ANALISIS DAN IMPLEMENTASI WORD SENSE DISAMBIGUATION MENGGUNAKAN ALGORITMA LESK DALAM METODE LEXICAL CHAIN PADA PERINGKASAN TEKS
IMPLEMENTATION AND ANALYSIS OF WORD SENSE DISAMBIGUATION USING LESK ALGORITHM IN LEXICAL CHAIN METHOD FOR TEXT SUMMARIZATI

Yusza Redityamurti[1], Kusuma Ayu Laksitowening[2], Yanuar Firdaus A.w.[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

---

Abstrak

Text Summarization adalah sebuah proses untuk menghasilkan summary atau ringkasan dari suatu artikel, tetapi tetap memiliki gambaran yang akurat dari isi suatu artikel. Text Summarization seringkali menghasilkan hasil ringkasan yang redundan karena adanya kata-kata bermakna ganda (ambigu). Word Sense Disambiguation adalah suatu proses untuk mengindentifikasi makna kata yang digunakan dalam kalimat tertentu, ketika kata memiliki sejumlah makna yang berbeda.

Sistem ini mengimplementasikan metode lexical chain with word sense disambiguation, yang merupakan pengembangan dari metode lexical chain. Lexical chain merupakan salah satu metode text summarization yang membentuk rantai leksikal berdasarkan hubungan semantik antar kata dalam teks. Metode lexical chain masih mempunyai kekurangan yaitu rantai leksikal yang terbentuk akan tidak akurat jika ada kata yang berambigu atau kata yang memiliki makna ganda. Oleh sebab itu maka metode lexical chain digabungkan dengan metode word sense disambiguation menggunakan lesk algorithm dengan knowledge source yaitu wordnet 3.0 untuk membantu menghilangkan ambiguitas kata dalam pembentukan rantai leksikal.

Hasil pengujian yang telah dilakukan pada Tugas Akhir ini menunjukkan bahwa, dengan menerapkan word sense disambiguation dalam metode lexical chain hasil ringkasan yang dihasilkan mengalami peningkatan performansi dibanding dengan metode original lexical chain. Penerapan metode word sense disambiguation dengan menggunakan algoritma lesk dapat membantu metode lexical chain dalam membentuk rantai leksikal dengan cara menghilangkan ambiguitas setiap candidate term yang akan membentuk rantai leksikal. Sehingga dapat meminimalisasi tingkat redundansi dalam pemilihan kalimat yang akan dijadikan hasil ringkasan.

Kata Kunci : Text Summarization, Lexical Chain, Word Sense Disambiguation, Lesk Algorithm, Wordnet.

Abstract

Text Summarization is a process to generate a summary from articles but it has accurate main points from the content of the articles. Text Summarization often produce summary with a redundant sentence because of the ambiguous words (ambiguity). Word Sense disambiguation is a process to identifying the meaning of words used in a particular sentence, when the word has several different meanings.

This system implements lexical chain with word sense disambiguation method, is an development from lexical chain method. Lexical chain is one kind of text summarization method, this method make form of lexical chains based on semantic relationships between words in the text. Lexical chain method still has a trouble, that is form of lexical chain can be inaccurate if there ambiguity word or words that have double meanings. Therefore, the lexical chain method combined with word sense disambiguation method using lesk algorithm with knowledge source that is WordNet 3.0 can helping to eliminate the ambiguity of words in lexical chain form.

The test results that has been done in this thesis shows that by applying the word sense disambiguation in lexical chain method can increased performance summary results if compared with the original lexical chain method. Use of the word sense disambiguation method using the lesk algorithm can help lexical chain method create lexical chains by removing the ambiguity of each candidate term. So can minimize the level of redundancy in the selection of a sentence that would be the result of summary.

Keywords : Text Summarization, Lexical Chain, Word Sense Disambiguation, Lesk Algorithm, Wordnet.

1. PENDAHULUAN

1.1 Latar Belakang Masalah

Era padat informasi dan teknologi saat ini, menciptakan generasi instan yang "menuntut" semua bergerak dengan cepat. Begitu juga dengan para pembaca yang menuntut semua informasi bisa didapatkan dengan cepat dan juga tidak harus membuang banyak waktu. Namun sering kali ada pembaca yang tidak ingin membaca keseluruhan isi artikel hanya untuk mengetahui isi garis besar artikel tersebut. Sehingga pembaca membutuhkan *summary* atau ringkasan dari artikel tersebut. Berdasarkan kebutuhan itu teknologi informasi menawarkan suatu solusi, sehingga memungkinkan pembaca dapat membuat ringkasan dari suatu artikel. Salah satu perkembangan dalam bidang teknologi informasi tersebut yaitu teknik *Text Summarization* atau peringkasan teks.

Text Summarization adalah sebuah proses untuk menghasilkan *summary* atau ringkasan dari suatu artikel, tetapi tetap memiliki gambaran yang akurat dari isi suatu artikel dan memiliki tidak lebih dari setengah artikel aslinya atau maksimal 50% [11]. Tujuannya adalah mengambil sumber informasi dengan mengutip sebagian besar isi yang penting dan menampilkan kepada pembaca dalam bentuk yang ringkas dan sesuai dengan kebutuhan pembaca. Teknik yang umum digunakan dalam peringkasan teks adalah mengambil kalimat yang paling penting dari sebuah dokumen (*extractive summary*) [11]. Dengan demikian teknologi ini diharapkan dapat membantu pembaca untuk menyerap informasi yang ada dalam artikel lewat *summary* karena akan menghasilkan suatu produk teks yang tetap memiliki atau mengandung bagian-bagian yang penting dari artikel aslinya.

Setiap bahasa alami memiliki kata yang dapat bermakna lebih dari satu, sesuai dengan konteks kalimat yang menyertainya. Kata bermakna lebih dari satu tersebut, dapat berpotensi menyebabkan keragu-raguan atau ambigu. Ambiguitas kata dapat menyebabkan suatu masalah dalam proses peringkasan teks. Karena dengan adanya ambiguitas akan berpotensi menyebabkan kalimat yang redundan sehingga mengganggu proses ekstraksi kalimat. Usaha untuk memilih makna dari kata tersebut berdasarkan konteks kalimat disebut *word sense disambiguation*.

Word sense disambiguation adalah suatu proses mengindentifikasi makna kata yang digunakan dalam kalimat tertentu ketika kata memiliki sejumlah makna yang berbeda [5]. Ada beberapa pendekatan untuk menghilangkan ambiguitas makna kata dalam kalimat berbahasa Inggris, yaitu *supervised learning* dan *unsupervised* [12]. Dalam tugas akhir ini, *word sense disambiguation* yang digunakan adalah pendekatan *unsupervised* menggunakan sumber informasi lain sebagai pengganti penandaan terhadap kata yang bermakna ambigu, yaitu *wordnet*. Pendekatan ini diterapkan dengan menggunakan algoritma *Lesk*. Algoritma *Lesk* bekerja berdasarkan intuisi bahwa kata yang bermakna ambigu

yang terdapat bersamaan dalam kalimat, digunakan untuk merujuk topik yang sama dan makna yang berhubungan dengan topik tersebut didefinisikan di dalam *wordnet* dengan menggunakan kata yang sama.

Tugas akhir ini merupakan pengembangan dari tugas akhir yang telah ada tentang *text summarization* menggunakan metode *lexical chain*. Pada tugas akhir sebelumnya terdapat kelemahan yaitu munculnya problem ambiguitas dalam pembentukan rantai leksikal (*Lexical Chain*). Rantai Leksikal (*Lexical Chain*) sendiri merupakan alat bantu yang baik untuk mengidentifikasi topik. Sehingga jika dalam pembentukan rantai leksikal terjadi ambiguitas, maka ringkasan yang dihasilkan menjadi kurang sesuai dengan topik dokumen (muncul kerancuan / redundan). Tugas akhir ini akan memperbaiki kelemahan tersebut dengan fokus kepada pembentukan rantai leksikal dengan menerapkan *word sense disambiguation* menggunakan algoritma *Lesk*, yang kemudian akan diimplementasikan dalam peringkasan teks.

1.2 Perumusan Masalah

Dengan mengacu pada latar belakang di atas, maka permasalahan yang akan dibahas dan diteliti pada tugas akhir ini adalah:

1. Bagaimana cara mendeteksi kata yang berambigu dan mencari makna yang tepat dari suatu kata.
2. Bagaimana menyelesaikan masalah ambiguitas makna kata dengan menerapkan *word sense disambiguation* dalam peringkasan teks yang menggunakan metode *lexical chain* pada artikel tekstual.
3. Bagaimana tingkat performansi metode *lexical chain* dengan menerapkan *word sense disambiguation* jika dibanding dengan metode *lexical chain* yang tanpa menerapkan *word sense disambiguation*.

Adapun Batasan masalah dalam tugas akhir ini adalah sebagai berikut :

1. Koleksi dokumen yang digunakan untuk pengujian berupa dokumen tekstual yang berbahasa Inggris.
2. Ringkasan hanya menangani input berupa dokumen tunggal atau *single document.*
3. Koleksi dokumen yang digunakan untuk pengujian dari koleksi dokumen DUC 2002.
4. Jenis ringkasan yang dihasilkan berupa *extractive summary* dengan kisaran panjang summary antara 10% - 50%.
5. Algoritma *Lesk* akan diimplementasikan bersama *WordNet 3.0* sebagai acuan istilah dan acuan *glossary* kata.
6. Untuk evaluasi, ringkasan referensi menggunakan Microsoft Word 2007 yang juga *automatic text summarization* bertipe *extractive summary* yang banyak dipakai oleh orang.

7. Evaluasi dilakukan dengan membandingkan *content overlap* antara hasil ringkasan sistem dan ringkasan referensi dengan menggunakan *ROUGE evaluation toolkit* (*Recall-Oriented Understudy for Gisting Evaluation*).
8. Jenis Parameter *ROUGE* yang digunakan yaitu *ROUGE-1, ROUGE-2, dan ROUGE-W.*

1.3 Tujuan

Adapun tujuan dari penelitian ini adalah sebagai berikut :

1. Mengimplementasikan metode *lexical chain* dengan menerapkan *word sense disambiguation* menggunakan algoritma *lesk* untuk melakukan peringkasan dokumen tekstual.
2. Menganalisis hasil ringkasan sistem dengan membandingkan terhadap hasil ringkasan dari aplikasi Microsoft Word 2007 dengan menggunakan tool ROUGE.
3. Membandingkan kinerja metode *lexical chain* tanpa menggunakan *word sense disambiguation* dengan metode *lexical chain* yang menggunakan *word sense disambiguation*. Dilihat dari segi akurasi.

1.4 Metodologi Penyelesaian Masalah

Metode yang digunakan dalam penyelesaian tugas akhir ini adalah menggunakan metode studi pustaka atau studi literatur dan analisis dengan langkah kerja sebagai berikut:

1. Studi Literatur.
   a. Studi Literatur dengan mempelajari literatur-literatur yang relevan dengan penelitian yang meliputi : *text mining, text summarization, lexical chain method, word sense disambiguation, lesk algorithm, lexical chain with word sense disambiguation method*.
   b. Mempelajari materi, pendalaman materi, dan memahami materi yang berhubungan dengan tugas akhir ini.
2. Analisis permasalahan
   a. Mempelajari konsep tentang *text summarization, lexical chain method, word sense disambiguation, lesk algorithm, lexical chain with word sense disambiguation method* yang akan diimplementasikan dalam perangkat lunak.
   b. Analisis mengenai permasalahan yang ada pada *text summarization, lexical chain method, word sense disambiguation, lesk algorithm, lexical chain with word sense disambiguation method*.
   c. Menganalisa metode *lexical chain* yang telah menerapkan *word sense disambiguation* untuk peringkasan teks.
   d. Simulasi dan analisis hasil peringkasan teks dengan metode *lexical chain* dengan menggunakan tool *ROUGE*.

3. Mengumpulkan requirement terhadap perangkat lunak yang akan dibangun.
   a. Mencari dan mengumpulkan data uji, data tersebut diperoleh dari dataset *DUC (Document Understanding Conference).*
   b. Mencari dan mengumpulkan dependensi dan tool *ROUGE* yang digunakan sebagai alat pengujian.
4. Melakukan desain atau perancangan sistem peringkasan teks yang akan dibangun.
5. Melakukan implementasi perancangan sistem peringkasan teks.
6. Melakukan pengujian dengan memasukkan data berupa dokumen artikel *.txt* serta menganalisis hasil keluaran program.
7. Menganalisis hasil ringkasan dengan menggunakan *tool ROUGE* dan yang dianalisis adalah *recall*, *precision* dan *f-measure*.
8. Pengambilan kesimpulan dan penyusunan laporan tugas akhir.

5. KESIMPULAN DAN SARAN

5.1. Kesimpulan

Beberapa kesimpulan yang dapat diambil dari tugas akhir ini yaitu :

1. Penerapan metode word sense disambiguation pada metode lexical chain mampu meningkatkan performansi pada sistem peringkasan teks berdasarkan pengujian terhadap 50 dokumen.
2. Penerapan metode *word sense disambiguation* dengan menggunakan algoritma *lesk* dapat membantu metode *lexical chain* dalam membentuk rantai leksikal dengan cara menghilangkan ambiguitas setiap *candidate term* yang akan membentuk rantai leksikal.
3. Penerapan metode *word sense disambiguation* dengan menggunakan algoritma *lesk* sudah baik dalam pemilihan makna kata, walaupun pada kasus tertentu hasilnya masih kurang akurat.
4. Semakin bertambahnya summary length, akan semakin baik hasil *summary* yang dihasilkan sistem. Hal ini dikarenakan metode lexical chain memiliki kekurangan yaitu tidak bisa mendeteksi keyphrase, sehingga jika summary length kecil akan mengalami kesulitan dalam pemilihan kalimat yang tepat.
5. Semakin banyak rantai leksikal yang terbentuk maka akan semakin baik hasil summary, dikarenakan nilai bobot antar kalimat akan semakin beragam sehingga dapat mengurangi tingkat redundansi antar kalimat.

5.2. Saran

Saran-saran yang dapat penulis uraikan untuk keperluan analisis selanjutnya adalah:

1. Mencoba algoritma yang lain untuk metode *word sense disambiguation*, lebih baik jika algoritma yang bersifat *supervised learning* sehingga dalam penghilangan ambiguitas kata tidak tergantung pada kamus.
2. Mengembangkan sistem peringkasan teks dengan metode *lexical chain with word sense disambiguation* menjadi peringkasan teks bertipe *abstractive summary*. Walaupun pada saat ini *abstractive summary* masih terus dikembangkan dan belum begitu banyak yang mengaplikasikan sebagai *automatic text summarization.*

DAFTAR PUSTAKA

[1] Banerjee, Satanjeev, 2002, *Adapting the Lesk Algorithm for Word Sense Disambiguation to WordNet*, Duluth U.S.A : Department of Computer Science University of Minnesota

[2] Banerjee, Satanjeev, Pedersen Ted. 2005. *"An Adapted Lesk Algorithm for Word Sense Disambiguation Using Wordnet"*. Duluth U.S.A : University of Minnesota

[3] Barzilay, Regina., Elhadad, Michael. *Using Lexical Chains for Text Summarization*. Mathematics and Computer Science Dept. Ben Gurion University in the Negev Beer-Sheva

[4] Brunn, Meru., Chali, Yllias., Pinchak, J. Christopher. *Text Summarization Using Lexical Chains*. Department of Mathematics and Computer Science University of Lethbridge

[5] Ekedahl, Jonas , Koraljka Golub, 2004, *Word Sense Disambiguation Using WordNet and the Lesk Algorithm* , Sweden : Dept. of IT, Lund Univ

[6] Hongyan Jing, et. al, 1998, "*Summarization Evaluation Methods: Experiments and Analysis*". New York U.S.A : Dept. of Computer Science Columbia University

[7] Lesk, Michael. 2001. *"Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone"*. Morristown : Bell Communications Research

[8] Lin, Chin-Yew. 2004. *"ROUGE: a Package for Automatic Evaluation of Summaries"*. Barcelona, Spain : In Proceedings of the Workshop on Text Summarization Branches Out

[9] Mani, I. 2001. "*Summarization Evaluation: An Overview*" Sunset Hills Road Reston U.S.A : The MITRE Corporation,

[10] Michel Galley, Kathleen McKeown, 2003, *Improving Word Sense Disambiguation in Lexical Chaining*, www.cs.columbia.edu/nlp/papers/2003/galley_mckeown_03.pdf, 03 Oktober 2009

[11] Oxford University Press. 2004. "*The Oxford Handbook of Computational Linguistics*". New York U.S.A : Oxford University

[12] Roberto Navigli, 2009, *Word sense disambiguation: A survey*, http://portal.acm.org, 03 Oktober 2009

[13] Wikipedia. 2009. *Word Sense Disambiguation*. http://en.wikipedia.org/wiki/word sense disambiguation, 03 Oktober 2009