# ABSTRAK

Nama : Achmad Fatchuttamam Abka

Program Studi : Doktor Ilmu Komputer

Judul : Peringkasan Lintas Bahasa Berbasis Transformer Menggunakan *Multilingual Word Embeddings* untuk Domain Bahasa Inggris-Indonesia

Pembimbing : Prof. Dr. Eng. Wisnu Jatmiko, S.T., M.Kom.

Peringkasan lintas bahasa adalah sebuah proses menghasilkan ringkasan dalam bahasa target dari dokumen sumber berbahasa lain. Peringkasan lintas bahasa merupakan permasalahan yang sangat menantang karena melibatkan dua bahasa yang berbeda. Secara tradisional, peringkasan lintas bahasa dilakukan dalam skema *pipeline* yang melibatkan dua langkah, yaitu penerjemahan dan peringkasan. Pendekatan ini memiliki masalah, yaitu munculnya *error propagation.* Untuk mengatasi masalah tersebut, penelitian ini mengusulkan peringkasan lintas bahasa abstraktif *end-to-end* tanpa secara eksplisit menggunakan mesin penerjemah. Arsitektur peringkasan lintas bahasa yang diusulkan berbasis Transformer yang sudah terbukti memiliki performa baik dalam melakukan *text generation*. Model peringkasan lintas bahasa dilatih dengan 2-*task learning* yang merupakan gabungan peringkasan lintas bahasa dan peringkasan satu bahasa. Hal ini dilakukan dengan menambahkan *decoder* kedua pada Transformer untuk menangani peringkasan satu bahasa, sementara *decoder* pertama menangani peringkasan lintas bahasa. Lebih lanjut, arsitektur peringkasan lintas bahasa juga ditambahkan komponen *multilingual word embeddings* untuk lebih meningkatkan performa model. Kedua bahasa, bahasa Inggris dan bahasa Indonesia, direpresentasikan oleh *multilingual word embeddings* yang nilai *embedding*-nya sudah dipetakan ke dalam ruang vektor yang sama. *Multilingual word embeddings* membantu model dalam memetakan relasi antara *input* dan *output* yang menggunakan bahasa berbeda. Evaluasi model dilakukan dengan menggunakan metrik ROUGE. Metrik pengukuran ini membandingkan ringkasan yang dihasilkan oleh sistem dengan ringkasan referensi. Skor ROUGE memiliki rentang nilai dari 0 hingga 100 dengan semakin besar nilai menandakan performa yang semakin baik. Hasil eksperimen menunjukkan model usulan mendapatkan kenaikan performa hingga +32,11 ROUGE-1, +24,59 ROUGE-2, +30,97 ROUGE-L untuk peringkasan lintas bahasa dari dokumen sumber berbahasa Inggris ke ringkasan berbahasa Indonesia dan hingga +30,48 ROUGE-1, +27,32 ROUGE-2, +32,99 ROUGE-L untuk peringkasan lintas bahasa dari dokumen sumber berbahasa Indonesia ke ringkasan berbahasa Inggris.

Kata kunci:
Peringkasan lintas bahasa, peringkasan otomatis, Transformer, *multilingual word embeddings*

# ABSTRACT

Name : Achmad Fatchuttamam Abka

Study Program: Doctor of Computer Science

Title : Transformer-Based Cross-Lingual Summarization Using Multilingual Word Embeddings for English-Indonesian Domain

Counsellor : Prof. Dr. Eng. Wisnu Jatmiko, S.T., M.Kom.

Cross-lingual summarization (CLS) is a process of generating summaries in the target language from source documents in other languages. Cross-lingual summarization is a challenging task because it involves two different languages. Traditionally, cross-lingual summarization is done in a pipeline scheme that involves two steps, namely translation and summarization. This approach has a problem, it introduces error propagation. To overcome this problem, this study proposes end-to-end abstractive cross-lingual summarization without explicitly using machine translation. The proposed cross-lingual summarization architecture is based on Transformer which has been proven to have good performance in text generation. The cross-lingual summarization model is trained with 2-task learning, which is a combination of cross-lingual summarization and monolingual summarization. This is accomplished by adding a second decoder to handle monolingual summarization, while the first decoder handles cross-lingual summarization. Furthermore, multilingual word embeddings component is also added to the cross-lingual summarization architecture to further improve the performance of the model. Both languages, English and Bahasa Indonesia, are represented by multilingual word embeddings whose embedding values have been mapped into the same vector space. Multilingual word embeddings help the model map the relation between input and output in different languages. Model evaluation is carried out using the ROUGE metric. This measurement metric compares system generated summaries to reference summaries. The ROUGE score has a range of values from 0 to 100 with the greater the value indicating the better the performance. The experimental results show that the proposed model achieves performance improvements of up to +32.11 ROUGE-1, +24.59 ROUGE-2, +30.97 ROUGE-L for cross-lingual summarization from English source documents to Indonesian summaries and up to +30,48 ROUGE-1 , +27.32 ROUGE-2, +32.99 ROUGE-L for cross-lingual summarization from Indonesian source documents to English summaries.

Key words:
Cross-lingual summarization, automatic summarization, Transformer, multilingual word embeddings