PENANDAAN GOLONGAN KATA BAHASA MELAYU MENGGUNAKAN PENDEKATAN BERASASKAN PERATURAN

AHMAD NUR HADI MUSTAQIM BIN ABDUL MUTALIB

NAZLIA OMAR

Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Penandaan Golongan Kata merupakan satu proses menandakan setiap kata dalam teks (korpus) bertepatan dengan bahagian tertentu ucapan berdasarkan definisi dan konteksnya dengan setiap kelas token seperti kata nama, kata kerja, kata sifat dan sebagainya. Penandaan Golongan Kata (GK) merupakan salah satu tugasan asas yang penting dalam proses capaian maklumat. Secara umumnya, terdapat tiga kaedah yang boleh digunakan dalam proses penandaan GK iaitu berasaskan linguistik, kaedah statistikal dan pendekatan berasaskan mesin. Masalah utama yang sering dihadapi dalam proses penandaan GK adalah kewujudan perkataan ketaksaan (ambiguity) dan perkataan tidak diketahui (unknown). Selain itu, kekurangan set peraturan sedia ada dan imbuhan sisipan yang tidak dinyatakan dalam kebanyakan pendekatan sedia ada juga menyukarkan proses penandaan GK dalam Bahasa Melayu. Objektif utama kajian ini adalah untuk membangunkan peraturan baru bagi penandaan GK Bahasa Melayu dan membangunkan alat penandaan GK berasaskan gabungan peraturan baru dan peraturan sedia ada. Proses ini bermula dengan pernyataan masalah, kajian literatur, pengumpulan data, pra-pemprosesan, pembangunan peraturan dan analisa keputusan yang dilakukan secara manual. Terdapat beberapa kekangan yang dihadapi sepanjang menghasilkan projek ini iaitu sumber rujukan untuk penandaan GK dalam Bahasa Melayu yang terhad, set peraturan dan alatan penandaan GK Bahasa Melayu yang terhad. Secara tuntasnya, hasil daripada kajian ini diharap dapat membantu para penyelidik dalam melaksanakan penandaan GK bagi korpus Bahasa Melayu dengan menghasilkan nilai ketepatan yang lebih tinggi melalui penghasilan peraturan baharu.

1 PENGENALAN

Pada masa kini, capaian maklumat adalah amat penting untuk mengumpul maklumat yang terdapat di sebalik setiap jenis dokumen yang tersedia atas talian. Capaian maklumat adalah

sains yang bertujuan untuk menyimpan dan membenarkan akses laju kepada satu jumlah informasi yang besar. Informasi ini terdapat dalam pelbagai jenis iaitu teks, visual atau pendengaran (Rani 2011). Terdapat beberapa tugasan yang perlu dilakukan dalam proses untuk capaian maklumat.

Salah satu tugasan asas dalam capaian maklumat ialah penandaan golongan kata. Penandaan golongan kata adalah proses menentukan kelas tatabahasa atau kelas perkataan setiap perkataan dengan tepat (Mohd Pouzi & Syarifah Fatem Na'imah 2014). Penandaan golongan kata memainkan peranan penting dalam sistem capaian maklumat terutama dalam ekstraksi maklumat dan penghuraian teks. Secara umumnya, proses penandaan golongan kata melibatkan tiga kaedah iaitu berasaskan linguistik, kaedah statistikal dan pendekatan pembelajaran mesin. Contoh linguistik ialah pendekatan berasaskan peraturan manakala contoh statistikal adalah Model Markov Tersembunyi (MMT). Contoh pendekatan pembelajaran mesin adalah Model Pepohon Keputusan (MPK).

Bagi Bahasa Inggeris, penandaan golongan kata telah digunakan secara meluas dan terdapat pelbagai algoritma dan peraturan yang tersedia untuk menghasilkan penandaan golongan kata. Ini mungkin kerana kaedah-kaedah frasa tatabahasa yang mudah dan tidak rumit untuk memahami dan digunakan (Alfred, Mujat & Obit 2013). Hal ini berbeza bagi Bahasa Melayu kerana Bahasa Melayu mempunyai tatabahasa yang pelbagai dan setiap perkataan boleh membawa maksud yang berbeza di dalam konteks ayat yang berbeza. Linguistik tradisional dibangunkan dengan baik untuk Bahasa Melayu tetapi terdapat sumber dan peralatan yang terhad tersedia untuk analisis linguistik komputer dalam Bahasa Melayu. Hal ini termasuklah sumber dan peralatan untuk penandaan golongan kata untuk Bahasa Melayu.

2 PERNYATAAN MASALAH

Masalah utama yang sering berlaku dalam proses penandaan golongan kata adalah kewujudan perkataan ketaksaan (ambiguity) dan perkataan tidak diketahui. Perkataan ketaksaan adalah perkataan sama yang membawa makna yang berbeza di dalam konteks yang berbeza. Misalnya perkataan *haus* yang membawa dua makna berbeza di dalam ayat berikut: "Saya hendak membeli air kerana *haus*." *Haus* disini membawa maksud tekak yang kering. Ayat berikutnya pula: "Tapak kasut adik sudah *haus*." *Haus* disini membawa maksud berkurangan ketebalannya kerana tergosok.

Seterusnya, untuk perkataan yang tidak diketahui pula ialah perkataan baru yang sama ada tidak terdapat dalam korpus latihan ataupun dalam kamus penandaan golongan kata

(Hassan Mohamed, Nazlia Omar & Mohd Juzaidin 2011). Kebiasaannya, jenis perkataan ini terhasil hasil daripada ejaan yang salah, singkatan yang digunakan untuk sesebuah perkataan ataupun perkataan daripada bahasa asing. Sebagai contoh, singkatan perkataan yang kepada "yg" dan "selfie".

Selain itu, masalah utama yang terdapat dalam proses penandaan golongan kata untuk Bahasa Melayu adalah kekurangan set peraturan sedia ada yang memerlukan kita membina secara manual dan tersendiri set peraturan untuk penandaan. Selain itu, wujud masalah imbuhan sisipan yang tidak dinyatakan dalam kebanyakan pendekatan sedia ada (Nur Ashikin & Nazlia Omar 2017). Tambahan pula, dari sudut bahasa, sesuatu perkataan asal atau perkataan akar yang ditambah imbuhan akan mengubah makna asal perkataan tersebut (Nik Safiah et al. 2015).

3 OBJEKTIF KAJIAN

Objektif kajian ini adalah seperti berikut:

i. Membangunkan peraturan baru bagi penandaan golongan kata (GK) Bahasa Melayu.

ii. Membangunkan alat penandaan GK berasaskan gabungan peraturan baru dan peraturan sedia ada.

4 METOD KAJIAN

Metod yang digunakan dalam kajian ini merangkumi keseluruhan proses kajian yang dijalankan bermula dari pernyataan masalah, kajian literatur, pengumpulan data, pra-pemprosesan, pembangunan peraturan dan analisis keputusan. Penggunaan model pembangunan yang sesuai penting untuk memastikan perjalanan kajian berjalan dengan lancar dan menjamin hasil kerja yang berkualiti. Proses kajian ini terbahagi kepada empat fasa iaitu Fasa Analisis, Fasa Penyediaan Data, Fasa Pembangunan dan Fasa Pengujian.

4.1 Fasa Analisis

Terdapat dua proses di dalam fasa analisis iaitu proses pertama adalah pernyataan masalah di mana masalah dicatat dan dinyatakan yang membawa kepada tercetusnya idea untuk menghasilkan kajian ini. Masalah ini hadir apabila terdapat informasi yang tersedia di dalam jumlah yang besar yang memerlukan pengekstrakan maklumat dilakukan menggunakan sistem capaian maklumat. Informasi ini pula tersedia di dalam Bahasa Melayu yang memerlukan

pembangunan alatan dan peraturan yang lebih banyak di dalam Bahasa Melayu untuk memudahkan proses pengekstrakan maklumat di dalam Bahasa Melayu. Proses seterusnya adalah proses kajian literatur di mana kajian dilakukan ke atas bahasa yang hendak digunakan di dalam kajian. Di dalam proses ini juga, pemilihan kaedah dan pendekatan yang akan digunakan disesuaikan ke atas bahasa yang digunakan untuk kajian.

4.2 Fasa Penyediaan Data

Fasa ini bermula dengan proses pengumpulan data yang dikumpul dan diasingkan kepada korpus latihan dan korpus ujian. Secara umumnya, data boleh dibahagikan kepada dua iaitu data primer dan data sekunder. Data yang digunakan di dalam kajian ini ialah data sekunder. Data yang telah dipilih untuk digunakan di dalam kajian ini ialah data daripada sumber terbuka seperti surat khabar atas talian iaitu Berita Harian dan Kosmo! Online. Data ini diekstrak ke dalam bentuk fail teks dan sebanyak 100 korpus telah digunakan hasil gabungan daripada data penyelidik terdahulu seperti Nur Ashikin & Nazlia Omar (2017) dan data yang telah diekstrak sendiri. Data ini telah diagihkan mengikut nisbah 80:20 di mana 80 peratus ialah korpus latihan dan 20 peratus ialah korpus ujian.

Setelah korpus dikumpul dan telah dipilih, korpus-korpus dalam bentuk teks mentah tersebut akan melalui pra-pemprosesan di mana korpus tersebut akan dibersihkan sebelum digunakan sebagai korpus latihan dan korpus ujian supaya hasil penetapan penandaan GK akan lebih tepat dan penanda GK perkataan akan lebih mudah dikenalpasti. Proses pra-pemprosesan bermula dengan korpus yang terdiri daripada teks mentah akan dipisahkan kepada bentuk ayat di mana setiap ayat akan dikenalpasti melalui tanda noktah, tanda seru atau tanda soal yang berada di akhir setiap ayat. Tanda-tanda akan menentukan pengakhiran setiap ayat untuk dipisahkan. Setelah dipisahkan ke dalam bentuk ayat, korpus tersebut akan melalui proses normalisasi di mana setiap tanda baca dan simbol yang tidak relevan akan disingkirkan daripada ayat untuk penyediaan perkataan sebelum melalui proses pentokenan. Proses terakhir di dalam pra-pemprosesan ialah proses pentokenan di mana setiap perkataan dipisahkan melalui ruang kosong (*white space*). Dalam proses ini, ruang kosong (*white space*) bertindak sebagai sempadan yang digunakan untuk memisahkan setiap perkataan.

4.3 Fasa Pembangunan

Di dalam kajian ini, fasa pembangunan terbahagi kepada dua iaitu pembangunan kamus GK dan set peraturan yang akan menetapkan penanda GK kepada setiap perkataan serta

pembangunan peralatan yang akan digunakan bagi menjalankan dan menyiapkan proses penandaan GK.

4.3.1 Pembangunan Kamus GK dan Set Peraturan

Di dalam kajian ini, kamus yang dibangunkan adalah secara manual berdasarkan set GK daripada kamus Hawkins (2008) adaptasi daripada set tanda GK dalam Dewan Bahasa dan Pustaka (DBP) kamus dwibahasa Hock (2009) di mana terdiri daripada senarai perkataan yang mempunyai kata akar dengan penanda GK masing-masing bagi membentuk leksikon. Proses mengenalpasti penandaan GK bagi perkataan-perkataan tersebut dilakukan dengan mencari satu persatu kata akar bagi setiap perkataan dan sekiranya ianya wujud di dalam kamus, penandaan GK akan ditetapkan bagi perkataan tersebut. Sekiranya perkataan tidak dijumpai di dalam kamus, perkataan tersebut akan disemak melalui set peraturan yang dibangunkan.

Set-set peraturan yang dibina di dalam kajian ini adalah berdasarkan gabungan peraturan sedia ada dan penambahan peraturan baharu seperti peraturan kata ganda dan sebagainya. Kajian ini menggunakan sebanyak 25 peraturan sedia ada yang diambil daripada peraturan yang dibangunkan oleh Alfred, Mujat dan Obit (2013) dan Nur Ashikin dan Nazlia Omar (2017) serta beberapa set peraturan baharu yang ditambah. Setiap set peraturan disusun mengikut aturannya tersendiri. Jadual 1.0 menunjukkan jenis GK di dalam set peraturan sedia ada ialah:

Jadual 1.0 Jenis GK di dalam set peraturan sedia ada

Bilangan	**Jenis GK**
1	Kata Nama
2	Kata Sifat
3	Kata Kerja
4	Kata Penegas
5	Kata Adverba
6	Kata Arah
7	Kata Sendi
8	Kata Bilangan
9	Kata Hubung
10	Kata Penguat
11	Kata Tanya

12	Kata Penentu
13	Penjodoh Bilangan
14	Kata Ganti Nama
15	Kata Pemeri
16	Kata Bantu Ragam
17	Kata Bantu Aspek
18	Kata Perintah
19	Kata Nama Khas
20	Kata Nafi
21	Kata Pembenar
22	Penanda Wacana
23	Kependekan
24	Gelaran
25	Kewujudan

Jadual 2.0 menunjukkan jenis GK di dalam set peraturan yang ditambah pula ialah:

Jadual 2.0 Jenis GK di dalam set peraturan sedia ada

Bilangan	**Jenis GK**
1	Kata Bilangan Pecahan
2	Kata Ganda Penuh
3	Kata Seru
4	Kata Pangkal Ayat
5	Kata Ganti Nama Tunjuk
6	Penjodoh Bilangan Tumbuhan
7	Kata Penguat Bebas
8	Kata Adjektif Waktu
9	Kata Adjektif Keadaan
10	Kata Adjektif Jarak
11	Kata Adjektif Ukuran
12	Kata Penguat Hadapan
13	Kata Penguat Belakang

Pembangunan jenis GK yang terdapat di dalam set peraturan untuk kajian ini menggunakan regex ataupun regular expression bagi mengenalpasti penandaan bagi perkataan tersebut. Operator untuk fungsi regex digunakan di dalam pembangunan peraturan ini seperti operator "||" iaitu atau, operator "&&" iaitu dan serta operator "?!" yang membawa maksud tidak bermula dengan.

4.3.2 Pembangunan Peralatan

Pembangunan peralatan yang digunakan di dalam kajian ini adalah menggunakan bahasa pengaturcaraan Python dan juga aplikasi Jupyter Notebook.

4.4 Fasa Pengujian

Di dalam fasa pengujian, kaedah yang digunakan untuk menguji keberkesanan kajian ini dalam menandakan GK Bahasa Melayu dibangunkan secara manual menggunakan bahasa pengaturcaraan Python dan aplikasi Jupyter Notebook. Proses penilaian akan dijalankan keatas hasil kejituan yang diperoleh daripada setiap ujian menggunakan kedua-dua korpus iaitu korpus latihan dan korpus ujian. Hasil daripada korpus latihan akan diuji sekali lagi menggunakan korpus ujian bagi menilai tahap kejituan penandaan GK. Formula untuk mendapatkan kejituan bagi proses penilaian adalah:

$$\text{Kejituan} = \frac{\text{bilangan perkataan yang ditanda dengan betul}}{\text{jumlah keseluruhan perkataan yang ditanda}} \text{ X } 100\%$$

Untuk mengetahui sama ada perkataan telah ditanda dengan betul, dokumen yang mengandungi perkataan yang telah ditanda menggunakan kamus GK ataupun set peraturan akan dibandingkan dengan dokumen yang mengandungi perkataan yang telah ditanda terlebih dahulu secara manual mengikut kamus tatabahasa dewan. Formula kejituan ini akan diguna untuk diuji keatas set latihan terlebih dahulu untuk mengenalpasti kesalahan yang terdapat di dalam peraturan dan mengenalpasti perkataan yang masih tidak dinyatakan peraturan untuk menandakannya di dalam set peraturan. Setelah pengujian ke atas set latihan mendapat hasil kejituan yang memuaskan, set peraturan yang telah ditambah baik akan diuji ke atas set ujian menggunakan formula kejituan ini. Hasil daripada kejituan yang diperoleh daripada set ujian

akan dibandingkan dengan hasil kejituan yang diperoleh menggunakan set peraturan daripada hasil kajian penyelidik terdahulu untuk melihat keberkesanan kajian ini.

5 HASIL KAJIAN

Hasil kajian ini terbahagi kepada tiga iaitu hasil kejituan yang diperoleh daripada set latihan, hasil kejituan yang diperoleh daripada set ujian dan yang terakhir ialah hasil perbandingan kejituan yang diperoleh daripada set ujian dengan set peraturan yang dibangunkan oleh penyelidik terdahulu iaitu Nur Ashikin (2017). Jadual 3.0 menunjukkan beberapa hasil kejituan yang diperoleh daripada penandaan perkataan menggunakan set latihan.

Jadual 3.0 Beberapa hasil kejituan untuk dokumen di dalam set latihan

Dokumen	Hasil Kejituan
1	93.38%
2	94.90%
3	92.61%
4	91.88%
5	92.36%
6	93.04%
7	92.71%
8	94.37%
9	92.22%
10	95.15%
Purata Kejituan	**93.26%**

Jadual 4.0 menunjukkan hasil kejituan yang diperoleh daripada penandaan perkataan menggunakan set ujian dan hasil kejituan daripada set ujian ini akan digunakan untuk membuat perbandingan dengan hasil kejituan yang diperoleh menggunakan set peraturan yang dibangunkan oleh Nur Ashikin (2017).

Jadual 4.0 Hasil kejituan untuk dokumen di dalam set ujian

Dokumen	**Hasil Kejituan**
1	89.93%
2	87.02%

3	88.65%
4	86.74%
5	84.76%
6	87.08%
7	91.21%
8	86.07%
9	89.43%
10	89.90%
Purata Kejituan	**88.08%**

Pengujian untuk set peraturan yang dibangunkan oleh Nur Ashikin (2017) juga menggunakan set ujian yang sama untuk melihat prestasi dan keberkesanan kajian ini dalam menandakan lebih terperinci golongan kata di dalam Bahasa Melayu. Jadual 5.0 menunjukkan perbandingan antara hasil kajian ini dengan kajian Nur Ashikin (2017).

Jadual 5.0 Perbandingan keputusan antara dua kajian

Dokumen	**Penanda GK Cadangan Kajian**	**Penanda GK Nur Ashikin (2017)**
1	89.93%	81.26%
2	87.02%	79.04%
3	88.65%	82.31%
4	86.74%	80.75%
5	84.76%	79.20%
6	87.08%	81.52%
7	91.21%	81.08%
8	86.07%	79.05%
9	89.43%	81.40%
10	89.90%	80.10%
Purata Kejituan	**88.08%**	**80.57%**

Berdasarkan Jadual 5.0, kebanyakan hasil pengujian kejituan antara kedua-dua penanda GK mempunyai perbezaan sebanyak 8% ke atas. Perbezaan ini mungkin terjadi disebabkan

terdapat kesalahan penandaan yang boleh didapati di dalam hasil kajian Nur Ashikin (2017) seperti perkataan *yang* yang ditanda sebagai GNR iaitu GK Ganti Nama Relatif di mana GK yang betul adalah KHR iaitu GK Kata Hubung Relatif. Selain itu, terdapat GK yang telah ditambah supaya lebih terperinci di dalam kajian ini seperti perkataan *awal* yang ditanda sebagai KAJW iaitu GK Kata Adjektif Waktu di mana di dalam kajian Nur Ashikin ditanda sebagai KAJ sahaja iaitu GK Kata Adjektif. Secara umumnya, GK tersebut adalah betul tetapi kerana pengujian ini menggunakan algoritma dan ia tidak dilakukan secara manual, ianya akan dinilai betul sekiranya tanda GK tersebut sama dengan tanda GK di dalam dokumen yang telah ditanda secara manual terlebih dahulu.

6 KESIMPULAN

Kesimpulannya, kajian ini dapat menambah baik penandaan sedia ada dan menghasilkan kejituan yang lebih tinggi berbanding kajian lepas. Di dalam proses kajian ini, pelbagai informasi diperoleh berkaitan penandaan golongan kata Bahasa Melayu menggunakan pendekatan peraturan yang menjadikan kajian ini lebih jelas dan mudah difahami tujuan penghasilannya. Penambahbaikan yang dapat dilakukan untuk kajian akan datang ialah dengan menyediakan set data berlabel untuk kegunaan umum bagi memudahkan proses kajian pemprosesan bahasa tabii berkaitan Bahasa Melayu dan menyediakan set data yang mempunyai hasil kejituan yang tinggi dan penandaan GK yang tepat untuk dijadikan sebagai penanda aras bagi proses perbandingan antara hasil kajian. Ianya juga dapat mengelakkan hasil kejituan yang berbeza bagi setiap kajian walaupun menggunakan pendekatan yang sama.

7 RUJUKAN

Alfred, Mujat, and Obit 2013. *A Ruled-Based Part Of Speech (RPOS) Tagger for Malay Text Articles*.

Hassan Mohamed, Nazlia Omar & Mohd Juzaidin 2011. Statistical Malay Part-of-Speech (POS) Tagger using Hidden Markov Approach, *2011 International Conference on Semantic Technology and Information Retrieval*.

Mohd Pouzi & Syarifah Fatem Na'imah 2014. Part Of Speech Tagger For Malay Language Based On Words Morphology, *International Symposium on Research in Innovation and Sustainability 2014.*

Nik Safiah, K., Farid M., O., Hashim, H. M. & Abdul Hamid, M 2015. Tatabahasa Dewan Edisi Ketiga 21, 23-25, 43.

Nur Asyikin & Nazlia Omar 2017. Penandaan Golongan Kata Bahasa Melayu Menggunakan Pendekatan Berasaskan Petua, Projek Ijazah Sarjana Teknologi Maklumat, Fakulti Teknologi Dan Sains Maklumat, Universiti Kebangsaan Malaysia.

Rani 2011. Importance of Information Retrieval, *Oriental Journey of Computer Science & Technology,* Vol. 4 (2), 459-462 (2011)