IMPLEMENTASI WORD SENSE DISAMBIGUATION DENGAN METODE MAXIMAL MARGINAL RELEVANCE PADA PERINGKASAN TEKS

IMPLEMENTATION OF WORD SENSE DISAMBIGUATION USING MAXIMAL MARGINAL RELEVANCE FOR TEXT SUMMARIZATION

Bening Suryani Pratiwi[1], Shaufiah[2], Moch. Arif Bijaksana [3]

[1]Prodi S1 Teknik Informatika, Fakultas Informatika, Universitas Telkom
[2]Fakultas Informatika, Universitas Telkom
[3]Fakultas Informatika, Universitas Telkom

[1]beningsuryani@gmail.com, [2]shaufiah@gmail, [3]arifbijaksana@telkomuniversity

Abstrak

Dalam meringkas sebuah teks terdapat permasalahan yang muncul dan mempengaruhi hasil dari peringkasan teks tersebut. Permasalahan yang muncul seperti ambiguitas kata dan redundansi. Untuk meningkatkan kualitas dari peringkasan teks tersebut maka, permasalahan ambiguitas dan redundansi harus diatasi. Sehingga pada tugas akhir ini dilakukan peringkasan teks pada single dokumen yang mengimplementasi Word Sense Disambiguation dengan metode Maximal Marginal Relevance. Tahapan yang dilakukan terdiri dari Preprocessing, Word Sense Disambiguation, perhitungan Cosine Similarity, perhitungan Maximal Marginal Relevance, dan evaluasi. Pada tahapan preprocessing dilakukan cleaning pada data seperti stopwords removal, tokenization, remove tag, lemmatization dan stemming. Proses Word Sense Disambiguation dipilih untuk mengatasi masalah ambigu pada term dan diganti dengan synset term pada peringkasan teks tersebut. Pada peringkasan ini akan menggunakan cosine similarity untuk mengukur kemiripan setiap kalimat dengan kalimat pada keseluruhan isi dokumen. Sedangkan metode Maximal Marginal Relevance digunakan untuk merangking ulang hasil dari perhitungan cosine similarity dan memilih kalimat dengan nilai MMR paling tinggi yang akan dijadikan summary dengan nilai compresion rate yang ditentukan. Metode MMR termasuk metode yang sederhana namun efisien untuk mengurangi redundansi. Hasil peringkasan teks otomatis ini selanjutnya dievalusi dan dianalisis dengan pengukuran precision, recall, dan F-Measure dan dilihat dari hasil survey pembaca terhadap summary yang dihasilkan. Dengan nilai Recall 35%, Precision 21%, dan F-Measure 25%.

Kata kunci : Word Sense Disambiguation, Maximal Marginal Relevance, Cosine Similarity.

Abstract

In summarizing a text there are problems that arise and affect the result of summary text. The problems that arise as the word ambiguity and redundancy. To improve the quality of the summary text then, ambiguity and redundancy issues must be addressed. So in this final summary text be on a single document that implements Word Sense disambiguation methods Maximal Marginal Relevance. Steps being taken consists of preprocessing, Word Sense disambiguation, Cosine Similarity calculation, calculation Maximal Marginal Relevance, and evaluation. In the preprocessing stage of cleaning performed on the data as stopwords removal, tokenization, remove tags, lemmatization and stemming. Word Sense disambiguation processes have to overcome the problem of ambiguity in the term and will be replaced with the term synset the summary text. In this summary use cosine similarity to measure the similarity of each sentence with the phrase on the entire contents of the document. While the Maximal Marginal Relevance method used to rank the results of a calculation cosine similarity and choose sentence with the highest MMR values that will be used as a summary to the value specified compresion rate. MMR methods is simple method but efficient to reduce redundancy. This automatic summary text results evaluated then and analyzed by measuring the precision, recall, and F-Measure and views of the reader survey results summary is generated. With Recall 35%, Precision 21%, dan F-Measure 25%.

Keyword : Word Sense Disambiguation, Maximal Marginal Relevance, Cosine Similarity.

1. Pendahuluan

Perkembangan informasi saat ini menyebabkan pembaca sukar menemukan dan menyaring informasi yang sesuai dalam suatu teks. Suatu teks dengan ukuran yang panjang dapat menyulitkan pembaca dalam membaca dan menyerap semua informasi dari teks tersebut. Sedangkan informasi dari suatu teks merupakan komponen yang penting dari keseluruhan isi teks. Permasalahan itu timbul karena informasi yang didapat sulit untuk dipahami dan tidak mudahnya mendapatkan intisari dari suatu informasi. Peringkasan teks dapat digunakan untuk mempermudah menemukan informasi yang relevan tanpa menghilangkan informasi penting dari teks tersebut.

Peringkasan teks adalah proses mengurangi dokumen teks dengan komputer untuk menciptakan sebuah ringkasan yang mempertahankan poin penting dari dokumen asli[1]. Dalam peringkasan teks tersebut akan menimbulkan permasalahan seperti ambiguitas kata dan redundansi yang mempengaruhi kualitas dari hasil peringkasan teks tersebut. Ambiguitas kata merupakan permasalahan yang memiliki sejumlah makna kata yang berbeda. Sedangkan redundansi adalah permasalahan yang muncul pada sejumlah kalimat yang berulang atau ganda. Sehingga dibutuhkan metode untuk mengurangi tingkat redundansi dan ambiguitas dari peringkasan teks tersebut.

Word sense disambiguation adalah suatu proses mengidentifikasi makna kata yang digunakan dalam kalimat tertentu ketika kata memiliki sejumlah makna yang berbeda[2]. Masalah ambiguitas kata pada proses peringkasan teks inilah yang akan diatasi. sehingga dibutuhkan proses untuk mengurangi masalah ambiguitas dengan menerapkan word sense disambiguation pada peringkasan teks.

Metode yang digunakan untuk menghasilkan ringkasan teks dengan redundansi minimum yaitu, maximal marginal relevance. Metode ini menggunakan teknik ekstraksi yang digunakan untuk mengurangi redundansi kalimat dengan cara menghitung similarity antar kalimat dan kalimat dengan kalimat lain yang terpilih sebagai ringkasan[3].

Tugas akhir ini merupakan pengembangan dari tugas akhir yang telah ada dengan menggunakan metode maximal marginal relevance. Pada tugas akhir sebelumnya, metode MMR digunakan untuk dapat mengurangi redundansi dan akan dikembangan dengan menerapkan word sense disambiguation dengan algoritma lesk pada peringkasan teks untuk mengurangi masalah ambiguitas. Untuk mengetahui kualitas dari hasil peringkasan teks tersebut akan dievaluasi menggunakan pengukuran F-Measure serta dilihat dari hasil survey pembaca terhadap hasil peringkasan teks dengan memberikan nilai. Dan melakukan analisis dengan membandingkan hasil peringkasan teks dengan menggunakan WSD dan hasil peringkasan teks tanpa menggunakan WSD .

2. Perancangan Sistem

2.1. Gambaran Umum Sistem

Pada Tugas Akhir ini dibangun sebuah sistem yang mampu meringkas sebuah dokumen tunggal artikel berita bahasa inggris CNN Corpus dengan format STORY yang menggabungkan word sense disambiguation dengan metode maximal marginal relevance.

Input sistem berupa sebuah teks dokumen artikel berita tanpa judul yang akan diekstrak menjadi teks yang lebih singkat dari teks inputan dengan mengambil informasi yang paling relevan dengan teks asli dan memiliki redundansi minimum serta mengurangi ambiguitas kata. Untuk meminmalisir redundansi dilakukan dengan metode maximal marginal relevance dan untuk mengurangi ambiguitas kata dilakukan proses word sense disambiguation menggunakan cosine lesk.

Hasil peringkasan teks yang dihasilkan, disesuaikan dengan nilai compression rate dan nilai parameter λ yang telah ditentukan. Kemudian dilakukan evaluasi terhadap hasil ringkasan sistem dengan highlight atau gold standard CNN corpus dengan perhitungan recall, precision, dan F-Measure menggunakan ROUGE Evaluation Toolkit.

Pembangunan sistem dibagi menjadi tiga tahap yaitu : (1) Word Sense Disambiguation, (2) Preprocessing, (3) Summarization.

2.2. Rancangan Sistem

Rancangan sistem yang dibangun sesuai dengan gambaran umum diatas akan dijelaskan dari setiap tahapannya sebagai berikut :

1. Word Sense Disambiguation

Word sense disambiguation merupakan proses untuk mengidentifikasi makna kata yang digunakan dalam kalimat tertentu, ketika kata memiliki sejumlah makna yang berbeda. Pada proses WSD ini menggunakan cosine lesk. Tahapan WSD mengubah term yang memiliki makna lebih dengan term yang ada pada WordNet.

Tahapan WSD yaitu Pos Tagging, Word Frequency, Simple Word. Tahapan untuk melakukan Word Sense Disambiguation :

- Pos Tagging : merupakan proses untuk memberikan label kata secara otomatis pada suatu kalimat. Word Frequency : merupakan proses untuk mencari kata yang sering digunakan pada suatu teks sehingga kata yang sering digunakan tersebut akan sering digunakan kembali. Dan sebaliknya, apabila semakin jarang kata yang digunakan maka menyebabkan kata tersebut semakin jarang untuk digunakan kembali. Word frequency ini menggunakan Zipf untuk melihat frekuensi dari term yang akan diganti.
- Simple Word : merupakan proses untuk memilih kata untuk menggantikan kata yang ambigu dengan melihat frekuensi kata yang paling tinggi dan paling banyak digunakan.

2. Preprocessing

Preprocessing merupakan proses mengolah teks untuk dimasukkan ke dalam sistem sehingga siap untuk diolah ke proses selanjutnya dan memiliki kualitas atau struktur yang baik sehingga hasil dari preprocessing sistem berupa teks yang sudah bersih.

Tahapan preprocessing dilakukan setelah teks diinputkan. Tahapan selanjutnya setelah menginputkan teks yaitu dilakukan Remove Tag, Case Folding, Tokenization, Stopwords Removal, dan Stemming. Berikut adalah penjelasan untuk setiap tahapan yang dilakukan pada preprocessing :

- Remove Tag : merupakan proses untuk mencocokan string teks yang ingin dihasilkan dari ringkasan.
- Case Folding : merupakan proses untuk mengubah huruf kapital menjadi huruf kecil
- Tokenization : merupakan proses untuk mememcahkan teks menjadi sebuah token atau term.
- Stopwords Removal : merupakan proses untuk menghilangkan kata yang tidak memiliki makna atau kata yang kurang berarti dan sering muncul dalam kumpulan kata.
- Stemming : merupakan proses untuk mengubah kata jamak menjadi kata tunggal dan menjadi kata dasar. Library pada stemming yang digunakan yaitu Porter Stemmer.

3. Summarization

Tahap selanjutnya yang dilakukan setelah word sense disambiguation adalah tahap summarization. Summarization merupakan proses untuk menghasilkan suatu ringkasan dengan mengambil informasi penting dari suatu teks. Informasi yang akan diambil untuk menjadi suatu ringkasan dilakukan dengan perhitungan maximal marginal relevance.

Tahapan summarization yaitu TF-IDF, Cosine Similarity, Maximal Marginal Relevance. Tahapan untuk melakukan Summarization :

- TF-IDF : Merupakan proses untuk memberi bobot dari jumlah kemunculan term dalam sebuah dokumen dan jumlah kemunculan term dalam koleksi dokumen.
- Cosine Somilarity : merupakan proses untuk kesamaan dokumen dengan dokumen.
- Maximal Marginal Relevance : merupakan proses untuk mengkombinasikan matriks cosine similarity untuk merangking kalimat-kalima dengan mencari nilai maximum dari nilai similarity.

2.3. Skenario Pengujian

Skenario pengujian pada penelitian tugas akhir ini terdiri dari:

1. Pengujian nilai compression rate dan nilai parameter λ pada perhitungan MMR
   Pengujian ini dilakukan untuk mengetahui hasil ringkasan dengan word sense disambiguation dan ringkasan asli dengan melihat nilai compression rate terhadap nilai recall, precision, dan f-measure. Nilai compression rate yang dilakukan pengujian dengan kenaikan yang konstan sebesar 10%, 20%, 30%, 40, dan 50%. Serta pengaruh hasil ringkasan dengan nilai parameter λ yang berbeda pada perhitungan maximal marginal relevance terhadap nilai recall, precision, dan f-measure. Nilai parameter λ berkisar 0-1. Nilai parameter λ ini untuk melihat seberapa relevan nilai MMR terhadap dokumen asal. Untuk pengujian ini digunakan dengan nilai parameter λ 0,1, 0.3, 0.5, 0.7, dan 0.9.
2. Pengujian expert Judgement

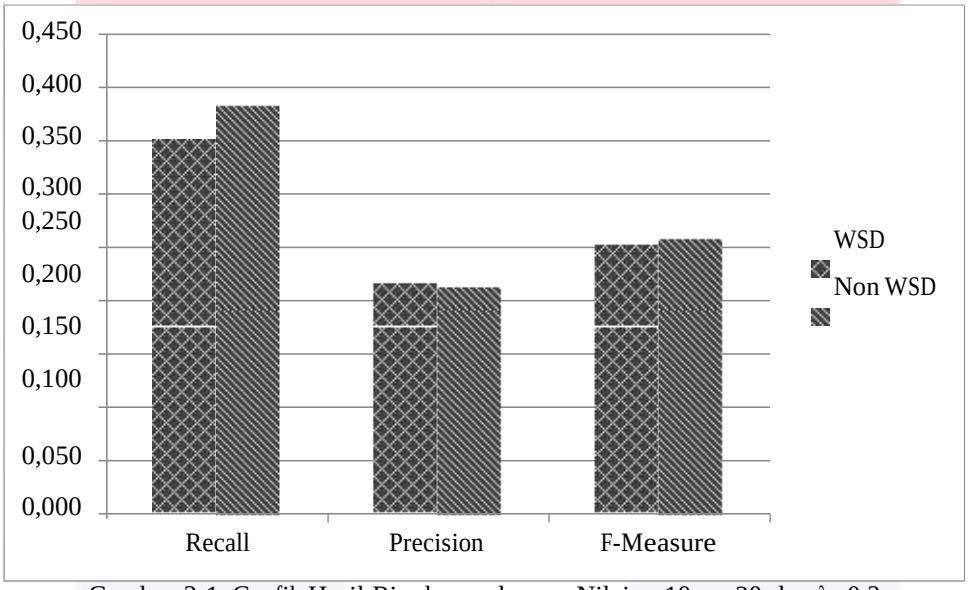
Pengujian ini dilakukan untuk melihat seberapa baik hasil ringkasan asli dan ringkasan dengan word sense disambiguation menurut ahli dengan membandingkan hasil ringkasan dengan teks asli. Penilaian diberikan dengan angka 1-5 untuk masing-masing hasil ringkasan.

2.4. Dataset

Dataset yang digunakan adalah CNN Corpus yang meliputi artikel berita dari seluruh dunia yang menyediakan 400 teks dengan kategori Afrika, Asia, bisnis, Eropa, Amerika Latin, Timur Tengah, AS, olahraga, teknologi, wisata, dan berita dunia. Corpus CNN ini adalah ringkasan berkualitas baik untuk setiap teks dengan "highlights" yang disediakan. Highlights ini terdiri dari beberapa kalimat yang penting untuk evaluasi. Highlights dapat diambil sebagai ringkasan dari referensi atau gold standard.

3. Pembahasan

3.1. Hasil Pengujian Nilai Compression Rate dan Nilai Parameter λ



Gambar 3-1. Grafik Hasil Ringkasan dengan Nilai n=10, n=20 dan λ=0.3

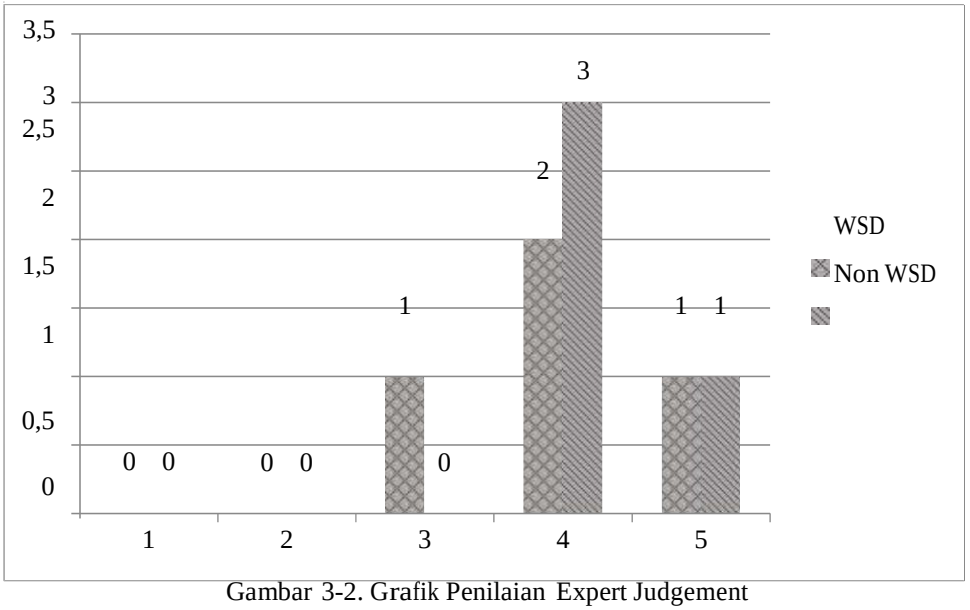
Berdasarkan hasil F-Measure paling baik dengan nilai compression rate 10%-20% serta nilai parameter λ = 0.3 untuk masing-masing hasil ringkasan. Maka didapatkan hasil ringkasan untuk nilai Recall, precision, dan F-Measure paling baik oleh unigram dengan F-Measure paling baik yaitu hasil ringkasan tanpa word sense disambiguation.

Untuk pengujian compression rate 10%-20% maka dapat menghasilkan ringkasan dengan nilai F-Measure yang paling baik untuk hasil ringkasan artikel berita bahasa inggris. Untuk pengujian nilai parameter λ=0.3 menghasilkan nilai F-Measure paling baik dari rata-rata keseluruhan. Hal ini karena semakin dekat nilai parameter λ dengan 0 maka nilai MMR yang diperoleh cenderung relevan terhadap kalimat yang relevan. Sedangkan apabila nilai parameter λ semakin mendakati 1 maka nilai MMR cenderung relevan dengan dokumen asal.

Nilai Recall yang dihasilkan berbanding terbalik dengan nilai Precision. Semakin besar n maka nilai Recall semakin besar. Sehingga apabila nilai Recall besar maka menyebabkan semakin kecil nilai Precision.

Untuk pengujian N-gram yang dihasilkan menunjukkan hasil yang paling baik adalah pengujian dengan unigram. Hal ini disebabkan karena terdapat banyak probability yang terjadi untuk kemunculan sebuah term dibandingkan dengan probability yang terjadi untuk kemunculan 2 term.

3.2. Hasil Expert Judgment



Gambar 3-2. Grafik Penilaian Expert Judgement

Berdasarkan Gambar 3-2 dapat disimpulkan bahwa hasil penilaiannya menunjukkan hasil ringkasan tanpa word sense disambiguation lebih baik dibandingkan hasil ringkasan dengan word sense disambiguation. Dengan nilai tertinggi untuk ringkasan asli 4 yang diberikan oleh 3 expert judgement dan nilai 4 sebanyak 2 expert judgement. Hasil ringkasan yang ditentukan untuk diberikan penilaian oleh expert judgement diperoleh dari hasil pengujian 1 pada Gambar 3-1.

4. Kesimpulan

Berdasarkan pengujian dan analisis pada bab 4 maka dapat diambil kesimpulan sebagai berikut:

1. Penggunaan Metode MMR dengan WSD sudah cukup baik untuk meminimalisirkan redundansi dan membantu mengurangi ambiguitas kata. Namun hasil ringkasan tidak sebaik hasil ringkasan tanpa WSD. Ini disebabkan karena setiap term yang diganti menyesuaikan dengan frekuensi yang diambil paling tinggi pada wordNet sehingga term yang diambil belum tentu sesuai dengan hasil ringkasan yang dihasilkan.
2. Hasil evaluasi didapat untuk hasil ringkasan artikel CNN Corpus bahasa inggris mencapai nilai paling baik pada nilai compression rate 10%-20% dan nilai parameter $\lambda = 0.3$ dengan nilai Recall 0.383, Precision 0.212, dan F-Measure 0.257 untuk peringkasan tanpa word sense disambiguation.
3. Dari hasil penilaian Expert Judgement menunjukkan bahwa hasil ringkasan tanpa WSD lebih baik dibandingkan dengan ringkasan WSD.

5. Saran

Saran yang diperlukan dari penelitian tugas akhir ini untuk pengembangan penelitian selanjutnya adalah sebagai berikut:

1. Mampu menangani Word Sense Disambiguation dengan term frequency
2. Menggunakan algoritma lesk yang lain.
3. Menggunakan metode maximal marginal relevance dengan tambahan. parameter penalty dan dikombinasikan dengan word sense disambiguation.

Daftar Pustaka

[1] Mustaqhfiri, M., Abidin, Z., & Kusumawati, R. (2011). Peringkasan Teks Otomatis Berita Berbahasa Indonesia Menggunakan Metode Maximal Marginal Relevance. 134-147.

[2] Redityamurti, Yusza; Laksitowening, Kusuma Ayu; Firdaus, Yanuar;. (2011). Implementation and Analysis of Word Sense Disambiguation using Lesk Algorithm in Lexical Chain Method for Text Summarization.

[3] Afifah, L. (2015). Peringkasan Dokumen Bahasa Indonesia Menggunakan Metode Maximum Marginal Relevance. 1-34.

[4] Golstein, J., & Carbonell, J. (1998). Summarization: Using MMR for Diversity Based-Rerangking and Evaluating Summaries. Language Technologies Istitute Carnegie Mellon University.

[5] Xie, S. (2010). Automatic Extractive Summarization Meeting Corpus. Dallas: The University of Texas at Dallas.

[6] Wicaksono, A., & Purwarianti, A. (2010). HMM Based POS Tagger for Bahasa Indonesia. Dalam Proceeding of 4th International MALINDO (Malay-Indonesian Language).

[7] Indriani, A. (2014). Maximum Marginal Relevance Untuk Peringkasan Teks Otomatis Sinopsis Buku Berbahasa Indonesia.

[8] Princeton University. (2005). WordNet A Lexical Database for English. Retrieved January 11, 2017, from http://wordnet.princeton.edu/

[9] Heuven, W. J. (t.thn.). A New and Improved Word Frequency Database for British English.

[10] Moral, Cristian;. (2014). A Survey of Stemming ALgorithms in Information Retrieval. Spain.