ANALISIS DAN IMPLEMENTASI PENGEMBANGAN METODE LONGEST COMMON SUBSEQUENCES DAN SKIP-BIGRAM CO-OCCURRENCE STATISTICS DALAM TEKNIK ROUGE UNTUK MENGEVALUASI MULTI-DOCUMENT SUMMARIZATION STUDI KASUS: RINGKASAN DARI MESIN PERINGKAS TEKS OTOMATIS

Tri Devi Arianti Sari[1], Warih Maharani[2], Moch Arif Bijaksana[3]

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

Abstrak

Evaluasi mesin peringkas teks otomatis adalah proses yang dilakukan untuk menentukan kualitas dari suatu mesin peringkas teks otomatis. Proses ini menghasilkan keluaran berupa nilai yang didapat dengan cara membandingkan antara ringkasan yang dihasilkan oleh mesin peringkas teks otomatis (disebut juga sebagai ringkasan kandidat) dengan ringkasan (ideal) lain yang dibuat oleh manusia (disebut juga sebagai ringkasan referensi) yang berasal dari teks sumber yang sama.

Pada Tugas Akhir penulis mengimplementasikan pengembangan dari metode Longest Common Subsequences (LCS) dan Skip Bigram Co Occurrences Statistics (SB) dalam teknik ROUGE untuk mengevaluasi mesin peringkas teks otomatis multi-document. Dimana pengembangan dari kedua metode tersebut bertujuan untuk menghasilkan suatu cara evaluasi yang memiliki tingkat keterhubungan yang baik dengan human judgement, karena informasi yang terambil menjadi semakin banyak.

Pengujian dilakukan dengan menggunakan Pearson's Coefficient of Correlation untuk mengukur hubungan linear antara nilai yang diberikan oleh aplikasi dan nilai rata-rata human judgement terhadap ringkasan yang sama. Hasil pengujian menunjukkan bahwa evaluasi dengan menggunakan algoritma LCS+SB mampu menghasilkan nilai evaluasi yang lebih tinggi dibandingkan algoritma LCS maupun SB, dengan hasil korelasi yang terkadang lebih tinggi dibandingkan kedua algoritma yang lain. Selain itu, korelasi dapat ditingkatan dengan melakukan penambahan jumlah ringkasan referensi yang digunakan.

Kata Kunci : Summarization Evaluation, Longest Common Subsequences, Skip Bigram Co Occurrences Statistics, ROUGE, Pearson's Coefficient of Correlation

Abstract

Evaluation of text summarization automatic machine is a process to determine the quality of an automatic text summarization. This process produces value, obtained with comparing the summaries generated by text summarization automatic machine (also known as a candidate summary) with a summary (ideally) made by other people (also known as a reference summary) originating from the same source.

In this final project, the author implements the development of Longest Common Subsequences (LCS) and Skip Bigram Co Occurrences Statistics (SB) methods in ROUGE techniques for evaluating multi-document text summarization automatic machine. The objective of this development of methods is to produce a way of evaluation that has a good level of corelation with the human judgment, because the information becomes more and more drawn.

Testing is done by using Pearson's Coefficient of Correlation to measure the linear relationship between the value given by the application and the average value of human judgment of the same summary. Test results showed that the evaluation by using LCS+ SB algorithm is capable of producing higher evaluation score than LCS or SB algorithms, with the correlation result that sometimes the is higher than the other two algorithms. In addition, the correlation can be better by adding the percentage of the number of used summary reference.

Keywords : Summarization Evaluation, Longest Common Subsequences, Skip Bigram Co Occurrences Statistics, ROUGE, Pearson's Coefficient of Correlation

1. Pendahuluan

1.1 Latar Belakang Masalah

Mengevaluasi ringkasan merupakan hal yang paling penting dalam perkembangan *Automatic Text Summarization* (ATS) [15]. Dewasa ini, telah banyak penelitian yang dilakukan untuk mengotomatisasi proses mengevaluasi ringkasan tersebut. Salah satu teknik mengevaluasi ringkasan yang telah menunjukkan keterhubungan yang baik dengan manusia adalah ROUGE.

ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) merupakan suatu teknik yang digunakan untuk menentukan kualitas dari mesin peringkas teks otomatis, dengan cara membandingkan ringkasan yang dihasilkan oleh mesin peringkas teks otomatis (*candidate summary*) dengan ringkasan (ideal) lain yang dibuat oleh manusia (*references summary*) [5].

ROUGE mengadopsi beberapa macam algoritma yang dapat digunakan untuk mencari kesamaan kata yang terdapat diantara ringkasan referensi dan ringkasan kandidat. Dua diantaranya adalah algoritma *Longest Common Subsequences* (LCS) dan algoritma *Skip Bigram Co Occurences Statistics* (SB). Berdasarkan [14], kedua algoritma ini memiliki kesamaan dalam proses pencocokannya, yaitu: tidak membutuhkan pencocokan yang berurutan, hanya menghitung kata-kata yang sama namun tetap mencerminkan urutan kata yang terdapat dalam kalimat tanpa harus memberikan standar panjang yang diperlukan.

Longest Common Subsequences (LCS) merupakan algoritma yang digunakan dalam teknik ROUGE, yang lebih dikenal sebagai ROUGE-L. Dimana metode ini memandang kalimat-kalimat penyusun ringkasan sebagai rangkaian kata-kata. Dan dalam pencocokannya kata-kata yang terdapat pada *candidate summary* dapat langsung dicocokkan dengan *references summary*. Namun, sesuai dengan namanya, algoritma LCS hanya menghitung kemiripan kata-kata yang terdapat pada rangkaian terpanjang saja, dan mengabaikan kata-kata yang mirip namun tidak berada pada rangkaian tersebut

Skip-Bigram Co-Occurrence Statistics (SB) merupakan algoritma yang digunakan dalam teknik ROUGE, yang lebih dikenal sebagai ROUGE-S. Dalam pencocokannya metode ini menggunakan seluruh pasangan kata-kata yang terdapat pada kalimat-kalimat yang ada pada *candidate summary* dengan *references summary*. Apabila dibandingkan dengan algoritma LCS, algoritma SB dapat menghitung semua urutan kata yang memiliki kesamaan, walaupun tidak sesensitif algoritma LCS.

Dengan adanya kelebihan dan kelemahan yang terdapat pada algoritma LCS dan SB dalam mengevaluasi mesin peringkas teks otomatis, maka berdasarkan [14] untuk memberikan nilai tambah terhadap kata-kata yang mirip namun tidak berada pada rangkaian LCS, dapat digunakan perhitungan dengan algoritma SB. Dengan harapan, melalui algoritma yang baru ini (algoritma LCS+SB), mampu menghasilkan suatu cara evaluasi yang memiliki tingkat keterhubungan yang baik dengan *human judgement*, karena informasi yang terambil menjadi semakin banyak.

1.2 Perumusan Masalah

Berdasarkan latar belakang tersebut, maka permasalahan yang dibahas adalah sebagai berikut:

1. Bagaimana cara mengimplementasikan algoritma LCS+SB dalam teknik ROUGE untuk mengevaluasi mesin peringkas teks otomatis.
2. Apakah algoritma LCS+SB dalam teknik ROUGE mampu menghasilkan metode yang lebih baik dalam mengevaluasi mesin peringkas teks otomatis, dengan melihat korelasinya terhadap penilaian manusia.
3. Faktor-faktor apa saja yang dapat mempengaruhi korelasi antara algoritma LCS+SB terhadap penilaian manusia seperti persentasi ekstraksi dan jumlah referensi.

Hipotesa awal dari penelitian ini adalah algoritma LCS+SB dalam teknik ROUGE dapat menghasilkan metode evaluasi yang lebih baik dalam mengevaluasi mesin peringkas teks otomatis.

Batasan masalah untuk penelitian ini adalah:

1. Dokumen yang dievaluasi adalah artikel berita berbahasa Indonesia dan Inggris.
2. Mesin peringkas teks otomatis yang digunakan sebanyak dua macam, yaitu: ATS Lexrank dan ATS Mead.
3. *Candidate* dan *reference summary* memiliki jumlah kalimat yang sama.
4. Untuk mengatasi ketidak-konsistenan manusia, maka maksimal *references summary* dan *human judgement* yang digunakan adalah sebanyak lima orang (dengan latar belakang pendidikan yang sama, yaitu: mahasiswa maupun alumni: IT TELKOM)
5. Output yang dihasilkan berupa nilai dari metode ROUGE yang digunakan.
6. Parameter yang dianalisis adalah faktor struktur informasi, cakupan informasi, keterhubungan informasi, dan redundansi informasi.

1.3 Tujuan

Tujuan dengan dilakukannya penelitian ini adalah:

1. Merancang dan membangun aplikasi yang menerapkan algoritma LCS+SB dalam teknik ROUGE untuk mengevaluasi mesin peringkas teks otomatis.
2. Menganalisis algoritma LCS+SB dalam teknik ROUGE untuk mengevaluasi mesin peringkas teks otomatis, dengan melihat korelasinya terhadap penilaian manusia.
3. Menganalisis faktor-faktor yang dapat mempengaruhi korelasi antara algoritma LCS+SB terhadap penilaian manusia seperti persentasi ringkasan dan jumlah referensi.

1.4 Metodologi Penyelesaian Masalah

Metodologi yang dilakukan untuk menyelesaikan permasalahan adalah sebagai berikut:

1. Studi literatur

   Melakukan pencarian serta mempelajari informasi dan pembelajaran tentang evaluasi mesin peringkas teks otomatis, khususnya mengenai konsep dan cara kerja ROUGE dalam mengevaluasi mesin peringkas teks otomatis.
2. Pengumpulan data-data

   Melakukan pencarian data yang digunakan untuk penelitian Tugas Akhir ini. Data yang dicari adalah artikel berita, baik artikel berita berbahasa Indonesia maupun berbahasa Inggris, yang nantinya diringkas dengan menggunakan mesin peringkas teks otomatis.
3. Analisis modifikasi ROUGE

   Melakukan analisis kemungkinan modifikasi atau pengembangan terhadap teknik ROUGE, berdasarkan kelebihan dan kekurangan yang terdapat pada metode ROUGE-L dan ROUGE-S.
4. Analisis dan perancangan aplikasi

   Melakukan analisis dan perancangan aplikasi teknik ROUGE berdasarkan algoritma LCS, SB, dan LCS+SB sehingga dapat digunakan untuk mengevaluasi mesin peringkas teks otomatis dengan pendekatan *summary-level*.
5. Implementasi aplikasi

   Melakukan implementasi aplikasi sesuai dengan hasil analisis dan perancangan dari teknik ROUGE berdasarkan algoritma LCS, SB, dan LCS+SB untuk mengevaluasi mesin peringkas teks otomatis dengan pendekatan *summary-level*.
6. Pengujian aplikasi

   Melakukan pengujian aplikasi dan menganalisis hasil keluaran aplikasi, sejauh mana hasil yang diberikan tersebut mampu menggambarkan kualitas dari mesin peringkas teks otomatis dengan melihat korelasinya terhadap penilaian manusia.
7. Pembuatan laporan Tugas Akhir

   Melakukan penyusunan laporan hasil penelitian yang telah dilakukan serta memberikan kesimpulan dari hasil penelitian tersebut.

5. Penutup

5.1 Kesimpulan

Berdasarkan analisis yang dilakukan terhadap hasil pengujian, diperoleh kesimpulan sebagai berikut:

1. Nilai evaluasi yang tinggi tidak menjamin mampu memberikan hasil korelasi yang tinggi dengan *human judgement*. Hal ini dikarenakan oleh faktor penilaian yang terdapat dalam *human judgement* seperti: struktur informasi, cakupan informasi, keterhubungan informasi dan redundansi informasi yang belum dapat diatasi oleh algoritma LCS, SB maupun LCS+SB.
2. Metode LCS+SB mampu menghasilkan nilai evaluasi yang lebih tinggi dibandingkan dengan metode LCS maupun SB dalam mengevaluasi mesin peringkas teks otomatis. Karena metode LCS+SB memberikan nilai tambah pada rangkaian kata yang sebelumnya diabaikan oleh metode LCS.
3. Korelasi dapat ditingkatkan dengan melakukan penambahan jumlah ringkasan referensi yang digunakan. Sedangkan dengan melakukan penambahan persentase ringkasan terhadap dokumen sumber tidak selalu dapat meningkatkan korelasi.

5.2 Saran

Berdasarkan hasil analisis dan kesimpulan, terdapat beberapa saran untuk perbaikan pada penelitian peringkasan teks sebagai berikut:

1. Mencoba menggunakan proses *preprocessing* data, seperti *stemming*, dalam mengevaluasi mesin peringkas teks otomatis.
2. Memperhitungkan urutan kalimat yang terdapat pada ringkasan dengan memberikan informasi tambahan mengenai id kalimat pada ringkasan.
3. Memberikan bobot pada token, untuk memastikan kata yang dipilih memiliki tingkat kepentingan yang tinggi.
4. Memperbanyak jumlah dan menggunakan *human judgement* dari kelompok dengan latar belakang pendidikan yang memiliki pengetahuan khusus mengenai kebahasaan, seperti, ahli bahasa.

Daftar Pustaka

[1] ___. Automatic summarization. http://en.wikipedia.org/wiki/Automatic_summarization.htm, didownload pada tanggal 20 Februari 2009.

[2] ___. *Correlation Coefficient*. http://mathbits.com/Mathbits/TISection/Statistics2/correlation.htm, didownload pada tanggal 20 Januari 2010.

[3] ___. Pearson product-moment correlation coefficient. http://en.wikipedia.org/wiki/Pearson_correlation, didownload pada tanggal 20 Februari 2009.

[4] Bonnie Dorr, Christof Monz, Stacy President, Richard Schwartz, and David Zajic. 2005. A Methodology for Extrinsic Evaluation of Text Summarization: Does ROUGE Correlate? In *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 1–8, Ann Arbor, Michigan. Association for Computational Linguistics.

[5] Chin-Yew Lin. 2004a. Looking for a few good metrics: Automatic summarization evaluation – how many samples are enough? In *Proceedings of NTCIR Workshop 4*, Tokyo, Japan.

[6] Chin-Yew Lin. Looking for a few good metrics: Rouge and its evaluation. In *Proceedings of the NTCIR Workshop 4*, Japan, 2004.

[7] City University of Hong Kong. 1998. *A Syntactic Marker-based Multi-level Discourse Analyzer with application to Chinese Full-text Abstraction*. [online]. (http://www.rcl.cityu.edu.hk/research/rp28.php, diakses tanggal 20 januari 2010)

[8] Donna Harman and Paul Over. 2004. The effects of human variation in duc summarization evaluation. In *Proceedings of the ACL-04 Workshop: Text Summarization Branches Out*, pages 10–17, Barcelona, Spain.

[9] Douthat, A. 1998. *The Message Understanding Conference Scoring Software User's Manual*. Proceedings of the 7th Message Understanding Conference (MUC-7).

[10] Eskanaluwa, Marissa Nur. 2009. *Analisis dan Implementasi Metoda BLEU (BiLingual Evaluation Understudy) dalam Mengevaluasi Hasil Terjemahan Machine Translation*. IT TELKOM.

[11] Hassel, Martin. 2007. *Resource Lean and Portable Automatic Text Summarization***.** Doctoral Thesis. KTH School of Computer Science and Communication. Stockholm, Sweden.

[12] Lin, C.Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, post-conference workshop of ACL 2004, Barcelona, Spain.

[13] Lin, C.-Y. and E. H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of 2003 Language Technology Conference* (HLT-NAACL 2003), Edmonton, Canada.

[14] Lin, C.-Y. and F. J. Och. 2004. Automatic evaluation of machine

translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of 42nd Annual Meeting of ACL* (ACL 2004), Barcelona, Spain.

[15] Liu, Feifan, and Y. Liu. "Correlation between ROUGE and Human Evaluation of Extractive Meeting Summaries." *Proceedings of ACL-HLT*. Columbus, OH, 2008.

[16] Purwasih, Nurzaitun. 2009. *Peringkasan Teks Otomatis Dokumen Tunggal Berbahasa Indonesia Menggunakan Graph-based Summarization Algorithm dan Similarity (Studi Kasus Artikel Berita)*. IT TELKOM.

[17] Radev, D., and D. Tam. 2003. Summarization evaluation using relative utility. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, 508-511.

[18] Sjöbergh, Jonas. Older versions of the ROUGEeval summarization evaluation system were easier to fool. *Information Processing and Management*, 2007, 43(6): 1500-1505.