metrics to report a high correlation with human judgments of quality. The metric is currently one of the most popular in the field. The central idea behind the metric is that "the closer a machine translation is to a professional human translation, the better it is".[8] The metric calculates scores for individual segments, generally sentences — then averages these scores over the whole corpus for a final score. It has been shown to correlate highly with human judgments of quality at the corpus level.[9] |
| NIST | The NIST metric is based on the BLEU metric, but with some alterations. Where BLEU simply calculates n-gram precision adding equal weight to each one, NIST also calculates how informative a particular n-gram is. That is to say, when a correct n-gram is found, the rarer that n-gram is, the more weight it is given.[11] For example, if the bigram "on the" correctly matches, it receives lower weight than the correct matching of bigram "interesting calculations," as this is less likely to occur. NIST also differs from BLEU in its calculation of the brevity penalty, insofar as small variations in translation length do not impact the overall score as much. |
| Word error rate | The Word error rate (WER) is a metric based on the Levenshtein distance, where the Levenshtein distance works at the character level, WER works at the word level. It was originally used for measuring the performance of speech recognition systems but is also used in the evaluation of machine translation. The metric is based on the calculation of the number of words that differ between a piece of machine-translated text and a reference translation.  A related metric is the Position-independent word error rate (PER), which allows for the re-ordering of words and sequences of words between a translated text and a reference translation. |
| METEOR | The METEOR metric is designed to address some of the deficiencies inherent in the BLEU metric. The metric is based on the weighted harmonic mean of unigram precision and unigram recall. The metric was designed after research by Lavie (2004) into the significance of recall in evaluation metrics. Their research showed that metrics based on recall consistently achieved higher correlation than those based on precision alone, cf. BLEU and NIST.[12]  METEOR also includes some other features not found in other metrics, such as synonymy matching, where instead of matching only on the exact word form, the metric also matches on synonyms. For example, the word "good" in the reference rendering as "well" in the translation counts as a match. The metric is also includes a stemmer, which lemmatises words and matches on the lemmatised forms. The implementation of the metric is modular insofar as the algorithms that match words are implemented as modules, and new modules that implement different matching strategies may easily be added. |
| LEPOR | A new MT evaluation metric LEPOR was proposed as the combination of many evaluation factors including existing ones (precision, recall) and modified ones (sentence-length penalty and n-gram based word order penalty). The experiments were tested on eight language pairs from ACL-WMT2011 including English-to-other (Spanish, French, German, and Czech) and the inverse, and showed that LEPOR yielded higher system-level correlation with human judgments than several existing metrics such as BLEU, Meteor-1.3, TER, AMBER and MP4IBM1.[13] An enhanced version of LEPOR metric, hLEPOR, is introduced in the paper.[14] hLEPOR utilizes the harmonic mean to combine the sub-factors of the designed metric. Furthermore, they design a set of parameters to tune the weights of the sub-factors according to different language pairs. The ACL-WMT13 Metrics shared task [15] results show that hLEPOR yields the highest Pearson correlation score with human judgment on the English-to-Russian language pair, in addition to the highest average-score on five language pairs (English-to-German, French, Spanish, Czech, Russian). The detailed results of WMT13 Metrics Task is introduced in the paper.[16] |
|  |  |