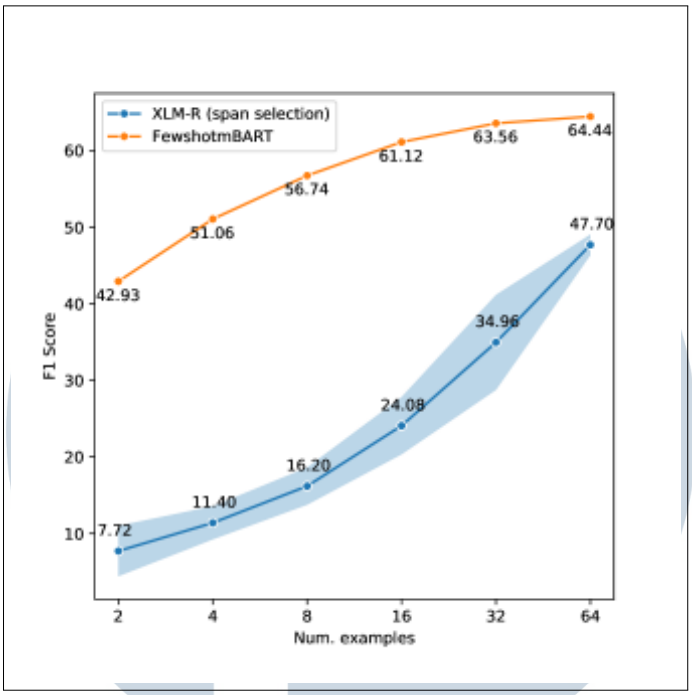
BAB 2
LANDASAN TEORI

Studi literatur seputar terkait dengan *fine tuning* model bahasa *pre-trained* berbasis arsitektur *transformer* yang memanfaatkan mekanisme *attention* telah menunjukkan hasil *state-of-the-art* terhadap berbagai NLP *tasks* seperti QA, *question generation*, parafrase, *text summarization*, analisis sentimen, *machine translation*, dan lainnya. Umumnya terdapat dua domain utama dalam NLP yang dikenal dengan NLG (*Natural Language Generation*) dan NLU (*Natural Language Understanding*). Tugas-tugas yang tergolong pada domain NLG meliputi proses generasi teks seperti translasi teks dari satu bahasa ke bahasa lain, parafrasi, *text summarization*, termasuk juga QA dan *question generation*. Sedangkan pada domain NLU dapat berupa tugas analisis sentimen, analisis leksikal, *spam filter*, dan berbagai *task* NLU lainnya.

Salah satunya pada studi terkait [11] yang merancang sebuah *framework* dalam *fine tuning* model bahasa MBART (*Multilingual* BART) [9] untuk *task* QA dan berhasil menghasilkan model yang mengungguli model bahasa XLM-Roberta [12] sebesar 41 skor F1 pada dataset TyDiQA [10] dengan konfigurasi jumlah sampel kecil seperti pada Gambar 2.1.



Gambar 2.1. Diagram perbandingan skor F1 XLM-Roberta dengan FewShotMBART
Sumber: [11]

Lalu juga terdapat studi terkait lainnya dengan pemanfaatan model bahasa BART untuk tugas *question-answering* dalam bahasa Inggris telah menunjukkan hasil *state-of-the-art* terbaru terhadap dataset ELI5 (*Explain Like I'm Five*) dengan 30.6 pada ROUGE-1, 6.2 pada ROUGE-2, serta 24.3 pada ROUGE-L secara abstraktif [13] yang dapat dilihat di Tabel 2.1. Permasalahan tugas *question-answering* dapat diatasi dengan berbagai model bahasa serta beragam pendekatan yang dapat dilakukan. Metode yang mendasari penelitian ini merupakan pendekatan *fine-tuning* yang berperan untuk mengerucutkan model bahasa NLG seperti BART dalam tugas *question-answering* terhadap bahasa yang bersifat *low resource* seperti bahasa Indonesia [14].

Tabel 2.1. Perbandingan model BART dengan penelitian lampau

	ELI5		
	R1	R2	RL
Best Extractive	23.5	3.1	17.5
Language Model	27.8	4.7	23.1

Seq2Seq	28.3	5.1	22.8
Seq2Seq Multitask	28.9	5.4	23.1
BART	**30.6**	**6.2**	**24.3**

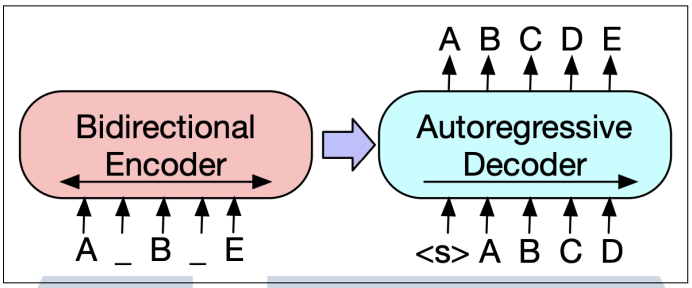
2.1 Transformer

Model bahasa berbasis arsitektur *transformer* dengan mekanisme *attention* memiliki dua komponen utama yang dikenal dengan *Self-Attention* dan *Feed Forward Neural Network* (FFNN) [4] atau *Multi Level Perceptron* pada studi GPT [6]. Komponen *Self-Attention* bertujuan untuk membaca seluruh *input sequence* sehingga dapat memahami konteks tekstual yang hasilnya akan diteruskan ke komponen FFNN. Sedangkan Komponen FFNN ini bertujuan untuk menentukan kata selanjutnya berdasarkan kata sebelumnya dengan memperhitungkan probabilitas kemunculan berdasarkan dataset yang di-*fit* terhadap model [15]. Secara umum terdapat 3 jenis model bahasa berbasis arsitektur *transformer*, yaitu *encoder-decoder* seperti BART [13], *encoder* seperti BERT [3], dan *decoder* seperti GPT [6].

Model bahasa GPT (*Generative Pre-Trained Transformer*) merupakan model bahasa *autoregressive* yang mampu dalam berbagai *task* generasi bahasa (NLG) seperti *named entity recognition* (NER), *part-of-speech tagging* (POS), *machine translation* (MT), dan berbagai *task* NLG lainnya. GPT bekerja dengan cara melakukan *embedding* terhadap *input* ke dalam matriks sehingga menghasilkan representasi vektor dari *input* bahasa aslinya dengan beberapa token. Kemudian *input* tersebut akan diteruskan ke 12 segmen *decoder* GPT [6].

Model bahasa BERT (*Bidirectional Encoder Representations from Transformers*) merupakan model bahasa *Natural Language Understanding* (NLU) yang melakukan *training* secara *bidirectional* (*left-to-right* dan *right-to-left*) agar dapat memiliki pemahaman yang lebih dalam terhadap suatu konteks bahasa dibandingkan dengan model bahasa satu arah. Teknik ini dikenal dengan *Masked Language Model* yang mengizinkan *Bidirectional Training* [3].

Model bahasa BART (*Bidirectional and Auto-Regressive Transformer*) yang merupakan model bahasa pada cabang *Natural Language Processing* (NLP) tepatnya pada NLG. Gambar 2.2 berikut merupakan arsitektur model bahasa BART yang terdiri dari *bidirectional encoder* (terdapat pada model bahasa BERT) dan disertai dengan *autoregressive decoder* (terdapat pada model bahasa GPT).



Gambar 2.2. Arsitektur BART secara keseluruhan

Sumber: [13]

Bidirectional encoder bekerja dengan cara menggantikan beberapa bagian dari *input* dengan *masked token* secara acak dan kemudian diikuti dengan prediksi token asli dari setiap *masked token* tersebut. Sehingga hal tersebut memberikan kemampuan dalam mendapatkan "pemahaman secara kontekstual" terhadap keseluruhan *input* ketika memprediksi token aslinya, karena model dapat "memahami" *input* sebelum dilakukan *masking* dan setelah dilakukan *masking* terhadap token tersebut. Kemudian *autoregressive decoder* bertujuan untuk memprediksi token selanjutnya berdasarkan dengan token yang ada (*input*) dan belum di *masked* serta *output* dari *bidirectional encoder*.

2.2 TyDiQA

Dataset *Typologically Diverse Question Answering* (TyDiQA) merupakan dataset *multilingual* yang meliputi sekitar 200 ribu pertanyaan dan disertai dengan jawabannya, yang terdiri dari 11 bahasa berbeda. Sebelas bahasa tersebut secara karakteristik dan *typological* berbeda yang diantaranya adalah bahasa Indonesia, Arab, Bengali, Finlandia, Jepang, Kiswahili, Korea, Rusia, Telugu, Thailand, dan Inggris. Pemanfaatan dataset dari bahasa yang bersifat *diverse* dengan tujuan untuk generalisasi terhadap berbagai bahasa di dunia. Seluruh pertanyaan yang terdapat pada dataset dituliskan oleh orang yang ingin tahu terhadap jawabannya, namun tidak tahu terhadap jawaban dari pertanyaan yang ditulisnya, dan data dikumpulkan secara langsung dari masing-masing bahasa tersebut, tanpa adanya keterlibatan translasi antarbahasa [10]. Struktur data beserta sampel datanya dapat dilihat pada Gambar 2.3 dan Gambar 2.4.

Gambar 2.3. Struktur data TyDiQA

Gambar 2.4. Sampel data TyDiQA

2.3 BLEU

Sistem evaluasi *Bi-Lingual Evaluation Understudy* (BLEU) adalah sistem evaluasi otomatis yang dirancang untuk *machine translation task*. BLEU dirancang sebagai bentuk sistem evaluasi yang tidak tergantung oleh bahasa tertentu dan dapat berkorelasi tinggi dengan evaluasi yang layaknya dilakukan oleh manusia. BLEU bekerja dengan cara membandingkan *candidate translation* (hasil translasi mesin yang akan dievaluasi) dengan *reference translation* yaitu berupa translasi yang sudah ada dan merupakan hasil translasi manusia. BLEU melakukan komputasi presisi terhadap fraksi setiap token dari hasil translasi kandidat yang muncul, atau yang tercakup oleh hasil translasi referensi secara n-gram. BLEU akan memberikan penalti terhadap kata pada hasil translasi kandidat yang tidak terdapat pada hasil translasi referensi serta terhadap kata pada hasil translasi kandidat yang muncul lebih sering dibandingkan dengan hasil translasi referensi. Metrik yang digunakan BLEU berada pada kisaran nol hingga satu [16].

2.4 ROUGE

Recall-Oriented Understudy for Gisting Evaluation atau yang umumnya dikenal dengan ROUGE adalah sistem evaluasi otomatis yang dapat dimanfaatkan terhadap berbagai *task summarization* [17]. ROUGE dapat secara otomatis menentukan kualitas suatu hasil kesimpulan dengan membandingkannya terhadap contoh ideal hasil kesimpulan karya manusia. Dalam perhitungan yang dilakukan ROUGE, terdapat beberapa pertimbangan yang memengaruhi perhitungan seperti n-gram, urutan kata, pasangan kata antara hasil kesimpulan dari komputer untuk dievaluasi dengan hasil kesimpulan buatan manusia. ROUGE sendiri terdiri dari 4 jenis, yaitu ROUGE-N, ROUGE-L, ROUGE-W, dan ROUGE-S yang setiap jenis-nya memiliki fokus masing-masing. ROUGE-N mengukur suatu kalimat teks menggunakan presisi berdasarkan nilai dari N. Sehingga apabila N = 1 (ROUGE-1) maka proses evaluasi / pengukuran berlangsung secara unigram di antara hasil mesin dengan hasil manusia, apabila N = 2 (ROUGE-2) maka *overlap* mengacu sebagai bigram terhadap hasil mesin dan manusia, serta berlaku untuk 3 (trigram) dan seterusnya. Sedangkan ROUGE-L mengukur berdasarkan susunan kata terpanjang menggunakan perhitungan *Longest Common Subsequence* (LCS). Kelebihan pemanfaatan LCS terlihat dari tidak diperlukan-nya hasil kecocokan secara *consecutive* melainkan secara *in-sequence* yang mencerminkan urutan kata pada tingkat

kalimat. Karena ROUGE-L secara otomatis termasuk n-gram *in-sequence* umum terpanjang sehingga tidak perlu spesifikasi dari panjang n-gram. Namun apabila terdapat beberapa kandidat kalimat sebagai komparasi yang salah satunya misal memiliki kecocokan secara *consecutive* dan satu lagi tidak *consecutive* tapi masih dalam *sequence* serta kedua kandidat memiliki jumlah kecocokan yang sama, maka skor ROUGE-L terhadap kedua kandidat akan sama. Hal tersebut dikarenakan ROUGE-L tidak mempertimbangkan aspek *consecutive* dalam suatu *sequence*. Sehingga ROUGE-W (weighted) mengatasi masalah tersebut dengan memperhitungkan *consecutive* sebagai bobot dari *sequence*, yang akan memberikan skor lebih besar terhadap kandidat consecutive pada skenario sebelumnya. Kemudian ROUGE-S (*Skip-Bigram Co-Occurrence Statistics*) yang mengukur secara bi-gram (dua kata) dan dilakukan lompatan maksimum sebanyak dua di antara kata pada setiap *sequence*.

2.5 HuggingFace

HuggingFace merupakan *library* yang menyediakan *package* dan berbagai *resource* NLP berupa dataset, model, *transformer*, *tokenizer*, serta *resource* NLP lainnya. Umumnya model-model tersebut berbasis arsitektur *transformer* seperti GPT [6], BERT [3], BART [13], T5 [18], dan beragam model yang menggunakan mekanisme *attention* lainnya. *Package transformers* yang disediakan oleh HuggingFace meliputi 30 *pre-trained model* dan 100 bahasa berbeda [19].

2.6 F1 *Score*

Metrik evaluasi yang memperhitungkan nilai rata-rata dari *precision* dan *recall*. Pada kasus *question answering* skor F1 dikomputasikan atas kata-kata individual dalam prediksi terhadap kata-kata yang ada di label. Jumlah kata yang terdapat antara prediksi dan label adalah dasar dari skor F1. *Precision* merupakan rasio jumlah kata yang terdapat di keduanya dengan jumlah total kata dalam prediksi, dan *recall* adalah rasio jumlah kata yang terdapat di keduanya dengan jumlah kata dalam *ground truth*, dengan *formula* sebagai berikut.

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.1}$$

2.7 *Exact Match* (EM)

Untuk setiap pasangan pertanyaan dan jawaban apabila seluruh karakter prediksi cocok dengan seluruh karakter label, maka nilai EM sama dengan satu, sebaliknya apabila tidak maka nilai EM sama dengan nol. Saat menilai contoh negatif, jika model memprediksi teks apa pun, secara otomatis menerima nol terhadap contoh itu. Kemudian nilai tersebut akan dirata-ratakan sehingga didapat skor EM sebagai performa model.

2.8 IndoNLG

Sebuah kumpulan sumber daya *Natural Language Generation* (NLG) bahasa Indonesia terhadap beberapa *downstream tasks* seperti *machine translation*, *question answering*, dan lain-nya. IndoNLG juga menyediakan *tokenizer* untuk tokenisasi bahasa Indonesia, sejumlah dataset dalam bahasa Indonesia seperti kumpulan berita Liputan6, *pre-trained language model* bahasa Indonesia seperti IndoBERT & IndoGPT, dan sumber daya NLP *benchmarking* lainnya.