Identifikasi Kata Majemuk Bahasa Indonesia

Fikri Haykal[1], Arie Ardiyanti Suryani[2], Sri Widowati[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]fikrihaykal@students.telkomuniversity.ac.id, [2]ardiyanti@telkomuniversity.ac.id,
[3]sriwidowati@telkomuniversity.ac.id

Abstrak

***Multi-word Expression* (MWE) tokenizer merupakan mesin untuk melakukan tokenisasi lebih dari dua kata, yang dapat digunakan untuk melakukan identifikasi kata majemuk. Pada tugas akhir ini dilakukan pembangunan mesin tersebut berbahasa Indonesia dengan metode berbasis aturan (*rule based*) berdasarkan pola kata majemuk dengan menggunakan tiga metode POS *tagger* yaitu, *Conditional Random Fields* (CRF) *tagger,* Bigram *tagger*, dan *Classifier Based tagger* dengan data latih sebanyak 226.328 kata dan data uji sebanyak 1.865 kata, lalu setelah melakukan uji coba dan evaluasi hasil, akurasi yang didapatkan dengan CRF *tagger* sebesar 77.97%, total kata yang didapat 295 kata kandidat kata majemuk, jumlah benar 230 kata dan jumlah salah 65 kata, lalu dengan Bigram *tagger* akurasi yang didapat sebesar 86,80%, total kata yang didapat sebanyak 144 kata kandidat kata majemuk, jumlah benar 125 kata dan jumlah salah 19 kata, dan yang terakhir menggunakan *Classifier Based tagger* akurasi yang didapat sebesar 82,13%, total kata yang didapat 235 kata kandidat kata majemuk, jumlah benar 193 kata dan jumlah salah 42 kata, jadi, jika menggunakan Bigram *tagger,* jumlah kata yang didapatkan sedikit tetapi akurasi yang didapatkan tinggi, sedangkan jika menggunakan CRF *tagger*, jumlah kata yang didapatkan banyak tetapi akurasi yang didapatkan rendah.**

Kata kunci: ***Multi-word Expression, Tokenizer, Rule Based, Conditional Random Fileds Tagger, Bigram Tagger, Classifier Based Tagger***

Abstract

Multi-word Expression (MWE) tokenizer is a machine to tokenize more than two words, which can be used to identify compound words. In this final project, the construction of the machine in Indonesian with a rule-based method based on compound word patterns using three POS tagger methods, namely, Conditional Random Fields (CRF) tagger, Bigram tagger, and Classifier Based tagger with as many as 226,328 training data. The word and test data were 1,865 words, then after testing and evaluating the results, the accuracy obtained with the CRF tagger was 77.97%, the total words obtained were 295 compound word candidate words, 230 correct words and 65 wrong words, then with Bigram The accuracy tagger obtained is 86.80%, the total words obtained are 144 compound word candidate words, 125 correct words and 19 incorrect words, and the last one using Classifier Based tagger, the accuracy obtained is 82.13%, the total words used There are 235 candidate compound words, 193 correct words and 42 incorrect words, so, if you use Bigram tagger, you get less words but the accuracy you get is high, whereas if you use the CRF tagger, you get more words but the accuracy you get is low.

Keywords: Multi-word Expression, Tokenizer, Rule Based, Conditional Random Fields Tagger, Bigram Tagger, Classifier Based Tagger

1. Pendahuluan

Latar Belakang

Multi-word expression merupakan sebuah kata yang menghasilkan makna baru yang terdiri dari setidaknya dua kata dasar yang digabungkan [7]. Dalam bahasa Indonesia contoh dari *Multi-word expression* adalah kata majemuk karena, gabungan morfem dasar yang seluruhnya berstatus sebagai kata yang mempunyai pola bunyi, gramatikal dan semantis yang khusus menurut kaidah bahasa yang bersangkutan. Pola khusus tersebut membedakannya dari gabungan morfem dasar yang bukan kata majemuk, atau dengan kata lain kata majemuk merupakan hasil proses perpaduan dua unsur kata yang mengandung satu makna baru [4, 5]. Pada tugas akhir ini dilakukan pembangunan mesin identifikasi kata majemuk dengan acuan aplikasi POS *tagger* yang dibangun oleh Universitas Indonesia, karena, pada POS *tagger* tersebut jika menggunakan *multi-word expression tokenizer* dengan CRF *tagger* akurasi yang didapat sebesar 70% sedangkan, jika tidak menggunakan *multiword expressions tokenizer* tersebut akurasi POS *tagger* sebesar 79% [1]. Oleh karena itu pada tugas akhir ini dilakukan pembangunan mesin identifikasi kata majemuk berbahasa Indonesia dengan menggunakan tiga POS *tagger* dengan metode yang berbeda, yaitu, CRF *tagger*, Bigram *tagger* dan *Classifier Based tagger* untuk mengetahui apakah POS *tagger* dengan tiga metode yang berbeda tersebut akan memengaruhi hasil akhir dari mesin identifikasi kata mejemuk berbahasa Indonesia yang dibangun ini dan akurasi hasil dapat mencapai >=80% atau tidak, dan juga dengan harapan dapat diimplementasikan pada POS Tagger bahasa Indonesia yang ada.

Topik dan Batasannya

Topik permasalahan pada tugas akhir ini adalah menurunnya akurasi ketepatan pada POS *tagger* yang dibangunn oleh Universitas Indonesia, jika menggunakan MWE *tokenizer* dengan metode CRF *tagger* akurasi ketepatan bernilai 70%, sedangkan jika tidak menggunakan MWE *tokenizer* tersebut akurasi ketepatan bernilai 79%. Hal tersebut berdampak pada hasil yang didapatkan akan menjadi kurang baik, karena, ada beberapa kata yang seharusnya merupakan kata majemuk tetapi mesin tidak dapat mengidentifikasi kata majemuk tersebut. Oleh karena itu pada tugas akhir ini akan dilakukan pembangunan mesin identifikasi kata majemuk dengan harapan akurasi hasil sebesar >=80%, adapun batasannya, mesin identifikasi kata majemuk ini hanya berbahasa Indonesia, hanya untuk kata majemuk *non*-senyawa, evaluasi dilakukan secara manual berlandaskan pola dan karakteristik kata majemuk yang sudah didefinisikan oleh ahli bahasa Indoensia, dan mengabaikan abiguitas kata majemuk tersebut.

Tujuan

Tujuan tugas akhir ini adalah mesin yang di bangun dapat menghasilkan akurasi hasil sebesar >=80% dengan harapan dapat digunakan untuk diimplementasikan pada POS Tagger berbahasa Indonesia yang ada.

Organisasi Tulisan

Pada jurnal ini dibagi menjadi 4 bagian, yaitu;

1. Studi Terkait

   Bagian ini berisi teori/studi/literatur yang mendukung (terkait erat) dengan topik TA yang dikerjakan. Bagian ini bisa bernama Tinjauan Pustaka atau Landasan Teori. Dalam bahasa Inggris disebut sebagai Related Work atau Literature Review.
2. Sistem yang Dibangun

   Pada bagian ini dijelaskan rancangan dan sistem yang dihasilkan
3. Evaluasi

   Pada bagian ini berisi dua sub-bagian, yaitu Hasil Pengujian dan Analisis Hasil Pengujian. Pengujian dan analisis yang dilakukan selaras dengan tujuan TA sebagaimana dinyatakan dalam Pendahuluan.
4. Kesimpulan

   Pada bagian ini berisi kesimpulan dan saran (f*uture works*) mengenai hasil dan apa yang bisa dilakukan berikutnya pada hasil pengerjaan ini.

2. Studi Terkait

2.1. Kata Majemuk

Kata majemuk adalah gabungan morfem dasar yang seluruhnya berstatus sebagai kata yang mempunyai pola bunyi, gramatikal dan semantis yang khusus menurut kaidah bahasa yang bersangkutan. Pola khusus tersebut membedakannya dari gabungan morfem dasar yang bukan kata majemuk, atau dengan kata lain kata majemuk merupakan hasil proses perpaduan dua unsur kata yang mengandung satu makna atau pengertian baru [4, 5].

Kata majemuk berbeda dengan idiom dan juga frasa, tetapi kata majemuk dan idiom hanya memiliki perbedaan yang sangat sedikit dibandingkan dengan frasa, karena kata majemuk dan idiom ketika dua leksem dipadukan akan membentuk makna baru, berbeda dengan frasa, ketika dua leksem dipadukan tidak membentuk makna baru [6].

Tabel 2.1 1 Perbedaan Kata Majemuk, Idiom, dan Frasa

	Kata Majemuk	**Idiom**	**Frasa**
Dominasi kata	Tidak ada	Melebur menjadi satu makna	Kata berdiri sendiri
Proses terbentuk	A+B = AB (membentuk makna baru)	A+B = C (membentuk makna baru)	A+B = AB (tidak membentuk makna baru)
Contoh kata	Sapu tangan	Buah bibir	Laut luas
Frekuensi kata	Sering berdampingan	Selalu berdampingan	Tidak selalu berdampingan

2.1.1. Karakteristik Kata Majemuk

Adapun karakterisitk dari kata majemuk, yaitu:

- Membentuk Makna Baru

  Contoh: rumah sakit, matahari, meja makan, kaki tangan, dan lain – lain
- Tidak Dapat Disisipi oleh Kata Lain

  Contoh: anak buah, pancaindera, hulubalang, tinggi hati, dan lain-lain
- Merupakan Kata Dasar Tanpa Imbuhan

  Contoh: sapu tangan, air mata, anak tiri, dan lain-lain

- Unsur Kata yang Membentuk Tidak Dapat Dibalik
  Contoh: orang tua, alam semesta, kaki tangan, dan lain - lain

2.1.2. Pola Kata Majemuk

Adapun pola kata majemuk, yaitu: [6]

- Dapat terbentuk dari gabungan kata benda + kata sifat, begitu sebaliknya.
- Dapat terbentuk dari gabungan kata benda + kata benda
- Dapat terbentuk dari gabungan kata benda + kata kerja
- Dapat terbentuk dari gabungan kata bilangan + kata benda
- Dapat terbentuk dari gabungan kata kerja + kata kerja
- Dapat terbentuk dari gabungan kata sifat + kata sifat

2.2. Conditional Random Fields (CRF) *Tagger*

CRF merupakan suatu *framework* untuk membangun model probabilistik diskriminatif untuk segmentasi dan pelabelan data [12, 13]. CRF adalah sebuah *undirected graphical model* di mana setiap titik mewakili variabel acak Yi yang distribusinya dikondisikan pada beberapa urutan pengamatan X, dan masing-masing *edge* merupakan ketergantungan antara dua variabel. Sehingga CRF *tagger* dapat diformulasikan sebagai berikut [12]:

$$P(Y|X,\lambda) = \frac{1}{Z(X)} \exp(\sum_{j} \lambda_j \sum_{i=1}^{n} f_j\,(Y_{i-1}, Y_i, X, i))$$

Dimana Z(X) adalah *normalization factor*. Pada pengaplikasiannya CRF *tagger* dalam melakukan POS *tagging* berbahasa Inggris menghasilkan akurasi 97% dan beberapa bahasa lain seperti Hindi sebesar 80.97%, Bengali sebesar 82.74%, dan Telugu sebesar 79.15% [12].

2.3. Bigram *Tagger*

Bigram *tagger* adalah sebuah *subclass* dari *sequential tagging method* dengan pendekatan *bigram method* dimana proses melabeli setiap kata dengan cara mempertimbangkan konteks setiap kata dengan menganalisisnya secara berpasangan sehingga bigram model diformulasikan sebagai berikut [14]:

$$P(w_i|w_1 w_2 \dots w_{i-1}) \approx P(w_i|w_{i-1})$$

Jadi, yang dimaksud dengan mempertimbangkan konteks setiap kata dengan menganalisisnya secara berpasangan adalah dengan mengetahui nilai peluang dari kata sebelumnya terhadap kata saat ini, lalu untuk mengestimasi probabilitas dari bigram diformulasikan sebagai berikut:

$$P(w_i|w_{i-1}) = \frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$$

Dengan w_i adalah kata saat ini dan w_{i-1} adalah kata sebelumnya dari w_i lalu didapat estimasi probabilitas bigram tersebut dengan menghitung total kemunculan w_i dengan w_{i-1} pada suatu kalimat dibagi dengan total kemunculan w_{i-1} pada suatu kalimat

2.4. Classifier Based Tagger

Classifier Based Tagger merupakan *subclass* dari sebuah *sequential tagging method* dengan pendekatan teorema *NaiveBayes Classifier* [15].

$$P(y|x_1, x_2, \dots, x_n) = \frac{P(x_1|y)P(x_2|y) \dots P(x_n|y)P(y)}{P(x_1)P(x_2) \dots P(x_n)}$$

Dimana y adalah kelas kata/label, x adalah kata dari hasil tokenisasi kalimat, $P(x_i|y)$ adalah probabilitas y terhadap x_i, P(y) probabilitas kemunculan y, $P(x_i)$ probabilitas kemunculan x_i [16].

2.5. Rule-Based Part-of-Speech Tagger

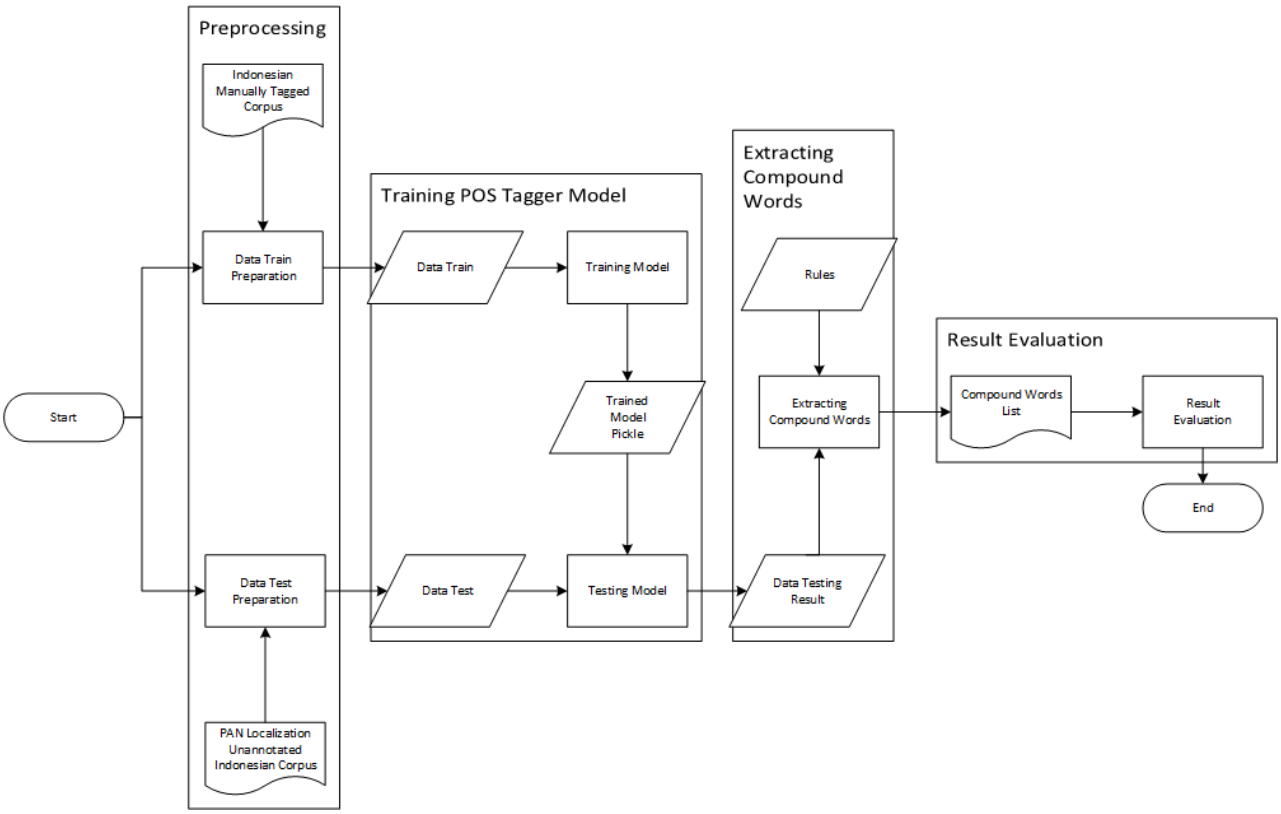
Rule-based part-of-speech tagger adalah sebuah aplikasi yang dapat melakukan pelabelan pada setiap kata dalam suatu kalimat dengan PoS atau tag yang sesuai dengan kelas katanya secara otomatis menggunakan pendekatan rule-based berdasarkan aturan tata bahasa Indonesia [1, 2].

Rule-based PoS tagger yang akan dibangun ini digunakan untuk mengidentifikasi kata majemuk bahasa Indonesia. Aturan utama yang akan diterapkan dalam rule-based PoS tagger ini adalah tokenisasi kalimat dengan Mulltiword Expression Tokenizer yang ditambahkan dengan pola dan karkteristik dari kata majemuk [1].

2.6. Korpus

Korpus adalah kumpulan ujaran yang tertulis atau lisan yang dipergunakan untuk menyokong atau menguji hipotesis tentang struktur bahasa [4]. Untuk membangun mesin identifikasi kata majemuk bahasa Indoneisa ini menggunakan dua jenis korpus, yaitu, korpus yang sudah berlabel dan korpus yang belum tidak berlabel, korpus yang sudah berlabel dijadikan sebagai data latih dengan jumlah kata sebanyak 226.328 kata dan digunakan untuk melatih model POS *tagger*, lalu untuk korpus yang tidak berlabel dijadikan sebagai data uji dengan jumlah kata sebanyak 1.865 kata dan digunakan untuk pengujian ekstraksi kata majemuk *non*-senyawa dengan *rules* yang sudah didefinisikan sebelumnya.

3. Sistem yang Dibangun



Gambar 1. Alur kerja mesin yang dibangun

Proses pembangunan mesin identifikasi kata majemuk bahasa Indoneisa ini dapat dilihat pada gambar 1 di atas, yang terdiri dari empat tahap utama, yaitu, *Preprocessing*, *Training POS Tagger Model*, *Extracting Compound Words*, dan *Result Evaluation*, tahap pertama dalam mesin ini yaitu melakukan *Preprocessing* yang dibagi menjadi dua proses, yaitu, proses *data train preparation* dan proses *data test preparation*, pada proses *data train preparation* dilakukan *converting data train* menjadi token-token perkalimat beserta kelas katanya pada setiap kata dalam kalimat tersebut, lalu untuk proses *data test preparation* prosesnya sama seperti *data train preparation* tetapi data yang digunakan berbeda, untuk *data train* menggunakan korpus berbahasa Indonesia yang sudah dianotasikan secara manual dengan jumlah kata sebanyak 226.328 kata dan untuk *data test* menggunakan korpus yang tidak dianotasikan dan dilakukan filtrasi kalimat mana saja yang sedikitnya mengandung satu kata majemuk *non*-senyawa sebanyak 1.865 kata, lalu hasil dari tahap *Preprocessing* berupa kumpulan kata yang akan digunakan untuk tahap berikutnya yaitu *Training* POS *Tagger Model*, pada tahap kedua ini, dibagi menjadi dua proses, yaitu, proses *training model* dan *testing model*, pada proses *training model*, data yang digunakan adalah *data train*, lalu hasil *training model* digunakan untuk melakukan *testing model* dengan menggunakan *data test* dengan *output data test* yang sudah berlabel/*tagged*, pada tahap ketiga hasil dari *testing model* akan diproses pada fungsi *extracting compound words* dengan menggunakan *pharse tree* beserta *rules* yang sudah dibuat berdasarkan pola dari kata majemuk, setelah melalui proses *extraction* maka didapatkan hasil berupa kumpulan kata majemuk berupa dokumen, lalu dokumen hasil dari tahap sebelumnya yang berisi kata majemuk akan dievalusi secara manual dengan rumus *mean* ($m = \frac{total\ benar}{total\ hasil}\ x\ 100\%$) digunakan sebagai nilai akurasi ketepatan hasil.

4. Evaluasi

4.1. Hasil Pengujian

Dari pengujian yang sudah dilakukan didapatkan hasil sebagai berikut:

Tabel 1. Hasil pengujian

No.	POS *Tagger*	Total Kata yang Didapat	Jumlah Benar	Jumlah Salah	Nilai Akurasi
1.	*Conditional Random Fields Tagger*	295	230	65	77,97%
2.	Bigram *Tagger*	144	125	19	86,80%
3.	*Classifier Based Tagger*	235	193	42	82,13%

4.2. Analisis Hasil Pengujian

Dari hasil pengujian dan evaluasi yang dilakukan secara manual dengan cara menyeleksi setiap kandidat kata majemuk yang didapat dengan menyamakan karakteristik dan pola kata majemuk yang sudah didefinisikan berdasarkan kaidah kata majemuk bahasa Indonesia dan melihat konteks kata majemuk tersebut dari kalimat yang terdapat pada data uji [6] didapatkan hasil seperti pada tabel 1 dapat dilihat nilai akurasi pada Bigram *Tagger* sebesar 86,80%, lalu pada *Classifier Based Tagger* nilai akurasi sebesar 82,13%, dan yang terakhir pada CRF *Tagger* nilai akurasi yang didapat sebesar 77,97%, dan juga ketiga POS *tagger* tersebut total kata yang didapatkan berbeda-beda,pada Bigram *Tagger* total kata yang didapatkan sebanyak 144 kata dengan 125 kata majemuk dan 19 kata bukan kata majemuk, pada *Classifier Based Tagger* total kata yang didapatkan sebanyak 235 kata dengan 193 kata majemuk dan 42 kata bukan kata majemuk, dan pada CRF *Tagger* total kata yang didapatkan sebanyak 295 kata dengan 230 kata majemuk dan 65 kata bukan kata majemuk.

5. Kesimpulan

Mesin identifikasi kata majemuk bahasa Indonesia berbasis aturan pada tugas akhir ini, selain *data train*, *data test*, dan *rules* yang digunakan dapat memengaruhi akurasi ketepatan, penggunaan metode POS *Tagger* juga dapat memengaruhi akurasi ketepatan mesin ini, dikarenakan setiap POS *Tagger* memiliki *output* yang berbeda-beda. Pada Bigram *Tagger* total kata yang didapatkan sebanyak 144 kata dengan 125 kata majemuk dan 19 kata bukan kata majemuk dengan akurasi ketepatan sebesar 86,80%, pada *Classifier Based Tagger* total kata yang didapatkan sebanyak 235 kata dengan 193 kata majemuk dan 42 kata bukan kata majemuk dengan akurasi ketepatan sebesar 82,13%, dan pada CRF *Tagger* total kata yang didapatkan sebanyak 295 kata dengan 230 kata majemuk dan 65 kata bukan kata majemuk dengan akurasi ketepatan sebesar 77,97%, jadi, jika menggunakan Bigram *tagger*, kandidat kata majemuk yang didapat lebih sedikit, tetapi nilai akurasi ketepatan yang didapatkan tinggi, sedangkan, jika menggunakan CRF *tagger* kandidat kata majemuk yang didapet lebih banyak, tetapi nilai akurasi ketepatan yang didapatkan rendah.

Daftar Pustaka

[1] F. Rashel, A. Luthfi, A. Dinakaramani, and R. Manurung, "Building an Indonesian rule-based part-of-speech tagger," Proc. Int. Conf. Asian Lang. Process. 2014, IALP 2014, pp. 70–73, 2014.

[2] R. Geometry and G. Analysis, "Penyelesaian Kata Ambigu Pada Proses POS Tagging Menggunakan Algoritma Hidden markov Model (HMM)," no. 978, pp. 347–358, 2017.

[3] A. Purwantiari and T. Suhardijanto, "INACL POS Tagging Convention Konvensi Pelabelan Kelas Kata INACL / MALKIN," 2017.

[4] H. Kridalasana, "Kamus Linguistik Keempat," 2009.

[5] KBBI, Kamus Bahasa Indonesia, vol. 1. 2008.

[6] M. A. Khak, "Idiom Dalam Bahasa Indonesia: Struktur Dan Maknai," Widyaparwa, vol. 39, no. 2, pp. 141–154, 2011.

[7] I. Ohnheiser and S. Olsen, "Multi-word expressions" no. 1969, pp. 1–16, 2011.

[8] Tim Pengembang Pedoman Bahasa Indonesia, Pedoman Umum Ejaan Bahasa Indonesia Edisi Keempat. 2016.

[9] V. C. M, J. Pragantha, and E. Purnamasari, "Implementasi Brill Tagger Untuk Memberikan POS-Tagging Pada Dokumen Bahasa Indonesia" Jakarta Jurnal Teknik dan Ilmu Komputer, pp. 301–315, 2012.

[10] E. Brill, "Some advances in transformation-based part of speech tagging," *Proceedings of the National Conference on Artificial Intelligence*, vol. 1, pp. 722–727, 1994.

[11] M. Alex and L. Q. Zakaria, "Brill's Rule-based Part of Speech Tagger for Kadazan," vol. 10, no. 1, 2014.

[12] Avinesh, P., and Karthik, G. "Part of Speech Tagging and Chunking using Conditional Random Fields and Transformation Based Learning". *Proceedings of IJ-CAI Workshop on "Shallow Parsing for South Asian Languages"*, 2007.

[13] Lafferty, J., McCallum, A., and Pereira, F., "Conditional Random Field: Probabilistic Model for Segmenting and Labeling Sequence Data". *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

[14] Jurafsky, D. 2019. "Language Modeling: Introducing to N-grams." [Online] Available at: https://web.stanford.edu/~jurafsky/slp3/slides/LM_4.pdf [Accessed 18 Agustus 2020]

[15] Steven, B., Edward, L., Ewan, K.. 2001. "NLTK 3.5 Documentation: nltk.tag package" [Online] Available at: http://www.nltk.org/api/nltk.tag.html#module-nltk.tag.sequential [Accessed 18 Agustus 2020]

[16] M. Rajasekar and A. Udhayakumar, "POS tagging using naïve bayes algorithm for tamil," Int. J. Sci. Technol. Res., vol. 9, no. 2, pp. 574–578, 2020.