Presentasi dengan LaTeX

The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics

Prames Ray Lapian

Teknik Informatika Universitas Padjadjaran

October 5, 2023

Daftar Isi

Deskripsi Judul

Makalah ini membahas pembuatan GEM Benchmark, sebuah tolok ukur hidup untuk Natural Language Generation (NLG) yang bertujuan untuk mengevaluasi dan mengukur kemajuan NLG. Makalah ini membahas keterbatasan tolok ukur yang ada saat ini dengan menyediakan lingkungan yang beragam dan multibahasa untuk evaluasi, mendorong penelitian yang bertanggung jawab melalui transparansi, dan bertujuan untuk meningkatkan evaluasi sistem NLG. Makalah ini menjelaskan proses pemilihan dataset, modifikasi yang dilakukan pada dataset, pengaturan eksperimental, dan hasil awal untuk berbagai tugas pembuatan bahasa.

Jurnal Pendukung

1. STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation.
2. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations.
3. Neural machine translation by jointly learning to align and translate.
4. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments.
5. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing.
6. The #benderrule: On naming the languages we study and why it matters.
7. Data statements for natural language processing: Toward mitigating system bias and enabling better science.
8. BERT: Pre-training of deep bidirectional transformers for language understanding.

Data yang Digunakan

GEM (Generation, Evaluation, and Metrics)

Generation, Evaluation, and Metrics (GEM) adalah lingkungan tolok ukur untuk Generasi Bahasa Alami dengan fokus pada Evaluasinya, baik melalui anotasi manusia dan Metrik otomatis.
GEM bertujuan untuk:

- Mengukur kemajuan NLG di 13 set data yang mencakup banyak tugas dan bahasa NLG.
- Memberikan analisis mendalam tentang data dan model yang disajikan melalui pernyataan data dan set tantangan.
- Mengembangkan standar untuk evaluasi teks yang dihasilkan dengan menggunakan metrik otomatis dan metrik manusia.

Merupakan tujuan kami untuk memperbarui GEM secara teratur dan mendorong praktik yang lebih inklusif dalam pengembangan dataset dengan memperluas data yang ada atau mengembangkan dataset untuk bahasa tambahan.

Data atau Informasi Pendukung

1. GLUE (General Language Understanding Evaluation benchmark)
2. CommonGen
3. E2E (End-to-End NLG Challenge)
4. ASSET
5. WikiLingua
6. TurkCorpus
7. MLSUM (MultiLingual SUMmarization)
8. ToTTo
9. DART
10. XSum

Terima Kasih

Terima kasih atas perhatiannya.