Pengecaman Entiti Nama Bahasa MelayuBerasaskan Peraturan

Ulfa Nadia
Nazlia Omar

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor Malaysia.

ulfa.nadia07@gmail.com

ABSTRAK

Pengecaman Entiti Nama (PEN) merupakan satu proses menganotasi atau memberi tanda nama dalam ayat untuk setiap kelas perkataan seperti nama individu, organisasi, tarikh, masa, dan lain-lain. Proses ini merupakan satu proses penting daripada sebahagian tugas asas dalam capaian maklumat, bagi prestasi pemprosesan teks. Masalah utama dalam PEN bahasa Melayu ialah penggunaan entiti nama yang mempunyai rujukan silang dengan entiti nama lain, pencampuran entiti nama yang berbeza dan pengulangan entiti nama. Objektif utama kajian ini adalah untuk membangun peraturan baru bagi PEN bahasa Melayu dan membandingkan prestasi PEN bahasa Melayu berasaskan peraturan dengan kajian lepas. Proses ini bermula dengan penyediaan korpus, pembangunan gazetir, pembangunan peraturan dan penilaian. Penyediaan korpus menggunakan fail teks yang diperoleh daripada berita atas talian meliputi pelbagai domain. Sebanyak 200 korpus telah dipilih dan 170 daripadanya dijadikan sebagai korpus latihan manakala baki selebihnya sebagai korpus ujian. Seterusnya data latihan diguna pada proses pembangunan gazetir dan proses pembangunan peraturan. Pembangunan peraturan merupakan proses yang melibatkan senarai gazetir. Pembangunan peraturan entiti nama dalam kajian ini memberi fokus kepada pengecaman entiti nama yang melibatkan 9 entiti iaitu nama individu, lokasi, organisasi, jawatan, tarikh, masa, kewangan, ukuran dan peratusan. Proses penilaian pula dilaksanakan bagi melihat keberkesanan peraturan PEN yang dibangunkan serta membuat perbandingan hasil PEN dengan kajian lepas. Secara keseluruhannya, pengujian ini memberikan hasil dengan nilai kejituan 90.23%, dapatan 92.13% dan ukuran-f 91.05% berbanding prestasi kajian lepas iaitu 94.44% nilai dapatan, 85% kejituan dan 89.47% ukuran-f. Hasil daripada kajian ini diharap dapat membantu penyelidik dalam melaksanakan PEN bahasa Melayu dengan menghasilkan nilai ketepatan yang lebih tinggi.

1. PENGENALAN

Peningkatan dokumen elektronik dari tahun ke tahun (Kanimozhi & Venkatesan 2015) menjadikan penyelidikan dalam Pemprosesan Bahasa Tabi'i (PBT) semakin penting dalam capaian semula maklumat (Jiang et al. 2016). PBT mempunyai potensi yang tinggi dalam kebanyakan aplikasi seperti rumusan artikel sistem soal jawab, sistem penilaian sebut harga dan sistem tutorial (Pallavi et al. 2018). PBT menjadikan maklumat tidak berstruktur kepada satu maklumat yang lebih bermakna. PBT boleh memproses artikel ke dalam bentuk yang lebih jelas dengan menghubungkan satu perkataan kepada perkataan yang lain. Antara cabang penyelidikan PBT ialah Pengecaman Entiti Nama (PEN).

PEN merujuk pada proses mencari sebahagian daripada teks yang mewakili kata nama khas dan kemudian mengklasifikasikannya dalam kategori yang sesuai. Sebahagian daripada teks tersebut boleh jadi sebagai indeks, link, dan dapat diguna juga dalam sistem soal jawab (Sazali 2016). PEN melakukan tugas tidak hanya mengenalpasti entiti nama dalam teks yang tidak terstruktur namun juga mengklasifikasikannya mengikut susunan jenis entiti yang telah ditentukan. Tugas PEN pertama kali dilakukan semasa persidangan MUC-6 ialah untuk menemukan jenis entiti, seperti orang, lokasi, dan organisasi serta masa, mata wang, dan peratusan dalam teks tidak terstruktur (Yong et al. 2011). Sebagai contoh terdapat nama individu dalam ayat "Tuanku Syed Faizuddin Putra Jamalullail berceramah di Merlimau, petang ini…". Untuk mengklasifikasikan nama orang iaitu "Tuanku Syed Faizuddin Putra Jamalullail" ke dalam entiti nama individu dan "Merlimau" ke dalam entiti lokasi

maka terdapat tiga pendekatan teknik yang perlu dilakukan seperti pendekatan berasaskan peraturan, pendekatan berasaskan statistik serta gabungan kedua-dua pendekatan ini (Alfred et al. 2014).

2. PENDEKATAN PEN BERASASKAN PERATURAN

Pendekatan berasaskan peraturan ialah peraturan yang dibangun oleh ahli linguistik untuk mengenalpasti entiti nama daripada segi morfologi, sintatik atau memerlukan kata kunci yang mencerminkan sifat-sifat teks (Aboaoga & Aziz 2013). Pendekatan ini menggunakan corak yang dibuat secara manual kepada perkataan dalam ayat dengan menggunakan satu set peraturan yang ditulis. Selain daripada mudah digunakan kerana ia menggunakan peraturan yang ringkas dan menjimatkan ruang (Mubarak et al. 2015). Penambahbaikan ke atas PEN juga mudah dilakukan kerana pendekatan ini menggunakan set kecil peraturan yang mudah dan kurang kompleks (Al-Olimat et al. 2017). Diantara contoh pendekatan berasaskan peraturan dalam PEN ialah "jika sesebuah kata nama khas yang memiliki ciri-ciri nama orang seperti adanya gelaran maka itu ialah termasuk entiti nama individu". Manakala statistik pula melibatkan penggunaan teknik pembelajaran mesin yang dibagi dalam 3 jenis iaitu dengan berselia, separa berselia dan tanpa selia (Alfred et al. 2014; Art et al. 2015; Sharum & Abdullah 2011).

Penyelidikan mengenai PEN telah berkembang, satu di antara bahasa dunia yang banyak digunakan seperti berbahasa Inggeris kerana korpus yang sudah tersedia sehingga memudahkan penyelidikan dibidang PEN. Manakala korpus bahasa Melayu tidak sebanyak sumber bahasa Inggeris dan tiada koleksi yang boleh diguna bagi PEN (Sazali 2016). Hal ini berlaku kerana sumber bahasa Inggeris tidak dapat diguna dalam bahasa Melayu. Jika PEN bahasa Melayu menggunakan korpus bahasa Inggeris akan memerlukan proses terjemah bahasa sehingga perlu masa yang lama untuk menyelesaikan pengecaman entiti nama. Sulaiman et al. (2017) mengecam entiti nama bahasa Melayu mengguna dua sistem yang sedia ada iaitu Stanford dan Illnois yang mengguna bahasa Inggeris. Hasil kajian menunjukkan terdapat kesalahan kerana perbezaan morfologi antara bahasa Inggeris dengan bahasa Melayu sehingga memperolehi ketepatan yang rendah. Namun PEN pada bahasa Melayu memiliki kesamaan dengan bahasa Inggeris seperti penggunaan huruf besar dan kata nama khas untuk mengecam entiti (Anthony 2013).

Sebahagian model pengecaman entiti nama menggunakan kamus sebagai sumber rujukan untuk menentukan entiti nama berdasarkan peraturan untuk menghasilkan prestasi yang tinggi dan optimal, sebahagian lagi menggunakan sebahagian kecil gazetir untuk menghasilkan nilai dapatan dan nilai kejituan yang tinggi. Pendekatan berdasarkan peraturan dalam mengenal pasti entiti nama bermaksud mengguna satu set peraturan yang terdahulu dan senarai kamus perkataan yang dibuat oleh manusia secara manual (Alfred et al. 2014).

Pada pendekatan berdasarkan peraturan, umumnya pembentukan perkataan dicapai berdasarkan morfologi dan struktur sintaks bahasa tertentu seperti menggunakan penandaan golongan kata, word precedence, ciri orthografik seperti penggunaan huruf besar dan kombinasi dengan menggunakan kamus (Budi et al. 2005). Pembangun sistem perlu mempunyai pengalaman dan kemahiran pengetahuan bahasa dan tata bahasa dalam menentukan ketepatan pengecaman entiti nama. Hal ini kerana proses pembangunan peraturan memerlukan pengujian yang dilakukan berulang kali bagi mendapatkan hasil pengecaman yang lebih tepat (Nadeau & Sekine 2006; Ranaivo-Malangon et al. 2015).

Budi et al. (2005) telah membangun satu model pengecaman entiti nama bagi bahasa Indonesia yang diberi nama InNER. Model ini menggunakan pendekatan berdasarkan peraturan terhadap kontekstual, morfologi dan penandaan golongan kata (POS tagger) bagi mengenalpasti entiti dalam bahasa Indonesia. Model ini digunakan untuk mencari entiti nama orang dalam sebuah ayat.

Morsidi et al. (2017) menggunakan regex untuk mengesan ciri istimewa daripada kata sendi nama dengan mengenal pasti struktur kata benda pada perkataan dalam suatu perenggan. Penyelidikan ini terhad berdasarkan anggapan bahawa huruf awalan perkataan kata nama khas dimula dengan huruf besar. Kata sendi nama diguna untuk mewakili kata benda sebagai objek, yang diguna bersama dengan morfologi atau anotasi perkataan seperti kata sifat dan preposisi (Abu Bakar et al. 2013).

Satu fenomena yang berlaku dalam PEN bahasa Melayu ialah domain tatabahasa seperti morfologi dan konteks (Awang 2010). Terdapat dua bahagian dalam tatabahasa iaitu konteks yang merupakan

bahagian pembentukan frasa dan ayat, dan morfologi iaitu satu bidang ilmu yang mengkaji perkataan dari segi struktur, bentuk dan penggolongan kata. Sebagai bahasa yang kaya dengan kosa katanya, bahasa Melayu mempunyai morfologinya yang tersendiri bagi menghasilkan perkataan lain (Mohamed et al. 2011).

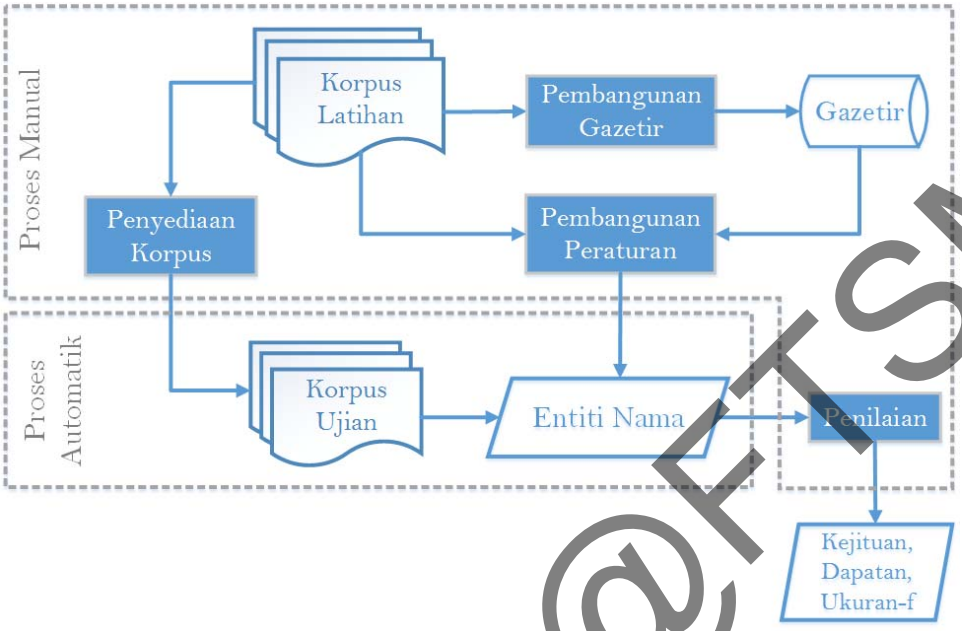
Kajian lepas bagi PEN bahasa Melayu kebanyakannya menggunakan pendekatan peraturan dan pendekatan sistem berselia termasuk teknik CRF. Pendekatan berselia pula sangat bergantung dengan banyaknya korpus latihan yang dapat mempengaruhi nilai ketepatan bagi PEN. Semakin banyaknya korpus latihan maka akan diperolehi hasil keputusan yang tinggi. Manakala korpus latihan yang sedikit dapat memperburuk hasil keputusan (Ulanganathan et al. 2017; Salleh et al. 2017). Selain itu, pendekatan berselia juga sangat bergantung pada ciri yang dipilih sebelum membangun model. Pemilihan ciri yang tidak sesuai dapat menjadi sebab kesalahan kategori entiti nama (Salleh et al. 2017).

Kaedah peraturan merupakan kaedah praktikal yang tidak memerlukan banyak data, mudah ditambah baik dan dianalisis hanya dengan menambahkan peraturan (Kumawat dan Jain 2015). Walaupun kajian sedia ada telah mengguna peraturan, namun jumlah penghasilan peraturan masih tidak mencukupi dan tidak menyeluruh kerana masih terdapat beberapa entiti nama yang tidak dicam secara sempurna, sebagai contoh Jenny @ Jita Eyir, Hospital Daerah Jasin. Bagi mengatasi struktur perkataan yang kompleks, maka dibangunkan beberapa peraturan bagi entiti nama. Justeru, kajian ini mengambilkira penambahan jumlah peraturan dan penggunaan korpus dari pelbagai domain agar PEN dapat dilihat secara menyeluruh. Kajian ini juga menambahkan beberapa jenis entiti lainnya seperti, masa, tarikh, peratusan, kewangan dan ukuran.

3. PEMBANGUNAN PERATURAN PEN

Terdapat 4 proses dalam menjalankan kajian PEN bahasa Melayu berasaskan peraturan iaitu penyediaan korpus, pembangunan gazetir, pembangunan peraturan dan penilaian. proses tersebut bermula dengan penyediaan korpus yang meliputi korpus latihan yang akan digunakan dalam pembangunan peraturan dan korpus ujian yang akan digunakan dalam proses penilaian. Seterusnya, proses pembangunan gazetir menghasilkan senarai sebahagian perkataan daripada entiti nama. Pembangunan peraturan dalam kajian ini dimulakan membina peraturan bagi 9 entiti nama iaitu individu, organisasi, jawatan, tarikh, masa, kewangan, ukuran dan peratusan. Peraturan bermula dengan memadankan perkataan pada gazetir dahulu, jika perkataan tersebut wujud dan perkataan selepasnya merupakan awalan huruf besar maka pengecaman entiti nama ditetapkan pada perkataan tersebut. Jika perkaatan tersebut tidak wujud dalam gazetir atau didapati mempunyai kurungan, atau titik, atau titik koma atau perkataan daripada gazetir entiti lain, maka pengecaman selesai.

Hasil PEN dipaparkan melalui penghasilan entiti nama dengan setiap perkataan dilabelkan dengan jenis PEN masing-masing. Proses penilaian kemudian dijalankan dengan membandingkan hasil pelabelan entiti nama tersebut dengan korpus latihan yang telah dilabelkan secara manual terlebih dahulu. Pada korpus ujian dan korpus kajian lepas pula dilakukan penilaian setelah mendapatkan hasil pelabelan entiti nama secara automatik daripada peraturan yang telah dibangunkan pada kajian ini. Keputusan akhir penilaian merupakan hasil perbandingan prestasi penandaan PEN dengan korpus



ujian dan korpus kajian lepas. Rajah 3 menunjukkan aliran proses kajian ini secara umum.

Rajah 1. Aliran metodologi penyelidikan

3.1 Penyediaan Korpus

Penyediaan korpus yang dilaksanakan ialah pengumpulan artikel berita yang akan diasingkan dan digunakan sebagai korpus latihan dan ujian bagi kajian ini. Artikel berita yang digunakan terdiri daripada pelbagai jenis berita nasional termasuklah kemalangan, sukan, bencana, semasa, perayaan dan politik. Pemilihan data daripada sumber berita, seperti Bernama, Berita Harian dan Malaysia Kini kerana sumber ini menggunakan bahasa formal dan telah disunting. Pengumpulan artikel berita diekstrak secara manual ke dalam bentuk fail teks pada bahagian isi.

Kumpulan artikel berita yang mengandungi fail teks disebut sebagai korpus. Penyediaan korpus meliputi korpus latihan yang akan digunakan dalam pembangunan peraturan dan korpus ujian yang akan digunakan dalam penilaian. Korpus latihan terdiri daripada 170 fail teks yang mengandungi 13987 patah perkataan dan korpus ujian sebanyak 30 fail teks dengan 7311 perkataan yang disediakan sebelum proses PEN dilaksanakan. Korpus latihan ialah korpus yang dijadikan dasar pembangunan peraturan dalam PEN, sedangkan korpus ujian ialah korpus yang digunakan bagi penilaian peraturan yang telah dibangun dalam PEN. Kandungan teks untuk dijadikan korpus sekurang-kurangnya perlu mencakupi pelbagai domain bagi mencipta variasi dalam pemilihan korpus latihan dan korpus ujian.

3.2 Pembangunan Gazetir

Alatan bantuan dalam pembangunan gazetir ini dibangun daripada Wikipedia dan 170 korpus latihan. Pembangunan gazetir merupakan proses membuat senarai sebahagian perkataan bagi setiap kategori entiti nama yang dikenalpasti. Gazetir dibangunkan mengikut keperluan setiap jenis. Antara gazetir tersebut ialah nama individu, nama jawatan, nama organisasi, dan nama syarikat. Sebagai contoh, gazetir untuk kategori individu ialah senarai nama awalan individu seperti Ahmad, Azlan, Nurul dan sebagainya dan juga digunakan nama gelaran, gelaran kehormat, gelaran diraja. Manakala gazetir yang digunakan untuk mengenal pasti organisasi ialah agensi, lembaga, persatuan, syarikat,

pertubuhan, kelab, hospital, sekolah, universiti, fakulti. Gazetir bagi sebahagian nama jawatan pula terdiri daripada senarai jawatan dan pekerjaan yang wujud di Malaysia. Antara senarai tersebut ialah pengurus, setiausaha dan jururawat. Manakala gazetir untuk kategori tempat ialah awalan nama tempat seperti Dewan, Taman, Jalan, dan sebagainya.

Kajian ini membangun gazetir yang telah dibina secara manual. Gazetir dibangunkan bagi PEN individu, organisasi, lokasi, jawatan dan ukuran. Sekiranya perkataan yang diproses mengandungi salah satu perkataan yang terdapat dalam gazetir kategori entiti nama tertentu, maka perkataan tersebut akan diberi label sebagai kategori tertentu.

3.3 Pembangunan Peraturan

Pembangunan peraturan dilakukan dengan menggunakan korpus latihan yang tediri daripada 170 artikel berita yang mengandungi 13987 patah perkataan dengan jumlah keseluruhan entiti sebanyak 3187 yang dikira secara manual. Bagi mengetahui corak, tatabahasa serta penggunaan huruf besar yang tepat pada entiti nama, pengiraan ini melibatkan salah seorang pensyarah bahasa Melayu sebagai pakar. Rajah 2 contoh fail teks daripada korpus latihan yang telah dianotasi secara manual. Pakar ini juga turut menyemak ketepatan entiti nama pada korpus latihan yang telah dikira secara manual pada fasa penilaian.

Rajah 2. Contoh korpus latihan hasil anotasi manual

Proses pembangunan peraturan ini dilakukan dengan merujuk kepada corak morfologi yang mengandungi perkataan ‘dan’ dan simbol koma serta padanan perkataan dalam senarai gazetir yang telah dibangunkan pada fasa pra-pembangunan peraturan. Rajah 3 menunjukkan pembangunan peraturan secara umum bagi PEN.

```
Input: Senarai ayat (A) yang terdiri daripada n baris artikel
Input: Gazetir, Konteks, Kamus
for i=0 to i=n
      Cari padaan kata daridapa gazetir dan morfologi guna regex
      if padan gazetir entiti L
         cari ayat (A) disekitar gazetir L yang berawalan huruf besar
         Ai-1 | Ai | Ai+1 lakukan pengecaman entiti L
      else if padan gazetir entiti Li
         cari ayat (A) disekitar gazetir Li yang berawalan huruf besar
         Ai-1 | Ai | Ai+1 lakukan pengecaman entiti Li
      else
         pengecaman entiti nama berakhir
```

Rajah 3. Pembangunan peraturan entiti nama

Contoh peraturan yang dibangun bagi individu:
Jika perkataan pertama (W) mengandungi perkataan gelaran daripada set A dalam gazetir dimana A= {Tengku, Nik, Haji, Tun, Tan Sri, Datuk, M., A. ...} dan perkataan seterusnya Wi+1 perkataan berawalan huruf besar maka perkataan tersebut ialah entiti nama individu. Jika Wi+1 didapati mempunyai kurungan, titik, koma atau titik koma atau perkataan daripada gazetir, konteks atau kamus entiti lain, maka pengecaman untuk entiti individu selesai. Sebagai contoh, pada ayat berikut "Tuanku Syed Faizuddin Putra Jamalullail mengajak umat Islam seluruh dunia bersatu…". Pada ayat diatas hanya perkataan 'Tuanku' yang terdapat dalam senarai gazetir. Selain itu, perkataan selepasnya memiliki ciri morfologi yang diawali oleh huruf besar.

3.4 Penilaian

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 yang dapat dirumuskan dengan menggunakan formula berikut (Alfred et al. 2014).

Kejituan= (tepat+(0.5*separa tepat))/(jumlah keseluruhan PEN oleh sistem)
Dapatan= (tepat+(0.5*separa tepat))/(jumlah keseluruhan PEN secara manual)
Ukuran-F= ((kejituan * dapatan))/(0.5*(kejituan + dapatan))

4. KESIMPULAN

Sebanyak 27 peraturan telah dibangunkan bagi PEN bahasa Melayu berasaskan peraturan. Termasuk kehadiran entiti nama bersama perkataan 'dan' dan simbol, pengulangan entiti nama pendek, penggunaan entiti nama yang mengandungi entiti lain dan percampuran entiti nama yang berbeza. Korpus yang diguna dalam kajian ini dapat dijadikan sebagai sumber korpus bagi bahasa Melayu dalam bidang capaian maklumat. Memandangkan korpus yang digunakan dalam kajian ini terdiri daripada pelbagai domain.

Di samping itu, hasil PEN ini juga dapat dijadikan sebagai sumber bagi penyelidikan lanjutan dalam bidang capaian maklumat. Memandangkan korpus yang digunakan terdiri daripada pelbagai domain, maka kajian-kajian selanjutnya dalam bidang capaian maklumat akan menjadi lebih mudah kerana terdapat variasi dalam penggunaan korpus. Selain itu, hasil dari kajian ini turut menyumbang pengetahuan mengenai keberkesanan teknik pengecaman entiti nama bahasa Melayu berdasarkan peraturan. Penggunaan PEN bahasa Melayu yang belum meluas berbanding bahasa lain kerana kekurangan sumber yang masih terhad. Oleh itu, dengan mengubah suai peraturan dalam kajian ini dan menambahkan peraturan baru diharapkan dapat membantu penyelidik lain dalam melaksanakan PEN bagi korpus bahasa Melayu untuk menghasilkan nilai ketepatan yang tinggi.

RUJUKAN

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

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.

Anthony, P. 2013. Advanced Data Mining and Applications 8346(1). doi:10.1007/978-3-642-53914-5

Anthony, P., Alfred, R., Leong, L. C., On, C. K. & Anthony, P. 2014. Malay Named Entity Recognition Based on Rule-Based Approach Malay Named Entity Recognition Based on Rule-Based Approach (1). doi:10.7763/IJMLC.2014.V4.428

Art, F., Industry, C., Pendidikan, U. & Idris, S. 2015. Malay Named Entity Recognition : A Review 2.

Aryoyudanta, B., Adji, T. B. & Hidayah, I. 2017. Semi-supervised learning approach for Indonesian Named Entity Recognition (NER) using co-training algorithm. *Proceeding – 2016 International Seminar on Intelligent Technology and Its Application, ISITIA 2016: Recent Trends in Intelligent Computational Technologies for Sustainable Energy* 7–12. doi:10.1109/ISITIA.2016.7828624

Awang, S. 2010. Pemartabatan bahasa kebangsaan dalam pembinaan negara bangsa. *Syarahan Bahasa anjuran Dewan Bahasa dan Pustaka Wilayah Timur dan IPG Tengku Ampuan Afzan* 1–12.

Budi, I., Bressan, S., Wahyudi, G., Hasibuan, Z. A. & Nazief, B. A. A. 2005. Discovery Science. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 3735(November 2015): 1–13. doi:10.1007/11563983