Bab I Pendahuluan

I.1 Latar Belakang

~~Dalam memahami sebuah bahasa, diperlukan sebuah model kecerdasan. Manusia memahami bahasa dengan menggunakan indra untuk menangkap representasi bahasa dalam bentuk teks maupun suara yang kemudian diproses oleh otak manusia untuk memahami pengertian tersebut. Pada *natural language processing* (NLP), diperlukan representasi yang dapat dipahami oleh sebuah model NLP. Bentuk representasi~~ teks yang dapat dipahami oleh indra dan otak manusia tidak dapat dipahami secara langsung oleh model NLP sehingga diperlukan sebuah bentuk representasi lain. Ada beberapa teknik dalam membuat representasi dari sebuah teks, seperti *bag of words*, *n-gram*, *word embedding*, *semantic role labeling* (SRL), dan *abstract meaning representation* (AMR). Dari representasi-representasi tersebut, dapat dilakukan berbagai macam *task* seperti peringkasan teks, klasifikasi sentimen, mesin translasi, *question-answering*, deteksi parafrasa, dan lain-lain. Teknik-teknik NLP terbaik saat ini umumnya menggunakan *language model* berbasis *transformer* yang diperkenalkan oleh Vaswani dkk. (2017). Self-supervised pre-training yang menggunakan *language model* berbasis *transformer* seperti BERT (Devlin dkk. 2018), MASS (Song dkk. 2019), BART (Lewis dkk. 2019), dan PALM (Bi dkk. 2020) telah menjadi teknik yang sangat baik dalam pemahaman dan generasi bahasa natural.

belum jelas relevansinya

AMR merupakan salah satu representasi yang memperhatikan semantik pada tingkat kalimat dalam bentuk struktur data graf berarah yang mempunyai akar (Banarescu dkk. 2013). AMR awalnya didesain untuk merepresentasikan kalimat berbahasa Inggris sehingga bentuk graf AMR juga dituliskan dengan Bahasa Inggris. Sebuah teks yang mengandung banyak kalimat dapat direpresentasikan menjadi beberapa graf AMR, dengan setiap graf merepresentasikan setiap kalimat. Graf AMR dapat ditulis dalam beberapa macam bentuk format yang dapat dipahami manusia maupun mesin. Notasi PENMAN merupakan salah satu bentuk format penulisan graf AMR dalam bentuk teks (Matthiessen dkk. 1991). Dalam membuat sebuah AMR dari sebuah kalimat, perlu dilakukan sebuah proses yang dinamakan AMR

kurang relevan di sini

parsing. Telah dikembangkan beberapa teknik dalam melakukan AMR *parsing*, seperti *AMR Parsing as Sequence-to-Graph Transduction* (stog) (Zhang dkk. 2019), *translate-and-parse* (Uhrig dkk. 2021), *Maximum Bayes Smatch Ensemble Distillation for AMR Parsing* (MBSE) (Lee dkk. 2021), dan *Graph Pre-training for AMR Parsing and Generation* (AMRBART) (Bai dkk. 2022).

sebaiknya bahas pendekatannya, baca Cai, D., & Lam, W. (2020). Amr parsing via graph-sequence iterative inference. Pendekatan AMR parsing:
Two-stage parsing
One-stage parsing.
Three types one-stage parsing methods:
Transition-based parsing
Seq2seq-based parsing
graph-based parsing

Dalam mengukur kelayakan hasil graf AMR yang dihasilkan dari suatu teknik, digunakan metrik *Evaluation Metric for Semantic Feature Structures* (SMATCH) (Cai dan Knight 2013). SMATCH mengukur derajat *overlap* antara dua struktur fitur semantik graf AMR. Teknik-teknik AMR *parsing* tersebut diukur kelayakannya menggunakan dataset pasangan teks berbahasa Inggris dengan graf AMR-nya. Dataset yang pertama kali dikeluarkan adalah AMR 1.0 (LDC2014T12). Dataset tersebut kemudian dikembangkan lagi dengan tambahan kalimat dan aturan anotasi menjadi AMR 2.0 (LDC2017T10) dan AMR 3.0 (LDC2020T02). Jenis dataset AMR yang lebih spesifik ke domain tertentu juga dibuat, seperti *The Little Prince* (TLP) yang berasal dari Novel The Little Prince dan *Bio AMR* (Bio) yang berasal dari beberapa artikel PubMed tentang kanker. Teknik AMR *parsing* terbaik saat ini adalah teknik MBSE dengan kinerja SMATCH 86.7 pada dataset AMR 2.0 dan 85.4 pada dataset AMR 3.0 (Lee dkk. 2021). Teknik AMR *parsing* yang bukan termasuk teknik *ensemble* terbaik saat ini adalah teknik AMRBART dengan kinerja SMATCH 85.4 pada dataset AMR 2.0 dan 84.2 pada dataset AMR 3.0 (Lee dkk. 2021).

terlalu detail di sini

rumit bacanya

AMR *parsing* awalnya digunakan untuk mengubah dari kalimat berbahasa Inggris menjadi sebuah graf AMR berbahasa Inggris. Namun, untuk merepresentasikan kalimat dari bahasa selain Bahasa Inggris, ada beberapa teknik lain yang dapat membaca kalimat bahasa lain menjadi graf AMR berbahasa Inggris. *Cross-lingual* AMR (XL-AMR) oleh Blloshmi dkk. (2020) merupakan salah satu teknik yang dapat mengubah kalimat dari bahasa selain Bahasa Inggris menjadi AMR berbahasa Inggris. Dalam teknik tersebut, dibutuhkan dataset pasangan teks berbahasa selain Bahasa Inggris yang dituju dengan graf AMR-nya. Pada Bahasa Indonesia, beberapa teknik AMR *parsing* untuk teks berbahasa Indonesia juga dikembangkan. Beberapa AMR *parsing* Bahasa Indonesia yang telah dikembangkan adalah

sebaiknya mulai dari cara konvensional ID-ID, sebelum masuk ke XLAMR.

AMR *parsing* berbasis aturan (Severina dan Khodra 2019), berbasis *dependency parser* (Ilmy dan Khodra 2020), dan berbasis lintas bahasa (Putra ~~dan Khodra~~ 2022). Teknik AMR *parsing* oleh Putra ~~dan Khodra~~ (2022) menggunakan teknik XL-AMR dan stog untuk melakukan AMR *parsing* dari kalimat berbahasa Indonesia menjadi graf AMR berbahasa Inggris. Model terbaik dari teknik tersebut menghasilkan kinerja SMATCH 51.0 yang masih kalah dengan *baseline*-nya yang menggunakan teknik *translate-and-parse* dengan kinerja SMATCH 62.5. Teknik AMR *parsing* stog mempunyai kinerja SMATCH 77.0 untuk dataset AMR 2.0.

Apakah tidak ada analisis kesalahan pada buku Tesisnya ?

Penggunaan XLAMR dan kenapa penting dikembangkan belum dibahas

I.2 Masalah Penelitian

Teknik *cross-lingual* AMR *parsing* terbaru untuk Bahasa Indonesia adalah teknik oleh Putra ~~dan Khodra~~ (2022) yang menggunakan teknik stog oleh (Zhang dkk. 2019). Sudah banyak teknik-teknik lain yang memiliki kinerja AMR *parsing* yang lebih baik, seperti AMRBART. Dataset yang digunakan oleh Putra dan Khodra (2022) juga belum menggunakan dataset terbaru, yakni AMR 3.0, yang memiliki kalimat dan aturan anotasi lebih banyak dari AMR 2.0.

Teknik AMR *parsing* yang bukan termasuk teknik *ensemble* terbaik saat ini adalah teknik AMRBART yang berbasis *language model* BART oleh Lewis dkk. (2019). Terdapat *language model* bernama PALM oleh Bi dkk. (2020) yang kinerjanya lebih baik dibandingkan BART untuk *task question-answering*, peringkasan teks, *question generation*, dan *conversational response generation*. *Language model* PALM belum pernah digunakan untuk AMR *parsing* berbahasa Inggris maupun Indonesia dan berpotensi untuk menghasilkan kinerja yang lebih baik dibandingkan teknik yang digunakan sebelumnya.

dijelaskan dulu di latar belakang

Problem tesis perlu dieksplisitkan ????

I.3 Tujuan

Tujuan tesis ini adalah untuk menghasilkan sebuah model *cross-lingual* AMR *parsing* dari kalimat berbahasa Indonesia menjadi AMR berbahasa Inggris serta melakukan evaluasi kelayakan model tersebut dalam merepresentasikan makna kalimat berbahasa Indonesia.

I.4 Hipotesis

Premis 1: AMR berbahasa Inggris dapat digunakan sebagai representasi kalimat dalam bahasa lain (Damonte dan Cohen 2017). Hal ini ditunjukkan dengan analisis kuantitatif berdasarkan nilai SMATCH (Cai dan Knight 2013).

Premis 2: Fitur *multilingual word embedding* dan pelatihan pada dataset silver dapat meningkatkan kinerja pembangkit AMR cross-lingual (Blloshmi dkk. 2020).

Premis 3: Teknik AMRBART oleh Bai dkk. (2022) yang menggunakan *language model* BART merupakan teknik selain *ensemble* yang menghasilkan kinerja AMR *parsing* terbaik berdasarkan metrik SMATCH.

Premis 4: *Language model* PALM oleh Bi dkk. (2020) menghasilkan kinerja lebih baik dibandingkan BART untuk *task question-answering*, peringkasan teks, *question generation*, dan *conversational response generation*.

Premis 5: Dataset pasangan kalimat Bahasa Indonesia dan graf AMR berbahasa Inggris dapat dihasilkan dari dataset pasangan teks berbahasa Inggris dan graf AMR yang teks berbahasa Inggrisnya ditranslasi menjadi Bahasa Indonesia serta korpus paralel Indonesia-Inggris yang teks berbahasa Inggrisnya dibangkitkan menjadi graf AMR (Putra dan Khodra 2022).

Premis 2 digabung saja disini

Berdasarkan premis-premis tersebut, disusun hipotesis sebagai berikut:

Hipotesis: Model AMR *parsing cross-lingual* untuk Bahasa Indonesia dapat dibangun dengan menggunakan teknik pre-training AMRBART menggunakan *language model* PALM dengan dataset yang dibangun dari dataset pasangan teks berbahasa Inggris dan graf AMR serta korpus paralel Indonesia-Inggris.

tidak cukup hanya dapat dibangun

I.5 Batasan Masalah

Batasan masalah dari tesis ini adalah:

1. ~~Sumber dataset yang digunakan sebagai dataset AMR adalah AMR 3.0 yang teks berbahasa Inggrisnya ditranslasi menjadi Bahasa Indonesia dan korpus paralel PANL-BPPT dan IWSLT17 yang teks berbahasa Inggrisnya~~

1,2,3 bukan batasan

dibangkitkan menjadi graf AMR.

2. *Language model* yang digunakan adalah PALM (Bi dkk. 2020).
3. Teknik *pre training* menggunakan teknik yang diadopsi pada teknik AMRBART (Bai dkk. 2022).
4. Evaluasi kuantitatif menggunakan metrik SMATCH (Cai dan Knight 2013) dan kualitatif menggunakan divergensi translasi (Dorr dkk. 2002).

I.6 Metodologi

Berikut adalah tahapan atau metodologi yang dilakukan dalam proses penulisan tesis ini.

1. Analisis Masalah

   Tahap awal dilakukan analisis permasalahan dalam melakukan AMR *parsing* dalam Bahasa Indonesia. Hal ini dilakukan dengan mendalami materi mengenai teknik-teknik AMR *parsing* yang ada sebelumnya dan membandingkan hasil kinerjanya.

2. Perancangan Solusi dan Implementasi

   Pada tahap ini dilakukan perancangan dan implementasi berdasarkan hasil analisis teknik-teknik AMR *parsing* sebelumnya. Dirancang sebuah solusi dengan menggunakan beberapa gabungan teknik dan model tersebut. Lalu dilakukan implementasi dari rancangan tersebut dengan melakukan *training* model terhadap dataset yang tersedia.

3. Pengujian dan Evaluasi

   Hasil implementasi sebelumnya diuji dan dievaluasi secara kuantitatif dengan menggunakan metrik tertentu untuk dibandingkan hasilnya dengan teknik sebelumnya. Evaluasi secara kuantitatif tersebut digunakan untuk kesimpulan kelayakan model AMR *parsing cross-lingual* untuk kalimat berbahasa Indonesia.

Bab II Tinjauan Pustaka

II.1 *Abstract Meaning Representation* (AMR)

Abstract Meaning Representation (AMR) merupakan sebuah bahasa yang merepresentasikan semantik suatu kalimat (Banarescu dkk. 2013). AMR dibentuk dalam bentuk struktur data graf yang mempunyai akar, berlabel, berarah, tanpa *cycle*, dan mencakup keseluruhan kalimat. AMR bertujuan untuk mengabstraksikan kalimat dari representasi sintaktiknya, di mana kalimat-kalimat yang memiliki pengertian yang sama seharusnya memiliki AMR yang sama, walaupun tidak direpresentasikan dengan kalimat yang sama. AMR awalnya dibuat untuk kalimat berbahasa Inggris dan tidak digunakan sebagai bahasa secara umum.

Daftar Pustaka

Bai, Xuefeng, Yulong Chen dan Yue Zhang (2022). "Graph Pre-training for AMR Parsing and Generation". Pada: *arXiv preprint arXiv:2203.07836*.

Banarescu, Laura dkk. (2013). "Abstract meaning representation for sembanking". Pada: *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, hal. 178–186.

Bi, Bin dkk. (2020). "Palm: Pre-training an autoencoding&autoregressive language model for context-conditioned generation". Pada: *arXiv preprint arXiv:2004.07159*.

Blloshmi, Rexhina, Rocco Tripodi dan Roberto Navigli (2020). "XL-AMR: Enabling cross-lingual AMR parsing with transfer learning techniques". Pada: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, hal. 2487–2500.

Cai, Shu dan Kevin Knight (2013). "Smatch: an evaluation metric for semantic feature structures". Pada: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, hal. 748–752.

Damonte, Marco dan Shay B Cohen (2017). "Cross-lingual abstract meaning representation parsing". Pada: *arXiv preprint arXiv:1704.04539*.

cari versi publishednya

Devlin, Jacob dkk. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". Pada: *arXiv preprint arXiv:1810.04805*.

Dorr, Bonnie J dkk. (2002). *Improved word-level alignment: Injecting knowledge about MT divergences*. techreport. MARYLAND UNIV COLLEGE PARK INST FOR ADVANCED COMPUTER STUDIES.

Ilmy, Adylan Roaffa dan Masayu Leylia Khodra (2020). "Parsing Indonesian Sentence into Abstract Meaning Representation using Machine Learning Approach". Pada: *2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA)*. IEEE, hal. 1–6.

Lee, Young-Suk dkk. (2021). "Maximum Bayes Smatch Ensemble Distillation for AMR Parsing". Pada: *arXiv preprint arXiv:2112.07790*.

Lewis, Mike dkk. (2019). “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension”. Pada: *arXiv preprint arXiv:1910.13461*.

Matthiessen, C. M. dkk. (1991). “Text Generation and Systemic-Functional Linguistics: Experiences from English and Japanese”. Pada: *Computational Linguistics* 19.1.

Putra, Aditya Rachman ~~dan Masayu Leylia Khodra~~ (2022). “Pembangkitan Abstract Meaning Representation Lintas Bahasa dari Kalimat Berbahasa Indonesia”. Tesis Program Magister. Institut Teknologi Bandung.

kalau tesis tidak memasukkan nama pembimbing

Severina, Verena dan Masayu Leylia Khodra (2019). “Multidocument abstractive summarization using abstract meaning representation for Indonesian language”. Pada: *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*. IEEE, hal. 1–6.

Song, Kaitao dkk. (2019). “Mass: Masked sequence to sequence pre-training for language generation”. Pada: *arXiv preprint arXiv:1905.02450*.

Uhrig, Sarah dkk. (2021). “Translate, then parse! A strong baseline for cross-lingual AMR parsing”. Pada: *arXiv preprint arXiv:2106.04565*.

Vaswani, Ashish dkk. (2017). “Attention is all you need”. Pada: *Advances in neural information processing systems* 30.

Zhang, Sheng dkk. (2019). “AMR parsing as sequence-to-graph transduction”. Pada: *arXiv preprint arXiv:1905.08704*.