Jurnal Ilmiah Informatika (JIF)
| ISSN (Print) 2337-8379 | ISSN (Online) 2615-1049 |
Jurnal online di akses di http://ejournal.upbatam.ac.id/index.php/jif

Segmentasi Dokumen Teks Dengan Metode Texttiling

Chintalya Magdalena[a)], *Bangun Hyolister Tambun*[b)]

[a)] [b]*Universitas Kristen Indonesia, www.uki.ac.id, Indonesia*

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 17 Oktober 2021
Revisi Akhir: 08 Januari 2022
Diterbitkan *Online*: 01 Maret 2022

KATA KUNCI

Segmentasi, Dokumen Teks, Metode, TextTiling

KORESPONDENSI

E-mail: magdalenachintalya@gmail.com

A B S T R A C T

In this paper, we will report our work on text segmentation on Indonesian speech documents. As a result of using Automatic Speech Recognition (ASR), the speech documents are transcribed into the text without any boundary for each document. The documents are certainly needed to be segmented regarding to its topics. We apply TextTiling method with various term weighted techniques such as TF-IDF, TF-IDF-Mutual Information, TF-IDF Mutual Information-Word Similarity, and TF-IDF-Word Frequency for measuring the similarity between segments. The result show TF-IDF-Mutual Information performed better in most of the collections.

1. PENDAHULUAN

Sejalan dengan perkembangan komponen-komponen elektronik. arsitektur komputer (prosesor) dan multimedia, maka berkembang pula topik-topik penelitian yang berhubungan dengan pengelolaan data multimedia. Saat ini, informasi yang tersedia, tidak hanya disimpan dalam bentuk dokumen teks, tetapi juga dapat dijumpai dalam bentuk sinyal wicara.

Informasi pada dokumen sinyal wicara dapat diubah menjadi informasi berbentuk dokumen teks, dengan menggunakan perangkat lunak *Automatic Speech Recognition* (ASR). Hasil keluaran dari ASR berupa dokumen teks yang tidak memiliki batasan akhir dan tidak tersegmentasi secara jelas, tentu menyulitkan dalam pengolahan data teks tersebut. Permasalahan ini merupakan salah satu alasan diperlukannya penelitian yang berhubungan dengan segmentasi dokumen teks.

Pada penelitian tentang segmentasi ini, hasil transkripsi berita suara diperoleh dengan mempergunakan ASR *Sphinx*, karena perangkat lunak ini telah digunakan untuk mentranskripsikan data suara berbahasa Indonesia menjadi dokumen teks. Metode segmentasi yang digunakan adalah jenis metode segmentasi dengan *lexical cohesion* yaitu metode *TextTiling*. Pada penelitian ini, metode pembobotan pada metode *TextTiling* akan dimodifikasi untuk meningkatkan hasil ketepatan segmentasi.

Masalah yang melandasi penelitian ini adalah hasil transkripsi yang dilakukan oleh *Automatic Speech Recognition* (ASR), pada umumnya tidak memiliki batas-batas akhir topik yang jelas, sehingga dibutuhkan teknik khusus untuk dapat melakukan segmentasi terhadap dokumen teks hasil transkripsi tersebut. Dalam rangka mengatasi masalah tersebut, maka pada penelitian ini akan dilakukan pengujian beberapa cara pembobotan pada metode segmentasi *TextTiling* untuk mendapatkan metode segmentasi dokumen teks hasil transkripsi berbahasa Indonesia yang terefektif.

Tujuan utama dalam penelitian ini adalah mendapatkan cara pembobotan yang efektif dalam melakukan segmentasi dokumen teks dengan metode *TextTiling*, khususnya untuk dokumen teks hasil transkripsi berita suara berbahasa Indonesia.

2. TINJAUAN PUSTAKA

2.1 *Topic Detection and Tracking* (TDT)

Topic Detection and Tracking adalah suatu penelitian yang difokuskan pada pengolahan hasil transkripsi berita suara. Penelitian ini timbul sebagai akibat adanya perkembangan teknologi pengenalan wicara yang pesat, sehingga dapat diperoleh transkripsi dari berita suara dengan tingkat akurasi pengenalan yang tinggi. Motivasi awal dan tujuan dari penelitian mengenai *Topic Detection and Tracking* berita suara adalah mendapatkan cara untuk memonitor berita suara dan memberitahukan kepada para analis tentang suatu kejadian baru dan menarik yang terjadi di dunia.

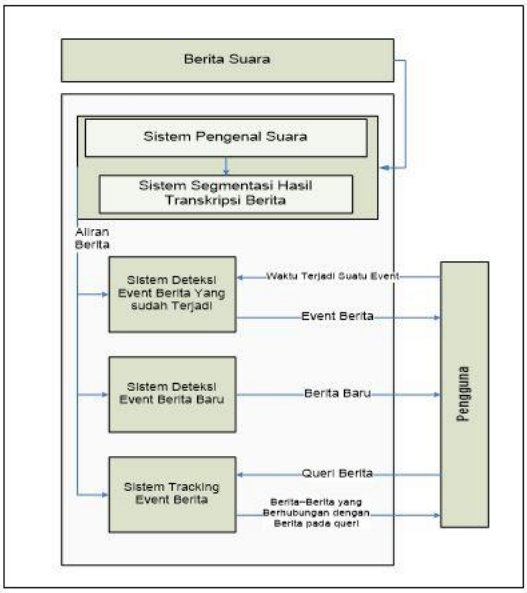
2.2 *Task pada Topic Detection and Tracking* (TDT)

Pada TDT istilah topik, *event* (kejadian) dan *activity* (aktifitas) umum digunakan. Pengertian dari ketiga istilah itu sesuai dengan TDT *Evaluation* pada tahun 1998 adalah sebagai berikut. *Event* (kejadian) adalah sesuatu kejadian yang terjadi pada waktu dan tempat tertentu. *Activity* (aktivitas) adalah serangkaian kegiatan yang memiliki tujuan dan fokus yang sama. Sedangkan topik adalah keseluruhan *event* atau *activity* yang saling berhubungan dan sedang berkembang.

Ada beberapa kegiatan (*task*) yang dimasukkan dalam topik TDT, kegiatan ini adalah :

1. *Story Segmentation*: Kegiatan yang melakukan proses segmentasi transkripsi berita suara menjadi berita-berita yang berbeda.
2. *Detection*: Kegiatan yang difokuskan untuk mengidentifikasikan keserupaan dan ketidakserupaan sebuah berita baru terhadap topik-topik berita yang sudah ada. Kegiatan *Detection* ini dapat dibagi menjadi 3 bagian yaitu :
   a. *First Story Detection* atau *New Event Detection*: pada task ini, akan dilakukan deteksi awal suatu kejadian dari hasil transkripsi berita suara.
   b. *Cluster Detection*: pada kegiatan ini akan dilakukan pengelompokkan hasil transkripsi berita berdasarkan topik-topik berita yang ada.
   c. *Story Link Detection*: proses pendeteksian keterkaitan antara beberapa hasil transkripsi berita suara yang dipilih secara acak.
3. *Tracking*: kegiatan yang melakukan proses pelacakan berita dari hasil transkripsi berita suara.

Hubungan antara kegiatan pada proses *detection* ini dapat dilihat pada arsitektur sistem deteksi dan pelacakan berita suara pada gambar 1.



Gambar 1. Arsitektur Sistem Topik *Detection and Tracking*

Algoritma *TextTiling* untuk menemukan struktur subtopik memiliki tiga bagian utama, yaitu [Hearst,1997] :

1. *Tokenization*
2. *Lexical Score Determination*
3. *Boundary Identification*

Pada bagian *Tokenization*, kata-kata pada dokumen teks akan diteliti. Bila kata-kata tersebut termasuk jenis kata yang sering muncul dan tidak memiliki makna penting (*stopword*s) maka, kata-kata tersebut tidak akan diikutsertakan pada proses tahap berikutnya. Kata-kata yang tidak termasuk *stopword*s akan dikelompokkan dan dianggap sebagai sebuah kalimat semu (*pseudosentence*). Pada tahap penentuan nilai leksikal (*Lexical Score Determination*) yang menunjukkan kedekatan antar segmen, maka terdapat dua metode yang dapat digunakan. Metode-metode tersebut antara lain : metode *block comparison* dan metode *vocabulary introduction*. Pada metode *block comparison*, blok yang terletak bersebelahan akan dibandingkan untuk melihat seberapa banyak kata yang sama di kedua blok tersebut, sedangkan pada metode *vocabulary introduction*, perbandingan antara blok yang berdekatan didasarkan pada jumlah kata-kata baru yang timbul di kedua blok.

Pada metode *block comparison*, nilai leksikal berupa *similarity* antara dua segmen (b_1 dan b_2) dirumuskan sebagai berikut :

$$score(i) = \frac{\sum_t w_{t,b1} w_{t,b2}}{\sqrt{\sum_t w^2_{t,b1} \sum_t w^2_{t,b2}}} \qquad (1)$$

Dimana :

$w_{t,b}$ = frekuensi kata t pada segmen b

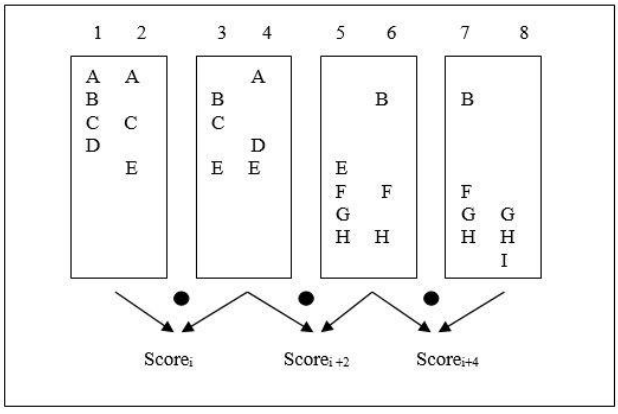
Nilai leksikal pada metode *vocabulary introduction*, berupa jumlah kemunculan kata-kata baru pada segmen, didapatkan dengan menghitung jumlah kemunculan kata pada segmen (b_1) dan (b_2) dibagi dengan dua kali jumlah seluruh kata pada kedua segmen tersebut, sehingga memenuhi persamaan 2 berikut :

$$score(i) = \frac{\sum katabaru_{b_1} + \sum katabaru_{b_2}}{w * 2} \qquad (2)$$

dimana :

b_1, b_2 = segmen-segmen yang berdekatan

w = jumlah seluruh kata pada segmen b_1, b_2



Gambar 2. Penentuan Nilai Leksikal *Score*

Gambar 2. diatas, menunjukkan proses penghitungan nilai leksikal antar blok pada metode *TextTiling*. Pada gambar tersebut, variabel *A* sampai dengan variabel *I* adalah variabel unit leksikal yang berisi kata-kata yang terdapat pada dokumen teks. Pada gambar tersebut juga dapat dilihat bahwa ukuran blok yang

digunakan adalah k = 2, sehingga masing-masing blok (*window*) akan berisi 2 *pseudosentence* atau 2 *token sequences*. Penghitungan nilai leksikal (*Lexical Score*) pada metode ini, akan dilakukan antara 2 blok yang berdekatan. Besar nilai *score* ini dihitung menggunakan persamaan 1 untuk metode *block comparison* dan persamaan 2 untuk metode *Vocabulary Introduction*.

2.3 Perolehan Informasi dan Metode Segmentasi Teks Berbahasa Indonesia

Sebuah sistem Perolehan Informasi pada dasarnya adalah sebuah sistem yang dapat membantu penggunanya untuk mendapatkan informasi yang diinginkan. Pada konteks diatas, maka sebuah sistem perolehan informasi *(Information Retrieval System)* dikatakan baik, bilamana informasi yang diperoleh dari sistem ini relevan dengan *query* penggunanya. Pada konteks itu pula, maka sebuah sistem perolehan informasi berbeda dengan bahasa perolehan data seperti *regular expression*.

Perkembangan teknologi komputer khususnya dibidang Multimedia, pada akhirnya juga mendorong Sistem Perolehan Informasi untuk berinteraksi dan memberikan informasi yang berhubungan dengan data multimedia seperti data suara, data image dan data biner lainnya. Pada kerangka itu, maka penelitian tentang segmentasi topik hasil transkripsi berita suara juga merupakan bagian dari sebuah sistem perolehan informasi dan merupakan salah satu perwujudan interaksi antara perolehan informasi dengan data multimedia. Pengukuran similarity antar dokumen merupakan bagian penting didalam sistem perolehan informasi karena menunjukkan hubungan antar dokumen teks . jenis-jenis pengukuran similarity antar dokumen dengan query, khususnya pada model *Vector Space Model* IR dapat dilihat pada tabel berikut.

Tabel 2. Jenis – jenis Similarity pada Vector Space Modell IR

Jenis Similarity	*Pengukuran Similarity*
Cosine Similarity	$Sim(Q,D_i)=\frac{\sum_{j=1}^{t} w_{qj} d_{ij}}{\sqrt{\sum_{j=1}^{t}(d_{ij})^2 \sum_{j=1}^{t}(w_{qj})^2}}$
Dice Coefficient Similarity	$Sim(Q,D_i)=\frac{2\sum_{j=1}^{t} w_{qj} d_{ij}}{\sqrt{\sum_{j=1}^{t}(d_{ij})^2 \sum_{j=1}^{t}(w_{qj})^2}}$
Jaccard Similarity	$Sim(Q,D_i)=\frac{\sum_{j=1}^{t} w_{qj} d_{ij}}{\sqrt{\sum_{j=1}^{t}(d_{ij})^2 +\sum_{j=1}^{t}(w_{qj})^2 -\sum_{j=1}^{t}(d_{ij})^2}}$

Dimana :

w_{qj} = bobot kata j pada query

d_{ij} = bobot kata j pada dokumen i

Q = vektor bobot kata pada query

D_i = vektor bobot kata pada dokumen i

3. METODOLOGI

Proses metodologi yang dilakukan, dibagi menjadi beberapa tahapan proses yaitu tahapan proses transkripsi, dan tahapan segmentasi dokumen teks hasil transkripsi.

Tahapan proses segmentasi akan menggunakan hasil dokumen teks yang dihasilkan oleh proses transkripsi dan membaginya menjadi dokumen-dokumen teks yang memiliki topik berbeda.

Pada tahapan segmentasi ini, metode yang digunakan metode *TextTiling* berbasis perbandingan blok dan berbasis *Vocabulary Introduction*.

Pada tahapan segmentasi dengan metode *TextTiling* berbasis perbandingan blok akan dilakukan eksperimen dengan model pembobotan *TF-IDF-Mutual Information, TF-IDF (World Frequency), TF-IDF-Mutual Information Word Similarity, TF-IDF dan Latent Semantic Analysis* (LSA). Pada bagian akhir eksperimen, seluruh hasil proses segmentasi dari metode *TextTiling* berbasis perbandingan blok dan berbasis *Vocabulary Introduction* akan dievaluasi dengan menghitung nilai *precision-recall* masing masing metode segmentasi.

Proses Segmentasi *TextTiling* dengan teknik pembobotan *TF-IDF-Mutual Information-Word Similarity* dilakukan dengan langkah-langkah sebagai berikut :

1. Hasil transkripsi akan dibagi menjadi beberapa segmen.
2. *TF,IDF,* dan *Mutual Information* dihitung
3. Matriks *Word Distance* dihitung
4. Bobot awal kata dari segmen akan dihitung dengan model *TF-IDF-Mutual Information.*
5. *Similarity* antar segmen yang berurutan dihitung.

6. Batas Segmentasi ditemukan bilamana, nilai *similarity* antara dua segmen berurutan lebih kecil dari pada nilai *similarity* dua segmen sebelum dan lebih kecil dari pada nilai *similarity* dua segmen sesudahnya.

4. HASIL DAN PEMBAHASAN

Tabel 3, dan 4 serta gambar 4 dan gambar 5 menampilkan seluruh hasil akurasi pengenalan wicara dan ketepatan proses segmentasi berupa *precision* dan *recall* proses segmentasi. Hasil pada tabel 3 dan tabel 4 menunjukkan, bahwa metode segmentasi yang konsisten berada di 3 urutan teratas perolehan nilai *precision* dan *recall* adalah metode segmentasi dengan pembobotan *TF-IDF-Mutual Information*, *TF-IDF-Mutual Information Word Similarity* dan *TF-IDF-Word Frequency*.

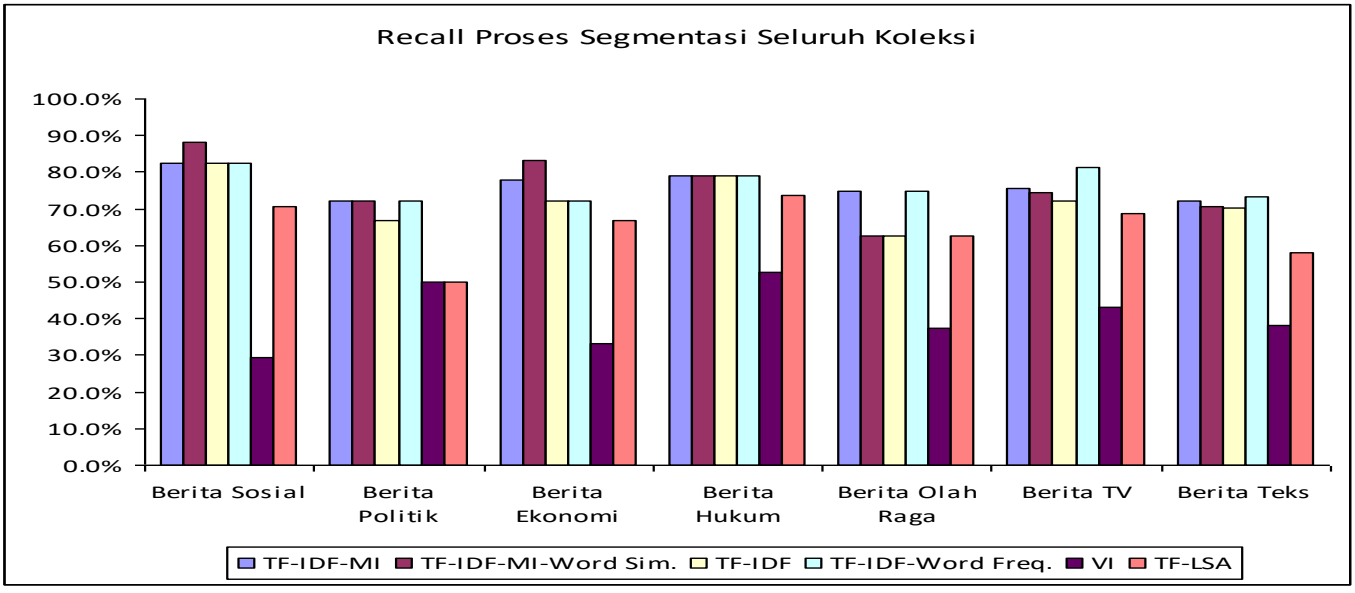
Tabel 3 dan 4, juga menunjukkan bahwa metode segmentasi dengan teknik pembobotan *TF-IDF-Mutual Information* adalah metode yang terbaik, karena pada metode ini, hasil segmentasi yang ditunjukkan baik pada saat tingkat akurasi perangkat pengenal suara (ASR) rendah (TV Olah Raga).

Pada metode segmentasi dengan pembobotan *TF-IDF-Mutual Information-Word Similarity*, bila kata-kata yang tidak tepat ditranskripsikan adalah kata-kata yang memiliki keterkaitan dengan kata-kata lain, maka nilai recall dari metode ini akan mengalami penurunan yang signifikan (TV Olah Raga), sebaliknya bila kata-kata yang salah ditranskripsikan bersifat independent, maka jika terjadi penurunan akurasi pada perangkat ASR, hasil dari metode segmentasi ini akan lebih baik bila dibandingkan metode yang lain (Koleksi TV Sosial).

Untuk koleksi berita yang bukan berasal dari berita suara, metode segmentasi yang memiliki nilai *recall* tertinggi adalah metode segmentasi dengan pembobotan *TF-IDF-Mutual Information Word Similarity* juga menunjukkan hasil yang baik.

Tabel 3. Hasil Recall Seluruh Model Segmentasi untuk Seluruh Koleksi

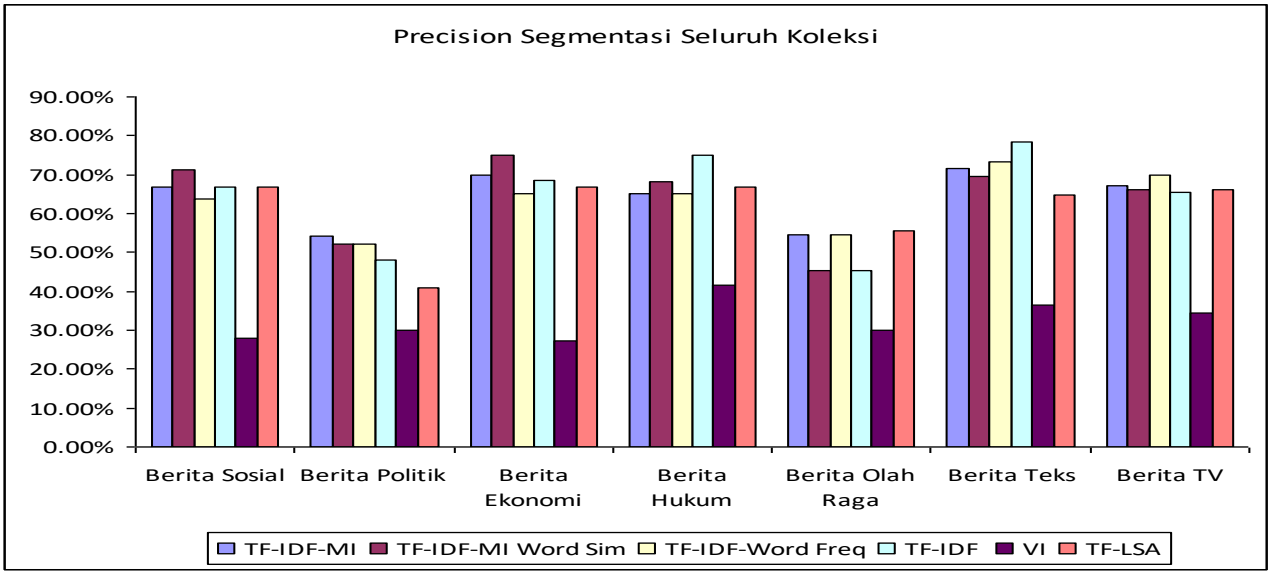
		TV Sosial	TV Politik	TV Ekonomi	TV Hukum	TV Olah Raga	TV	Teks
Akurasi Pengenalan Kata		69,5%	76,3%	77,5 %	79,0%	58,8%	73,8%	-
Metode Segmentasi	TF-IDF-MI	82.4%	72.2%	77.8%	78.9%	75.0%	75.6%	72.2%
	TF-IDF-MI Word Sim.	88.2%	72.2%	83.3%	78.9%	62.5%	74.4%	70.5%
	TF-IDF Wrd. Freq.	82.4%	72.2%	72.2%	78.9%	75.0%	81.4%	73.2%
Metode Segmentasi	TF-IDF	82.4%	66.7%	72.2%	78.9%	62.5%	72.1%	70.3%
	VI	29.4%	50.0%	33.3%	52.6%	37.5%	43.0%	38.1%
	TF-LSA	70.6%	50.0%	66.7%	73.7%	62.5%	68.6%	57.9%



Gambar 4. Grafik Persentase *Recall* Hasil Segmentasi Seluruh Koleksi

Tabel 4. Hasil Precision Seluruh Model Segmentasi untuk Seluruh Koleksi

		TV Sosial	TV Politik	TV Ekonomi	TV Hukum	TV Olah Raga	Teks	TV
Akurasi Pengenalan Kata		69,5%	76,3%	77,5 %	79,0%	58,8%	-	73,8%
Metode Segmentasi	TF-IDF-MI	66.70%	54.20%	70.00%	65.20%	54.50%	71.70%	67.00%
	TF-IDF-MI Word Sim	71.40%	52.00%	75.00%	68.20%	45.50%	69.40%	66.00%
	TF-IDF-Word Freq	63.60%	52.00%	65.00%	65.20%	54.50%	73.30%	70.00%
	TF-IDF	66.70%	48.00%	68.40%	75.00%	45.50%	78.40%	65.30%
	VI	27.80%	30.00%	27.30%	41.70%	30.00%	36.60%	34.60%
	TF-LSA	66.70%	40.90%	66.70%	66.70%	55.60%	64.70%	66.30%



Gambar 5. Grafik Persentase *Precision* Hasil Segmentasi Seluruh Koleksi

Tabel 5. Precision-Recall Segmentasi Untuk Ukuran Segmen Berbeda

Teknik Pembobotan	Precision			Recall		
	30 Kata Segmen	60 Kata Segmen	90 kata Segmen	30 Kata Segmen	60 Kata Segmen	90 kata Segmen
TFIDF-Word Freq	36.1%	73.3%	84.9%	77.9%	72.2%	59.5%
TFIDFMI	35.5%	71.7%	84.5%	77.4%	71.2%	59.8%
TFIDFMI Word Similarity	34.8%	69.4%	83.0%	76.4%	69.4%	58.6%
TFIDF	42.1%	78.4%	87.4%	76.5%	69.3%	58.3%
LSI	34.7%	64.6%	72.9%	60.7%	57.0%	45.4%
VI	18.4%	36.7%	52.4%	37.8%	37.7%	36.2%

Pengaruh ukuran-ukuran segmen terhadap perolehan *precision-recall* dari proses segmentasi ditampilkan pada Tabel 6. Hasil pada Tabel 6 tersebut menunjukkan, bahwa nilai *precision* akan bertambah baik bila ukuran segmen bertambah besar, sebaiknya perolehan *recall* akan bertambah kecil bilamana ukuran segmen bertambah besar.

Hasil eksperimen menunjukkan bahwa, ukuran segmen yang terbaik adalah pada ukuran segmen 60 kata per segmen, dimana pada ukuran tersebut kombinasi nilai *precision-recall* mencapai nilai yang terbaik.

5. KESIMPULAN

Kesimpulan-kesimpulan yang didapat dari hasil penelitian mengenai segmentasi dokumen teks hasil transkripsi berita suara berbahasa Indonesia adalah :

1. Hasil evaluasi dari beberapa teknik segmentasi yang digunakan menunjukkan teknik segmentasi dengan model perbandingan blok dan mempergunakan metode pembobotan *TF-IDF-Mutual Information*, *TF-IDF(Word Frequency)*, *TF-IDF-Mutual Information Word Similarity*, dan *TF-IDF* dapat melakukan segmentasi dengan ketepatan antara 66,7%-81,4% jika dibandingkan dengan teknik segmentasi metode pembobotan *LSA* yang memiliki ketepatan segmentasi antara 50%-73,7% dan *Vocabulary Introduction* yang memiliki ketepatan segmentasi antara 29,4%-52,6 %
2. Berdasarkan hasil evaluasi maka dapat disimpulkan bahwa metode segmentasi dengan teknik pembobotan *TF-IDF*(*Word Freqyuency*) merupakan metode segmentasi yang memiliki ketepatan segmentasi lebih tinggi baik pada dokumen teks hasil transkripsi (81,4%) ataupun pada dokumen teks (73,3%)
3. Ukuran Segmen mempengaruhi hasil perolehan *precision-recall* dari proses segmentasi. Berdasarkan ujicoba maka ukuran segmen yang terbaik adalah 60 kata per segmen
4. Hasil analisa kesalahan yang telah dilakukan menunjukkan bahwa faktor-faktor yang menyebabkan terjadinya kesalahan pada proses segmentasi adalah adanya akurasi pengenalan kata yang tidak baik, ukuran segmen yang tidak tepat sehingga menimbulkan segmen-segmen yang terlihat memiliki topik berbeda walaupun berada pada berita yang sama, dan faktor ketiga yang dapat menyebabkan kesalahan dalam melakukan segmentasi adalah adanya dua berita berurutan dengan topik yang sama
5. Peningkatan perolehan nilai ketepatan segmentasi metode pembobotan *TF-IDF-Mutual Information*, dan *TF-IDF*(*Word Frequency*) dapat mencapai 12,5 % bila dibandingkan dengan metode pembobotan *TF-IDF* pada segmentasi berita suara
6. Untuk segmentasi berita teks, teknik pembobotan *TF-IDF berbasis Word Frequency* dan *TF-IDF-Mutual Information* menunjukkan hasil baik dengan ketepatan segmentasi 73,2% dan 72,2% jika dibandingkan dengan ketepatan segmentasi teknik pembobotan TF-IDF (70,3%), teknik pembobotan Latent Semantic Analysis (57,9 %), dan Vocabulary Introduction (38,1%).

DAFTAR PUSTAKA

[1] Allan J., "Topic Detection and Tracking: Event-based Information Organization", Springer, 2002.

[2] Allan,J.,Carbonell, J., Doddington G., Yamron, J., Yang Y. "Topic Detection and Tracking Pilot Study Final Report", In Proceedings of the DARPA Broadcast,1998

[3] Arman A.A, "Proses Pembentukan dan Karakteristik Sinyal Ucapan", Departemen Teknik Elektro ITB, http://indotts.melsa.net.id/, access date : July 7, 2008

[4] Choi.,F.Y.Y, "Advances in domain independent linear text segmentation.", ACM International Conference Proceeding Series, Vol. 4, pp 26 - 33, 2000.

[5] Deerwester,S., Dumais,S.T., Harshman,R., "Indexing by Latent Semantic Analysis", JASIS, 1990.

[6] Gilbert, D. "JfreeChart Class Library",2006

[7] Grossman, D.A., Frieder, O., "Information Retrieval",Springer,2004

[8] Hauptmann,A.G., Witbrock,M.J., "Story Segmentation and Detection of Commercials In Broadcast News Video", Advances in Digital Libraries Conference,1998.

[9] Hearst, M. "TextTiling : Segmenting Text into Multi-paragraph SubTopic Passages", Computational Linguistics Volume 23 , Issue 1 (March 1997, pp 33 - 64, 1997.

[10] Kozima, H., "Text segmentation based on similarity between words" Proceedings of the 31st annual meeting on Association for Computational Linguistics, pp 286–288, 1993.

[11] Kozima, H., Furugori, T., "Segmenting Narrative Text Into Coherent Scenes", Lit Linguist Computing, pp 13-19, 1994.

[12] Jurafsky, D., Martin, H.M., "Speech and Language Processing", Prentice-Hall,2000.

[13] Lamprier, S., Amghar, T., Levrat B., Saubion, F., "SegGen: a Genetic Algorithm for Linear Text Segmentation", IJCAI, 2007.

[14] Lestari, D.P., Iwano, K., Furui.S Large Vocabulary Continous Speech Recognition System for Indonesian Language, 15th Indonesian Scientific Conference In Japan, 2006

[15] Microsoft Corporation, "HTK Book", Cambridge University-Engineering Department, 2001-2005.

[16] Pangaribuan, H., & Simanjuntak, P. (2021). Analisis Kualitas Perbandingan Citra Dengan Metode Segmentasi Citra. Jurnal Teknik Informatika UNIKA Santo Thomas, 289-297.

[17] Ponte,J.,Croft, B., "Text segmentation by topic", In the proceedings of the first European Conferenceon research and advanced technology for digital lib, U. Mass Technical Report TR-97-18, 1997

[18] Reynar, J.C., "Topic Segmentation: Algorithms and applications.", PhD Thesis, University of Pennsylvania, Seattle, WA, 2000.

[19] Setihajid, S.B., "Analisis Kamus Fonetik, Model Bahasa, dan Konfigurasi Properti pada Sistem Pengenal Suara untuk Bahasa Indonesia Menggunakan Sphinx-4", Skripsi, Fasilkom Universitas Indonesia, 2008

[20] Stokes, N. "Applications of Lexical Cohesion Analysis in the Topic Detection and Tracking Domain", Doctor Thesis, National University of Ireland, 2004

[21] Shoup,J.E., "Phonological Aspects of Speech Recognition", In Lea, W.A. (Ed.) Trends in Speech Recognition, pp 125-138, Prentice-Hall,1980

[22] Takao, S., Ogata, J., Ariki, Y., "Topic Segmentation of News Speech Using Word Similarity", Proc. of the 8th ACM international conference on Multimedia, pp 442 – 444, 2000.

[23] Takao, S., Ogata, J., Ariki, Y.,"Study On New Term Weighting Method and New Vector Space Model Based On Word Space in Spoken Document Retrieval", Proc. of the Int'l Conf. on Recherche d'Informations Assistee par Ordinateur (RIAO'00), pp.116-131, 2000.

[24] Utiyama, M.,Isahara, H, "A statistical model for domain-independent text segmentation", Proc. of the 39th Annual Meeting on Association for Computational Linguistics, Toulouse, France, pp 499 – 506, 2001.

[25] Walker, W., et al. "Sphinx 4: A Flexible Open Source Framework for Speech Recognition", Sun Microsystems Laboratories Technical Reports, TR-2004-139, 2004.

[26] Xuedong, Huang. "Spoken Language Processing: a guide to teory, algorithm and system development", Prentice Hall, 2001

[27] Yates, B., Neto,R. "Modern Information Retrieval.ACM", Press/Addison-Wesley, 1999.

[28] Zahra, A., "Penyusunan Kamus Fonetik dalam Pengembangan Sistem Pengenalan Suara Otomatis untuk Bahasa Indonesia", Skripsi, Fasilkom Universitas Indonesia, 2008

BIODATA PENULIS

Chintalya Magdalena
Mahasiswa Pascasarjana Program Studi Teknik Elektro Universitas Kristen Indonesia Jakarta. www.uki.ac.id

Bangun Hyolister Tambun
Mahasiswa Pascasarjana Program Studi Teknik Elektro Universitas Kristen Indonesia Jakarta. www.uki.ac.id