Prosiding
ANNUAL RESEARCH SEMINAR 2016
6 Desember 2016, Vol 2 No. 1

ISBN : 979-587-626-0 | UNSRI http://ars.ilkom.unsri.ac.id

Studi Awal Peringkasan Dokumen Bahasa Indonesia Menggunakan Metode *Latent Semantik Analysis* dan *Maximum Marginal Relevance*

Santun Irawan[1], Hermawan[2]
[1,2]STMIK GI MDP
[1,2]Magister Teknik Informatika Universitas Sriwijaya
Palembang, Sumatera Selatan, Indonesia
[1]santunirawan@mdp.ac.id, [2]hermawanmdp@gmail.com

Samsuryadi[3]
[3]Teknik Informatika, Fasilkom, Universitas Sriwijaya
Jl. Raya Palembang–Prabumulih Km 32 Indralaya, OI
Sumatera Selatan, Indonesia 30662
[3]syamsuryadi@unsri.ac.id

***Abstrak*—Pertumbuhan informasi *online* mengalami peningkatan yang signifikan pada dewasa ini. Peningkatan informasi ini memerlukan suatu mekanisme yang mampu menyajikan informasi secara efektif. Salah satu solusi yang ditawarkan adalah melakukan peringkasan teks secara otomatis tanpa menghilangkan konten dan makna dari dokumen. Pada penelitian ini dilakukan penggabungan metode *Latent semantic analysis* dan metode *Maximum marginal relevance* untuk proses peringkasan multi-dokumen, sehingga menghasilkan suatu ringkasan dokumen yang tetap mengandung informasi yang dianggap penting dan mewakili topik dari dokumen yang diringkas tersebut.**

Keywords—informasi online, *Latent semantic analysis, Maximum marginal relevance*, peringkasan multi-dokumen.

I. PENDAHULUAN

Dewasa ini, Informasi *online* mengalami peningkatan yang sangat pesat. Berdasarkan kenyataan ini, diperlukan suatu mekanisme yang mampu menyajikan informasi secara efektif [1]. Oleh karena itu, dalam rangka mengatasi masalah ini, penelitian tentang peringkasan teks otomatis tidak terstruktur telah meningkat dan menerima banyak perhatian dalam beberapa tahun terakhir. Selain itu, pekerjaan penelitian saat ini untuk peringkasan teks otomatis cenderung fokus pada peringkasan multi-dokumen dari pada peringkasan satu dokumen [2]. Umumnya, tujuan peringkasan teks otomatis untuk menyingkat teks sumber dalam rangka untuk membuatnya lebih pendek namun tanpa kehilangan konten dan makna dari dokumen [3]. Peringkasan dapat diekstraksi berasal dari satu dokumen atau beberapa dokumen. Peringkasan yang diekstrak dari satu dokumen, dinamakan peringkasan tunggal dokumen, sedangkan peringkasan yang diekstrak dari beberapa dokumen yang membahas topik yang sama, dinamakan peringkasan multi-dokumen [4]. Peringkasan teks biasanya diklasifikasikan menjadi 2 teknik, yaitu peringkasan abstraktif dan ekstraktif [5]. Suatu peringkasan abstraktif memperoleh pemahaman tentang konsep utama dalam dokumen [6].

Pada penelitian ini dilakukan penggabungan metode *Latent semantic analysis* (LSA) dan Metode *Maximum marginal relevance* (MMR) untuk proses peringkasan multi-dokumen, sehingga menghasilkan suatu peringkasan yang dapat mengandung informasi yang dianggap penting dan mewakili topik dari dokumen yang akan diringkas. Metode LSA dapat membentuk ruang semantik yang berisi kata-kata yang penting dalam dokumen yang saling berdekatan satu sama lain, sehingga informasi yang penting bisa didapatkan. Metode *latent semantic analysis* dijalankan terlebih dahulu sebelum dilakukan proses peringkasan dengan menggunakan metode *maximum marginal relevance.* Hasil metode *Latent semantic analysis* merupakan *score* tiap kalimat yang akan digabungkan saat perhitungan *score* awal pada metode *maximum marginal relevance*. Setelah seluruh proses peringkasan dokumen dijalankan, maka dihasilkan sebuah ringkasan sebagai *output* dari sistem yang telah dijalankan. Ringkasan yang dihasilkan akan dievaluasi dengan mengukur seberapa relevannya ringkasan yang dihasilkan dengan ringkasan referensi yang dibuat oleh ahli Bahasa Indonesia.

Prosiding
ANNUAL RESEARCH SEMINAR 2016
6 Desember 2016, Vol 2 No. 1

ISBN : 979-587-626-0 | UNSRI http://ars.ilkom.unsri.ac.id

II. PENELITIAN TERKAIT

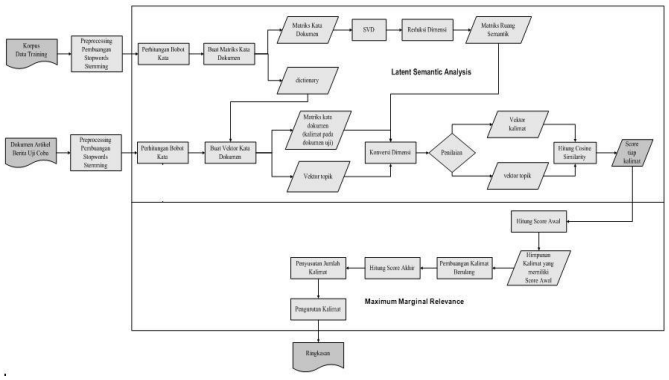
Sebuah ringkasan menurut Radev [7] dapat diartikan sebagai sebuah teks yang dihasilkan dari satu atau lebih kalimat yang menyampaikan informasi penting dari dokumen asli. Panjang dari sebuah ringkasan tidak lebih dari setengah panjang dokumen asli dan biasanya lebih pendek lagi.

Beberapa peneliti telah melakukan dengan Pendekatan peringkasan dokumen tunggal untuk meringkas multi-dokumen. Misalnya, Steinberger et al. menggunakan Metode *Latent Semantic analisis* (LSA) untuk meringkas multi-dokumen [8]. Meskipun mereka menerapkan Teknik peringkasan yang canggih, namun terdapat beberapa kelemahan dari teknik ini seperti proses dalam menyortir kalimat, ketika dimasukkan ke dalam ringkasan dari dokumen yang berbeda dan anaforis ekspresi yang tidak bisa diselesaikan dalam konteks ringkasan.

Selain itu, Goldstein et al. melakukan pendekatan ekstraksi kalimat untuk peringkasan multi-dokumen [1]. Pendekatan mereka memiliki beberapa keuntungan seperti waktu proses cepat dan domain-independen. Namun, Karena mereka tidak menerapkan teknik pemahaman bahasa alami, metode mereka menghasilkan ringkasan yang memiliki bagian-bagian yang terpisah satu sama lain. Sedangkan Tardan et al. menggunakan pendekatan analisis semantik untuk melakukan peringkasan dokumen tunggal dalam Bahasa Indonesia [9].

III. METODOLOGI PENELITIAN

Kerangka kerja (*framework*) yang diusulkan untuk peringkasan multi-dokumen disajikan pada Gambar 1.



Gambar 1. Rancangan Kerangka Kerja Proses Peringkasan Multi-Dokumen

Pada penelitian ini, proses peringkasan multi-dokumen Berbahasa Indonesia akan dilakukan dengan menggabungkan metode *latent semantic analysis* dengan metode m*aximum marginal relevance.* Tujuan utama dari penelitian ini adalah untuk menghasilkan suatu ringkasan yang baik dan isinya mewakili seluruh informasi yang ada pada dokumen tersebut, ringkasan yang lebih singkat daripada dokumen aslinya sehingga memudahkan pembaca dalam mengetahui isi dan garis besar dokumen tanpa perlu membaca keseluruhan isi dokumen.

Data yang digunakan adalah dokumen artikel berita yang diperoleh dari artikel *online,* dan situs-situs berita di internet lainnya. Seluruh dokumen yang ada akan dilakukan tahap *preprocessing*, yaitu proses pemformatan dokumen, pembuatan indeks dokumen, dan tokenisasi. Tahap selanjutnya adalah pembuangan *stopwords* dan proses *stemming*. Proses *stemming* ini bertujuan untuk mempermudah dalam melakukan proses perhitungan bobot kata, dan menghitung persamaan antar kalimat. Perhitungan bobot kata dilakukan dengan menggunakan perhitungan TF*IDF, sehingga menghasilkan himpunan kata yang sudah diberi bobot.

Metode LSA ini akan dijalankan terlebih dahulu sebelum dilakukan peringkasan menggunakan metode m*aximum marginal relevance.* Data dari korpus data *training* akan dibuat matriks kata dokumen, sehingga menghasilkan matriks kata dokumen dan *dictionary*. Matriks kata dokumen akan direduksi dimensinya oleh SVD (*Singular Value Decompotion*) ke dimensi 300, sehingga menghasilkan matriks ruang semantic. Hasil yang didapat pada metode *latent semantic analysis* merupakan *score* setiap kalimat (nilai skalar *cosine similarity* antara vektor kalimat dan vektor topik).

Untuk menghasilkan ringkasan, maka dilakukan penyusutan kalimat berdasarkan tingkat kompresi (*compression rate*) yang telah ditentukan, yaitu 10% dan 20%.Kalimat-kalimat yang akan diambil untuk dibuat ringkasan adalah kalimat-kalimat yang memiliki peringkat teratas. Hasil ringkasan yang didapat setelah proses peringkasan multi-dokumen selesai dilakukan akan dievaluasi untuk dapat mengetahui kualitas dari ringkasan yang dihasilkan dari metode ringkasan dokumen yang digunakan.

3.1. Pre-processing

Information Retrieval System atau temu kembali informasi merupakan proses pencarian tidak tersrtuktur

ISBN : 979-587-626-0 | UNSRI http://ars.ilkom.unsri.ac.id

untuk memenuhi kebutuhan informasi dari sekumpulan koleksi yang besar [12]. Secara garis besar ada dua pekerjaan yang ditangani oleh sistem temu kembali informasi (TKI), yaitu melakukan *pre-processing* terhadap dokumen dan kemudian menerapkan metode tertentu untuk menghitung kedekatan (relevansi atau *similarity*) antara dokumen dengan *query* [11].

Tugas pokok pada tahapan *pre-processing* di dalam TKI adalah membangun *index* dari koleksi dokumen. *Index* adalah himpunan *term* yang menunjukkan isi atau topik yang dikandung oleh dokumen. Ekstraksi *term* biasanya melibatkan tiga operasi utama, yaitu pemisahan rangkaian term (tokenisasi), penghapusan *stop-words*, dan *stemming* [11].

Tokenisasi adalah tugas memisahkan deretan kata di dalam kalimat, paragraph atau halaman menjadi *token* atau potongan kata tunggal atau *termed word*. Tahapan ini juga menghilangkan karakter-karakter tertentu seperti tanda baca dan mengubah semua token ke bentuk huruf kecil (*lower case*) [11].

Stop word didefinisikan sebagai *term* yang tidak berhubungan (*irrelevant*) dengan subyek utama dari database meskipun kata tersebut sering kali hadir di dalam dokumen. Contoh *stop words* adalah "ada", "adalah", "adanya", "adapun", "agak", dll [11].

Kata-kata yang muncul di dalam dokumen sering mempunyai banyak varian morfologik. Karena itu, setiap kata yang bukan *stop-words* direduksi ke *stemmed word* (*term*) yang cocok yaitu kata tersebut distem untuk mendapatkan bentuk akarnya dengan menghilangkan awalan atau akhiran [13].

3.2. Latent Semantik Analysis

Latent Semantic Analysis (LSA) merupakan suatu teknik untuk mengekstraksi, menyajikan, serta menganalisis makna kontekstual dari kata yang muncul dengan pendekatan matematis dan statistik yang diaplikasikan pada korpus yang besar. LSA menggunakan model vektor seperti *Vector Space Model* (VSM), namun penilaiannya tidak hanya berdasarkan frekuensi kemunculan kata, namun tergantung pada analisis matematika yang mampu menganalisis serta menyimpulkan relasi yang lebih dalam [10].

Langkah awal adalah data *training* dibuat menjadi matriks kata dokumen, sehingga menghasilkan matriks kata dokumen dan *dictionary*. Matriks kata dokumen direduksi dimensi oleh SVD (*Singular Value Decompotion*) ke dimensi

300 [15], sehingga menghasilkan matriks ruang semantik, yang di dalamnya berisi vektor-vektor kata. Data dari dokumen artikel uji coba dibuat menjadi vektor kata dokumen, sehingga menghasilkan matriks kalimat dan matriks topik sesuai dengan *dictionary* dari data *training*, yang di dalamnya berisi vektor-vektor kalimat yang mewakili setiap kalimat yang ada di dalam dokumen, dan vektor topik yang mewakili topik inti dokumen. Topik adalah kalimat yang ada pada semua dokumen yang digabung menjadi satu.

Penilaian yang ada pada metode LSA adalah melakukan perhitungan *cosine similarity* antara vektor kalimat dengan vektor topik, sehingga menghasilkan *score* setiap kalimat yang akan digabungkan pada saat menghitung nilai *score* awal pada tahap m*aximum marginal relevance.*. Asumsi bahwa dokumen uji akan mencakup beberapa kata yang ada pada korpus. Pengguna dapat mengambil asumsi bahwa ada berapa banyak kata pada korpus yang harus dicakup pada dokumen berita, misalkan ada sebanyak *n*. Untuk mendapatkan *score* setiap kalimat, sistem akan mengukur seberapa dekat makna semantik antara kalimat dengan seluruh kata yang ada pada topik. Kalimat yang dianggap dicakup oleh topik dalam dokumen uji adalah kalimat yang menghasilkan *cosine similarity* tertinggi. Hal ini dilakukan dengan membandingkan vektor kalimat dengan vektor topik, kemudian akan didapatkan nilai akhir yang dilakukan dengan cara mengambil *cosine similarity* dengan nilai tertinggi.

3.3. Maximum Marginal Relevance

Metode *maximum marginal relevance* (MMR) merupakan salah satu metode ekstraksi ringkasan (*extractive summary*) yang digunakan untuk meringkas dokumen tunggal atau multi-dokumen. MMR meringkas dokumen dengan menghitung kesamaan (*similarity*) antara bagian teks [11].

Pada peringkasan dokumen dengan metode MMR dilakukan proses segmentasi dokumen menjadi kalimat dan dilakukan pengelompokkan sesuai dengan gender kalimat tersebut. MMR digunakan dengan mengkombinasikan matriks *cosine similarity* untuk merangking kalimat-kalimat sebagai tanggapan pada *query* yang diberikan oleh *user*. Penghitungan MMR dinyatakan dalam Persamaan (1) sebagai berikut.

$$MMR = argmax[\lambda * Sim_1 (Si,Q) - (1 - \lambda) * maxSim_2 (S_i,S')] \quad (1)$$

S_i adalah kalimat di dokumen, sedangkan S' adalah kalimat yang telah dipilih atau diekstrak. Koefisien λ digunakan

ISBN : 979-587-626-0 | UNSRI http://ars.ilkom.unsri.ac.id

untuk mengatur kombinasi nilai untuk memberi penekanan bahwa kalimat tersebut dan untuk mengurangi redudansi. Nilai parameter λ adalah mulai dari 0 sampai dengan 1 ([0,1]). Untuk peringkasan small dokumen, seperti pada berita, menggunakan nilai parameter λ = 0.7 atau λ = 0.8, karena akan menghasilkan ringkasan yang baik. Sim_1 adalah matriks *similarity* kalimat S_i terhadap *query* yang diberikan oleh *user* sedangkan Sim_2 adalah matriks *similarity* kalimat S_i terhadap kalimat yang telah diekstrak sebelumnya [3].

IV. HASIL AWAL

4.1. Dokumen Uji Coba

Dokumen uji coba akan terdiri dari 5 kategori berita, yaitu ekonomi, hukum, internasional, olahraga, dan teknologi. Dokumen uji coba pada metode LSA akan dibedakan menjadi dua jenis data untuk setiap kategori berita, yaitu :

Data pertama merupakan file kalimat yang berisi kalimat-kalimat yang ada pada setiap dokumen berdasarkan kategorinya dokumennya.

Data keduamerupakan file topik yang berisi kalimat-kalimat yang ada pada setiap dokumennya dimana kalimat tersebut akan digabungkan menjadi 1 dokumen saja.

4.2. Hasil Uji Coba

Keluaran dari sistem berupa dokumen yang berisi hasil ringkasan.Hasil ringkasan yang dihasilkan dari sistem akan dibandingkan dengan ringkasan referensi yang dibuat oleh ahli Bahasa Indonesia dengan cara menghitung korelasi antara keduanya. Rata-rata korelasi yang disebutkan dalam hasil penelitian didapat dengan menghitung akurasi dari jumlah kalimat yang sama muncul pada hasil ringkasan dari sistem dan ringkasan referensi.

Output dari ringkasan yang akan dihasilkan oleh sistem akan dibuat berdasarkan tingkat kompresi yang berbeda, yaitu 10% dan 20%, dapat dilihat pada Tabel 1.

Tabel 1. *Output* Ringkasan Berdasarkan Tingkat Kompresi

Kategori	Jumlah Kalimat	*Output* Ringkasan Berdasarkan Tingkat Kompresi	
		10%	20%
Ekonomi	96	9 kalimat	19 kalimat
Hukum	51	5 kalimat	10 kalimat
Internasional	47	4 kalimat	9 kalimat
Olahraga	41	4 kalimat	8 kalimat
Teknologi	81	8 kalimat	16 kalimat

V. KESIMPULAN

Hasil awal berupa kerangka kerja untuk peringkasan multi-dokumen dengan metode *latent semantic analysis* dan metode *Maximum Marginal Relevance*. Kerangka kerja ini dapat diterapkan untuk menghasilkan beberapa faktor yang sangat menentukan terhadap hasil ringkasan referensi yang dibuat oleh ahli Bahasa Indonesia, yaitu: variasi dari pembobotan untuk setiap skenario, banyaknya data *training* yang digunakan, dan variasi dari tingkat kompresi. Variasi pembobotan setiap skenario sangat menentukan hasil akurasi dari ringkasan yang dihasilkan terhadap ringkasan referensi.

REFERENSI

1. J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz. Multi-Document Summarization By Sentence Extraction. In Proceedings of ANLP/NAACL-2000 Workshop on Automatic zummarization, 2000.
2. T. Hirao, T. Fukusima, M. Okumura, C. Nobata and H. Nanba. Corpus and Evaluation Measures for Multiple Document Summarization with Multiple Sources. In Proceedings of the Twentieth International Conference on Computational Linguistics (COLING), pp. 535–541, 2004.
3. Y.J. Kumar and N. Salim. Automatic Multi Document Summarization Approaches. Journal of Computer Science Vol. 8 No.1, pp. 133-140, 2012.
4. M.G. Ozsoy, I. Cicekli, and F.N. Alpaslan. Text Summarization of Turkish Texts using Latent Semantic Analysis. In Proceedings of the 23rd International Conference on Computational Linguistics, COLING '10, pp. 869–876, Stroudsburg, PA, USA. Association for Computational Linguistics, 2010.
5. V. Gupta and G.S. Lehal. A Survey of Text Summarization Extractive Techniques. Journal of Emerging Technologies in Web Intelligence, Vol. 2, No. 3, August 2010.
6. G. Erkan and D.R. Radev. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research, Vol. 22, pp. 457-479, 2004.
7. Radev, D. R., Hovy, E. H., & McKeown, K., "Introduction to the Special Issue on Summarization", Computational Linguistics, vol. 28, no. 4, hal. 399- 408, 2002.Ferreira, R., Freitas, F., Cabral, L. d., Lins, R. D., Lima, R., Franc¸a, G., . . . Favaro, L., "A Context Based Text Summarization System", 11th IAPR International Workshop on Document Analysis Systems, IEEE, 2014
8. J. Steinberger and M. K˘rišt'an. LSA-Based Multi-Document Summarization. In 8th International PhD Workshop on Systems and Control, a Young Generation Viewpoint, pp. 87–91, Balatonfured, Hungary, 2007

ISBN : 979-587-626-0 | UNSRI http://ars.ilkom.unsri.ac.id

9. P.P.D. Tardan, A. Erwin, K.I. Eng, and W. Muliady. Automatic Text Summarization Based on Semantic Analysis Approach for Documents in Indonesian Language. In Information Technology and Electrical Engineering (ICITEE), 2013 International Conference on, pp. 47-52, 2013.
10. Launder et al., "An Introduction to Latent Semantic Analysis," pp. 259 –284. 1998.
11. Indriani,Aida.,*Maximum marginal relevance untuk peringkasan teks otomatis sinopsis buku berbahasa indonesia, Seminar Nasional Teknologi Informasi dan Multimedia 2014, STMIK AMIKOM Yogyakarta, 8 Februari 2014*
12. C.D. Manning, P. Raghavan dan H. Schütze, " An Introduction to Information Retrieval", in Cambridge University Press, pp.26, April 1, 2009.
13. Husni Ilyas, "Unified Messaging System Information Retrieval & Klasifikasi Teks", in Komputasi | Suatu Permulaan Data Mining & IR, pp.6-8, Januari 22, 2010.
14. D. Kalman, "A Singularly Valuable Decomposition: The SVD of a Matrix," The American University, Washington DC. Februari, 2002.
15. M. Mustaqhfiri, Z. Abidin, R. Kusumawati, "Peringkasan Teks Otomatis Berita Berbahasa Indonesia Menggunakan Metode Maximum Marginal Relevance", in n nitro PDF professional, pp. 134-147.