ROUGE-NW SEBAGAI PENGGABUNGANROUGE-N DAN ROUGE-W UNTUK MENGEVALUASI AUTOMATIC TEXT SUMMARIZATION (ATS)

Ali Marjan[1], Moch Arif Bijaksana[2], Warih Maharani[3]

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

Abstrak

Mengevaluasi ringkasan merupakan salah satu permasalahan yang penting dalam perkembangan Automatic Text Summarization (ATS). Evaluasi ini merupakan suatu proses yang bertujuan untuk menilai suatu sistem peringkas teks otomatis yang dilihat dari hasil atau keluaran suatu sistem peringkas teks otomatis tersebut.

Pada Tugas Akhir ini diimplementasikan pengembangan variasi metoda ROUGE (Recall-Oriented Understudy for Gisting Evaluation) yang merupakan metoda untuk mengevaluasi sistem peringkasan teks otomatis (ATS), yaitu ROUGE-NW yang merupakan penggabungan dari metode ROUGE-1(n-gram 1) dengan metode ROUGE-W. ROUGE merupakan suatu teknik evaluasi yang digunakan untuk menentukan kualitas dari sebuah ringkasan secara otomatis dengan cara membandingkan ringkasan yang dihasilkan oleh ATS dengan ringkasan (ideal) lain yang dibuat oleh manusia [5]. Selain itu, sebagai pembanding diterapkan juga metode asli ROUGE yaitu ROUGE-1, ROUGE-2, dan ROUGE-W.

Pengujian dilakukan dengan melihat korelasi atau keterhubungan score yang dihasilkan oleh masing-masing metode ROUGE dengan penilaian yang dihasilkan oleh manusia(human judgement). Hasil pengujian menunjukkan bahwa perhitungan score ROUGE-2 dan ROUGE-NW cenderung memiliki korelasi yang lebih tinggi dengan penilaian manusia dibandingkan dengan metode ROUGE-1 dan ROUGE-W.

Kata Kunci : evaluasi, metode ROUGE, n-gram(s), korelasi

Abstract

Evaluation of summaries is an important problem in development of automatic text summarization (ATS). Evaluation is a process that purpose to assess a automatic text summarization based on its output.

In this assignment implemented method variation ROUGE (Recall-Oriented Understudy for Gisting Evaluation) which is a method for evaluating automatic Text Summarization system (ATS), that is method of ROUGE-NW representing merger of method of ROUGE-1(N-GRAM 1) with method of ROUGE-W. ROUGE is an evaluation technique used to determine the quality of an automatic summary by comparing the summary produced by the ATS with another summary (ideal) created by human[5]. Moreover, as comparator applied also original method of ROUGE that is ROUGE-1, ROUGE-2, and ROUGE-W.

Evaluation is done by looking at the correlation or the association of score generated by each assessment method ROUGE with the produced by human (human judgment). The result of this experiment shows that the computation of ROUGE-2 and ROUGE-NW has more correlation with human judgment rather than ROUGE-1 and ROUGE-W.

Keywords : evaluation, ROUGE method, n-gram(s), correlation

1. Pendahuluan

1.1 Latar Belakang

S*ummary* atau ringkasan didefinisikan sebagai sebuah teks yang dihasilkan dari satu atau lebih teks, mengandung informasi dari teks sumber dan panjangnya tidak lebih dari setengah teks sumber. Kita mengetahui bagaimana membuat ringkasan dan mengetahui seperti apakah seharusnya sebuah ringkasan yang baik. Relitasnya setiap orang dapat meringkas dan setiap orang mempunyai ringkasan yang menurut mereka baik. Masalah yang dihadapi saat ini adalah apakah mungkin evaluasi ringkasan tersebut dinilai objektif, jika setiap orang mempunyai anggapan ringkasan sendiri mengenai ringkasan yang baik[11].

Awalnya evaluasi peringkasan teks dilakukan secara manual. Walaupun saat ini masih dilakukan tetapi perlahan-lahan digantikan dengan proses *automatic*. Alasan untuk mengotomatisasi metode evaluasi peringkasan teks adalah jika menggunakan evaluasi secara manual akan menghabiskan waktu yang lama, mengeluarkan dana yang mahal, dan hasilnya tidak dapat digunakan kembali. Sedangkan metode evaluasi peringkasan teks otomatis tidak mengeluarkan biaya yang banyak, berlangsung cepat, tidak tergantung pada bahasa dan hubungan dengan pendapat manusia[1].

Dalam perkembangan *Automatic Text Summarization* (ATS), salah satu hal yang penting adalah mengevaluasi ringkasan [12]. Saat ini, telah banyak penelitian yang dilakukan untuk mengautomatisasi proses mengevaluasi ringkasan tersebut. Salah satu teknik mengevaluasi ringkasan yang telah menunjukkan keterhubungan yang baik dengan manusia adalah ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*).

ROUGE merupakan suatu teknik evaluasi yang digunakan untuk menentukan kualitas dari sebuah ringkasan secara automatis dengan cara membandingkan ringkasan yang dihasilkan oleh ATS dengan ringkasan (ideal) lain yang dibuat oleh manusia [8]. ROUGE menggunakan ukuran yang berbeda-beda. ROUGE juga telah digunakan pada *Document Understanding Conference* (DUC) sebagai sebuah metric evaluasi untuk mengevaluasi peringkasan teks. DUC menyediakan dataset dalam jumlah yang banyak dengan berbagai macam ringkasan yang dibuat oleh manusia pada tiap-tiap dokumennya yang kemudian dibandingkan dengan ringkasan yang dibuat menggunakan mesin. ROUGE memiliki lima macam ukuran berbeda, yaitu: ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S, ROUGE-SU [10]. Pada tugas akhir ini, menerapkan metode ROUGE-N, ROUGE-W dan pengembangan variasi ROUGE dengan menggabungkan dua metode tersebut.

N-gram Co-Occurrence Statistics dalam teknik ROUGE disebut juga sebagai ROUGE-N. Dimana metode ini menggunakan setiap kata yang cocok antara *candidate summary* (ringkasan yang dihasilkan ATS) dengan *references summary* (ringkasan ideal yang dibuat oleh manusia) sebagai ukurannya dan tidak memperhatikan urutan kata-kata yang terdapat pada ringkasan.

Weighted Longest Common Subsequence Statistics dalam teknik ROUGE disebut juga sebagai ROUGE-W. Metode ini merupakan perkembangan dari metode ROUGE-L dengan memandang kalimat-kalimat yang ada pada *candidate*

summary sebagai rangkaian kata-kata yang nantinya akan dicocokkan pada *references summary* menggunakan dynamic programming.

Gagasan utama dalam tugas akhir ini adalah bagaimana cara mengevaluasi ringkasan dengan mempertimbangkan *score* variasi ROUGE. Dalam tugas akhir ini, akan dihitung *score* ROUGE-N, ROUGE-W, dan perpaduan antara ROUGE-N dan ROUGE-W untuk menentukan kualitas dari sebuah ringkasan secara automatis dengan cara membandingkan ringkasan yang dihasilkan oleh ATS dengan ringkasan (ideal) lain yang dibuat oleh manusia. Alasan pemilihan metode pengembangan dengan menggabungkan dua metode ROUGE yaitu ROUGE-N dan ROUGE-W adalah karena pada metode ROUGE-N hanya membandingkan kata-kata yang sama antara kandidat ringkasan dengan referensi ringkasan dan urutan kesamaan katanya tidak berpengaruh, sedangkan pada metode ROUGE-W perhitungan dilakukan dengan cara mencocokkan urutan rangkaian kata yang sama dengan bobot tertentu antara kandidat ringkasan dengan referensi ringkasan sedangkan kata yang sama dan tercecer pada rangkaian kata tidak diperhitungkan.

Score ROUGE dapat dengan mudah dihitung untuk beberapa system atau untuk versi baru didalam pengembangan system ATS. Ini membuat ROUGE menjadi sebuah alat (*tool*) yang sangat berguna untuk pengembangan system peringkasan teks otomatis. *Score* ROUGE juga dapat menunjukkan kecocokan antara *Automatic Text Summarization* dengan *human judgments* [14].

1.2 Perumusan Masalah

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

a. Bagaimana cara mengimplementasikan variasi dari teknik ROUGE dari penggabungan antara metode *N-gram Co-occurrence Statistics* dan *Weighted Longest Common Subsequence Statistics*.
b. Bagaimana pengaruh variasi teknik ROUGE dari penggabungan antara metode *N-gram Co-occurrence Statistics* dan *Weighted Longest Common Subsequence Statistics* dalam mengevaluasi *Multidocument Summarization.*
c. Apakah pengembangan variasi teknik ROUGE tersebut akan menghasilkan metode yang lebih baik dalam mengevaluasi *Multidocument Summarization*, dilihat dari korelasinya terhadap penilaian manusia.

Batasan masalah untuk penelitian ini adalah :

1. Dokumen berupa multi-dokumen *summary*.
2. Dokumen yang dievaluasi adalah ringkasan berita berbahasa Indonesia dan Inggris.
3. Mesin peringkas teks otomatis yang digunakan dua macam.
4. Untuk mengatasi ketidak-konsistenan manusia, maka *reference summary* yang digunakan sebanyak lima buah *reference*. Berdasarkan Nenkova and Passonneau (2003) 5 referensi sudah cukup mewakili.
5. *Candidate* dan *reference summary* memiliki jumlah kalimat yang sama.
6. Pada proses *proprocessing* data hanya menggunakan *stopwords*.
7. Output yang dihasilkan berupa *score* berdasarkan metode ROUGE yang digunakan.

8. Parameter ROUGE yang dianalisis adalah *recall, precision*, dan *F-Measure.*
9. Menggunakan *human judgement* dalam mengevaluasi metode ROUGE yang digunakan.

1.3 Tujuan

Tujuan dari dilakukannya penelitian ini adalah:

a. Mengimplementasikan metode ROUGE asli untuk mengevaluasi *multi-document Summarization.*
b. Mengimplementasikan pengembangan variasi dari teknik ROUGE dengan menggabungkan antara metode *N-gram Co-occurrence Statistics* dan *Weighted Longest Common Subsequence Statistics*.
c. Menganalisis pengaruh variasi teknik ROUGE dari penggabungan antara metode *N-gram Co-occurrence Statistics* dan *Weighted Longest Common Subsequence Statistics* dalam mengevaluasi *multi-document summarization*.
d. Menganalisis pengembangan variasi teknik ROUGE tersebut dalam mengevaluasi *multi-document summarization* dibandingkan dengan metode ROUGE asli, dilihat dari tingkat korelasinya dengan *human judgement*.

1.4 Metodologi

Adapun metodologi yang digunakan dalam penulisan Tugas Akhir ini adalah sebagai berikut:

a. Melakukan studi literatur khususnya mengenai evaluasi peringkasan teks otomatis, *Recall-Oriented Understudy for Gisting Evaluation, N-gram Co-Occurrence Statistics,* dan *Weighted Longest Common Subsequences.*
b. Menganalisa kekurangan dan kelebihan dari Metode *N-gram Co-occurrence Statistics* dan *Weighted Longest Common Subsequence Statistics* untuk Mengevaluasi Multi-Document Summarization serta kemungkinan variasi dari penggabungan kedua metode tersebut.
c. Menganalisa kebutuhan perangkat lunak dan merancang perangkat lunak yang mampu mengimplementasikan variasi metode *Rouge* dari penggabungan Metode *N-gram Co-occurrence Statistics* dan *Weighted Longest Common Subsequence Statistics* untuk Mengevaluasi Multi-Document Summarization.
d. Melakukan implementasi perangkat lunak sesuai dengan perancangan yang telah dilakukan.
e. Melakukan pengujian pada implementasi variasi metode Rouge dari penggabungan metode *N-gram Co-occurrence Statistics* dan *Weighted Longest Common Subsequence Statistic.*
f. Menganalisis hasil pengujian.
g. Mengambil kesimpulan dari hasil analisis yang telah dilakukan, serta mendokumentasikanhasil perancangan, implementasi, pengujian, dan analisis kedalam suatu bentuk laporan yang telah disusun sejak awal.

5 Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan pengujian dan analisis yang telah dibahas dan dilaksanakan pada bab 4, maka dapat disimpulkan bahwa dari perhitungan korelasi dengan *Pearson's*, ROUGE-2 dan ROUGE-NW cenderung memiliki korelasi yang lebih tinggi terhadap penilaian manusia dibandingkan dengan ROUGE-1 dan ROUGE-W. Selain itu juga didapatkan s*core* dari ROUGE-2 dan modifikasi variasi penggabungan ROUGE yaitu ROUGE-NW, cenderung memiliki *score* yang lebih tinggi dibandingkan *score* ROUGE-1 dan ROUGE-W. Hal ini dikarenakan pada metode ROUGE-NW tidak hanya menghitung kesamaan kata yang terdapat pada rangkaian kata WLCS saja tetapi memperhitungkan kesamaan kata yang tidak terdapat pada rangkaian kata WLCS.

5.2 Saran

Berikut ini saran-saran yang dapat dipertimbangkan untuk pengembangan Tugas Akhir di masa mendatang:

1. Data uji (baik kandidat maupun referensi) agar diperbanyak jumlahnya.
2. Mencoba menambahkan variasi jumlah referensi pada setiap percobaan.
3. Mencoba menggabungkan metode variasi ROUGE lainnya untuk mendapatkan korelasi yang tinggi terhadap penilaian manusia.
4. Mencoba menggunakan *stemming* dalam *preprocessing* data.

Daftar Pustaka

[1] Bildner, Sebastian, 2009, *Automatic Evaluation Metrics for Summarisation and Machine Translation*, Workshop On Motivation Introducing Two Methods for Evaluation Evaluation of Metrics, Dikutip: 23 Maret 2009, [online].

[2] Bonnie J. Dorr, Christof Monz, Stacy President, Richard Schwartz, and David Zajic, 2005, "A Methodology for Extrinsic Evaluation of Text Summarization: Does ROUGE Correlate?", In Proceedings of the Association for Computational Linguistics Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, Ann Arbor, MI, pp. 1--8, Dikutip: 15 Desember 2008, [online]. Available: http://www.aclweb.org/anthology/W/W05/W05-0901.pdf

[3] Chin-Yew Lin, Eduard Hovy, Liang Zhou, and Junichi Fukumoto. *"Automated Summarization Evaluation with Basic Elements"*. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy, May 24 - 26, 2006. Dikutip: 19 November 2009, [online]. Available:

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.67.7845&rep=rep1&type=pdf

[4] Computing Pearson's Correlation Coefficient, http://davidmlane.com/hyperstat/A51911.html, Dikutip: 25 Maret 2009 [Online].

[5] Gildea, Daniel & Liu, Ding, 2006. *"Stochastic Iterative Alignment for Machine Translation Evaluation"*, In Proceedings of the COLING/ACL on Main conference, Dikutip: 02 Januari 2009, [online]. Available: http://www.cs.rochester.edu/~gildea/pubs/liu-gildea-acl06.pdf

[6] Harman, Donna & Over, Paul, 2004. *"The Effects of Human Variation in DUC Summarization Evaluation"*, Workshop On Text Summarization Branches Out, Dikutip: 20 Desember 2008, [online]. Available: http://www.aclweb.org/anthology/W/W04/W04-1003.pdf

[7] Lin, Chin-Yew & Hovy, Eduard, "Automatic Evaluation Of Summaries Using N-Gram Co-Occurrence Statistics", In Proceedings of the 2003 Conference of the North American. Dikutip: 20 Desember 2008, [online]. Available: http://www.isi.edu/natural-language/people/hovy/papers/03HLT-NAACL-ROUGE-eval.pdf

[8] Lin, Chin-Yew. *"Looking for a Few Good Metrics: Automatic Summarization Evaluation - How Many Samples Are Enough?"*. In Proceedings of the NTCIR Workshop 4, Tokyo, Japan, June 2 - June 4, 2004. Dikutip: 15 Desember 2008, [online]. Available: http://research.microsoft.com/en-us/people/cyl/ntcir4.pdf

[9] Lin, Chin-Yew, *"Looking for a few good metrics: ROUGE and its evaluation"*, In: Proc. Working Notes of NTCIR-4 (vol. Supl. 2). Dikutip: 15

Desember 2008, [online]. Available: http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/OPEN/OPENSUB_Chin-Yew_Lin.pdf

[10] Lin, Chin-Yew. 2004. *"ROUGE: a Package for Automatic Evaluation of Summaries"*. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, July 25 - 26, Dikutip: 20 Desember 2008, [online] Available: belobog.si.umich.edu/clair/anthology/query.cgi?type=Paper&id=W04-1013.

[11] Lin, Chin-Yew. *"ROUGE: a Package for Automatic Evaluation of Summaries"*. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, July 25 - 26, 2004. Dikutip: 20 Desember 2008, [online] Available: www.law.kuleuven.ac.be/icri/conferences/Lin.pdf

[12] Liu, Feifan & Liu, Yang, *"Correlation between ROUGE and Human Evaluation of Extractive Meeting Summaries"*, In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies. Dikutip: 20 Desember 2008, [online]. Available: www.aclweb.org/anthology-new/P/P08/P08-2051.pdf

[13] Radev, Dragomir R. & Tam, Daniel, *"Summarization evaluation using relative utility"*, In Proceedings of the twelfth international conference on Information and knowledge management. Dikutip: 19 November 2009, [online]. Available: http://portal.acm.org/citation.cfm?id=956863.956960

[14] Sj¨obergh, Jonas, *"Older Versions of the ROUGEeval Summarization Evaluation System were Easier to Fool"*, Pergamon Press, Inc. Tarrytown, NY, USA. Dikutip: 20 Desember 2008, [online]. Available: http://dr-hato.se/research/abuserouge.pdf

[15] Zajic, David M. *"Multiple Alternative Sentence Compressions (MASC) as a Tool for Automatic Summarization Tasks,"* Ph.D. Thesis, University of Maryland, College Park, 2007. Dikutip: 23 Maret 2009, [online]. Available:

http://www.umiacs.umd.edu/~dmzajic/dissertation/diss.pdf

[16] Ziheng, Lin, *"Lead Classification for Automatic Text Summarization"*, In Undergraduate Research Opportunity Program (UROP). Dikutip: 23 Maret 2009, [online]. Available:

http://aye.comp.nus.edu.sg/publications/theses/zihengLinUROPThesis.pdf