

# 1. Pendahuluan

## Latar Belakang

Al-Quran merupakan kitab suci yang menjadi panutan dan rujukan seluruh umat Muslim di dunia. Dalam era modern sekarang ini, Al-Quran dapat dijumpai dalam berbagai macam media digital. Al-Quran terdiri dari 114 Surat, 6236 ayat, dan 77845 kata. Dikarenakan isi Al-Quran yang banyak, pembaca relatif susah untuk mempelajari Al-Quran secara menyeluruh dan memiliki kemungkinan untuk melewatkan informasi penting yang ada di Al-Quran. Salah satu upaya untuk mempermudah pembaca dalam mempelajari dan menyaring informasi dari Al-Quran adalah dengan melakukan pencarian kesamaan redaksional.

Proses pencarian kesamaan redaksi dimulai dengan melakukan identifikasi terhadap terjemahan ayat-ayat Al-Quran yang mempunyai kemiripan. Selanjutnya redaksi terjemahan ayat-ayat Al-Quran yang mempunyai kemiripan itu dipilah lagi untuk menentukan jenis kemiripan yang terkandung di dalamnya: apakah kemiripan lafal (redaksi) atau makna [1]. Untuk mencari kesamaan redaksi pada penelitian ini, telah dilakukan pemrosesan *Longest Common Substring* dan *Longest Common Subsequence*. Akan tetapi hasil dari pemrosesan tersebut belum sesuai dengan ekspektasi yang diharapkan, pemotongan kalimat masih tidak sempurna dan sering kali artinya tidak sesuai dengan kalimat sebenarnya. Sehingga dibutuhkan proses lebih lanjut yaitu melalui *shallow parsing*. Dengan *shallow parsing*, pemotongan kalimat dilakukan dalam bentuk himpunan frasa sehingga kalimat dipotong dengan tepat dan tidak mengubah arti sebenarnya.

*Shallow parsing* biasa disebut juga dengan *chunking*. Metode *chunking* memiliki dua pendekatan yaitu pendekatan *machine learning* dan pendekatan *rule-based* [2]. Dalam penelitian ini, penulis menggunakan pendekatan *rule-based chunking*. *Rule-based chunking* dapat menyaring suatu kalimat dengan mengelompokkan kata-kata yang dianggap penting menggunakan aturan-aturan berdasarkan aturan tata bahasa sehingga kalimat-kalimat tersebut dapat dimengerti dengan mudah.

Proses penyaringan pada penelitian ini dilakukan melalui beberapa tahap, yaitu membagi kata pada masing-masing kalimat, memprediksi *Part of Speech* (POS) dari kata-kata tersebut, mencari dan mengelompokkan kata sesuai dengan aturan *chunker*. Sistem memiliki input *data set* berupa terjemahan ayat-ayat Al-Quran yang sudah dicari *Longest Common Substring* [3] dan *Longest Common Subsequence* [4] oleh penelitian sebelumnya dengan output berupa himpunan frasa.

## Topik dan Batasannya

Berdasarkan latar belakang tersebut, topik yang berkaitan dengan penelitian ini diantaranya adalah:

### 1. Parafrasa

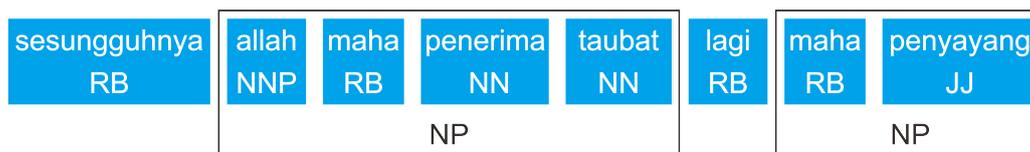
Himpunan parafrasa adalah hasil dari penelitian ini, hasil tersebut diperoleh dari penyaringan menggunakan Rule-Based Chunking.

### 2. Input dan Output Data

Input data yang akan di saring berupa kumpulan frasa yang telah diproses menggunakan metode *Longest Common Substring* dan *Longest Common Subsequence*. Data tersebut kemudian diproses dalam beberapa tahap yaitu membagi kata pada setiap kalimat atau *tokenization*, menandai setiap kata dengan aturan tata bahasa atau *part of speech (POS) tagging*, kemudian diproses kedalam *chunking*.

Sebagai contoh pada arti surat Al-Baqarah ayat 37 (2:37), *kemudian adam menerima beberapa kalimat dari tuhannya maka allah menerima taubatnya sesungguhnya allah maha penerima taubat lagi maha penyayang*. setelah dicari *Longest Common Substring* ditemukan frasa *sesungguhnya allah maha penerima taubat lagi maha penyayang* yang sama pada surat An-Nisa' ayat 16 dan 64.

Input terdapat pada kotak kecil berwarna biru yang berupa frasa hasil dari proses pencarian *Longest Common Substring*, yang kemudian setiap kata dibagi, dan ditandai dengan Tagset. Output adalah hasil dari proses *chunking* yang terdapat pada kotak besar yang diberi label NP; diilustrasikan pada Gambar 1.



Gambar 1. Ilustrasi frasa yang sudah ditandai.

## Tujuan

Penelitian ini bertujuan untuk membangun himpunan parafrasa menggunakan algoritma *rule-based chunking* sehingga kalimat dipersingkat dengan tepat dan tidak mengubah arti yang sebenarnya, serta mengetahui tingkat

*precision, recall* dari data yang telah didapatkan agar dapat dibandingkan dengan pendekatan *shallow parsing* yang lain. Sehingga nantinya dapat digunakan untuk mempermudah proses dalam mempelajari Al-Qur'an.