Prediksi Jeda dalam Ucapan Kalimat Bahasa Melayu Pontianak Menggunakan Hidden Markov Model Berbasis Part of Speech

Article *in* Jurnal Teknologi Informasi dan Ilmu Komputer · August 2020

DOI: 10.25126/jtiik.2020742166

CITATIONS
0

READS
34

3 authors, including:

Arif Bijaksana Putra Negara
Tanjungpura University
19 PUBLICATIONS **37** CITATIONS

SEE PROFILE

Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.202072166
Vol. 7, No. 4, Agustus 2020, hlm. 755-764 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 30/E/KPT/2018 e-ISSN: 2528-6579

PREDIKSI JEDA DALAM UCAPAN KALIMAT BAHASA MELAYU PONTIANAK MENGGUNAKAN HIDDEN MARKOV MODEL BERBASIS PART OF SPEECH

Arif Bijaksana Putra Negara[1], Hafiz Muhardi [2], Evi Fathiyah Muniyati*[3]

[1,2,3]Program Studi Informatika Fakultas Teknik, Universitas Tanjungpura, Indonesia
Email: [1]arifbpn@informatika.untan.ac.id, [2]hafiz.muhardi@informatika.untan.ac.id,
[3]evifathiyahmuniyati@gmail.com
*Penulis Korespondensi

(Naskah masuk: 1 Juli 2019, diterima untuk diterbitkan: 7 Oktober 2019)

Abstrak

Informasi jeda adalah salah satu faktor pendukung dari ucapan berkualitas yang dihasilkan oleh sistem *Text to Speech*. Sebelumnya sudah ada penelitian untuk memprediksi jeda pada bahasa Melayu Pontianak menggunakan metode lain, namun masih belum mendapatkan hasil yang baik. Penelitian ini bertujuan untuk memprediksi jeda pada ucapan kalimat bahasa Melayu Pontianak berbasis *part of speech* dengan menggunakan *tools Hidden Markov Model* (HMM). HMM akan menghitung nilai probabilitas dari setiap kemungkinan yang ada. Penelitian ini menggunakan data berupa file rekaman ucapan penutur yang membacakan 500 kalimat berbahasa Melayu Pontianak dan *set* PoS baru yang dikembangkan dari beberapa *set* PoS yang telah ada. Hasil yang didapatkan dari sistem ini yaitu teks kalimat bahasa Melayu Pontianak beserta prediksi jedanya. Indeks jeda dikategorikan menjadi 5 kategori yaitu indeks jeda "0" menandakan tidak ada jeda, "1" menandakan jeda singkat, "2" menandakan jeda panjang, "," menandakan tanda baca koma, dan "." menandakan akhir kalimat. Hasil prediksi kemudian diuji menggunakan pengujian akurasi kecocokan jeda ucapan dalam satu kalimat penuh dan pengujian *precision*, *recall* dan *f-measure*. Frasa jeda ucapan yang diuji yaitu frasa jeda 1+2 dan frasa jeda 2. Pengujian dilakukan dengan membandingkan hasil model bigram dan trigram. Berdasarkan pengujian yang telah dilakukan, model trigram lebih baik dalam menghasilkan prediksi jeda ucapan pada kalimat bahasa Melayu Pontianak.

Kata kunci: *prediksi jeda, Hidden Markov Model, part of speech, bigram, trigram*

PAUSE PREDICTION IN PONTIANAK MALAY LANGUAGE SENTENCES USING HIDDEN MARKOV MODEL BASED ON PART OF SPEECH

Abstract

Pause information is one of the supporting factors of quality speech produced by the Text to Speech system. Previously there had been research to predict pauses in Pontianak Malay language using other methods, but it still did not get good results. This study aims to predict pauses in Pontianak Malay language sentences using the Hidden Markov Model (HMM) tools based on part of speech. HMM will calculate the probability value of each possibility. This research uses recording file of speeches from speakers who read 500 Pontianak Malay sentences and a new PoS set developed from several existing PoS sets. The results are Pontianak Malay language sentence along with the pause prediction. The pause indices are categorized into 5 categories, the pause index "0" indicates that there is no pause, "1" indicates a short pause, "2" indicates a long pause, "," indicates the comma punctuation, and "." indicates the end of the sentence. The prediction results are then tested using a speech pause match accuracy test in one full sentence and testing of precision, recall and f-measure. The speech pause phrases that are tested are the pause phrase 1+2 and the pause phrase 2. The test is done by comparing the results of the bigram and trigram models. Based on the tests that have been done, the trigram model is better at producing predictions of speech pauses in Pontianak Malay language sentences.

Keywords: *pause prediction, Hidden Markov Model, part of speech, bigram, trigram*

1. PENDAHULUAN

Bahasa merupakan salah satu alat komunikasi yang digunakan oleh manusia dalam kehidupan sehari-hari. Kemampuan penguasaan bahasa yang baik dapat mempermudah proses interaksi dengan orang lain. Bahasa Melayu Pontianak merupakan

bahasa yang dituturkan oleh masyarakat Pontianak dalam kehidupan sehari-hari. Hasil Sensus Penduduk pada tahun 2010 menunjukkan bahwa dari berbagai bahasa daerah yang terdapat di Indonesia, persentase penggunaan bahasa Melayu yang digunakan oleh masyarakat Kalimantan Barat mencapai 20.45% (1.615.978 juta jiwa) dari total penduduk Kalimantan Barat (Na'im dan Syahputra, 2011).

Saat ini, pelestarian kebudayaan bahasa Melayu Pontianak sangat diperlukan untuk mencegah ancaman dari globalisasi yang berdampak pada berkurangnya penutur dalam menggunakan bahasa Melayu Pontianak. Salah satunya yaitu dengan memanfaatkan teknologi *Text To Speech* (TTS). *Text To Speech* (TTS) adalah suatu sistem yang dapat mengkonversi teks menjadi ucapan. Kualitas ucapan pada TTS dapat dinilai dari kejelasan dan kealamian ucapan yang dihasilkan. Untuk menghasilkan ucapan yang berkualitas tersebut, TTS membutuhkan informasi jeda. Penentuan jeda dalam ucapan kalimat sangatlah penting, karena jeda dapat memperjelas informasi mengenai makna atau maksud dari suatu kalimat yang disampaikan.

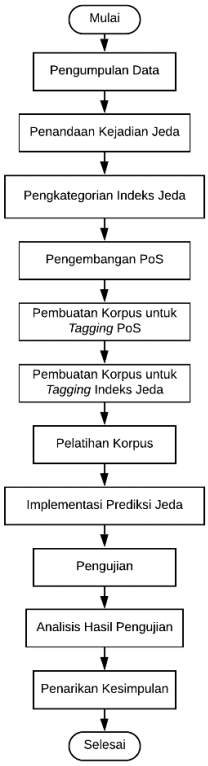
Berbagai penelitian telah dilakukan untuk menghasilkan prediksi jeda ucapan dengan menggunakan berbagai macam pendekatan atau metode, diantaranya penelitian yang dilakukan oleh Yu dan Tao (2005) yang menggunakan metode *Classification and Regression Tree* (CART), penelitian oleh Do et al. (2015) menggunakan CRF-*based*, penelitian Kamiludin (2017) menggunakan *rule based*, dan penelitian yang dilakukan oleh Nugraha (2014) menggunakan *Hidden Markov Model* (HMM).

Hidden Markov Model (HMM) adalah pemodelan statistik di mana sebuah sistem "menghasilkan" (*emit*) urutan simbol yang dapat diamati (*observation symbols*) berdasarkan sebuah proses probabilistik yang parameternya tidak diketahui (*hidden parameters*) (Manurung, 2016). *Hidden Markov Model* memiliki 2 macam bagian, yaitu *observed state* dan *hidden state*. *Observed state* merupakan bagian yang dapat diamati secara langsung dan *hidden state* merupakan bagian yang tidak dapat diamati (Wibisono, 2008). HMM juga sukses diterapkan untuk menangani masalah-masalah yang ada pada *Natural Language Processing* seperti *part of speech tagging*, *phrase chunking*, *speech recognition*, *text to speech*, dan mengambil informasi dari sekumpulan dokumen. Namun, HMM memerlukan data yang besar. Maka dari itu, untuk mengurangi jumlah korpus yang besar, maka korpus berupa rangkaian kalimat yang bersumber dari ucapan asli penutur direpresentasikan ke dalam beberapa rangkaian *Part of Speech* (PoS). *Part of speech* atau kelas kata adalah kelompok klasifikasi kata-kata yang sesuai fungsinya dalam suatu konteks, seperti kata benda, kata kerja, kata sifat, kata keterangan, kata penghubung, dan lain-lain. Kemudian, rangkaian *part of speech* tersebut akan ditandai. Penandaan *part of speech* adalah suatu proses yang memberikan label kelas kata secara otomatis pada suatu kata dalam kalimat. Beberapa penggunaan PoS-*tagging* adalah untuk menghapus perbedaan yang tidak relevan, menghapus ambiguitas, membantu *stemming*, dan membantu pencarian kata benda (Christianti M, Pragantha dan Victor, 2016). Penelitian yang dilakukan oleh Wicaksono dan Purwarianti (2010) mengembangkan PoS-*tagger* untuk bahasa Indonesia yaitu IPOSTagger dengan menggunakan himpunan tipe PoS yang dikembangkan sebanyak 35 tipe PoS.

Mengacu pada penelitian-penelitian tersebut, maka penelitian yang akan dilakukan adalah memprediksi jeda pada ucapan kalimat bahasa Melayu Pontianak dengan menggunakan *Hidden Markov Model* (HMM) berbasis *Part of Speech* (PoS). Penelitian ini menggunakan data yang berasal dari buku Sepok Satu karangan Pay Jarot Sujarwo. Data tersebut merupakan korpus yang akan dilatih dan ditandai menggunakan HMM sehingga menghasilkan kalimat bahasa Melayu Pontianak beserta prediksi jedanya.

2. METODE PENELITIAN

Beberapa tahapan yang dilakukan pada penelitian ini seperti yang terlihat pada Gambar 1.



Gambar 1. Diagram alir penelitian

2.1 Pengumpulan Data

Data yang digunakan merupakan *file* rekaman suara dari teks kalimat yang ada pada buku Sepok

Satu karangan Pay Jarot Sujarwo dan dilakukan oleh penutur ahli. Jumlah kalimat yang dibacakan pada *file* rekaman suara tersebut sebanyak 500 kalimat. *File* suara ini berformat .wav dengan resolusi 16bit dan *sampling rate* 44100 Hz.

2.2 Penandaan Kejadian Jeda

File suara yang telah disiapkan kemudian diolah menggunakan aplikasi WaveSurfer untuk ditandai kejadian jeda dari setiap kata. Setiap kejadian jeda diberi tanda "sil". Durasi jeda yang dihasilkan berada pada *file* dengan format *.breaks.

2.3 Pengkategorian Indeks Jeda

Penelitian ini menggunakan 5 kategori indeks jeda, antara lain indeks jeda "0" menandakan tidak ada jeda, indeks jeda "1" menandakan jeda singkat, indeks jeda "2" menandakan jeda panjang, indeks jeda "," menandakan tanda baca koma, dan indeks jeda "." menandakan akhir kalimat.

2.4 Pengembangan PoS

Pengembangan *set* PoS dilakukan agar memperjelas penandaan tipe kata dalam kalimat sehingga memudahkan HMM untuk memprediksi jeda. Pengembangan PoS dilakukan dengan mengelompokkan kata-kata sesuai dengan tipe PoS yang telah dikembangkan oleh Kamiludin (2017), kemudian untuk memperjelas kelas kata lainnya maka ditambahkan tipe PoS dari penelitian Adriani (2009) dan Setyaningsih (2017).

2.5 Pembuatan Korpus untuk *Tagging* PoS

Korpus latih yang dibuat berisi 500 teks kalimat yang berasal dari buku Sepok Satu yang diberikan *tag* PoS secara manual sesuai dengan *set* PoS yang telah dikembangkan.

2.6 Pembuatan Korpus untuk *Tagging* Indeks Jeda

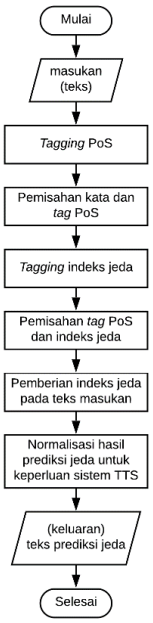
Rangkaian *tag* PoS yang ada pada korpus latih diambil dan diberi indeks jeda secara manual sesuai dengan nilai durasi pada setiap kejadian jeda, sehingga didapat 500 rangkaian teks tipe PoS beserta indeks jedanya yang kemudian digunakan untuk korpus latih prediksi indeks jeda.

2.7 Pelatihan Korpus

Pada proses pelatihan korpus untuk *tagging* PoS, korpus latih yang berisi "kata/PoS" disimpan dalam folder IPOSTagger dengan ekstensi .crp. Kemudian proses pelatihan dilakukan melalui *command prompt* pada folder IPOSTagger. Pada proses pelatihan korpus untuk *tagging* indeks jeda, korpus latih yang berisi "PoS/Indeks Jeda" disimpan dalam folder JEDATagger dengan ekstensi .crp. Kemudian proses pelatihan dilakukan melalui *command prompt* pada folder JEDATagger.

2.8 Implementasi Prediksi Jeda

Tahapan untuk proses prediksi jeda seperti yang terlihat pada Gambar 2.



Gambar 2. Diagram alir proses prediksi jeda

2.8.1 *Tagging* PoS

Sebelum memulai proses *tagging* PoS, dilakukan penambahan spasi pada token terlebih dahulu. Penambahan spasi pada token dilakukan untuk kalimat masukan yang tidak memiliki spasi baik sebelum dan setelah tanda baca koma serta sebelum tanda baca yang menandai akhir kalimat seperti tanda titik, tanda seru dan tanda tanya. *Tagging* PoS adalah proses yang dilakukan untuk mendapatkan *tag* PoS atau kelas kata dari setiap kata yang ada. Kalimat masukan ditandai untuk mendapatkan *tag* PoS dengan menggunakan *tools* HMM yaitu IPOSTagger yang telah dilatih sebelumnya dan hasilnya disimpan pada *file* dengan format *.hsl yang berisikan "kata/PoS".

2.8.2 Pemisahan Kata dan *Tag* PoS

Hasil dari proses *tagging* PoS sebelumnya yang berisi "kata/PoS" diolah lagi untuk memisahkan antara kata dengan *tag* PoS dan disimpan dalam *file* keluaran yang berbeda.

2.8.3 *Tagging* Indeks Jeda

Tagging indeks jeda merupakan proses yang dilakukan untuk mendapatkan indeks jeda dari setiap

tag PoS. *File* teks yang berisikan *tag* PoS dari hasil *tagging* PoS selanjutnya diproses menggunakan JEDATagger yang telah dilatih untuk mendapatkan indeks jeda pada setiap kelas kata yang ada.

2.8.4 Pemisahan *Tag* PoS dan Indeks Jeda

Hasil dari proses *tagging* indeks jeda yang berisi "PoS/Indeks Jeda" selanjutnya diolah untuk memisahkan *tag* PoS dengan indeks jedanya dan indeks jeda yang didapat disimpan ke dalam sebuah *file* teks baru.

2.8.5 Pemberian Indeks Jeda pada Teks Masukan

Indeks jeda yang telah didapat dari proses *tagging* indeks jeda selanjutnya digabungkan dengan teks kalimat masukan.

2.8.6 Normalisasi Hasil Prediksi Jeda untuk Keperluan TTS

Hasil akhir prediksi jeda yang berisi kata beserta indeks jedanya selanjutnya diolah lagi untuk keperluan implementasi sistem *text to speech* (TTS) dimana indeks jeda "1" untuk jeda singkat diubah menjadi tanda "|" dan indeks jeda "2" untuk jeda panjang diubah menjadi tanda "||", serta menghilangkan tiga kategori indeks jeda lainnya yaitu indeks jeda "0" yang menyatakan tidak ada jeda, indeks jeda "," untuk tanda baca koma dan indeks jeda "." yang menyatakan akhir kalimat.

2.8.7 Rancangan Antarmuka

Antarmuka sistem prediksi jeda berupa halaman *web* yang berisi *label*, sebuah tombol untuk memulai proses prediksi jeda serta lima buah kolom yang digunakan untuk memasukkan teks yang akan diprediksi, menampilkan hasil *tagging* PoS yang berisi "kata/PoS", menampilkan hasil *tagging* indeks jeda yang berisi "PoS/Indeks Jeda", menampilkan hasil akhir prediksi jeda yang berisi "kata/Indeks Jeda", dan menampilkan hasil akhir prediksi jeda untuk keperluan implementasi pada sistem TTS.

2.8.7.1 Pengujian

Pengujian dilakukan dengan membandingkan jeda yang dihasilkan oleh sistem dengan jeda ucapan penutur menggunakan pengujian akurasi kecocokan frasa jeda ucapan dalam satu kalimat penuh dan pengujian *precision, recall, f-measure* yang terbagi menjadi tiga, yaitu pengujian *baseline*, peningkatan jumlah korpus dan *K-fold cross validation*. Sistem ini diujikan menggunakan dua model n-gram, yaitu bigram dan trigram. Frasa jeda ucapan yang diujikan yaitu frasa jeda 1+2 dan jeda 2.

2.8.7.2 Analisis Hasil Pengujian

Pada tahap ini, hasil pengujian sistem dianalisis dan direpresentasikan ke dalam grafik untuk mempermudah penarikan kesimpulan.

2.8.7.3 Penarikan Kesimpulan

Kesimpulan dibuat berdasarkan tahapan-tahapan yang telah dilakukan sebelumnya dengan melihat apakah penggunaan *Hidden Markov Model* berbasis *part of speech* dapat memprediksi jeda sesuai dengan yang diharapkan.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pengkategorian Indeks Jeda

Durasi tiap indeks jeda yang diperlukan untuk proses penandan jeda sesuai hasil perhitungan kejadian jeda dapat dilihat pada Tabel 1.

Tabel 1. Kategori Indeks Jeda

Indeks Jeda	Keterangan
0	Tidak ada jeda (0 ~ < 0.025 s)
1	Jeda singkat (0.025 ~ <= 0.333 s)
2	Jeda panjang (> 0.333 s)
,	Tanda baca koma (,)
.	Akhir kalimat (?!.)

3.2 Hasil Pengembangan PoS

Hasil dari pembuatan PoS bahasa Melayu Pontianak yaitu tabel PoS dengan jumlah sebanyak 46 *set* PoS seperti pada Tabel 2.

Tabel 2. PoS Bahasa Melayu Pontianak

No	PoS	Deskripsi	Contoh Kata
1	VBR	Verba Reduplikasi	Jalan-jalan, poto-poto
2	VBK	Verba Berkonjugasi	Bersalam-salam, berputar-putar
3	VBT	Verba Transitif	Makai, nenggek, njajah
4	VBI	Verba Intransitif	Betanyak, balek, nuron
5	IN	Kata Depan	di, ke, dari , pade
6	UH	Kata Seru	Oi, woi, alamak
7	AR	Artikulus	Sang, si
8	RP	Partikel	pon, lah, jak
9	JJ	Kata Sifat	kaye, lawar, pandai, budoh
10	CON	Konjungsi	dan, kalok
11	OP	Open Parenthesis	({ [
12	CP	Close Parenthesis	) }]
13	.	Sentence Terminator	.! ? …
14	.	Koma	,
15	:	Colon	: :
16	SYM	Simbol	*%#&@
17	CR	Currency	Rp, $
18	MD	Modal	nak, haros
19	NEG	Negation	bukan, jangan , tadak
20	SL	Slash	/
21	DS	Dash	-
22	QT	Quotation	" '

No	PoS	Deskripsi	Contoh Kata
23	WP	WH-Pronoun	Ape, siape, berape
24	WDT	WH-Determiner	Ape, siape, barangsiape
25	DT	Determiner	ini, ni , tu, tu, tuh
26	FW	Foreign Word	wonderful, story
27	US	Unit Symbol	Gr, Kg, Cm
28	CDP	Primary Numeral	Satu, duak, tige
29	CDO	Ordinal Numeral	Kesatu, Keduak, ketige
30	CDI	Irregular Numeral	Beberape, segale, semue
31	CDF	Fraction Numeral	Setengah, seperempat
32	CDA	Kata Bantu Bilangan	biji, ekor, buah, orang
33	CDC	Collective Numeral	ratusan, ribuan, pulohan
34	RB	Adverb	paleng, sementara
35	WPRB	WH-Adverb	Cemane, ngape
36	FRB	Adverb of Frequency	jarang, sering, kadang-kadang
37	DRB	Adverb of Degree	agak, hamper, cukop
38	TRB	Adverb of Time	udah, belom, dulok, sekarang
39	PRP	Personal Pronoun	aku, saye, kau, die
40	PRL	Locative Pronoun	sanak, sine, situk
41	PRN	Number Pronoun	satu-satunye, dua-duanye
42	NNP	Proper Noun	Eropa, Indonesia, Belanda
43	NNG	Genitive Common Noun	bukunye, rumahnye
44	NNC	Countable Common Noun	buku, rumah, karyawan
45	NNU	Uncountable Common Noun	aek, gula, nasi, ujan
46	NN	Common Noun	Martabat, janji

3.3 Hasil dan Analisis Pengujian Kecocokan Frasa Jeda Ucapan dalam Satu Kalimat Penuh

Pengujian untuk menilai tingkat akurasi kecocokan frasa jeda ucapan dilakukan dengan menggunakan 500 kalimat sebagai korpus latih dan 500 kalimat yang sama digunakan sebagai korpus uji. Hasil dari pengujian ini dapat dilihat pada Tabel 3.

Tabel 3. Hasil Pengujian Kecocokan Frasa Jeda Ucapan dalam Satu Kalimat Penuh

Ket.	Akurasi	
	Frasa jeda 1+2	Frasa Jeda 2
Bigram	30.8%	80%
Trigram	30.8%	81.4%

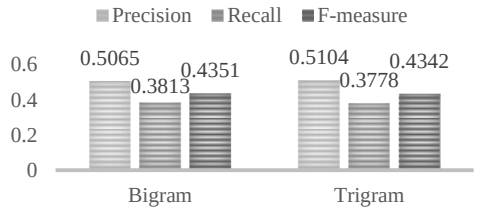
Hasil pengujian kecocokan frasa jeda ucapan dalam satu kalimat penuh menunjukkan bahwa nilai akurasi terbaik dihasilkan oleh model trigram sebesar 30.8% untuk frasa jeda 1+2 dan 81.4% untuk frasa jeda 2. Model bigram juga menghasilkan nilai yang sama untuk frasa jeda 1+2, namun menghasilkan nilai akurasi frasa jeda 2 yang lebih rendah daripada model trigram. Rendahnya nilai akurasi pada frasa jeda 1+2 disebabkan oleh banyaknya frasa jeda ucapan yang terbentuk oleh indeks jeda "1" dan jeda singkat dan indeks jeda "2" atau jeda panjang pada hasil prediksi jeda yang tidak sesuai dengan jeda ucapan penutur. Sedangkan untuk frasa jeda 2, sistem prediksi jeda dapat memprediksi jeda ucapan yang sesuai dengan jeda ucapan penutur sehingga menghasilkan nilai akurasi yang baik. Hasil dari penggunaan HMM pada pengujian ini lebih tinggi daripada menggunakan metode *Shallow Parsing* yang menghasilkan nilai akurasi sebesar 10.6%.

3.4 Hasil Pengujian *Precision, Recall* dan *F-measure*

3.4.1 Hasil dan Analisis Pengujian *Baseline*

Pengujian *baseline* menggunakan 500 kalimat sebagai korpus latih, kemudian 500 kalimat tersebut juga digunakan sebagai korpus uji. Hasil pengujian baseline dapat dilihat pada Tabel 4. Berdasarkan hasil pengujian untuk frasa jeda 1+2, model trigram menghasilkan nilai *precision* tertinggi sebesar 0.5104 dan model bigram menghasilkan nilai *recall* tertinggi sebesar 0.3813 dan 0.4351 untuk *F-measure*.

Sedangkan untuk frasa jeda 2, nilai *precision*, *recall* dan *F-measure* yang terbaik dihasilkan oleh model trigram dengan nilai *precision* sebesar 0.8713, *recall* sebesar 0.756 dan *F-measure* sebesar 0.8095. Hasil perhitungan untuk frasa jeda 1+2 dapat dilihat pada Gambar 3.



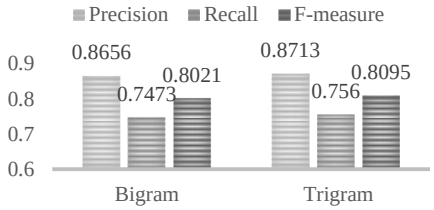
Gambar 3. Grafik pengujian *baseline* terhadap frasa jeda 1+2

Tabel 4. Hasil Pengujian *Baseline*

Keterangan	Relevan Terambil (a)	Relevan Tak Terambil (b)	Tak Relevan Terambil (c)	Precision $(\frac{a}{a+c})$	Recall $(\frac{a}{a+b})$	F-Measure $(2.\ \frac{p\ .\ r}{p+r})$
Bigram						
Frasa Jeda 1+2	543	881	529	0.5065	0.3813	0.4351
Frasa Jeda 2	689	233	107	0.8656	0.7473	0.8021
Trigram						
Frasa Jeda 1+2	538	886	516	0.5104	0.3778	0.4342
Frasa Jeda 2	697	225	103	0.8713	0.756	0.8095

Pada grafik tersebut terlihat adanya perbedaan yang sangat kecil antara nilai *precision*, *recall* dan *F-measure* untuk frasa jeda 1+2 yang dihasilkan oleh model trigram dan model bigram. Hal ini

menunjukkan bahwa prediksi jeda yang dihasilkan oleh bigram dan trigram tidak jauh berbeda.



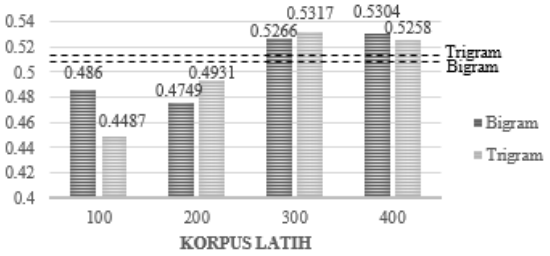
Gambar 4. Grafik pengujian *baseline* terhadap frasa jeda 2

Hasil perhitungan untuk frasa jeda 2 dapat dilihat pada Gambar 4. Pada grafik tersebut terlihat adanya perbedaan kecil antara hasil dari kedua model n-gram, namun masih dapat terlihat jelas bahwa nilai *precision, recall* dan *F-measure* untuk frasa jeda 2 yang dihasilkan oleh model trigram lebih baik daripada nilai yang dihasilkan oleh model bigram.

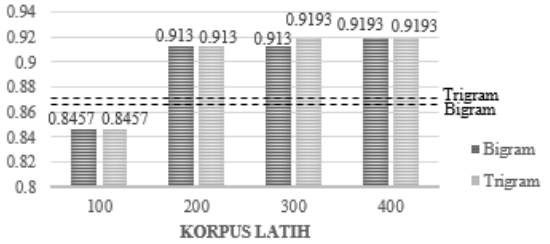
3.4.2 Hasil dan Analisis Pengujian Peningkatan Jumlah Korpus

Pengujian ini dilakukan dengan meningkatkan jumlah korpus latih. Korpus uji yang digunakan sebanyak 100 kalimat dan 400 kalimat lainnya dijadikan korpus latih. Berdasarkan hasil pengujian, nilai rata-rata *precision* frasa jeda 1+2 tertinggi dihasilkan oleh model bigram dengan nilai sebesar 0.5045. Sedangkan nilai rata-rata *recall* dan F-measure tertinggi dihasilkan oleh model trigram dengan nilai sebesar 0.379 untuk *recall* dan 0.4306 untuk *F-measure*. Hasil pengujian untuk frasa jeda 1+2 dapat dilihat pada Tabel 5.

menunjukkan terjadinya penurunan nilai pada penggunaan korpus 200 kalimat meskipun jumlah korpus telah ditingkatkan dari penggunaan korpus sebanyak 100 kalimat. Hal ini dapat terjadi karena bertambahnya jumlah prediksi jeda singkat dan jeda panjang yang tidak sesuai dengan ucapan penutur pada penggunaan korpus yang lebih banyak. Pada grafik tersebut juga dapat dilihat bahwa hanya penggunaan korpus sebanyak 300 dan 400 kalimat yang dapat mencapai nilai *baseline*.



Gambar 5. Grafik *precision* frasa jeda 1+2



Gambar 6. Grafik *precision* frasa jeda 2

Hasil pengujian precision untuk frasa jeda 2 dapat dilihat pada Gambar 6. Grafik tersebut menunjukkan

Tabel 5. Hasil Pengujian untuk Frasa Jeda 1+2

Korpus Uji	Korpus Latih	Bigram Precision	Bigram Recall	Bigram F-measure	Trigram Precision	Trigram Recall	Trigram F-measure
100	100	0.486	0.3728	0.4219	0.4487	0.3763	0.4093
100	200	0.4749	0.3728	0.4177	0.4931	0.3835	0.4315
100	300	0.5266	0.3907	0.4486	0.5317	0.3907	0.4504
100	400	0.5304	0.3441	0.4174	0.5258	0.3656	0.4313
Rata-Rata		**0.5045**	**0.3701**	**0.4264**	**0.4998**	**0.379**	**0.4306**

Untuk frasa jeda 2, model trigram menghasilkan nilai rata-rata *precision, recall* dan *F-measure* yang lebih baik daripada model bigram. Nilai rata-rata yang dihasilkan yaitu sebesar 0.8993 untuk *precision*, 0.8192 untuk *recall* dan 0.8574 untuk *F-measure*. Hasil pengujian untuk frasa jeda 2 dapat dilihat pada Tabel 6.

adanya kenaikan nilai pada peningkatan jumlah korpus. Hal ini dapat terjadi karena bertambahnya hasil prediksi jeda panjang yang sesuai dengan jeda ucapan penutur. Hasil dari penggunaan 100 korpus latih tidak dapat mencapai nilai *baseline*, namun penggunaan 200, 300 dan 400 kalimat sebagai korpus latih dapat melebihi nilai *baseline*.

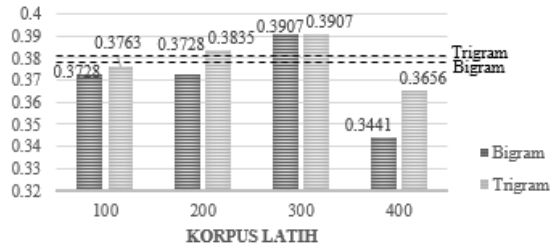
Tabel 6. Hasil Pengujian untuk Frasa Jeda 2

Korpus Uji	Korpus Latih	Bigram Precision	Bigram Recall	Bigram F-measure	Trigram Precision	Trigram Recall	Trigram F-measure
100	100	0.8457	0.774	0.8083	0.8457	0.774	0.8083
100	200	0.913	0.8305	0.8698	0.913	0.8305	0.8698
100	300	0.913	0.8305	0.8698	0.9193	0.8362	0.8757
100	400	0.9193	0.8362	0.8757	0.9193	0.8362	0.8757
Rata-Rata		**0.8978**	**0.8178**	**0.8559**	**0.8993**	**0.8192**	**0.8574**

Hasil pengujian *precision* untuk frasa jeda 1+2 dapat dilihat pada Gambar 5. Grafik tersebut

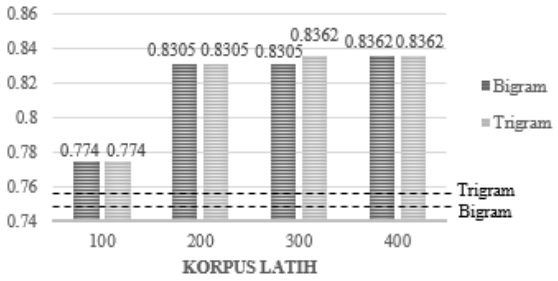
Hasil pengujian *recall* untuk frasa jeda 1+2 dapat dilihat pada Gambar 7. Grafik ini

menunjukkan kenaikan nilai hingga penggunaan korpus latih sebanyak 300 kalimat, namun terjadi penurunan pada penggunaan korpus latih sebanyak 400 kalimat. Hal ini menunjukkan bahwa sistem prediksi jeda masih belum dapat memprediksi jeda yang sesuai dengan ucapan penutur dan nilai *recall* yang dihasilkan belum semuanya dapat mencapai nilai *baseline*.

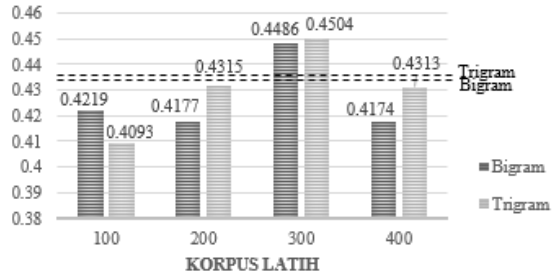


Gambar 7. Grafik *recall* frasa jeda 1+2

Hasil pengujian *recall* untuk frasa jeda 2 dapat dilihat pada Gambar 8. Grafik ini menunjukkan nilai *recall* yang meningkat ketika jumlah korpus ditingkatkan. Hal ini membuktikan bahwa semakin banyak jumlah korpus, maka semakin baik pula hasil prediksi jeda. Selain itu, semua nilai *recall* yang dihasilkan sudah dapat melebihi nilai *baseline*.



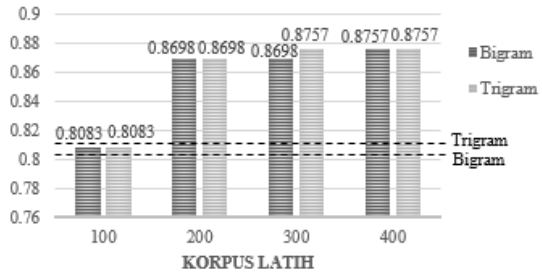
Gambar 8. Grafik *recall* frasa jeda 2



Gambar 9. Grafik *F-measure* frasa jeda 1+2

Hasil pengujian *F-measure* untuk frasa jeda 1+2 dapat dilihat pada Gambar 9. Grafik ini menunjukkan kenaikan nilai akurasi yang tidak konsisten. Hal ini dikarenakan adanya nilai *precision* dan *recall* yang menurun karena banyaknya jeda yang dihasilkan oleh sistem yang tidak sesuai dengan jeda ucapan penutur sehingga nilai akurasi yang diperoleh juga mengalami penurunan. Selain itu, hanya akurasi pada penggunaan 300 kalimat sebagai korpus latih yang dapat mencapai nilai *baseline*.

Hasil pengujian *F-measure* untuk frasa jeda 2 dapat dilihat pada Gambar 10. Grafik ini menunjukkan kenaikan nilai yang stabil ketika jumlah korpus ditingkatkan. Hal ini terjadi karena banyaknya jumlah prediksi jeda yang sesuai dengan ucapan penutur ketika korpus ditingkatkan sehingga menghasilkan nilai *precision* dan *recall* yang baik dan berdampak pada nilai akurasi yang juga meningkat. Nilai akurasi yang dihasilkan sudah dapat melebihi nilai baseline untuk model bigram. Untuk model trigram, semua nilai dapat melebihi nilai *baseline* kecuali pada penggunaan korpus latih sebanyak 100 kalimat.



Gambar 10. Grafik *F-measure* frasa jeda 2

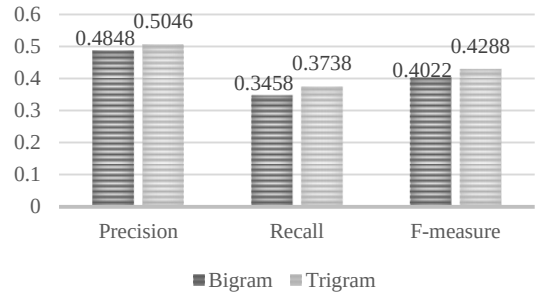
3.4.3 Hasil dan Analisis Pengujian K-*Fold Cross Validation*

Pengujian K-*fold cross validation* menggunakan lima *fold* dengan pembagian 20% korpus uji dan 80% korpus latih. Hasil pengujian untuk frasa jeda 1+2 dapat dilihat pada Tabel 7. Tabel 7 menunjukkan bahwa model trigram menghasilkan nilai rata-rata terbaik sebesar 0.5046 untuk *precision*, 0.3738 untuk *recall* dan 0.4288 untuk *F-measure*. Nilai rata-rata *precision*, *recall* dan *F-measure* yang dihasilkan belum dapat mencapai nilai *baseline*.

Tabel 7. Hasil Pengujian untuk Frasa Jeda 1+2

Korpus Uji	Bigram			Trigram		
	Precision	Recall	F-measure	Precision	Recall	F-measure
1-100	0.5304	0.3441	0.4174	0.5258	0.3656	0.4313
101-200	0.3557	0.2066	0.2614	0.4057	0.2575	0.315
201-300	0.4792	0.3251	0.3874	0.5187	0.3922	0.4467
301-400	0.5895	0.4839	0.5315	0.6009	0.4803	0.5339
401-500	0.4694	0.3695	0.4135	0.4721	0.3735	0.417
Rata-Rata	**0.4848**	**0.3458**	**0.4022**	**0.5046**	**0.3738**	**0.4288**

Hasil pengujian untuk frasa jeda 1+2 dapat direpresentasikan kedalam grafik seperti pada Gambar 11. Berdasarkan grafik tersebut, dapat dilihat dengan jelas bahwa model trigram menghasilkan nilai rata-rata yang lebih tinggi daripada model bigram. Nilai yang dihasilkan oleh kedua model n-gram relatif rendah karena masih terdapat banyak frasa yang terbentuk oleh indeks jeda "1" atau jeda singkat dan indeks jeda "2" atau jeda panjang yang tidak sesuai dengan frasa jeda ucapan penutur.



Gambar 11. Grafik frasa jeda 1+2 pengujian K-*Fold cross validation*

dengan jelas bahwa model trigram juga menghasilkan nilai rata-rata yang lebih tinggi daripada model bigram. Nilai rata-rata yang dihasilkan oleh kedua model n- gram adalah nilai yang baik karena terdapat banyak frasa yang terbentuk oleh indeks jeda "2" atau jeda panjang yang sesuai dengan frasa jeda ucapan penutur.

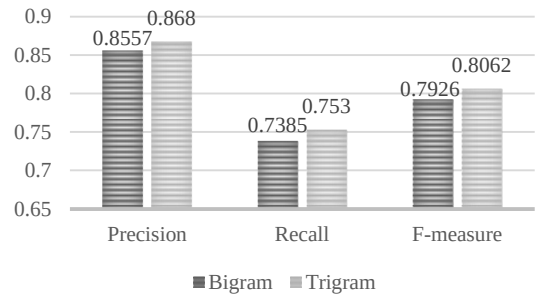
4. KESIMPULAN

Berdasarkan hasil analisis dan pengujian yang telah dilakukan terhadap sistem prediksi jeda dalam ucapan kalimat bahasa Melayu Pontianak menggunakan *Hidden Markov Model* berbasis *part of speech* yang dibuat, maka dapat diambil kesimpulan bahwa *Hidden Markov Model* (HMM) dapat menghasilkan prediksi jeda ucapan pada kalimat bahasa Melayu Pontianak dengan menggunakan *part of speech* (PoS). Selain itu, berdasarkan pengujian yang telah dilakukan, masih terdapat beberapa hasil pengujian yang belum dapat mencapai nilai *baseline*. Berdasarkan hasil dari keseluruhan pengujian yang telah dilakukan, model trigram merupakan model yang cenderung lebih baik daripada model bigram dalam memprediksi jeda

Tabel 8. Hasil Pengujian untuk Frasa Jeda 2

Korpus Uji	Bigram			Trigram		
	Precision	Recall	F-measure	Precision	Recall	F-measure
1-100	0.9193	0.8362	0.8757	0.9193	0.8362	0.8757
101-200	0.7871	0.6455	0.7093	0.8129	0.6667	0.7326
201-300	0.859	0.7444	0.7976	0.8805	0.7778	0.826
301-400	0.8636	0.7415	0.7979	0.8636	0.7415	0.7979
401-500	0.8493	0.7251	0.7823	0.8639	0.7427	0.7987
Rata-Rata	**0.8557**	**0.7385**	**0.7926**	**0.868**	**0.753**	**0.8062**

Hasil pengujian untuk frasa jeda 2 dapat dilihat pada Tabel 8. Tabel 8 menunjukkan bahwa nilai rata-rata terbaik untuk frasa jeda 2 juga dihasilkan oleh model trigram dengan nilai rata-rata *precision* sebesar 0.868, *recall* sebesar 0.753 dan *F-measure* sebesar 0.8062. Nilai rata-rata *precision, recall* dan *F-measure* yang dihasilkan belum dapat mencapai nilai *baseline*.



Gambar 12. Grafik frasa jeda 2 pengujian K-*Fold cross validation*

Hasil pengujian untuk frasa jeda 2 dapat direpresentasikan kedalam grafik seperti pada Gambar 12. Berdasarkan grafik tersebut dapat dilihat

ucapan pada kalimat bahasa Melayu Pontianak. Sistem prediksi jeda pada penelitian ini belum dapat menghasilkan nilai yang cukup baik untuk pengujian frasa jeda 1 yang digabung dengan frasa jeda 2 atau frasa jeda 1+2 karena hasilnya masih dibawah 70%. Hasil dari semua skenario pengujian yang telah dilakukan menunjukkan bahwa sistem prediksi jeda dapat memprediksi indeks jeda "2" atau jeda panjang dengan baik dan ditunjukkan dengan nilai *precision, recall* dan *F-measure* serta nilai akurasi frasa jeda ucapan dalam satu kalimat penuh untuk frasa jeda 2 yang cukup tinggi, sehingga jeda panjang dapat diimplementasikan untuk sistem *text to speech* (TTS). Perlu adanya penelitian lebih lanjut agar dapat meningkatkan keakuratan prediksi jeda pada bahasa Melayu Pontianak dengan menambahkan file ucapan dari sumber penutur ahli yang berbeda. Selain itu, penelitian selanjutnya juga dapat dilakukan dengan menambah jumlah n-gram untuk dibandingkan hasilnya atau membandingkan nilai akurasi translasi penggunaan *set* PoS yang telah dikembangkan pada penelitian ini dengan *set* PoS lainnya yang telah dikembangkan pada penelitian lain.

DAFTAR PUSTAKA

ADRIANI, M. 2009. Developing Postag for Bahasa Indonesia. Jakarta: PAN Localization Project.

CHRISTIANTI M, V., J. Pragantha dan Victor. 2016. Part-of-Speech Tagging untuk Bahasa Indonesia Menggunakan Stanford POS-Tagging. Jakarta: Universitas Tarumanegara.

DO, Q. T., et al. 2015. Improving Translation of Emphasis with Pause Prediction in Speech-to-speech Translation Systems. Japan: Nagoya University.

KAMILUDIN, M. I. 2017. Prediksi Jeda pada Ucapan Bahasa Melayu Pontianak dengan Menggunakan Metode Shallow Parsing. Pontianak: Universitas Tanjungpura.

MANURUNG, R. 2016. Tutorial: Pengenalan terhadap POS tagging dan Probabilistic Parsing. Workshop Nasional INACL.

NA'IM, A. dan Syahputra, H. 2011. Hasil Sensus Penduduk 2010: Kewarganegaraan, Suku Bangsa, Agama dan Bahasa Sehari-Hari Penduduk Indonesia, Jakarta: Badan Pusat Statistik.

NUGRAHA, A. T. 2014. Prediksi Jeda dalam Ucapan Kalimat Bahasa Indonesia dengan Hidden Markov Model. Pontianak: Universitas Tanjungpura.

SETYANINGSIH, E. 2017. Part of Speech Tagger untuk Bahasa Indonesia dengan Menggunakan Modifikasi Brill. Surabaya: Sekolah Tinggi Teknik Surabaya.

WIBISONO, Y. 2008. Penggunaan Hidden Markov Model untuk Kompresi Kalimat. Bandung: Institut Teknologi Bandung.

WICAKSONO, A. F. dan Purwarianti, A. 2010. HMM Based Part-of-Speech Tagger for Bahasa Indonesia. Proceedings of 4th International MALINDO (Malay and Indonesian Language) Workshop.

YU, J. dan Tao, J. 2005. The Pause Duration Prediction for Mandarin Text-to-Speech System. China: Chinese Academi of Science.

Halaman ini sengaja dikosongkan