

# USING CROSS-LINGUAL EMBEDDING FOR INDONESIAN-ENGLISH LANGUAGE PLAGIARISM

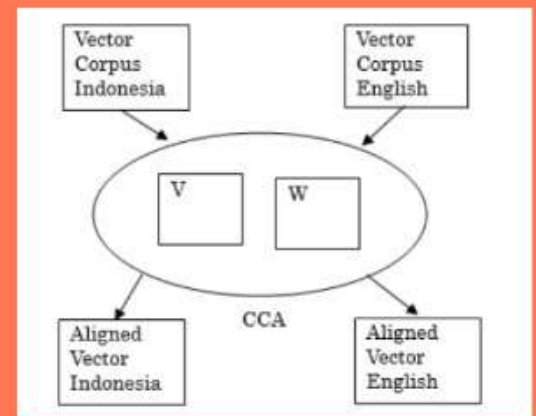
Yulyani Arifin,  
Sani M. Isa, Lili Ayu Wulandhari,  
Edi Abdurachman

## BACKGROUND

The increasing plagiarism by student or research due to the easier get information from internet even through same language or cross language. Also the accuracy from machine translation more increasing. There are many research about detecting plagiarism for mono language or cross language. But Mostly for European Language, It is still opportunity research for Indonesian Language. Usually for detecting cross language plagiarism need machine translation, or Parallel Corpora or Comparable Corpora. It is not many parallel corpora provided especially for Indonesian Language.

## PURPOSE

- 1 To develop Parallel Corpus for Indonesia English Language especially for Academic Research
- 2 To implement the cross lingual embedding to analyze the similarity between words in Indonesia and English and document level



## CONCLUSION

This research is still going on. With parallel corpora Indonesia English can be use in cross language detecting plagiarism especially in academic context. Using this CCA-Word2Vec can be seen in Word in English and Indonesia is near with the same context



**YULYANI ARIFIN**

DCS of Bina Nusantara University  
yulyaniarifin@binus.ac.id