PENGEMBANGAN MODEL *DEEP LEARNING* UNTUK PEMBANGKIT SOAL OTOMATIS MENGGUNAKAN *RECURRENT NEURAL NETWORK*

DEEP LEARNING MODEL BUILDING FOR AUTOMATIC QUESTION GENERATOR USE RECURRENT NEURAL NETWORK

Danang Wijaya, Handaru Jati
Fakultas Teknik Universitas Negeri Yogyakarta,
danangwijaya.2018@student.uny.ac.id

ABSTRACT

The application of Artificial Intelligence in the field of education, especially learning evaluation is still lacking. There are no tools for teachers in developing questions for learning evaluation. The aims of this study are: (1) To develop a deep learning model for automatic question generation using a recurrent neural network. (2) Guarantee the accuracy and performance obtained from the deep learning that has been developed with an evaluation using untrained automatic metrics. The method used is Research and Development with deep learning model development stages consisting of translating datasets, preparing datasets, building deep learning models, and evaluating models. The model was developed with the SQuAD2.0 dataset which was translated into Indonesian and built using OpenNMT. The subject of this research is the result of the evaluation of the model's performance using untrained automatic metrics. The results of this study are: (1) A deep learning model for automatic question generation using a recurrent neural network, which consists of several configurations of the BiGRU and BiLSTM models with variations of cased and uncased training datasets. (2) The best model evaluation results were obtained on the Uncased BiLSTM model with a score of BLEU-1 38.10, BLEU-2 20.69, BLEU-3 10.58, BLEU-4 5.78, ROUGE-L 42.98, METEOR 18.31.

Keywords*: Artificial Intelligence, Deep Learning, Automatic Question Generator, Recurrent Neural Network,* SQuAD v2.0, OpenNMT, *Untrained Automatic Metrics*

ABSTRAK

Penerapan *Artificial Intelligence* di bidang pendidikan khususnya evaluasi pembelajaran masih kurang. Selain itu belum adanya alat bantu untuk guru dalam mengembangkan soal untuk evaluasi pembelajaran. Tujuan dari penelitian ini adalah: (1) Mengembangakan model *deep learning* untuk pembangkit soal otomatis menggunakan *recurrent neural network.* (2) Menjamin akurasi dan kinerja yang diperoleh dari model *deep learning* yang telah dikembangkan dengan evaluasi menggunakan *untrained automatic metric*. Metode yang digunakan adalah *Research and Development* dengan tahapan pengembangan model *deep learning* yang terdiri dari menerjemahkan dataset, menyiapkan dataset, membangun model *deep learning*, evaluasi model. Model dikembangkan dengan dataset SQuAD2.0 yang diterjemahkan ke bahasa Indonesia dan dibangun menggunakan OpenNMT. Subjek dari penelitian ini adalah hasil dari evaluasi kinerja model yang dilakukan dengan menggunakan *untrained automatic metrics*.Hasil dari penelitian ini adalah: (1) Model *deep learning* untuk pembangkit soal otomatis menggunakan *recurrent neural network*, yang terdiri dari beberapa konfigurasi model BiGRU dan BiLSTM serta variasi dataset latih *cased* dan *uncased*. (2) Hasil evaluasi model terbaik diperoleh pada model BiLSTM *Uncased* dengan skor BLEU-1 38.10, BLEU-2 20.69, BLEU-3 10.58, BLEU-4 5.78, ROUGE-L 42.98, METEOR 18.31.

Kata Kunci: *Artificial Intelligence, Deep Learning*, Pembangkit soal otomatis, *Recurrent Neural Network*, SQuAD2.0, OpenNMT, *Untrained Automatic Metrics.*

PENDAHULUAN

Teknologi saat ini sudah banyak dimanfaatkan untuk membantu pekerjaan manusia disegala aspek kehidupan, seperti pada bidang kesehatan, industri, bahkan pendidikan. Salah satu teknologi yang paling banyak penerapannya dalam kehidupan sehari-hari adalah teknologi kecerdasan buatan atau *Artificial Intelligence (AI).*Kecerdasan Buatan yang sering disebut dengan *Artificial Intelligence* atau AI adalah bidang ilmu komputer yang berfokus pada penciptaan mesin cerdas yang dapat bekerja dan bereaksi seperti manusia. Beberapa kegiatan komputer yang memiliki kecerdasan buatan adalah seperti pengenalan ucapan, belajar, perencanaan dan pemecahan masalah (Habeeb, 2017). Penerapan AI sendiri sudah banyak, akan tetapi penggunaan AI dalam bidang pendidikan masih kurang.

Pendidikan khususnya di Indonesia sendiri masih memiliki banyak masalah terkait dengan akses dan kualitasnya. Secara umum, AI dapat dikatakan

sebuah ilmu yang digunakan untuk meniru kecerdasan yang dimiliki oleh makhluk hidup untuk diterapkan didalam sebuah mesin dengan tujuan untuk memecahkan suatu masalah, dengan adanya AI dapat dimungkinkan permasahalan didalam bidang pendidikan dapat kurangi. Salah satu permasalahan dibidang pendidikan Indonesia adalah pada proses evaluasi pembelajaran, hal ini terkait dengan belum adanya pemanfaat teknologi informasi secara maksimal. Evaluasi pembelajaran adalah suatu kegiatan yang penting untuk dilakukan, karena merupakan sebuah alat untuk tolak ukur hasil dari sebuah proses pembelajaran yang telah dilakukan sebelumnya. Menurut Gronlund (1976) dalam (Purwanto, 2009) evaluasi adalah suatu proses yang sistematis untuk menentukan atau membuat keputusan sampai sejauh mana tujuan-tujuan pembelajaran telah dicapai oleh siswa. Evaluasi pembelajaran yang sering dilakukan adalah dengan memberikan soal yang terkait dengan materi yang diberikan saat proses pembelajaran berlangsung. Tugas seorang pendidik atau guru dalam proses evaluasi pembelajaran adalah membuat soal-soal yang digunakan dalam evaluasi pembelajaran. Soal-soal yang dibuat oleh guru juga dapat digunakan untuk latihan siswa.

Permasalahan yang muncul pada saat proses evaluasi pembelajaran adalah ketika guru harus membuat butir soal yang jumlahnya tidak sedikit, sedangkan seorang guru juga memiliki tugas lain yang berkaitan dengan administrasi sekolah dan lain sebagainnya, serta waktu yang dimiliki oleh seorang guru juga sedikit. Selain itu kendala lain yang dimiliki oleh seorang guru dalam proses evaluasi pembelajaran adalah guru kesulitan dalam mengembangkan instrumen dalam membuat soal tes (Asarina, 2014). Banyaknya tugas seorang guru, dan guru juga sering mendapatkan kesulitan ketika mengembangankan soal untuk evaluasi pembelajaran, maka sering sekali soal-soal yang digunakan pada evaluasi pembelajaran tahun sebelumnya digunakan kembali dan membuat siswa hanya meghafalkan jawaban yang benar dari soal-soal tersebut, maka diperlukannya alat bantu untuk guru mengembangkan dalam mengembangkan soal yang digunakan untuk proses evaluasi pembelajaran. Alat bantu yang sudah ada belum memiliki tingkat akurasi dan kinerja yang baik, dari masalah-masalah tersebut untuk dapat mengurangi dan mengatasinya, maka dapat diimplementasikannya sebuah teknologi dari bidang AI yaitu *Natural Language Processing* (NLP).

NLP adalah sebuah ilmu gabungan antara AI dan linguistik pada tahun 1950-an. Sebelumnya NLP berbeda dengan *Text Information Retrival* (IR), yang menggunakan teknik statistika yang skalabel untuk melakukan pengindeksan dan pencarian teks dalam volume yang besar secara efisien (Garbade, 2018). NLP yang populer dan saat ini banyak pengaplikasiannya seperti untuk *text summarization, machine translation, question answering,* dan *automatic question generator*. Teknologi NLP yang penulis usulkan adalah *Automatic Question Generator* (AQG). AQG merupakan sebuah sistem yang dapat membuat pertanyaan dari informasi yang ada berupa teks dengan menggunakan algoritma tertentu dan pola tertentu. AQG termasuk salah satu tugas NLP yaitu *Natural Language Generation* (NLG) yang implementasinya dapat dilakukan menggunakan *Machine Learning* (ML) dengan algoritma *Deep Learning* (DL) seperti *Sequence-to-Sequence* yaitu *Recurrent Neural Network* (RNN) yang terdiri dari *Long Short Term Memory* (LSTM) dan *Gated Recurrent Unit* (GRU). Hasil model dari NLG dapat dilakukan pengukuran untuk mengetahui tingkat performanya menggunakan metode *Untrained Automatic Metric* (UAM). Standar UAM yang digunakan untuk evaluasi otomatis saat ini adalah *Bilingual Evaluation Understudy* (BLEU) (Papineni dkk., 2001), selain itu metrik selain BLEU adalah *Metric for Evaluation of Translation with Explicit Ordering* (METEOR) (Lavie & Agarwal, 2007) dan *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) (Lin, 2004).

Dengan adanya AQG diharapkan dapat bekerja seperti halnya seorang pengajar yang dapat membuat sebuah soal dari pengetahuan yang dimiliki dan teks yang dibaca dan dipahaminya, sehingga dapat membantu tugas seorang guru dalam membuat soal yang akan digunakan untuk proses evaluasi pembelajaran atau digunakan untuk memberikan stimulus siswa dan mengetahui kemampuan siswa terhadap materi yang telah diberikan oleh guru. Pada teknologi pembelajaran *Intelligent Tutoring System* (ITS) sistem AQG menjadi sebuah komponen yang penting (Fattoh dkk., 2015), yaitu dalam penerapan pada salah satu module ITS pada tahap *domain knowledge*, dalam tahap ini AQG dapat digunakan sebagai latihan soal yang dibuat berdasarkan konsep pengetahuan.

METODE PENELITIAN

Penelitian "Pengembangan model *deep learning* untuk pembangkit soal otomatis menggunakan *recurrent neural network*" ini menggunakan metode *Research and Development* (R&D) dengan tahapan pengembangan model *deep learning* yang terdiri dari menerjemahkan dataset, menyiapkan dataset, membangun model *deep learning*, evaluasi model.

Waktu dan Tempat Penelitian

Penelitian ini dimulai pada bulan Mei 2021 sampai dengan bulan Januari 2022. Tempat penelitian dilaksanakan di Program Studi Pendidikan Teknik Informatika, serta tempat-tempat lain di lingkungan Universitas Negeri Yogyakarta yang diperlukan untuk mendukung pelaksanaan penelitian.

Sumber Data/Subjek Penelitian

Subjek dari penelitian ini adalah akurasi dari produk yang dihasilkan atau prototipe model *deep learning* yaitu berasal dari hasil evaluasi mesin syaraf penerjemah yang dilakukan dengan *automated measures* menggunakan *untrained automatic metrics* yaitu metrik evaluasi BLEU, ROUGE dan METEOR. Objek dari penelitian ini adalah prototipe dari model *deep learning* itu sendiri.

Prosedur Pengembangan

Prosedur pengembangan model *Deep Learning* terdiri dari 4 tahapan yaitu: Menerjemahkan dataset, Menyiapkan dataset, Membangun model *deep learning*, Evaluasi model. Prosedur pengembangan dapat dilihat pada Gambar x.

Gambar 1. Tahapan Pengembangan

Fase tersebut dijabarkan sebagai berikut:

1. Menerjemahkan Dataset

Dataset yang digunakan dalam penelitian ini adalah dataset SQuAD v2.0 yang merupakan dataset pemahaman bacaan (*reading comprehension*) yang digunakan untuk *machine reading comprehension* yang berisi paragraf artikel dan pasangan pertanyaan-jawaban dalam bahasa inggris. Pada tahap ini dilakukan penerjemahan dataset dari bahasa inggris ke bahasa Indonesia dengan bantuan google translate.

2. Menyiapkan Dataset

Dataset SQuAD v2.0 yang telah diterjemahkan kedalam bahasa Indonesia selanjutnya dilakukan pemrosesan untuk dapat digunakan didalam proses pembuatan atau melatih model. Fokus pada tahap ini adalah menentukan posisi jawaban didalam paragraf artikel dari pertanyaan yang ada. Selain itu juga dilakukan penyiapan fitur linguistik untuk *POS tagging* dan *NER*, selanjutnya teks paragraf digunakan untuk input dan pertanyaan untuk target. Pada tahap ini juga membuat variasi input untuk *uncased* dan *cased* input. Selanjutnya membuat representasi untuk setiap input, fitur, target. Dalam tahap ini juga dilakukan pemisahan dataset yang digunakan untuk data latih dan data uji.

3. Membangun Model *Deep Learning*

Tahap membangun model *deep learning* dilakukan dengan proses latih model menggunakan *library* OpenNMT-py yang berbasis *framework* Py-Torch dan mengimplementasikan RNN model yaitu *Bidirectional Recurrent Unit* (BiGRU) dan - *Bidirection Long-Short Term Memory* (BiLSTM) dengan variasi dataset *cased* dan *uncased.* Pembuatan dan pelatiahan model menggunakan *pre-trained word embedding* dari FastText Indonesia.

4. Evaluasi Model

Tahap evaluasi model dilakukan dengan metode *automatic measures* menggunakan *Untrained Automatic Metrics* (UAM) terhadap model yang telah dibuat menggunakan metrik BLEU, ROUGE, dan METEOR.

Data, Instrumen, dan Teknik Pengumpulan

Setelah semua model dengan berbagai konfigurasi selesai dibangun kemudian dilakukan proses evaluasi dengan menggunakan beberapa *untrained automatic metrics.* Hasil evaluasi dari setiap konfigurasi model dengan setiap metrik dikumpulkan untuk dibandingkan. Metrik yang digunakan adalah BLEU, ROUGE, dan METEOR.

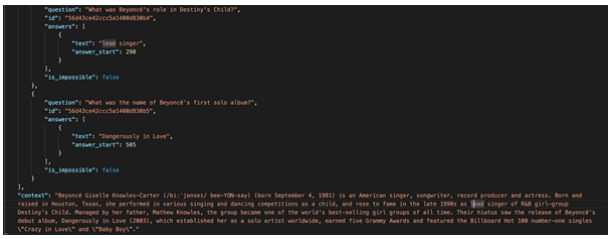
Teknik Analisis Data

Teknik analisis data adalah sebuah cara untuk mengolah dan menganalisis data yang telah dikumpulkan, berdasarkan instrumen penelitian yang digunakan. Analisis data yang dilakukan dari hasil evaluasi setiap konfigurasi model yang telah melalui tahap latih menggunakan *untrained automatic metric* adalah melakukan perbandingan kinerja hasil evaluasi dari proses latih model untuk kemudian diputuskan manakah konfigurasi model yang memiliki kinerja terbaik.

HASIL DAN PEMBAHASAHAN

1. Menerjemahkan Dataset

Pada tahap ini dilakukan penerjemahan dataset dari bahasa inggris ke bahasa Indonesia dengan bantuan *google translate* API V2. Hasil dari penerjemahan dataset menggunakan *google translate* belum dilakukan perbaikan, karena banyaknya jumlah data yang ada di dataset. Berikut adalah dataset latih dan uji asli dari SQuAD2.0.

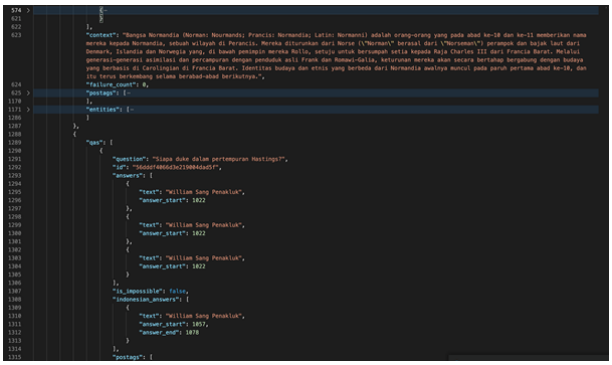


Gambar 2. Dataset latih asli SQuAD2.0

Gambar 3. Dataset uji asli SQuAD2.0

Setelah dilakukan penerjemahan menggunakan bantuan google transalate, dataset berubah menjadi seperti berikut.

Gambar 4. Dataset latih terjemahan



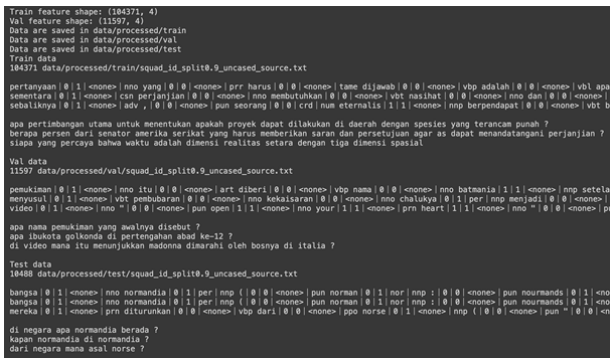
Gambar 5. dataset uji terjemahan

2. Menyiapkan Dataset

a. *Data Preparation*

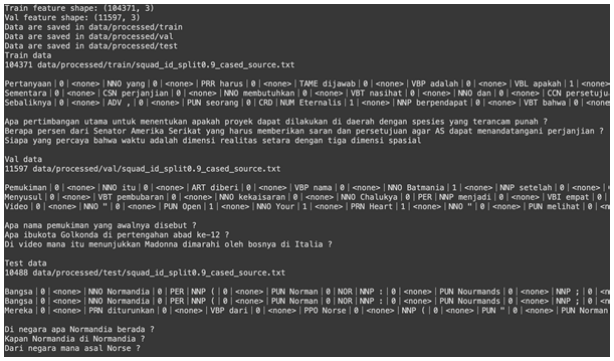
Dataset SQuAD2.0 yang telah diterjemahkan kedalam bahasa Indonesia selanjutnya dilakukan beberapa tahap pemrosesan yaitu : (1) Penghapusan data topik dan pertayaan yang tidak dapat terjawab atau tidak ditemukan jawabannya didalam paragaraf, (2) Melakukan proses *tokenization* dan *stemming* pada paragraf, pertanyaan, dan jawaban dilakukan untuk selanjutnya dilakukan penyiapan fitur linguistik seperti *POS-Taging*, dan *Named Entity Recognition.* (3) Menentukan posisi dari jawaban pertanyaan didalam paragraf, Sebelumnya indikator posisi jawaban dalam dataset SQuAD adalah berbasis karakter, pada penelitian kali ini kami merubah menjadi berbasis kata, (4) Pembuatan fitur linguistik untuk input konteks yang berupa lokasi jawaban (ans), indikator (case), *part-of-speech* (POS), *Named Entity* (NE), (5) Langkah selanjutnya adalah melakukan pembagian dataset yang akan digunakan sebagai dataset latih, validasi, dan uji.

Pembagian ini dilakukan pada setiap variasi dataset yaitu *uncased* dan *cased*. Setiap data latih, validasi dan uji terdapat dua jenis data yang disimpan dalam file yang berbeda, yaitu file *source* yang berisi data paragraf yang telah diberikan fitur linguistik dan dipisahkan dengan spasi dan file target yang berisi pertanyaan yang dijadikan target dari file *source*. Jumlah data yang tersisa untuk dataset variasi *uncased* adalah 104.371 data latih, 11.597 data validasi, dan 10.488 untuk data uji, detail dari data *uncased* dapat dilihat pada gambar berikut.



Gambar 6. Pembagian dataset varian *uncased*

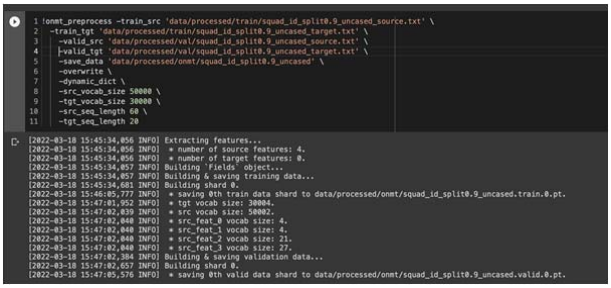
Kemudian untuk dataset dengan variasi *cased* tedapat 104.371 data latih, 11.597 data validasi, dan 10.488 data uji, detail data dapat dilihat pada gambar berikut.



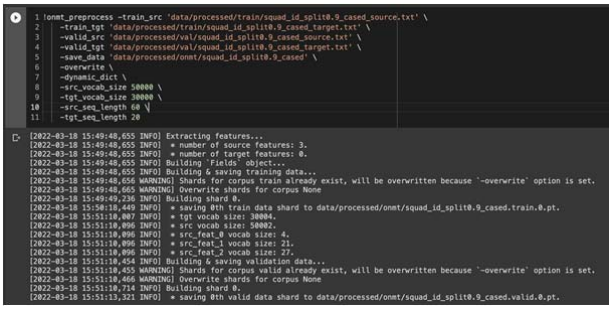
Gambar 7. Pembagian dataset varian *cased*

b. *Data Preprocessing*

Tahap *preprocessing* ditujukan untuk membuat data *vocab dictionary* yang diperlukan untuk membuat *embedding* dari *pretrained* model FastText Indonesia. Data yang digunakan pada tahap *preprocessing* adalah data latih dan data validasi dari setiap variasi dataset.



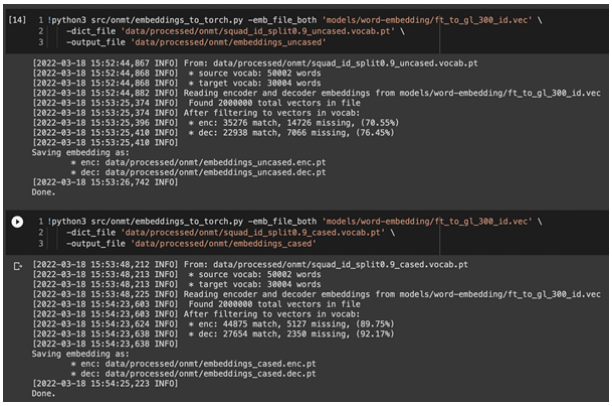
Gambar 8. *Preprocessing* varian *uncased*



Gambar 9. *Preprocessing* varian *cased*

c. *Embedding to torch*

Pada tahap ini dilakukan pembuatan *word embeeding* yaitu pembuatan representasi dari kata ke dalam bentuk vektor dari data latih dan data validasi dengan *pretrained* model FastText Indonesia yang diubah ke format GloVe dengan *dictionary vocab* dari data hasil *preprocessing* dataset



Gambar 10. Pembuatan *Word Embedding* setiap varian dataset

3. Membangun Model *Deep Learning*

Proses membangun model adalah proses latih model yang dilakukan didalam google colab dengan tipe *runtime* GPU dan menggunakan beberapa *toolkit, library*, dan *framework* seperti OpenNMT-py, dan PyTorch. Arsitektur RNN yang digunakan adalah menggunakan arsitektur dari (Du dkk., 2017) dengan menambahkan fitur linguistik pada konteks input. Berikut adalah hasil dari proses latih model dari setiap konfigurasi model dan data latih.

Nama Model	*Validation Accuracy*	*Step*	*Time (s)*
BiGRU *Uncased*	50.15	32100	6136
BiGRU *Cased*	49.63	32100	3105
BiLSTM *Uncased*	50.91	16050	1888
BiLSTM *Cased*	50.50	16050	1861

4. Evaluasi Model

Tahap evaluasi model dilakukan dengan metode *automatic measures* terhadap prototipe model yang telah dibuat menggunakan metrik BLEU, ROUGE, dan METEOR. Sebelum melakukan evaluasi setiap model, dilakukan inferensi yaitu pembuatan soal menggunakan setiap model yang telah dibangun dari dataset uji. Inferensi dilakukan dengan menggunakan data uji dari setiap variasi dataset. Berikut ini adalah beberapa sampel hasil inferensi yang telah dilakukan.

Keterangan	Kalimat
1	
Input	Teori kompleksitas komputasi adalah cabang dari teori komputasi

	dalam ilmu komputer teoretis yang berfokus pada pengelompokan masalah komputasi sesuai dengan kesulitan yang melekat, dan menghubungkan kelas-kelas tersebut satu sama lain.
Jawaban	Cabang dari teori komputasi dalam ilmu komputer teoretis yang berfokus pada pengelompokan masalah komputasi sesuai dengan kesulitan yang melekat, dan menghubungkan kelas-kelas tersebut satu sama lain
Target	Apa yang dimaksud dengan teori kompleksitas komputasi ?
BiGRU *Cased*	Apa yang dimaksud dengan teori kompleksitas komputasi ?
BiGRU *Uncased*	apa yang dimaksud dengan teori kompleksitas komputasi ?
BiLSTM *Cased*	Teori kompleksitas komputasi adalah cabang dari teori apa ?
BiLSTM *Uncased*	teori kompleksitas komputasi berfokus pada apa ?
2	
Input	Mesin uap adalah mesin pembakaran eksternal, di mana fluida kerja terpisah dari produk pembakaran.
Jawaban	Mesin pembakaran eksternal
Target	Apa itu mesin uap?
BiGRU *Cased*	Mesin uap apa yang digunakan untuk membuat motor uap ?
BiGRU *Uncased*	apa itu mesin uap ?
BiLSTM *Cased*	Apa itu mesin uap ?
BiLSTM *Uncased*	apa itu mesin uap ?
3	
Input	Siklus termodinamika ideal yang digunakan untuk menganalisis proses ini disebut siklus Rankine.
Jawaban	Siklus Rankine
Target	Apa nama siklus termodinamika dalam proses mesin uap?
BiGRU *Cased*	Apa nama lain untuk siklus termodinamika ideal ?
BiGRU *Uncased*	apa siklus termodinamika ideal yang digunakan untuk menganalisis proses ini ?
BiLSTM *Cased*	Siklus termodinamika ideal disebut siklus apa ?
BiLSTM *Uncased*	apa siklus termodinamika ideal yang digunakan untuk menganalisis proses ini ?
4	
Input	Oksigen adalah unsur kimia dengan simbol O dan nomor atom 8.
Jawaban	8
Target	Nomor atom oksigen adalah?
BiGRU *Cased*	Berapa angka atom ?
BiGRU *Uncased*	berapa nomor atom yang oksigen ?
BiLSTM *Cased*	Berapa nomor atom untuk Oksigen ?
BiLSTM *Uncased*	berapa nomor atom oksigen ?

Evaluasi dilakukan terhadap hasil inferensi dari setiap model pada setiap varian dataset dengan cara membandingkan hasil inferensi dengan data uji *target* menggunakan *tools* nlg-eval dari (Sharma dkk., 2017). Berikut adalah hasil evaluasi yang telah dilakukan menggunakan *untrained automatic metrics* dengan bantuan *tools* nlg-eval.

Nama Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
BiGRU *Uncased*	37.58	20.21	10.19	5.47	42.80	17.97
BiGRU *Cased*	33.92	17.47	8.64	4.72	39.63	17.54
BiLSTM *Uncased*	38.10	20.69	10.58	5.78	42.98	18.31
BiLSTM *Cased*	34.37	17.70	8.93	4.91	39.62	17.75

SIMPULAN DAN SARAN

Simpulan

Berdasarkan hasil penelitian dapat disimpulkan bahwa Pengembangan Model *Deep Learning* untuk Pembangkit Soal Otomatis Menggunakan *Reccurent Neural Network* selesai dilakukan. Pengembangan model menggunakan dataset dari SQuAD v2.0 yang diterjemahkan ke bahasa Indonesia. Pengembangan model *deep learning* menggunakan metode yang terdiri dari (1) Menerjemahkan dataset, (2) Menyiapkan dataset, (3) Membangun model *deep learning*, (4) Evaluasi model.

Evaluasi untuk mengetahui kinerja model telah dilakukan terhadap seluruh model yang dihasilkan. Terdapat empat model *deep learning* yang dihasilkan dari penelitian ini yaitu model BiGRU dan BiLSTM dengan kedua model menggunakan dua variasi dataset *cased* dan *uncased*. Proses evaluasi dilakukan dengan menggunakan *untrained automatic metrics* yang terdiri dari BLEU-1,2,3,4, METEOR, dan ROUGE-L mendapatkan hasil terbaik yaitu pada model BiLSTM *Uncased* dengan skor BLEU-1 38.10, BLEU-2 20.69, BLEU-3 10.58, BLEU-3 5.78, ROUGE-L 42.98, METEOR 18.31..

Model *deep learning* yang dihasilkan sudah dapat digunakan untuk membuat soal pertanyaan dari input materi yang diberikan walaupun masih memiliki keterbatasan dan dapat ditingkatkan kembali untuk penelitian selanjutnya.

Saran

Dari hasil uraian simpulan serta masih adanya keterbatasan produk dari penelitian, terdapat saran bagi penelitian mendatang adalah sebagai berikut:

1. Melakukan penggujian lebih lanjut dengan metode *Human Subjective Judgment* untuk mendapatkan penilaian dari sudut pandang pemahaman manusia.
2. Melakuakn proses latih model dengan mekanisme model lain seperti *Transformer*.
3. Melakukan perbaikan terhadap hasil terjemahan dataset latih dan uji.

DAFTAR PUSTAKA

Asarina, R. (2014). *STUDI EKSPLORASI KENDALA-KENDALA GURU DALAM PEMBELAJARAN IPS DI SMP WILAYAH KECAMATAN MOYUDAN*.

Bangor, A., Kortum, P. T., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies*, 114–223.

Du, X., Shao, J., & Cardie, C. (2017). *Learning to Ask: Neural Question Generation for Reading Comprehension* (arXiv:1705.00106). arXiv. http://arxiv.org/abs/1705.00106

Fattoh, I. E., Aboutabl, A. E., & Haggag, M. H. (2015). *Semantic Question Generation Using Artificial Immunity*. 9.

Garbade, D. M. J. (2018, Oktober 15). *A Simple Introduction to Natural Language Processing*. Medium. https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32

Habeeb, A. (2017). *Introduction to Artificial Intelligence*. https://doi.org/10.13140/RG.2.2.25350.88645/1

Lavie, A., & Agarwal, A. (2007). Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments. *Proceedings of the Second Workshop on Statistical Machine Translation - StatMT '07*, 228–231. https://doi.org/10.3115/1626355.1626389

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*, 74–81. https://aclanthology.org/W04-1013

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 311. https://doi.org/10.3115/1073083.1073135

Purwanto, P. (2009). *Evaluasi Hasil Belajar*. Pustaka Pelajar.

Sudaryono, D. (2015). *Metodologi Riset di Bidang TI: (Panduan Praktis, Teori dan Contoh Kasus)*. Penerbit Andi.