SISTEM MODEL PENGECAMAN ENTITI NAMA UNTUK TEKS MEDIA SOSIAL BAHASA MELAYU

Muhammad Haziq bin Mohd Khairi
Sabrina Binti Tiun

Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Pengiktirafan entiti nama (PEN) adalah tugas mengklasifikasikan atau melabel elemen dalam teks ke dalam kategori seperti Orang, Lokasi atau Organisasi. Bagi bahasa Melayu, mengenali entiti yang dinamakan adalah tugas yang mencabar kerana slang yang unik dan kompleks dari segi mencari makna. Berdasarkan pemerhatian, projek sebelumnya lebih kepada model konseptual Automated Malay Named Entity Recognition (AMNER) menggunakan kaedah medan rawak bersyarat untuk bahasa Melayu untuk mengenali entiti daripada data teks tidak berstruktur. Analisis ini memfokuskan pada model pengembangan berdasarkan ciri bahasa Melayu yang memandu proses pengecaman entiti dari dokumen teks tidak berstruktur. Walau bagaimanapun, mengenali entiti yang dinamakan di media sosial semakin menarik hari ini. Berdasarkan pemerhatian, projek, Sulaiman et al. (2017) mengecam entiti nama bahasa Melayu mengguna dua sistem sedia ada iaitu Stanford dan Illnois yang mengguna bahasa Inggeris. Hasil kajian memperoleh nilai ketepatan yang rendah karena terdapat kesalahan bagi morfologi antara bahasa Inggeris dengan bahasa Melayu. PEN dalam Bahasa Inggeris dapat menunjukkan result yang baik memandangkan banyak data yang ada untuk digunakan dalam penyelidikan. Walaubagaimanapun, PEN untuk Bahasa Melayu tidak banyak diketengahkan. Ini kerana dari segi dasar Bahasa Melayu itu sendiri, terlalu banyak morphologi dari sejarah dahulu yang banyak perkataan diubah, dibaiki dan ditambah. Atas sebab-sebab tersebut, penyelidikan dalam Bahasa Melayu menjadi lebih sukar. Metod kajian ini menggunakan kaedah *Conditional Random Field* (CRF). Keputusan daripada kajian ini berhasil pada perkataan yang tertentu sahaja iaitu Bahasa Melayu formal. Kesimpulannya, semua PEN Bahasa Melayu adalah dalam Bahasa formal. Peraturan dan kaedah pembelajaran mesin terselia yang digunakan untuk Bahasa Melayu formal tidak sesuai digunakan untuk teks Bahasa Melayu tidak formal. Oleh itu, memerlukan korpus anotasi PEN yang terdiri dari teks media sosial dan kaedah yang sesuai untuk melatih PEN menggunakan korpus teks tersebut.

1 PENGENALAN

Pengecaman entiti nama (PEN) adalah tugas mengklasifikasikan atau melabel elemen dalam teks ke dalam kategori seperti ‘Orang’, ‘Lokasi’ atau ‘Organisasi’. Bagi bahasa Melayu, mengenalpasti entiti nama adalah tugas yang mencabar kerana slang yang unik dan kompleks dari segi mencari makna. Berdasarkan pemerhatian, projek, Sulaiman et al. (2017) mengecam entiti nama bahasa Melayu mengguna dua sistem sedia ada iaitu Stanford dan Illnois yang mengguna bahasa Inggeris. Hasil kajian mereka memperoleh nilai ketepatan yang rendah karena terdapat kesalahan bagi morfologi antara bahasa Inggeris dengan bahasa Melayu. PEN dalam Bahasa Inggeris dapat menunjukkan keputusan yang baik memandangkan banyak data yang ada untuk digunakan dalam penyelidikan. Walaubagaimanapun, PEN untuk Bahasa Melayu tidak banyak diketengahkan. Ini kerana dari segi dasar Bahasa Melayu itu sendiri, terlalu banyak morphologi dari sejarah dahulu yang banyak perkataan diubah, dibaiki dan ditambah. Jadi dengan sebab itu penyelidikan PEN dalam Bahasa Melayu agak sukar. (Sazali, 2016).

Banyak percubaan untuk melakukan penyelidikan terhadap Bahasa Melayu dalam media sosial untuk mendapat keputusan baik dengan algoritma bahasa mesin (Alfred et al. 2014). Walaubagaimanapun, ianya sukar kerana Bahasa Melayu mempunyai banyak terjemahan mesin kurang efektif untuk Bahasa Melayu selain Bahasa Inggeris. Sebagai contoh, di antara pendekatan yang dilakukan bagi kajian PEN ialah pendekatan berasaskan peraturan. Peraturan dibangun oleh pakar linguistik yang mampu mengecam entiti nama dipandang daripada aspek morfologi, sintaktik atau kata kunci yang menjelaskan sifat teks (Aboaoga & Aziz 2013).

Penyelidikan yang lepas banyak tentang PEN Bahasa Melayu kebanyakannya menggunakan kaedah peraturan dan pendekatan pembelajaran mesin berselia. Pendekatan pembelajaran mesin berselia bergantung dengan jumlah korpus latihan yang dapat mempengaruhi nilai ketepatan bagi PEN. Semakin banyak korpus latihan, maka akan hasil keputusan yang tinggi diperoleh dan sebaliknya (Ulanganathan et al. 2017; Salleh et al. 2017). Selain itu, pendekatan pembelajaran mesin berselia juga bergantung pada ciri sebelum membangunkan model. Ciri yang salah atau tidak sesuai akan menyebabkan berlakunya kesalahan kategori entiti nama (Salleh et al., 2017).

2 PERNYATAAN MASALAH

Dalam kajian PEN, antara kekangan yang mungkin dialami adalah data pendua (duplicate). Hal ini kerana mungkin apabila data diperoleh apabila dilaksanakan data tersebut kemungkinan mesin akan keliru dengan ayat Bahasa Inggeris seperti “bestlah”, “terrorlah” (power), dan “viralkan”(viral). Oleh itu, data tersebut mungkin tidak menjadi tersusun dan teratur

Kekangan kedua adalah kesukaran mengecam Bahasa dialek atau rojak. Malaysia mempunyai 14 negeri dan setiap dialek adalah berbeza dari segi setiap negeri. Contohnya perkataan “Henfon(handphone) aku lawa”, “aku dah makan oredi(already)” dan “Tak apa, cincai (whatever) la”, “Walao weh! (surprise)” and “Aduh, kau ni potong stim la”. Kemungkin Bahasa mesin tidak dapat mengecam satu persatu ayat yang digunakan dan jika dapat mengecam mungkin tidak diambil perkataan tersebut.

Kesimpulannya, semua PEN Bahasa Melayu adalah dalam Bahasa formal. Peraturan dan kaedah pembelajaran mesin terselia yang digunakan untuk Bahasa Melayu formal tidak sesuai digunakan untuk teks Bahasa Melayu tidak formal. Oleh itu, memerlukan korpus anotasi PEN yang terdiri dari teks media sosial dan kaedah yang sesuai untuk melatih PEN menggunakan korpus teks tersebut.

3 OBJEKTIF KAJIAN

Projek ini bertujuan untuk membina sistem pengecaman entiti nama (PEN) bagi teks media sosial bahasa Melayu menggunakan kaedah peraturan dan pembelajaran mesin berselia CRF.

Di samping itu, dapat membuat perbandingan sistem PEN Bahasa Melayu teks media sosial yang dibina dengan sistem yang terdahulu.

4 METOD KAJIAN

Metodologi adalah amat diperlukan untuk menampakan gambaran yang jelas tentang apa yang dibuat untuk sesuatu kaedah yang dijalankan. Proses ini mempunyai tiga fasa iaitu fasa penyediaan korpus, kaedah peraturan dan penilaian.

Fasa penyediaan korpus adalah satu fasa dimana membantu menyelesaikan sejumlah masalah linguistik, seperti struktur ayat, lemma, carian terbalik(reversed search) mengikut sifat morfologi, pemilihan konteks, n-gram untuk pelbagai pertanyaan carian. Sumber media sosial seperti artikel didalam Twitter, dalam tajuk kemalangan, sukan, bencana, hal-hal semasa, perayaan dan politik.

Kaedah peraturan telah ditunjukkan dalam prosesnya bagi pengecaman bagi setiap perkataan dalam artikel berita yang diproses menggunakan ungkapan nalar dengan corak tertentu dan melibatkan gazetir yang telah dibangun. Ungkapan nalar (regex) iaitu rentetan (string) teks khas bagi menentukan corak carian (Morsidi et al. 2017). Ungkapan nalar dibangun berdasarkan ciri khusus yang 42 diambil daripada perkataan yang cenderung sering muncul untuk menentukan penyelesaian terhadap kekangan string atau perkataan (Salleh et al. 2017).

Fasa terakhir adalah penilaian Pengujian dan penilaian dilakukan terhadap hasil dapatan kajian yang diperoleh dari fasa sebelumnya. Setiap peraturan pengecaman entiti nama yang dibangunkan diuji dengan menggunakan kedua-dua jenis korpus. Korpus latihan digunakan semasa pembangunan peraturan dan hasilnya diuji dengan menggunakan korpus ujian bagi menilai tahap ketepatan PEN. Keputusan dinilai dari segi kejituan, dapatan dan ukuran-f.

5 HASIL KAJIAN

Model pembangunan sistem pengecaman entiti nama ini ini telah berjaya dibangunkan berdasarkan keperluan dan objektif yang telah dinyatakan. Terdapat beberapa kekuatan dan kelemahan yang dikenalpasti terdapat di dalam model ini untuk dijadikan panduan kepada kajian-kajian akan datang. Penyelidik berharap bahawa kajian ini boleh dijadikan rujukan kepada pembangun-pembangun aplikasi mudah alih untuk membangunkan aplikasi yang tidak sahaja menyumbang kepada bidang teknologi maklumat, malah juga kepada bidang yang lain supaya setiap sisi masyarakat boleh mendapat manfaat daripada teknologi yang dibangunkan.

6 KESIMPULAN

Kesimpulannya, model ini membentangkan tentang pembangunan dan pengaturcaraan yang telah dilakukan terhadap sistem ini. Sistem pengecaman entiti nama ini telah dibangunkan

dengan perisian Python kerana Bahasa pengaturcaraan ini lebih sesuai digunakan untuk model Bahasa seperti ini. Walaubagaimanapun, dataset yang baru dibuat hanyalah untuk mengecam beberapa entiti sahaja jadi akan berlaku masalah apabila mengecam dengan lebih banyak entiti selepas ini. Akan tetapi, ia akan dapat diselesaikan dengan banyak membuat koding yang bersesuaian dengan dataset yang kita lakukan. Di samping itu, pembangunan ini telah mengambil masa yang lama untuk menyelesaikan masalah dalam sebuah kod. Oleh hal yang demikian, pembangunan mendapat kekangan untuk emnyiapkan projek ini dalam masa yang ditetapkan.

7 RUJUKAN

1. Chekima, K. & Rayner Alfred. (2017). Sentiment Analysis of Malay Social Media Text. 4th International Conference on Computational Science and Technology Proceedings, 29- 30 November, Kuala Lumpur, Malaysia

2. Li, D. (1998) The Plight of the Purist. In Pennington, M. (Ed.). Language in Hong Kong at Century's End (pp. 161-190). Hong Kong: Hong Kong University Press.

3. Tagging with QTAG. (2007). Retrieved 15 May, 2018 from https://www1.essex.ac.uk/linguistics/research/resgroups/clgroup/Resources/Nugues/Q TAG/

4. Anbananthen, K. S. M., Krishnan, J. K., Sayeed, M. S. & Muniapan, P. (2017). Comparison of Stochastic and Rule-Based POS Tagging on Malay Online Text. American Journal of Applied Sciences. Vol. 14(9), 843-851.

5. Alfred, R., Leong, L. C., On, C. K. & Anthony, P. 2014. Malay Named Entity Recognition Based on Rule-Based Approach 4(3). doi:10.7763/IJMLC.2014.V4.428

6. Alfred, R., Mujat, A. & Obit, J. H. 2013. A Ruled-Based Part of Speech (RPOS) tagger for Malay text articles 7803 LNAI(PART 2), 50–59.

7. Salleh, M. S., Asmai, S. A., Basiron, H. & Ahmad, S. 2017. A Malay Named Entity Recognition Using Conditional Random Fields. International Conference on Information and Communication Technology (ICoICT) A 0(c).

8. Charles Sutton and Andrew McCallum. Piecewise training of undirected models. In 21st Conference on Uncertainty in Artificial Intelligence, 2005.

9. Using GATE as an Annotation Tool by Tom Kenter, Diana Maynard. (https://gate.ac.uk/sale/am/annotationmanual-gate2.pdf)

10. Sazali, S. S. B. 2016. Information Extraction : Evaluating Ner From Classical Malay Documents. International Conference on Information Retrieval and Knowledge Management Information 48– 53.

11. Ulanganathan, T., Ebrahimi, A., Chu, B., Xian, M., Bouzekri, K., Mahmud, R. & Hoe, O. H. 2017. Benchmarking Mi-NER: Malay Entity Recognition Engine Department of Artificial Intelligence University of Malaya (c): 52–58.

12. Alfred, R., Leong, L. C., On, C. K. & Anthony, P. 2014. Malay Named Entity Recognition Based on Rule-Based Approach 4(3). doi:10.7763/IJMLC.2014.V4.428

13. Alfred, R., Mujat, A. & Obit, J. H. 2013. A Ruled-Based Part of Speech (RPOS) tagger for Malay text articles 7803 LNAI(PART 2), 50–59.

14. Matt Richardson and Shawn Wallace. " Getting started with Raspberry pi " Published byMaker Media Inc. First Edition Dec 2012.

15. Udayakumar G. Kulkarni "Arduino : A beginner's Guide" 11-07-2017.

16. R. Grangel and C. Campos, "Agile Model-Driven Methodology to Implement Corporate Social Responsibility," *Comput. Ind. Eng.*, vol. 127, no. November 2018, pp. 116–128, 2019

17. Farouzan "Data Communication and Networking SE" Global Edition McGraw-Hill 5th Edition 2013.

18. Kimmo Karvinen, Tero Karvinen, Getting Started with Sensors: Measure the World withelectronics, Arduino, and Raspberry Pi, 2014, Maker Media, Inc

19. The Official Raspberry Pi Projects Book, The MagPi , 2016.

20. Jody Culkin, Eric Hagan, Learn Electronics with Arduino: An Illustrated Beginner's Guide toPhysical Computing, Maker Media, Inc, 2017.

21. Mockflow (2019). Perisian pembangunan antara muka pengguna. Diambil dariwww.mockflow.com

22. Lucidchart (2010). Perisian pembangunan carta alir. Diambil dari www.lucidchart.com

23. The Python Tutorial (2001-2021), Python Software Foundationhttps://docs.python.org/3/tutorial/index.html

24. Rob Zwetsloot, "The Official Raspberry Pi Handbook" 2021