

Penyederhanaan kalimat dalam dokumen menggunakan Metode A noisy-channel

Markus Bernard¹, Ibnu Asror, S.T.,M.T.², Indra Lukmana Sardi, S.T., M.T.³

¹bernardmarkus@students.telkomuniversity.ac.id, ²iasror@telkomuniversity.ac.id,

³indraluk@telkomuniversity.ac.id

Abstrak

Informasi adalah sekumpulan data yang telah diolah dan mempunyai arti, salah satu informasi ada dalam sebuah dokumen. Dokumen tentang sains, teknologi, dan kedokteran terdapat kata asing diikuti dengan deskripsi dari kata-kata asing. Deskripsi dari kata-kata asing biasanya dituliskan dalam kalimat-kalimat yang panjang yang orang-orang awam dalam bidang sains, teknologi, dan kedokteran sulit untuk memahami kalimat tersebut. Penyederhanaan kalimat bertujuan untuk menyederhanakan leksikal, gramatikal, atau kompleksitas struktural teks sementara tetap mempertahankan nya makna semantiknya. Penyederhanaan kalimat pada bagian ini bertujuan untuk mempersingkat kalimat-kalimat mengenai sains, teknologi, dan kedokteran dalam bahasa inggris. Sentence Compression berfungsi untuk mengompres suatu kalimat untuk mendapatkan kalimat yang lebih pendek dari pada kalimat aslinya. Dari Penyederhanaan kalimat dapat memberikan hasil kalimat yang lebih pendek dengan tetap mempertahankan makna dari kalimat aslinya. Hasil Penelitian diperoleh sebuah kalimat yang lebih pendek dari kalimat aslinya, dengan pengurangan kata dalam sebuah kalimat adalah 15% dari kalimat aslinya. Dan hasil kalimat yang kurang baik untuk kalimat yang mempunyai jumlah kata yang sedikit (dibawah 9 kata).

Kata kunci : Informasi, Penyederhanaan Kalimat

Abstract

Information is a collection of data that has been processed and has meaning, one of the information is in a document. Documents about science, technology, and medicine have foreign words followed by descriptions of foreign words. Descriptions of foreign words are usually written in long sentences that lay people in the fields of science, technology, and medicine find it difficult to understand the sentence. The simplification of the sentence aims to simplify the lexical, grammatical, or structural complexity of the text while retaining its semantic meaning. The simplification of the sentence in this section aims to shorten the sentences about science, technology, and medicine in English. Sentence Compression serves to compress a sentence to get a sentence that is shorter than the original sentence. From simplification of sentences can give shorter sentences while retaining the meaning of the original sentence. Research results obtained a sentence that is shorter than the original sentence, with the reduction of words in a sentence is 15% of the original time. And the result is less good sentences for sentences that have a small number of words (under 9 words).

Keywords: Information, Simplification Sentences

1. Pendahuluan

Latar Belakang

Website adalah suatu halaman web yang saling berhubungan yang umumnya berisikan kumpulan informasi yang disediakan secara perorangan, kelompok, atau organisasi [1]. Sebuah situs web biasanya ditempatkan setidaknya pada sebuah server web yang dapat diakses melalui jaringan seperti Internet, ataupun jaringan wilayah lokal (LAN) melalui alamat Internet yang dikenali sebagai URL.

Dalam era teknologi ini, kehidupan manusia semakin bergantung pada teknologi, teknologi menjadi hal yang harus dipelajari oleh semua kalangan masyarakat, salah satu media pembelajaran yang dapat digunakan dalam mempelajari teknologi adalah melalui website ataupun melalui dokumen-dokumen, namun dalam bidang sains, teknologi, dan kedokteran terdapat banyak kata-kata asing yang sulit dipahami seperti: bandwidth, domain dll. Di dalam kata-kata asing terdapat penjelasan yang panjang sehingga menyulitkan dalam pengertiannya dan kurangnya fitur penyederhanaan kata kata artikel pada sebuah website mengakibatkan masyarakat yang awam akan teknologi tidak memahami maksud dari artikel tersebut [2].

Berdasarkan masalah tersebut maka diusulkan untuk menggunakan Text Simplification yang dapat menyederhanakan sebuah kalimat dalam suatu dokumen. Text Simplification mempunyai beberapa metode yaitu Lexical Simplification, Syntactic Simplification dan Sentence Compression. Lexical Simplification yang berfungsi untuk mengenali kata yang susah atau tidak sering digunakan dan menggantikannya dengan kata-kata yang lebih mudah dimengerti tanpa mengubah makna dalam kata-kata tersebut. Syntactic Simplification yang berfungsi untuk

penyederhanaan dalam suatu kalimat untuk mendapatkan kalimat yang lebih sederhana [3]. Pada pengerjaan ini dibangun dengan pendekatan. Sentence Compression bertujuan untuk mendapatkan rangkuman atau summary dari sebuah teks dengan menghapus, mengganti atau menambahkan kata pada kalimat asli untuk mendapatkan kalimat yang lebih sederhana, metode yang digunakan adalah A-noisy channel model untuk sentence compression memiliki mekanisme untuk mendapatkan kata yang lebih sederhana dari pada kalimat aslinya dengan cara mencari kata yang ditambahkan atau opsional pada suatu kalimat tersebut [2].

Topik dan Batasannya

Berdasarkan Latar Belakang yang telah didefinisikan di atas maka dapat dirumuskan perumusan masalah sebagai berikut :

1. Bagaimana penerapan A noisy-channel untuk menentukan kalimat pendek dalam sebuah kalimat?
2. Bagaimana penerapan A noisy-channel untuk memanipulasi struktur kalimat?

Sedangkan batasan masalah yang terdapat pada tugas akhir ini yaitu: sistem dibangun hanya untuk melakukan penyederhanaan teks dari sebuah artikel yang berkaitan dengan sains, teknologi, dan kedokteran yang menggunakan bahasa inggris.

Tujuan

Tujuan yang ingin dicapai dalam tugas akhir ini adalah merancang dan membangun sebuah sistem yang memiliki fitur untuk melakukan penyederhanaan kata kata pada sebuah kalimat, dengan menerapkan metode A noisy-channel.

Organisasi Tulisan

Organisasi tulisan yang terdapat pada buku ini adalah sebagai berikut:

- a. 1 Pendahuluan
Pada bab ini terdapat beberapa unsur yaitu: latar belakang masalah, perumusan masalah, batasan masalah, tujuan, metode penelitian masalah dan sistematika penulisan.
- b. 2 Studi Terkait
Bab ini menjelaskan teori-teori yang mendasari penyelesaian masalah yang diangkat dalam penelitian tugas
- c. 3 Sistem yang Dibangun
Pada bab ini menjelaskan tentang alur perancangan sistem yang dibangun, gambaran setiap proses yang ada dalam sistem, dan mendefinisikan setiap kebutuhan dari sistem.
- d. 4 Evaluasi
Bab ini menjelaskan tentang pengujian, skenario pengujian, dan analisis pengujian.
- e. 5 Kesimpulan dan Saran
Pada bab ini merupakan bab terakhir yaitu berisi kesimpulan dari pembahasan yang telah diuraikan dan saran guna perbaikan penelitian kedepannya.

2. Studi Terkait

2.1 Text Mining

Text Mining atau disebut mining data text, yang textnya dapat berasal dari dokumen maupun kutipan dalam website. Text Mining bertujuan untuk mencari pola dalam teks, yaitu proses penganalisisan teks guna menyarikan informasi yang bermanfaat untuk tujuan tertentu, Text mining juga dapat bertujuan untuk mencari topik utama dalam suatu dokumen, untuk merangkum suatu dokumen, dan juga dapat digunakan untuk pengorganisasian dokumen [4].

2.1.1 Text Pre-processing

Text Pre-processing merupakan tahap awal dalam mengolah text mining, yang merupakan proses pengubahan bentuk data yang belum terstruktur menjadi data yang terstruktur sesuai dengan kebutuhan yang nanti akan digunakan untuk proses text mining pada tahap selanjutnya [5], adapun proses Text Pre-processing yang akan digunakan adalah Tokenization/Tokenisasi.

a. Tokenization

Tokenization atau tokenisasi bertujuan untuk memisahkan atau memotong string masukan berdasarkan tiap kata penyusunnya, dan menghilangkan angka, tanda baca dan karakter selain huruf alfabet, pada tahap ini juga

akan di lakukan case folding dan cleaning untuk menjadikan semua huruf menjadi huruf kecil dan menghilangkan link, tag dan lainnya [5]. Hasil tokenization adalah berupa tokens. Tokens adalah kata-kata yang dipisahkan dari teks aslinya tanpa mempertimbangkan adanya duplikasi.

2.1.2 N-gram

N-gram merupakan salah satu proses yang digunakan dalam text mining. N-gram adalah potongan karakter N dari string yang lebih panjang. Meskipun dalam literatur, istilah tersebut dapat mencakup pengertian kumpulan karakter yang ada bersamaan dalam sebuah string [6].

a. Bigram

2-Gram atau dikenal juga Bi-gram (bigram) dengan melakukan penggerakan 1(satu) kata maju kedepan sehingga menghasilkan 2 kata [7]. contoh bigram dari kalimat "The number of threads per processor is not infinite" adalah :

EOS the
the number
number of
of threads
threads per
per processor
processor is
is not
not infinite
Infinite EOS

b. Skip-gram

Skip-gram merupakan teknik yang digunakan di area speech processing, dimana n-gram yang dibentuk kemudian ditambahkan juga dengan tindakan "skip" pada token-tokennya [8]. contoh skip-gram dari kalimat "The number of threads per processor is not infinite" adalah :

Contoh menggunakan skip-1-gram
number of threads per processor is not infinite
The of threads per processor is not infinite
The number threads per processor is not infinite
The number of per processor is not infinite
The number of threads processor is not infinite
The number of threads per is not infinite
The number of threads per processor not infinite
The number of threads per processor is infinite
The number of threads per processor is not

Contoh menggunakan skip-2-gram
of threads per processor is not infinite
The threads per processor is not infinite
The number per processor is not infinite
The number of processor is not infinite
The number of threads is not infinite
The number of threads per not infinite
The number of threads per processor infinite
The number of threads per processor is

2.1.3 Text Simplification

Text simplification (TS) atau Penyederhanaan text adalah teknik mengurangi text yang bermakna leksikal dan kompleksitas sintaksis teks. Penyederhanaan teks (TS) bertujuan untuk menyederhanakan leksikal, gramatikal, atau kompleksitas struktural teks sementara tetap mempertahankan makna semantiknya dengan cara menambahkan kata, menghapus kata, merubah kata atau memanipulasi struktur kalimat [9].

a. Lexical Simplification

Tujuan dari Lexical Simplification atau Penyederhanaan leksikal adalah untuk menggantikan kata-kata kompleks dan ekspresi dalam teks yang diberikan, dengan memberikan alternatif sederhana yang setara makna dalam konteks dalam suatu teks tersebut, dengan pengganti yang lebih sederhana dapat membuat teks lebih mudah dimengerti oleh masyarakat umum tanpa merubah konteks dalam suatu teks tersebut [10].

b. Syntactic Simplification

Syntactic Simplification atau penyederhanaan sintaksis adalah tugas Natural Language Processing, yang tujuannya adalah untuk menulis ulang kalimat untuk mengurangi kompleksitas tata bahasa mereka sambil menjaga makna dan informasi konten mereka [3].

c. Sentences Compression

Sentences Compression simplification merupakan salah satu metode simplification yang bertujuan untuk mendapatkan rangkuman atau summary dari sebuah teks. Sentences Compression simplification dapat menghapus, mengganti atau menambahkan kata pada kalimat asli untuk mendapatkan kalimat yang lebih sederhana [2].

2.2 Natural Language Processing

Sedangkan data mining bertujuan menemukan pola dalam data numerik dan kategorikal terstruktur, Text Mining dan Natural Language Processing (NLP) melakukannya dengan berfokus pada teks manusia yang ditulis, secara tidak terstruktur atau semi-terstruktur. NLP telah mengembangkan berbagai teknis yang secara tata bahasanya terinspirasi dari teks sintaksis yang diurai menggunakan informasi dari tata bahasa formal dan kamus, informasi yang dihasilkan kemudian ditafsirkan secara semantik dan digunakan itu ekstrak informasi tentang apa yang tertulis dalam teks tersebut [11].

a. A noisy-channel model for sentence compression

A noisy-channel model banyak digunakan dalam aplikasi Natural Language Processing (NLP), termasuk pengenalan suara, mesin terjemahan, partof-speech tagging, transliterasi, dan pengambilan informasi [2].

Dalam A noisy-channel model, kita melihat string panjang dan membayangkan bahwa itu awalnya string pendek, dan kemudian seseorang menambahkan beberapa tambahan teks opsional. Kompresi adalah masalah mengidentifikasi string pendek yang asli. Hal ini tidak penting apakah "original" string adalah asli atau hipotesis [2].

Dalam A noisy-channel model harus memecahkan tiga masalah [2]:

- Source model. Kita harus menetapkan setiap string s probabilitas dari $P(s)$, yang memberikan kesempatan s dihasilkan sebagai "original short string" dalam proses hipotesis di atas.
- Channel model. Kita menetapkan untuk setiap pasangan string $\langle s, t \rangle$ probabilitas dari $P(t | s)$, yang memberikan kesempatan bahwa ketika string pendek s diperluas, menghasilkan string panjang t .
- Decoder. Ketika kita mengamati string panjang t , kita mencari string pendek s yang memaksimalkan $P(s | t)$. Hal ini setara dengan mencari s yang memaksimalkan $P(s) \cdot P(t | s)$.

2.3 Context-free grammar

Context-free Grammar (CFG) adalah sebuah tata bahasa dimana tidak terdapat pembatasan pada hasil produksinya, Context-free Grammar (CFG) mempunyai 4-tuple $G = (N, \Sigma, R, S)$ dimana [12]:

- N adalah simbol non-terminal sebagai contoh S adalah singkatan untuk "sentence", NP singkatan untuk "noun phrase", VP singkatan untuk "verb phrase", dan seterusnya.
- Σ adalah simbol terminal yang berisikan kumpulan kata dalam kosakata.
- R adalah rules atau aturan tata bahasa seperti $S \rightarrow NP VP$ atau $NN \rightarrow man$. dimana aturan pertama menentukan bahwa sebuah S (kalimat) dapat terdiri dari sebuah NP yang diikuti oleh seorang VP dan aturan kedua menentukan bahwa NN (kata benda tunggal) dapat terdiri dari kata man.
- S adalah simbol yang mewakili simbol awal.

a. Probabilistic Context-Free Grammars (PCFGs)

Probabilistic Context-Free Grammars (PCFGs) didefinisikan sebagai berikut [12]:

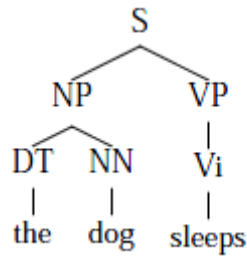
- Context-free Grammar $G = (N, \Sigma, R, S)$
- Parameter $q(\alpha \rightarrow \beta)$ untuk setiap aturan $\alpha \rightarrow \beta \in R$. Parameter $q(\alpha \rightarrow \beta)$ bisa diartikan adalah sebuah peluang bersyarat dalam memilih aturan $\alpha \rightarrow \beta$ dari perluasan simbol non-terminal α . Sebagai contoh $X \in N$ maka menghasilkan rumus sebagai berikut :

$$\sum_{\alpha \rightarrow \beta \in R: \alpha = X} q(\alpha \rightarrow \beta) = 1 \quad \dots\dots\dots(1)$$

Dimana $q(\alpha \rightarrow \beta) \geq 0$ untuk semua $\alpha \rightarrow \beta \in R$. Sehingga dalam parser-tree (t) yang berisi aturan $\alpha_1 \rightarrow \beta_1, \alpha_2 \rightarrow \beta_2, \dots, \alpha_n \rightarrow \beta_n$, menghasilkan peluang dari pcfg sebagai berikut :

$$p(t) = \prod_{i=1}^n q(\alpha_i \rightarrow \beta_i) \quad \dots\dots\dots(2)$$

Sebagai contoh pencarian peluang dalam sebuah parser-tree dibawah ini :



Gambar 1 Contoh Parser-tree

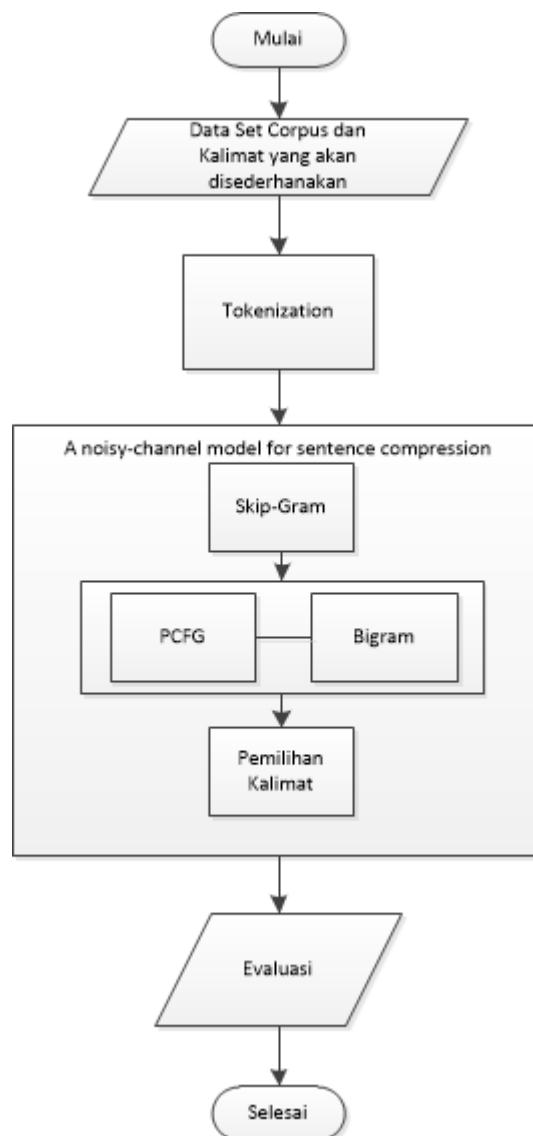
Dari hasil parser-tree pada gambar 1 diatas dihasilkan perhitungan P(t) dibawah ini :

$$p(t) = q(S \rightarrow NP VP) \times q(NP \rightarrow DT NN) \times q(VP \rightarrow Vi) \times q(DT \rightarrow the) \times q(NN \rightarrow dog) \times q(Vi \rightarrow sleeps)$$

3. Sistem yang Dibangun

3.1 Sistem Penyederhanaan Kalimat

Sistem penyederhanaan kalimat yang akan dibuat adalah seperti pada gambar 2 dibawah ini :



Gambar 2 Sistem Penyederhanaan Kalimat

3.1.1 Data Set

Data set dalam penyederhanaan kalimat berupa Corpus dan kalimat mengenai sains, teknologi, dan kedokteran yang menggunakan bahasa inggris yang akan disederhanakan. Sebagai contoh kalimat yang digunakan adalah “that uses repeated matrix multiplication to calculate all pairs shortest path”. Sumber corpus dan teks yang akan disederhanakan terdapat pada <http://elsevierlabs.github.io/OA-STM-Corpus/>.

3.1.2 Text Pre-Processing

Text Pre-Processing merupakan metode yang sangat penting pada text mining. Adapun langkah untuk melakukan text pre-processing yang akan digunakan adalah Tokenization.

a. Tokenization

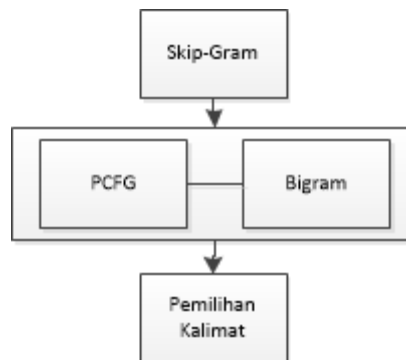
Hasil tokenization adalah berupa tokens. Dari kata “that uses repeated matrix multiplication to calculate all pairs shortest path” dihasilkan token pada Tabel 1 dibawah ini :

Tabel 1 Hasil Tokenization

Kalimat	Token
that uses repeated matrix multiplication to calculate all pairs shortest path	that
	uses
	repeated
	matrix
	multiplication
	to
	calculate
	all
	pairs
	shortest
	path

3.1.3 A noisy-channel model for Sentence Compression

A noisy-channel model for sentence compression bertujuan untuk menghasilkan sebuah string pendek dengan langkah pada Gambar 3 dibawah ini :



Gambar 3 A noisy-channel model for sentence compression

a. Skip-gram

Skip-gram digunakan untuk menghasilkan kalimat yang baru. Skip-gram dilakukan dari skip-1-gram atau pemotongan 1 kata hingga pemotongan N-1 (N adalah banyaknya kata dalam suatu kalimat), ini dilakukan untuk menghasilkan kalimat-kalimat yang baru. Beberapa peluang yang dihasilkan dari skip gram dari kalimat “that uses repeated matrix multiplication to calculate all pairs shortest path” adalah sebagai berikut :

Contoh menggunakan skip-1-gram

uses repeated matrix multiplication to calculate all pairs shortest path
 that repeated matrix multiplication to calculate all pairs shortest path
 that uses matrix multiplication to calculate all pairs shortest path
 that uses repeated multiplication to calculate all pairs shortest path
 that uses repeated matrix to calculate all pairs shortest path
 that uses repeated matrix multiplication calculate all pairs shortest path
 that uses repeated matrix multiplication to all pairs shortest path
 that uses repeated matrix multiplication to calculate pairs shortest path
 that uses repeated matrix multiplication to calculate all shortest path
 that uses repeated matrix multiplication to calculate all pairs path
 that uses repeated matrix multiplication to calculate all pairs shortest

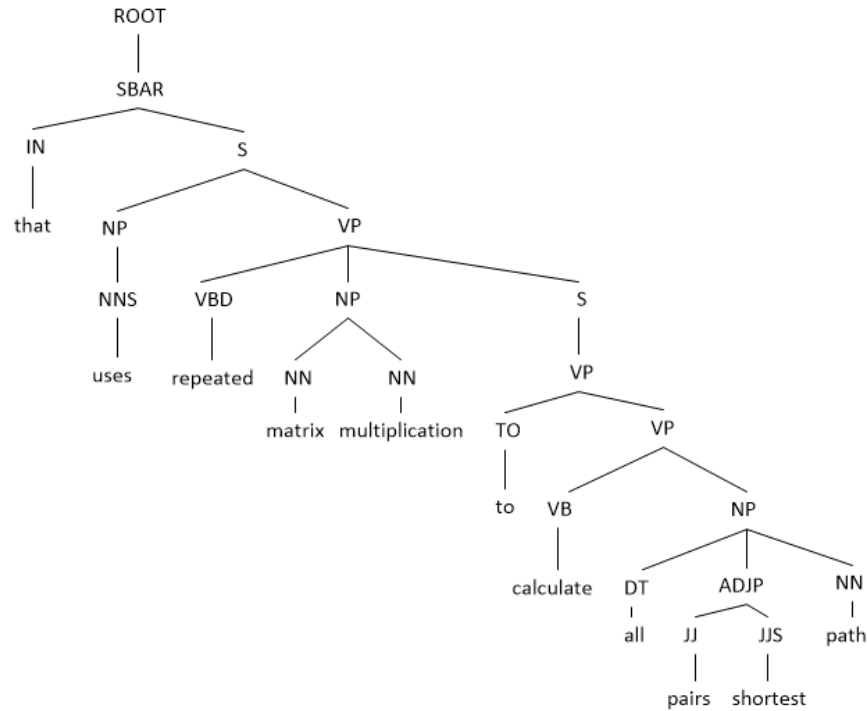
Contoh menggunakan skip-2-gram

repeated matrix multiplication to calculate all pairs shortest path
 that matrix multiplication to calculate all pairs shortest path
 that uses multiplication to calculate all pairs shortest path
 that uses repeated to calculate all pairs shortest path
 that uses repeated matrix calculate all pairs shortest path
 that uses repeated matrix multiplication all pairs shortest path
 that uses repeated matrix multiplication to pairs shortest path

that uses repeated matrix multiplication to calculate shortest path
 that uses repeated matrix multiplication to calculate all path
 that uses repeated matrix multiplication to calculate all pairs

b. Probabilistic Context-Free Grammars (PCFGs)

Sebuah teks inputan akan di jadikan sebuah parser-tree terlebih dahulu yang kemudian akan dihitung nilai PCFG-nya. Dari sebuah teks “that uses repeated matrix multiplication to calculate all pairs shortest path” akan menghasilkan parser-tree sebagai berikut :



Gambar 4 parser-tree that uses repeated matrix multiplication to calculate all pairs shortest path

Dari hasil parser-tree pada gambar 4 diatas dihasilkan perhitungan $P(t)$ dibawah ini :

$$\begin{aligned}
 P(s) = & P_{cfg}(ROOT \rightarrow S | ROOT) \times P_{cfg}(SBAR \rightarrow IN \ S | SBAR) \times P_{cfg}(S \rightarrow NP \ VP | S) \times P_{cfg}(NP \rightarrow NNS | NP) \times \\
 & P_{cfg}(VP \rightarrow VBD \ NP \ S | VP) \times P_{cfg}(NP \rightarrow NN \ NN | NP) \times P_{cfg}(S \rightarrow VP | S) \times P_{cfg}(VP \rightarrow TO \ VP | VP) \times \\
 & P_{cfg}(VP \rightarrow VB \ NP | VP) \times P_{cfg}(NP \rightarrow DT \ ADJP \ NN | NP) \times P_{cfg}(ADJP \rightarrow JJ \ JJS | ADJP) \times P_{cfg}(IN \rightarrow that | IN) \times \\
 & P_{cfg}(NNS \rightarrow uses | NNS) \times P_{cfg}(VBD \rightarrow repeated | VBD) \times P_{cfg}(NN \rightarrow matrix | NN) \times \\
 & P_{cfg}(NN \rightarrow multiplication | NN) \times P_{cfg}(TO \rightarrow to | TO) \times P_{cfg}(VB \rightarrow calculate | VB) \times P_{cfg}(DT \rightarrow all | DT) \times \\
 & P_{cfg}(JJ \rightarrow pairs | JJ) \times P_{cfg}(JJS \rightarrow shortest | JJS) \times P_{cfg}(NN \rightarrow path | NN)
 \end{aligned}$$

c. Bigram

Bigram berujuan untuk menghitung nilai dari dua buah kata, dari kata pertama dengan satu kata didepannya. Dari sebuah teks “that uses repeated matrix multiplication to calculate all pairs shortest path” akan menghasilkan peluang sebagai berikut :

$$\begin{aligned}
 P(s) = & P_{bigram}(that | EOS) \times P_{bigram}(uses | that) \times P_{bigram}(repeated | uses) \times P_{bigram}(matrix | repeated) \times \\
 & P_{bigram}(multiplication | matrix) \times P_{bigram}(to | multiplication) \times P_{bigram}(calculate | to) \times P_{bigram}(all | calculate) \\
 & \times P_{bigram}(pairs | all) \times P_{bigram}(shortest | pairs) \times P_{bigram}(path | shortest) \times P_{bigram}(EOS | path)
 \end{aligned}$$

d. Pemilihan Kalimat

Pemilihan kalimat dilakukan untuk memilih kalimat yang lebih pendek dari pada kalimat aslinya. Pemilihan kalimat yang akan dimabil adalah 60% hasil dari $P(s)$ kalimat aslinya, kebanyakan penyederhanaan kalimat menggunakan kompresi sebesar 60%, maka dalam sistem ini juga menggunakan 60% kompresi. Hasil penilaian kalimat $P(s)$ dari kalimat-kalimat pendek dihitung selisihnya yang mendekati dengan 60% dari nilai $P(s)$ kalimat aslinya. Untuk pemilihan kalimat yang lebih detail terdapat pada Lampiran 1 Pemilihan Kalimat.

3.1.4 Evaluasi

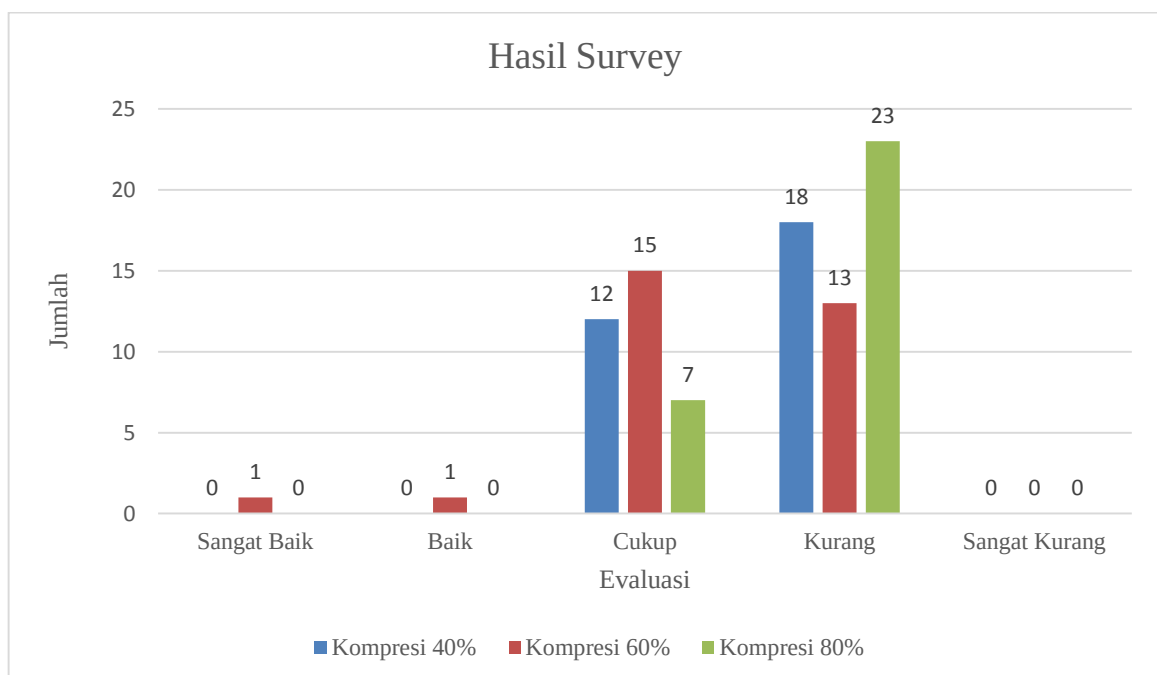
Evaluasi terhadap hasil yang didapat dari Sentences Compression simplification dengan evaluasi manual yang dilakukan, untuk mengetahui perbandingan kesederhanaan teks yang dihasilkan dan penggunaan kuesioner yang ditujukan kepada guru/dosen bahasa inggris untuk mengetahui apakah kata yang dipersingkat masih menyampaikan makna yang sama dari kalimat aslinya.

4. Evaluasi

Evaluasi yang dilakukan adalah dengan membandingkan kalimat penjelasan dengan kalimat sederhana dan penggunaan kuesioner.

4.1 Hasil Pengujian

Hasil kompresi dari 10 kalimat yang dilampirkan pada Lampiran 3 : Hasil Kompresi, dengan kompresi sebesar 40%, 60% dan 80%. Survey kuesioner yang ditujukan kepada guru/dosen bahasa inggris untuk mengetahui apakah kata yang dipersingkat masih menyampaikan makna yang sama dari kalimat aslinya.



Gambar 5 Diagram hasil survey

4.2 Analisis Hasil Pengujian

Hasil kompresi dari 10 kalimat yang dilampirkan pada Lampiran 3 : Hasil Kompresi. Jika diberikan SB = 5, B = 4, C = 3, K = 2, SK = 1 maka dihasilkan evaluasi kalimat pada Tabel 2 dibawah ini :

Tabel 2 Evaluasi Kalimat

Kata	Kalimat Penjelasan	Evaluasi 40%	Evaluasi 60%	Evaluasi 80%
dynamic programming algorithm	that uses repeated matrix multiplication to calculate all pairs shortest path	Cukup	Kurang	Kurang
Parallelizing PRAM model	parallelizing the matrix multiplication algorithm such that each element in the matrix is calculated in parallel	Cukup	Baik	Kurang
pluripotent cells	the unique characteristic to self renew in vitro while maintaining the capacity to differentiate into a broad number of cell types	Cukup	Kurang	Kurang
The Indus Shelf	provides a potential archive for tracking changes in sediment discharge from the river and relating	Kurang	Cukup	Kurang

	this to changing conditions in the onshore flood plain			
The shelf	dominated by the submarine Indus delta and a large sinuous canyon structure	Kurang	Cukup	Cukup
Smectite	product of chemical weathering and its presence is typically associated with alteration of volcanic minerals	Kurang	Kurang	Kurang
Tephrochronology	employed as a stratigraphic correlation tool in a broad range of disciplines	Cukup	Cukup	Kurang
self-assembly	distinctive method to construct chiral materials from chiral or achiral constituents	Kurang	Cukup	Cukup
Soil phosphorus	an essential element for plant growth but is often slowly available to plants within the soil environment	Kurang	Cukup	Kurang
Lophelia pertusa	the most widespread cold water coral species frequently found in the North Atlantic	Kurang	Kurang	Cukup

Hasil survey dari 10 kalimat pada gambar 5 jika diberikan SB = 5, B = 4, C = 3, K = 2, SK = 1 maka dihasilkan kompresi 40% menghasilkan hasil Kurang, kompresi 60% menghasilkan hasil Cukup dan kompresi 80% menghasilkan hasil Kurang. Untuk hasil kompresi menggunakan A-noisy channel beberapa menghasilkan nilai Cukup untuk kompresi sebesar 60%.

5. Kesimpulan

Kesimpulan yang didapat dari hasil survey untuk penyerdehanan kalimat dalam dokumen menggunakan metode A-noisy channel adalah sebagai berikut :

- Kalimat yang dihasilkan cukup dengan kompresi 60%
- Penyerdehanaan kalimat dapat merubah arti dalam suatu kalimat
- Sistem tidak dapat menentukan suatu kata yang merupakan kata penting dalam suatu kalimat

Saran untuk pengerjaan lebih lanjut penyerdehanan kalimat dalam dokumen menggunakan metode A-noisy channel adalah sebagai berikut .:

- Perlunya perubahan sistem dalam pencarian kalimat yang lebih pendek untuk mendapatkan hasil yang maksimal
- Perlunya sistem menentukan kata yang merupakan kata penting dalam suatu kalimat

Daftar Pustaka

- [1] "Wikipedia," [Online]. Available: https://id.wikipedia.org/wiki/Situs_web#cite_note-1. [Accessed 30 11 2016].
- [2] D. M. Kevin Knight, "Summarization beyond sentence extraction: A probabilistic approach to sentence compression," 2002.
- [3] A. Siddharthan, "Syntactic Simplification for Improving Content Selection in Multi-Document Summarization".
- [4] C. Z. Charu C. Aggarawal, Mining Text Data, London: Springer, 2012.
- [5] M. J. I. M. N. Dr. S. Vijayarani, "Preprocessing Techniques for Text Mining - An Overview".
- [6] W. B. Cavnar and J. M. Trenkle, "N-Gram-Based Text Categorization".
- [7] D. Jurafsky and J. H. Martin, "N-Grams," in *Speech and Language Processing*, 2014.
- [8] B. A. W. L. L. G. Y. W. David Guthrie, "A Closer Look at Skip-gram Modelling".
- [9] G. Attardi, A. Chanev and A. V. M. Barone, "A Dependency Based Statistical Translation Model," 2011.
- [10] J. H. B. A. Yea-Seul Kim, "SimpleScience : Lexical Simplification of Scientific Terminology".
- [11] C.-N. Ziegler, Mining for Strategic Competitive Intelligence, Berlin: Springer, 2012.
- [12] M. Collins, Probabilistic Context-Free Grammars (PCFGs), 2001.
- [13] E. Labs. [Online]. Available: <http://elsevierlabs.github.io/OA-STM-Corpus/>.

