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

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

Peringkasan Otomatis Dengan Ekstraksi Informasi Untuk Dokumen Berita Ter-*cluster*

Ridwan Ilyas
Program Studi Informatika
Universitas Jenderal Achmad Yani
Cimahi, Indonesia
ilyas@lecture.unjani.ac.id

Fajri Umbara
Program Studi Informatika
Universitas Jenderal Achmad Yani
Cimahi, Indonesia
fajri.umbara@gmail.com

A***bstract—Keterbukaan dan kemudahan mengakses informasi membuat jumlah informasi menjadi sangat banyak. Banyaknya informasi untuk satu hal yang sama menimbulkan information overload. Masalah tersebut muncul dalam berbagai bidang seperti berita, dokumen karya ilmiah dan media sosial. Dibutuhkan sistem yang mampu membantu pengguna untuk menghasilkan berita yang lengkap dengan cara membangun sistem peringkasan otomatis. Pada penelitian ini diajukan membentuk serangkayan standar dalam tahapan peringkasan berita dengan konfirgurasi dinamis pada masing-masing tugas (clustering, ekstraksi informasi dan peringkasan). Dengan membangun sistem peringkasan dari mulai proses clustering, ekstraksi informasi dan peringkasan diharapkan menghasilkan hasil ringkasan yang utuh, lengkap dan memiliki tingkat keterbacaan tinggi.***

Keywords: peringkasan otomatis, ekstraksi informasi, clustering, berita online

I. LATAR BELAKANG

Peringkasan otomatis merupakan bagian yang tidak terpisahkan dari ilmu Pemrosesan Bahasa Alami/ *Natural Language Processing*(NLP). Ringkasan didefinisikan sebagai sebuah teks yang dihasilkan dari satu atau lebih teks asal dan mengandung informasi penting dari teks asal dengan panjang tidak lebih dari setengah teks aslinya atau lebih sedikit dari itu [1]. Dengan definisi tersebut, maka peringkasan memiliki tiga karakteristik yaitu sumber peringkasan bisa satu atau lebih dokumen, peringkasan harus mengandung infomasi penting dan hasilnya relatif singkat.

Secara umum peringkasan terbagi atas dua teknik yaitu ekstraksi dan abstaksi, dilihat dari bagaimana data hasil peringkasan disajikan. Peringkasan dengan teknik ekstraksi menghasilkan informasi-informasi penting yang merupakan bagian dari sumber asal berita. Peringkasan dengan teknik abstraksi menghasilkan struktur bahasa yang baru dari teks asal seperti halnya kita menceritakan ulang suatu kejadian. Teknik yang paling populer/banyak digunakan adalah teknik ekstraksi.

Terdapat banyak metode peringkasan otomatis dengan teknik ekstraksi. Awalnya penelitian berpusat pada teknik mengelola dokumen dengan beberapa pendekatan seperti berdasarkan frekuensi dari kata-kata yang ada pada teks asal [2] atau berdasarkan dari posisi kalimat [3]. Setelah itu berkembang teknik peringkasan menggunakan pembelajaran mesin/*machine learning* yang digunakan untuk menilai kalimat-kalimat berdasarkan fitur tertentu seperti panjang kalimat, keberadaan kata-kata penting, keberadaan kata-kata untuk tema tertentu dan fitur dari paragraph [4].

Dengan alasan meningkatkan akurasi dan hasil yang lebih spesifik, maka dilakukan eksperimen untuk melakukan peringkasan otomatis dengan memakai teknik Ekstraksi Informasi/*Information Extraction*(IE). Dikembangkan sebuah sistem yang mengadopsi ekstraksi informasi untuk peringkasan otomatis yang diberi nama RIPTIDES yang bekerja peringkasan berita berdasarkan *scenario template* yang dipilih oleh pengguna [5]. Penelitian lain mencoba melakukan peringkasan kumpulan dokumen dengan pendekatan *novel* berdasarkan *cross-document Information Extraction* [6] dengan hasil peningkatan pada ROUGE-2 recall dan ringkasan yang lebih diterima oleh pembaca (0.78 lebih baik *TAC Content score* dan 0.11 lebih baik *Readability/Fluency score*).

Pada penelitian dengan sistem RIPTIDES [5], peringkasan sangat bergantung pada *template* yang dipilih oleh pengguna, jika ingin merangkum untuk topik yang lain diperlukan proses untuk membentuk template baru. Belum dimungkinkan agar sistem mampu memilih *template* otomatis atau adanya template generik tapi dengan konstrain tertentu.

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

II. KAJIAN PUSTAKA

2.1. *Peringkasan Otomatis*

Peringkasan otomatis adalah sekumpulan *task* menghasilkan dokumen berisi informasi-informasi penting dari satu atau lebih dokumen yang panjangnya tidak lebih dari setengah dokumen asal [2]. Kunci utama dari peringkasan adalah sumbernya bisa satu atau lebih dokumen, mengandung informasi-informasi penting dan ukurannya ralatif singkat. Perinngkasan otomatis dalam diklasifikasikan dalam dua jenis yaitu *single document* dan *multi document* (Suneetha 2001). Dua klasifikasi tersebut terkait dari sumber data yang dipakai.

Model proses dari peringkasan otomatis dibagi ke dalam tiga tahapan. Tahapan pertama adalah praproses teks untuk mengubah interpretasi teks ke bentuk representasi. Tahapan kedua adalah mengubah sumber representasi teks asal ke representasi peringkasan menggunakan algoritma tertentu. Tahapan terakhir adalah menghasilkan ringkasan dari representasi sebelumnya [8].

2.2. *Teknik Peringkasan*

Teknik peringkasan otomatis dapat dikelompokan dalam beberapa pendekatan. Pendekatan ini dilihat dari kompleksitas dan juga berhubungan dengan *time line* penelitian dalam peringkasan otomatis [9]. **Pendekatan klasik** dimana peringkasan difokuskan dari sumber asal seperti judul, frekuensi kata [2] atau posisi kalimat [3] Pada awal permulaan penelitian tentang peringkasan otomatis banyak ditemukan pendekatan sejenis.

Pendekatan pembelajan mesin dimana sebelumnya sistem diberi pelatihan tentang kalimat utama atau bukan dengan beberapa fitur seperti posisi kata, keberadaan kata kapital, jenis kalimat langsung atau tidak langsung dan kalimat mengandung kata penting atau tidak [4]. Selain fitur-fitur tersebut juga digunakan fitur yang lebih mendalam seperti *idf, tf, signature words*[1], gabungan dari dua kata, satu kata tunggal dan jenis kata, dengan algoritma *Naive-Bayes* [10]. Algoritma pembelajaran mesin lain seperti *Decition Tree* [11], *Hidden Markov Models* [12], *Log-Linear Models* [13] dan *Neural Networks* [14] dipakai dalam peringkasan dengan maksud meningkatkan akurasi.

Pendekatan analisa bahasa natural lebih mendalam yang banyak memodelkan *text's discourse structure* [9]. Usaha untuk meningkatkan performansi dari peringkasan adalah dengan memperbanyak jumlah analisa bahasa [15]. Digunakan keterhubungan antar teks secara sekuensial, singkatan dan jarak setiap teks (*lexical chain*). Salah satu cara untuk menemukan *lexical chain* adalah dnegan Wordnet [16], dengan tahapan memilih set kandidat kata, lalu setiap kandidat dicari rantai yang paling tepat tergantung kriteria keterkaitannya dan jika ditemukan kata dimasukkan ke dalam rantai.

Pendekatan berbasis pengetahuan dimana cara melakukan peringkasan dilakukan sebagaimana pakar melakukan peringkasan. Dalam pendekatan ini ada beberapa teknik seperti melakukan pemilihan kata berdasarkan fekuensi, keberadaan istilah dan lokasi kalimat, teknik peringkasan berdasarkan rule dengan pelatihan terlebih dahulu dengan teks hasil peringkasan pakar dan teknik peringkasan dengan membangun stuktur pengetahuan pakar [17]. Pendekatan berbasis pengetahuan digunakan juga untuk menentukan topik utama pada teks dengan pendekatan membangun struktur semantik teks yang lebih umum pada sebuah kalimat berdasarkan pengetahuan pakar [18].

2.3. *Ekstraksi Informasi*

Ekstraksi Informasi adalah suatu teknik yang digunakan untuk menghasilkan informasi yang relevan dari dokumen berskala besar dengan hasil berupa informasi yang tekstruktur *Resolution*. Hasil dari penelitian ini menunjukan bahwa isi dan keterbacaan ringkasan lebih baik jika menggunakan ekstraksi informasi.

III. GAGASAN

Sistem peringkasan berita otomatis yang diajukan merupakan sekumpulan tugas berantai dari beberapa yaitu *clustering* berita, ekstraksi informasi dan peringkasan dokumen. Tugas *clustering* berita diperlukan agar kumpulan berita yang masuk adalah berita dengan topik yang sama. Ekstraksi informasi digunakan untuk menjaring informasi utama dari berita sehingga mempermudah proses peringkasan. Peringkasan dokumen bekerja untuk membangun hasil ekstraksi informasi menjadi dokumen baru yang utuh dengan informasi yang lengkap dan memiliki tingkan keterbacaan tinggi.

Gambar 1 Rangkaian Tugas Sistem Peringkasan Berita

1. Clustering Berita. Teknik mengumpulkan berita dengan topik yang sama, secara khusus digunakan untuk peringkasan dokumen, perlu dikaji karena faktor ini dapat berakibat terhadap hasil akhir peringkasan. Pada bagian ini dilakukan konfigurasi terhadap atribut data, jumlah *clustering* dan pemilihan *centroid* (pusat *cluster*).
2. Ekstraksi Informasi. Menghasilkan informasi utama dari sekumpulan berita, menjadi bagian yang paling penting karena dari sinilah bahan beringkasan dihasilkan. Diperlukan analisis khusus agar teknik ekstrasi yang dipakai mampu secara penuh mendukung tugas peringkasan menjadi lebih mudah. Pada bagian ini dilakukan konfigurasi terhadap teknik ekstraksi, batasan

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

objek ektraksi (kata, prase, kalimat atau paragraf), jumlah class dan algoritma.

3. Peringkasan Otomatis. Hasil dari esktraksi informasi belum menjadi keluaran yang dapat disajikan kepada pengguna. Dipelukan proses pemilihan, pembanguanan ulang dan analisa keterbacaan. Pada proses akhir ini, analisa hubungan antar kalimat dan paragraf yang dibangun harus memenuhi standar bahasa yang benar. Pada bagian ini dilakukan konfirgurasi terhadap besarnya keluaran (headline, singkat atau panjang), bentuk (poin, paragraf, atau dokumen) dan kekhususan yang dinginkan (pilihan antara 5w+1h).

Gagasan dari penelitian ini adalah membentuk serangkayan standar dalam tahapan peringkasan berita dengan konfirgurasi dinamis pada masing-masing tugas (clustering, ekstraksi informasi dan peringkasan). Dengan adanya standar alur maka pengguna akan dimudahkan mendapatkan ringkasan baik dengan konfigurasi yang diinginkan atau pun konfigurasi yang disarankan dari hasil penelitian. Hasil analisis keterkaitan atar tahap, dapat menjadi acuan awal untuk terus mengembangkan sistem peringkasan berita otomatis.

Hasil dari penelitian diharapkan berupa sistem baru yang dapat beradaptasi terhadap konfigurasi peringkasan yang diinginkan baik berupa aturan-aturan baru atau pun sebuah algoritma baru yang bisa muncul setelah proses penelitian dilaksanakan. Selain itu, hasil dari penelitian ini berita perangkat lunak yang mengimplementasikan skema sistem peringkasan untuk digunakan pengguna umum yaitu masyarakat dan pengguna khusus yaitu peneliti pada bidang teknologi bahasa.

IV. MANFAAT PENLITIAN

Penelitian dalam bidang peringkasan otomatis dengan objek masalah *information overload* memiliki beberapa manfaat dalam untuk berbagai sumber dokumen seperti:

a. Abstraksi makalah ilmiah. Peringkasan otomatis dapat digunakan untuk membangun abstrak dari makalah ilmiah. Abstrak merupakan sekumpulan informasi utama dari makalah yang ditulis, dengan peringkasan otomatis dapat digunakan untuk membantu abstrak.
b. Ringkasan studi literatur. Ringkasan makalah pendukung dari penelitian yang sedang dikerjakan perlu disertakan dalam publikasi penelitian. Makalah yang menjadi acuan perlu diringkas, dengan peringkasan otomatis dapat digunakan untuk membantu membangun paragraph-paragraf pendukung pada bagian studi literatur sebuah makalah.
c. Ringkasan posting media sosial. Besarnya data pada media sosial menjadi potensi untuk dilakukan analisis sentiment terhadap objek tertentu. Informasi yang tersebar dapat menjadi bahan menentukan penilaian untuk kasus, prodak atau objek tertentu, dengan peringkasan otomatis maka hasil yang didapatkan menjadi lebih efisien.
d. Ringkasan banyak berita. Banyaknya portal berita yang menerbitkan satu objek berita yang sama membuat informasi terhadap satu berita menjadi sangat banyak namun tersebar dalam berbagai halaman. Dengan peringkasan otomatis maka berita-berita yang sama dapat diringkas menjadi satu berita dengan informasi yang utuh dan lengkap.

Fokus penelitian yang diajukan dalam penelirian ini adalah peringkasan otomatis untuk kumpulan berita online bahasa Indonesia. Sekumpulan berita dengan topik yang sama menjadi objek dari sistem peringkasan dengan hasil berupa berita baru. Hasil dari peringkasan berupa berita baru yang memiliki unsur kelengkapan isi berita 5w1h (who what where when why + how), unsur waktu penerbitan (*time series*), unsur sumber portal berita dan analisis objek berita.

Dengan adanya sistem peringkasan yang akan dibangun dalam penlitian ini, maka hasilnya bermanfaat sebagai alat bantu dalam melakukan analisis berita lebih efisien dibandingkan cara membaca seluruh berita yang ada. Hasil dari peringkasan dapat menjadi keluaran dengan manfaat langsung atau pun tidak langsung. Manfaat langsung seperti yang telah disebutkan sebelumnya yaitu membantu membaca berita lebih efektif. Manfaat tidak langsung dari keluaran sistem peringkasan dapat menjadi bahan untuk sistem analisis dokumen lainnya seperita sentiment analisis, sosial media analisis, review produk dan lain-lain.

V. KESIMPULAN

1. Dengan membangun sistem peringkasan dari mulai proses clustering, ekstraksi informasi dan peringkasan diharapkan menghasilkan hasil ringkasan yang utuh, lengkap dan memiliki tingkat keterbacaan tinggi.

2. Dengan melakukan seluruh penelitian diharapkan menghasilkan model baru yang lengkap dalam kajian peringkasan otomatis dalam lingkup berita elektronik.

Hasil dari peringkasan otomatis dihapkan dapat membantu pengguna mendapatkan berita yang lengkap untuk satu kejadian tanpa perlu membaca seluruh terbitan dari portal berita online.

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

DAFTAR PUSTAKA

[1] D. R. Radev, E. Hovy and K. McKeown, "Introduction to the special issue on summarization," *Computational Linguistics,* vol. 28, pp. 399-408, 2004.

[2] H. P. Luhn, "The automatic creation of literature abstracts.," *IBM Journal of Research Development,* vol. 2, no. 2, pp. 159-165, 1958.

[3] P. Baxendale, "Machine-made index for technical literature - an experiment," *IBM Journal of Research Development,* vol. 2, no. 4, pp. 354-361, 1958.

[4] J. Kupiec, J. Pedersen and F. Chen, "A trainable document summarizer," *In Proceedings SIGIR '95,* pp. 68-73, 1995.

[5] W. Michael, K. Tanya, C. Claire, N. Vincent, P. David and W. Kiri, "Multidocument Summarization via Information Extraction," Department of Computer Science Cornell University, Ithaca, New York, 2001.

[6] Heng-Ji, L. Juan, F. Benoit, G. Dan and H.-T. Dilek, "Re-ranking Summaries based on Cross-document Information Extraction," Berkeley, 2012.

[7] S. Suneetha, "Automatic Text Summarization: The Current State of the art," *International Journal of Science and Advanced Technology,* vol. 1, pp. 283-293, November 2001.

[8] K. S. Jones, "Automatic summarizing: factors and directions, In: Inderjeet Mani and Mark T. Maybury (Eds.), Advances in Automatic Text Summarization," pp. 1-12, 1999.

[9] D. Das and A. F. T. Martin, A Survey on Automatic Text Summarization, Language Technologies Institute, Carnegie Mellon University, 2007.

[10] C. Aone, M. E. Okurowski, J. Gorlinsky and B. Larsen, "A trainable summarizer with knowledge acquired from robust nlp techniques," *Advances in Automatic Text Summarizatio,* pp. 71-80, 1999.

[11] C.-Y. Lin and E. Hovy, "Identifying topics by position,"

[12] *Proceedings of the Fifth conference on Applied natural language processing,* pp. 283-290, 1997.

[13] J. M. Conroy and D. P. O'leary, "Text summarization via hidden markov models," *Proceedings of SIGIR '01,* pp. 406-407, 2001.

[14] M. Osborne, "Using maximum entropy for sentence extraction," *Proceedings of the ACL'02 Workshop on Automatic Summarization,* pp. 1-8, 2002.

[15] K. Svore, L. Vanderwende and C. Burges, "Enhancing single-document summarization by combining RankNet and third-party sources," *Proceedings of the EMNLP-CoNLL,* pp. 448-457, 2007.

[16] R. Barzilay and M. Elhadad, "Using lexical chains for text summarization," *Proceedings ISTS'97,* 1997.

[17] G. A. Miller, "Wordnet: a lexical database for english," *Commun. ACM,* vol. 38, no. 11, pp. 39-41, 1995.

[18] U. Habn and I. Mani, "The Challenges of Automatic Summarization," *0018-9162/00/IEEE,* November 2000.

[19] Y.-C. Lin, "Knowledge-based Automatic Topic Identification," *CA 90089-2562,* 1991.

[20] K. Kaiser and S. Miksch, Information Extraction A Survey, Vienna: Vienna University of Technology, 2005.

[21] M. Hearst, "What is text mining," 2004. [Online]. Available: http://www.sims.berkeley.edu/~hearst/textmining.html.