UNJUK PERFORMA METODE INFORMATION EXTRACTION : LINGUISTIK PADA DOMAIN PARIWISATA

Fahmi Lutfiansyah Moechtar
Program Studi Sistem Informasi Universitas Bakrie
E-mail : f.l.moechtar@gmail.com

Guson Kuntarto
Program Studi Sistem Informasi Universitas Bakrie
E-mail : gusonkuntarto@bakrie.ac.id

Abstract

The Increasing growth of information on the Internet requires ontology for always up-to-date. Ontology is a concept of domain that represented by classes, relationships, properties and instances. One way to keep ontology up-to-date is by populating the ontology. The most important aspect of the ontology population process is the information from the web (semi-structure) that will be added to ontology. The process of adding instance or ontology population has shift from manual to automatic process. With the condition of ontology population process in Indonesia that still manual, especially in Bali tourism domain, this research focus on shows performance of automatic extraction engine methods using linguistic methods on the Bali tourism domain. In order to measure relevancy of linguistic method result the precision, recall and f-measure value was used. The result shows that the linguistic method gives the relevant results with highest score at 0.61. This score is limited with the definition of the term based on the DWIPA ontology.

.

Keywords: *Ontology population*, *extraction engine, linguistic method, Part of Speech*

Abstrak

Semakin cepatnya perkembangan informasi di dalam Internet menuntut suatu ontologi untuk selalu *up-to-date*. Ontologi merupakan suatu konseptualisasi dari suatu domain yang direpresentasikan dalam *class*, *relation*, *property* dan *instance*. Salah satu cara untuk menjaga ontologi selalu *up-to-date* adalah dengan melakukan populasi terhadap ontologi tersebut. Aspek terpenting dalam proses populasi tersebut adalah informasi yang ditambahkan yang didapat dari web (*semi-structure*). Proses ekstraksi informasi itupun kini mulai beralih dari manual ke otomatis. Dengan kondisi proses populasi ontologi di Indonesia yang masih manual khususnya pada domain pariwisata Bali, penelitian ini berfokus pada unjuk performa metode *automatic extraction engine* dengan menggunakan metode linguistik pada domain pariwisata Bali. *Precision recall* dan *f-measure* digunakan untuk mengukur relevansi dari metode *extraction engine* linguistik. Hasil penelitian menunjukkan bahwa metode *linguistik* memberikan hasil yang cukup relevan dengan nilai *precision* tertinggi 0.61. Nilai *precision* yang didapat pada penelitian ini terbatas pada definisi *term* yang digunakan berdasarkan pada DWIPA *ontology*.

Kata kunci: *Ontology population*, *extraction engine, linguistic method, Part of Speech*

PENDAHULUAN

Sebagai suatu negara kepulauan yang memiliki luas lebih dari 1.000.000 km² Indonesia memiliki berbagai macam hal yang menarik mulai dari tempat tempat bersejarah yang banyak, kekayaan alam yang melimpah, budaya yang beraneka ragam serta tempat tempat wisata yang sangat menarik untuk dikunjungi. Hal ini menjadi daya tarik tersendiri baik bagi warga negara Indonesia dan bagi warga negara asing terutama pada bidang pariwisata Indonesia. Dan Bali merupakan salah satu tempat tujuan utama dalam berwisata terutama untuk turis asing, terbukti pada tahun 2013 Bali menyumbang 40% atau sekitar 40 triliun rupiah dari total pendapatan nasional yang didapat dari hasil kalkulasi pendapatan jasa penerbangan, agen perjalanan, serta pendapatan hotel dan restoran di Bali [1].

Pencapaian ini tidak datang dengan sendirinya, salah satu cara yang digunakan adalah dengan memberikan informasi tentang Bali itu sendiri di Internet. Hal ini didukung pula oleh jumlah

pengguna Internet di dunia yang mencapai lebih dari 3 miliar orang [2]. Perkembangan teknologi informasi yang sangat pesat memicu kebutuhan akan teknologi informasi itu sendiri di berbagai bidang seperti bidang pariwisata, karena pariwisata adalah kegiatan yang didasari oleh informasi [3]. Pariwisata dengan berbasis teknologi informasi dikenal dengan sebutan *E-tourism*. Salah satu bentu bentuk dari *e-tourism* itu sendiri adalah penggunaan *website* untuk memasarkan suatu situs pariwisata. Beberapa *website* yang memberikan informasi pariwisata di Indonesia antara lain : www.wisatanet.com , www.indonesia-tourism.com ,www.tourismindonesia.com, www.indonesiatourism.com [4].

Namun pada kenyataannya dalam pemanfaatan *e-tourism* masih terdapat tantangan terutama pada mesin pencari yang terkadang memberikan informasi yang kurang relevan dengan kata kunci yang dicari. Tantangan ini juga didasari oleh sistem yang digunakan untuk mengkases informasi masih belum berbasis pengetahuan (*semantics*) [5], Oleh karena itu mulailah diterapkan semantics dalam pencarian seperti Atqiya yang menciptakan semantik web yang digunakan untuk mencari data mobil [5] serta Fadhilah yang menerapkan semantik web untuk pencarian buku perpustakaan [6]. Dalam pembangunan sematik web terdapat suatu pilar yang mendukung pembangunan semantik web yang disebut dengan ontologi [7]. Ontologi merupakan definisi konsep yang berada dalam suatu ranah serta hubungan yang berlaku diantaranya [8]. Dalam pengembangan ontology sendiri terdapat beberapa proses, salah satunya adalah Ontology Population. *Ontology Population* merupakan suatu suatu proses menambahkan *instance* baru dari sumber data yang beragam kedalam suatu ontologi [9].

Dalam melakukan *ontologi population* terutama dengan metode semi otomatis dan otomatis terdapat beberapa tahapan penting antara lain pengumpulan data, ektraksi *instance*, lalu proses penambahan dalam ontologi [9]. Proses pengumpulan data adalah proses dimana data yang tersebar di internet dikumpulkan yang disebut dengan *multimedia corpus.* Dalam melakukan ontology population sendiri dapat dilakukan melalui beberapa cara antara lain secara manual, semi otomatis, dan otomatis [10].

Dalam melakukan *ontologi population* terutama dengan metode semi otomatis dan otomatis terdapat beberapa tahapan penting antara lain pengumpulan data, ektraksi *instance*, lalu proses penambahan dalam ontologi [9]. Proses pengumpulan data adalah proses dimana data yang tersebar di internet dikumpulkan yang disebut dengan *multimedia corpus.* Proses selanjutnya adalah proses ektraksi *instance* dari *multimedia corpus* yang telah dikumpulkan sebelumnya. Dikarenakan jumlah *multimedia corpus* yang bertambah setiap waktu berakibat proses ekstraksi *instance* dari *multimedia corpus* harus selalu dilakukan untuk menjaga kebaruannya. Oleh karena itu harus diterapkan metode untuk melakukan ekstraksi secara otomatis untuk memudahkan dalam proses tersebut dengan menggunakan *extraction engine*.

Hingga saat ini terdapat banyak *extraction engine* yang telah dikembangkan oleh manusia, secara umum *extraction engine* tersebut dapat dibagi menjadi 3 tipe berdasarkan dari metode yang digunakan untuk menemukan *term* dalam suatu teks yaitu *linguistic*, *statistic* dan *hybrid* [11]. Kualitas dari *term* yang dihasilkan dari *extraction engine* juga mempengaruhi hasil dari proses *ontology population*, apabila *term* yang dihasilkan adalah *term* yang berkualitas buruk maka akan menghasilkan ontologi yang buruk juga.

Tujuan dari penelitian ini adalah untuk mengevaluasi salah satu dari ketiga metode yaitu linguistic (POS) dengan menggunakan *multimedia corpu*s domain pariwisata pada daerah Bali serta mengukur relevansi dari *term* yang dihasilkan dari metode *linguistic* tersebut dengan menggunakan metode pengukuran *Precision*, *Recall* dan *F-Measure*.

Lopez dkk [12] pada tahun 2010 melakukan penilitan untuk menguji metode ekstraksi term yaitu menggunakan dua metode yaitu linguistik dan statistik. Pada penelitiannya Lopez dkk [12] menggunakan corpus Journal de Pediatria untuk melakukan evaluasi. Untuk metode linguistik Lopez dkk menggunakan sofware ExATOLP (*Automatic Extractor of Term for Ontologies Postuguese Language*) yang merupakan suatu software yang secara otomatis mengekstrak seluruh *noun* dan mengklasifikasikannya berdasarkan jumlah kata. Sedangkan untuk metode statistik Lopez dkk menggunakan NSP yang merupakan *software* untuk mengekstrak n-gram dari suatu *corpus*. Pada fase pengukuran Lopez dkk menggunakan metode pengukuran *recall, precision* dan *f-measure*. Hasil pengukuran menunjukkan bahwa linguistik memberikan hasil yang lebih baik dibandingkan dengan statistik. Namun hal tersebut tidak menunjukkan bahwa statistik lebih buruk dari linguistik karena pada metode stastistik yang diterapkan memiliki kelebihan yaitu lebih mudah untuk diadaptasikan tergantung dari *corpus* yang digunakan.
Selain Lopez dkk, Zhang dkk [13] juga melakukan penelitian untuk menguji dua metode ekstraksi term yaitu statistik dan hybrid dalam lima algoritma berbeda. Statistik meggunakan TF-IDF

dan wierdness. Sedangkan *hybrid* menggunakan C-Value, Glossex dan TermExtractor. Pada penelitian ini Zhang menggunakan dua *corpus*, yaitu *corpus* dari MEDLINE dan wikipedia. Dari penelitian yang dilakukan Zhang dkk menunjukkan bahwa *hybrid* menunjukkan performa yang lebih baik dibandingkan dengan statistik. Pada *corpus* MEDLINE C-value menunjukkan performa terbaik sedangkan pada *corpus* wikipedia algoritma Termex performa yang terbaik. kesimpulan ini diambil setelah dilakukannya perhitungan precision dan UAP *metrix* pada setiap metode pada setiap *corpus*.

Dengan terus berkembangnya penelitian yang berfokus pada evaluasi dari metode ekstraksi *instance* seperti yang telah dilakukan oleh Conrado dkk [11], Lopes dkk [12] dan Zhang dkk [13] akan sangat mendukung bagi para peneliti untuk melakukan pengembangan aplikasi yang menerapkan metode metode tersebut. Akan tetapi, minimalnya evaluasi yang dilakukan dengan menggunakan *multimedia corpus* yang berasal dari Indonesia menimbulkan pertanyaan cukup kompetenkah metode metode tersebut diterapkan dengan *multimedia corpus* yang berasal dari Indonesia? seberapa besar relevansi kandidat *instance* hasil penerapan dari metode tersebut apabila digunakan dengan *multimedia corpus* yang berasal dari Indonesia?

Berdasarkan dari fakta yang ada, terdapat peluang melakukan pengujian terhadap metode ektraksi *instance* terutama metode linguistik dengan menggunakan *multimedia corpus* Indonesia pada domain pariwisata. Diharapkan akan didapatkan hasil evaluasi performa dari metode linguistik (POS) terhadap *multimedia corpus* Indonesia pada domain pariwisata untuk digunakan pada penelitian selanjutnya. Dataset (*corpus*) yang dikumpulkan dapat digunakan sebagai acuan pada penelitian selanjutnya.

TINJAUAN PUSTAKA

1. Extraction Engine

Semantic Pada dasarnya istilah “ontologi” berasal dari studi filosofi yang berfokus pada keberadan atau eksistensi. Pada bidang filosofi ontologi dibicarakan sebagai suatu teori dari eksistensi atau keberadaan yang pada akhirnya istilah ini diadopsi oleh pada peneliti *Artificial Intelligent* untuk diterapkan pada bidang ilmu komputer. Ontologi pada awalnya diperkenalkan oleh Thomas Gruber pada tahun 1992 yang didefinisakannya sebagai spesifikasi dari suatu konseptual [15]. Namun berdasarkan dari pengertian Russel dan Norvig, Ontologi lebih mengarah pada teori tentang keberadaan. Dan pada konteks penelitian ini ontologi lebih mengarah pada definisi dari bidang *Artificial Intelligent* yang menyebutkan bahwa ontologi merupakan suatu spesifikasi formal dari konsep pada suatu domain dimana hubungan, batasan dan aksioma ditunjukkan [11].

Konsep semi-otomatis *annotation* diterapkan salah satunya dalam *Ontologi Learning*. *Ontologi Learning* merupakan suatu bentuk penggunaan dukungan otomatis atau semi otomatis dalam membangun suatu ontologi [13]. Menurut Georgios Petasis dalam penelitiannya terdapat beberapa proses utama dalam *ontology learning* yaitu *ontology population*, *ontology enrichment*, *ontology inconsistency resolution* dan *ontology evealuation* [9].

Menurut George Petasis [9] dalam proses ontologi populasi dibagi menjadi tiga tahapan utama yaitu pengumpulan data, ektraksi konsep/*instance,* lalu proses penambahan dalam ontologi dan menghasilkan *populated ontology*. Proses pengumpulan data adalah proses dimana data yang tersebar di internet dikumpulkan yang disebut dengan multimedia corpus. Mengingat kecepatan dari bertambahnya data di internet menyebabkan proses ini perlu dilakukan secara berulang ulang dan terus menerus agar menjaga agar informasi yang ada dalam suatu ontologi terjaga ke-*update*-annya dan mengikuti perkembangan dari informasi yang ada. Proses selanjutnya adalah proses ektraksi instance dari multimedia corpus yang telah dikumpulkan sebelumnya. Setelah dihasilkan kandidat instance yang sudah siap ditambahkan barulah tahap penambahan dalam ontologi dilakukan.

Georgios Petasis [9] menyebutkan bahwa dalam proses *ontology population* dibutuhkan 2 hal penting yaitu inisial ontologi yang akan dilakukan *population* dan suatu mesin *instance extraction* yang mana bertugas untuk menemukan *instance* dari *concepts* dan *relation* dari *multimedia corpus.* Hasil *instance* yang telah diektrak dari *multimedia corpus* selanjutnya akan digunakan untuk ditambahkan dalam inisial ontologi [9]. Proses pencarian *instance* itu sendiri bisa dibagi menjadi 3 yaitu manual, semi otomatis atau otomatis. Manual apabila seseorang melakukan penambangan *instance* dari suatu teks atau data secara manual dengan membaca dan menentukan terminologi apa yang pas mewakili dari teks tersebut lalu barulah diubah menjadi kandidat *instance*. Namun melakukan proses itu secara manual akan menjadi suatu tugas yang akan memakan waktu yang sangat lama mengingat jumlah dari informasi yang tersebar di internet sangat cepat bertambah [9].

Oleh karenanya digunakanlah *extraction engine* untuk memotong waktu pengerjaan.

Seperti yang telah dijelaskan sebelumnya bahwa secara umum metode ekstraksi *instance* dapat dibagi menjadi 3 yaitu *linguistic, statistic dan hybrid* [11]. Kualitas dari *instance* yang dihasilkan dari *extraction engine* juga mempengaruhi hasil dari proses *ontology population*, apabila *instance* yang dihasilkan adalah *instance* yang berkualitas buruk maka akan menghasilkan ontologi yang buruk juga. Oleh karenanya perlu adanya evaluasi untuk menentukan metode mana yang tepat untuk digunakan untuk suatu *multimedia corpus* agar dapat menghasilkan *instace* yang berkualitas baik.

Salah satu dari penerapan metode linguistic adalah POS (*Part of Speech*). POS atau banyak dikenal dengan POSTagger merupakan suatu skema yang membaca suatu teks dan menandai tiap kata yang ada sesuai dengan jenis kata tersebut seperti *noun, verb, adjective* dll [16]. Namun biasanya pada penerapannya kebanyakan aplikasi POSTagger menggunakan *POS tag* yang lebih spesifik seperti '*noun-plural*'. Secara garis besar metodologi dalam menghasilkan suatu *automatic POS Tagger* dibagi menjadi *statistical* dan *rule based* [17]. Dalam hal akurasi yang dihasilkan pendekatan secara *statistic* menghasilkan nilai 95-97% dan *rule based* menghasilkan 97% berdasarkan dari penelitian yang dilakukan oleh Eric Brill [18]. Hingga saat ini kedua metodologi tersebut telah diaplikasikan dalam berbagai macam *system* seperti Brill *rule based tagger* [18], *stochastic tagger* [19] dan hingga kini telah diaplikaskan di berbagai domain untuk melakukan *POS Tagging*

2. *Precision Recall* dan *F-measure*

Untuk melakukan evaluasi terhadap hasil ekstraksi dari metode yang diuji (*linguistic, statistical* dan *hybrid*) terdapat berbagai macam salah satunya adalah yang telah diterapkan oleh Lopez dkk [12] dimana mereka menggunakan perhitungan *Precision, Recall* dan *F-measure* untuk menentukan relevansi dari kandidat *term* yang dihasilkan dari penerapan metode (*linguistic* dan *statistical)* yang mereka gunakan.

Precision (P) adalah rasio jumlah dari *term* relevan yang diterima (A) dengan jumlah *term* relevan dan tidak relevan yang di terima (B) :

$$P = \frac{A}{A+B} \times 100\%$$

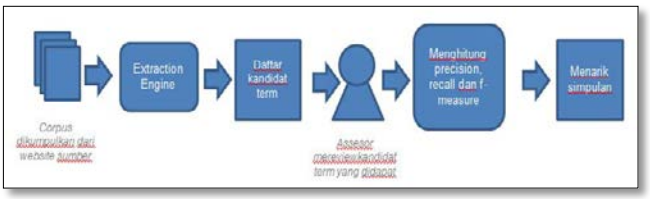
Recall (R) adalah rasio jumlah *term* relevan yang diterima (A) dengan jumlah *term* relevan yang seharusnya diterima (C) :

$$R = \frac{A}{A+C} \times 100\%$$

Dan *f-measure* (F) merupakan equilibrium antara kedua indeks yang telah dihitung sebelumnya dengan nilai β. B adalah nilai yang digunakan untuk menentukan kecenderungan suatu penelitian, apabila lebih cenderung ke recall maka β<1 sedangkan bila cenderung ke precision β>1 dan apabila seimbang maka β = 1 [36]

$$F = 2 \times \frac{P \cdot R}{P+R} \times 100\%$$

Apabila nilai *precision* yang didapat tinggi maka nilai *recall* yang didapat cenderung rendah dan begitu pula sebaliknya [40]. Salah satu tantangan yang dihadapi dalam perhitungan *Precision, Recall* dan *F-measure* adalah dikarenakan proses validasi untuk menentukan mana *term* yang relevan dan mana yang tidak dilakukan oleh manusia, tidak bisa dipungkiri bahwa hasil yang didapat bebas dari subjektifitas penilai itu sendiri [20]. Oleh karenanya digunakan lebih dari satu penilai untuk mendapatkan hasil yang tidak subjektif.



Gambar 1.Metode Penelitian

Metodologi

Kerangka kerja dari penelitian ini seperti dijelaskan pada gambar [1]. Pertama tama multimedia corpus dikumpulkan dari Internet. Setelah itu metode *linguistic* (*Part-of-Speech*) diaplikasikan pada *multimedia corpus* yang telah dikumpulkan sebelumnya. Metode *linguistic* (*Part-of-Speech*) yang digunakan dalam penelitian ini diaplikasikan dalam bentuk script PHP.

Daftar kandidat term yang didaptkan dari hasil ekstraksi kemudian dilakukan perhitungan akurasi berdasarkan kandidat term yang dihasilkan. Perhitungan keakuratan dari hasil ekstraksi ini dibagi menjadi dua bagian : 1) Review yang dilakukan oleh assesor dan 2) Perhitungan *precision, recall* dan f-*measure*. Assesor yang digunakan untuk melakukan *review* berjumlah 2 orang untuk menghindari subjektifitas assesor.

Ketika seluruh hasil ekstraksi *term* yang didapat telah selesai direview oleh seorang assesor kemudian hasil tersebut akan digunakan untuk

menghitung *precision*, *recall* dan f-*measure*. Perhitungan inilah yang nantinya akan digunakan sebagai dasar analisis untuk membandingkan mana metode yang dapat menghasilkan kandidat term yang lebih bagus.

HASIL DAN ANALISIS

Multimedia Corpus

Multimedia corpus yang dipakai dalam penelitian ini didapat dari tiga website yang menyediakan artikel tentang pariwisata bali, antara lain :

1. http://www.baligoldentour.com/bali-interest-place,
2. http://www.balistarisland.com/Bali-Interesting-Place/,
3. http://www.indonesia.travel/en/discover-indonesia/region-detail/35/bali.

Keseluruhan artikel yang didapat dari web tersebut dikumpulkan secara otomatis dengan menggunakan *web crawler* yang dieksport dalam bentuk file txt. Jumlah artikel yang didapat dari hasil ekstraksi ditunjukkan dalam tabel 1.

Hasil Ekstraksi

Dengan berdasar pada multimedia corpus yang sudah dikumpulkan secara otomatis kemudian dilakukanlah proses ekstraksi term dengan menerapkan metode *linguistic* (*Part-of-Speech*). Pada penelitian ini term yang dicari hanyalah noun dengan tanpa dilakukannya perhitungannya pada prosesnya sehingga seluruh noun yang ditemukan dalam proses akan dianggap sebagai suatu kandidat term. Hasil ekstraksi term dapat dilihat pada tabel 2.

Table 1 Jumlah term yang diterima

	jumlah artikel	jumlah kata
BGT	50	26750
BS	64	24701
IT	46	39643
total	**160**	**91094**

Table 2 Jumlah term diterima

Metode	Jumlah term diterima		
	BGT	BS	IT
Linguistic	5595	5384	7595

Analisis

Pada penelitian ini precision, recall dan f-measure dihitung menggunakan teknik pooling yang ditentukan oleh dua assesor yang memiliki cukup pengalaman dan pengetahuan tentang pariwisata Bali. Hasil perhitungan recall, precision dan f-measure dari kedua metode terdapat pada tabel 3.

Secara keseluruhan nilai hasil perhitungan precision yang didapat dari metode linguistic menunjukkan rata rata sebesar 0.48 dengan nilai maksimal sebesar 0.61 dan nilai terendah sebesar 0.41. Walaupun metode linguistik yang digunakan menghasilkan nilai yang masih dibawah dari 0.7 dimana itu masih jauh dari nilai yang ditemukan pada literatur yaitu sebesar 0.8 namun bukan berarti bahwa metode linguistik merupakan metode yang buruk untuk ekstraksi term. Salah satu faktor yang menyebabkan hal ini terjadi adalah definisi dari "*term*" yang digunakan pada penelitian ini masih terbatas pada pada *class* dan *atribute* yang terdapat pada *ontology* DWIPA versi 1, yang mana hanya terbatas pada *Attraction*, *Events*, *Accomodation* dan *Regency*. Keterbatasan ini menyebabkan nilai *precision* yang dihasilkan menjadi kecil.

Terlepas dari itu terdapat faktor lain yaitu metode penilaian yang digunakan pada penelitian ini menggunakan pooling yang memberikan daftar kandidat *term* dari masing masing metode dan direview oleh assesor. Hal ini menyebabkan assesor tidak mengetahui secara sepenuhnya kata tersebut berasal dari teks seperti apa sehingga assesor hanya terpaku pada batasan *class* dan *attribute* yang ada pada dwipa versi satu untuk menilai kandidat *term* yang dihasilkan.

Table 3 Hasil perhitungan precision, recall dan f-measure

		Precision	Recall	F-Measure
BGT	**Assesor 1**	0.405	0.422	0.408
	Assesor 2	0.437	0.401	0.407
BS	**Assesor 1**	0.511	0.386	0.406
	Assesor 2	0.613	0.394	0.425
IT	**Assesor 1**	0.418	0.392	0.397
	Assesor 2	0.474	0.391	0.405

KESIMPULAN

Dalam penelitian ini telah dibangun 3 data set yang didapat dari 3 website. Selain itu ditemukan juga bahwa metode ekstraksi linguistik masih memberikan hasil dibawah 0.7, namun tidak dapat diartikan bahwa metode ini buruk. Hal tersebut disebabkan oleh beberapa faktor, salah satunya adalah adanya keterbatasan dari kandidat *term* yang didapat pada *class* dan *attribute* yang ada pada ontology DWIPA versi 1. Untuk penelitian selanjutnya diharapkan untuk memperluas cakupan definisi dari "*term*" yang digunakan sehingga dapat

meningkatkan nilai *precision*, *recall* dan *f-measure* yang dihasilkan.

DAFTAR PUSTAKA

[1] I. W. Stats, "*Asia marketing research. internet usage. population statistic, and facebook information*," http://www.internetworldstats.com/stats.htm/, 2014, diakses pada 13 Januari 2015.

[2] Internet World Stats. 2014. *Asia Marketing Research. Internet Usage. Population Statistic, and Facebook Information*. http://www.internetworldstats.com/stats.htm. Diakses pada 13 Januari 2014.

[3] R. Winaga, "Basis pengetahuan pariwisata berbasis ontologi yang memodelkan waktu valid," Master's thesis, Universitas Indonesia, Jakarta, 2013.

[4] L. Y. Banowosari, I. W. S. W, S. Wirawan, and T. J. Dewi, "Memperkaya instances pada ontologi pariwisata dengan sumber dari internet," dalam Konferensi Nasional Sistem Informasi 2012. STMIK-STIKOM, 2012, pp. 214–219.

[5] A. Nurul Atqiya and T. Hariguna, "Semantik web untuk pencarian data mobil," http://digilib.ump.ac.id/files/disk1/23/jhptump-ump-gdl-akhlisnuru-1108-1-b.22.pdf/,2013, diakses pada 01 April 2015.

[6] C. N. Fadilah Navaatul and J. Herlina, "Penerapan teknologi semantic web pada aplikasi pencarian koleksi perpustakaan (studi kasus: Perpustakaan fti upn veteran yogyakarta.yogyakarta," dalam Proceedings: Seminar Nasional Sistem Informasi 2010, 2010, pp. D118–D128.

[7] M. Krtzsch, "Ontology," http://semanticweb.org/wiki/Ontology/, 2012, Diakses pada January 13, 2015.

[8] T. Berners-Lee, J. Hendler, and O. Lassila, "*The semantic web*," Scientific American, May 2001, accessed January 13, 2015.

[9] G. Petasis, V. Karkaletsis, G. Paliouras, A. Krithara, and E. Zavitsanos, "*Ontology population and enrichment: State of the art*," in Knowledge-driven multimedia information extraction and ontology evolution. Springer-Verlag, 2011, pp. 134–166.

[10] P. A. Kogut and W. S. Holmes III, "*Aerodaml: Applying information extraction to generate daml annotations from web pages*." in Semannot@K-CAP 2001, 2001.

[11] M. S. Conrado, R. G. Rossi, T. Pardo, S. O. Rezende et al., "*Applying transductive learning for automatic term extraction: the case of the ecology domain*," in Informatics and Applications (ICIA), 2013 Second International Conference on. IEEE, 2013, pp. 264–269.

[12] L. Lopes, L. H. Oliveira, and R. Vieira, "*Portuguese term extraction methods: Comparing linguistic and statistical approaches*," in International Conference on Computational Processing of Portuguese Language, PROPOR, vol. 6, 2010.

[13] Z. Zhang, J. Iria, C. Brewster, and F. Ciravegna, "A comparative evaluation of term recognition algorithms." in LREC, 2008.

[14] Drumond.Lucas dan Girardi.Rosario.2008.*A Survei of Ontology Learning Procedures.*Federal University of Maranhao. Sao Luis.

[15] H. Park, A. Yoon, and H.-C. Kwon, "*Task model and task ontology for intelligent tourist information service*," International Journal of u-and e-Service, Science and Technology, vol. 5, no. 2, pp. 43–58, 2012.

[16] S. N. Group et al., "*Stanford log-linear part of speech tagger*," URL: http://nlp. stanford. edu/software/tagger. shtml.

[17] M. Alex and L. Q. Zakaria, "*Brill's rule-based part of speech tagger for kadazan*," Int. J. on Recent Trends in Engineering and Technology, vol. 10, no. 1, 2014.

[18] E. Brill, "*A simple rule-based part of speech tagger*," in Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics, 1992, pp. 112–116.

[19] K. W. Church, "*A stochastic parts program and noun phrase parser for unrestricted text*," in Proceedings of the second conference on Applied natural language processing. Association for Computational Linguistics, 1988, pp. 136–143.

[20] P. Drouin, "*Term extraction using non-technical corpora as a point of leverage*," Terminology, vol. 9, no. 1, pp. 99–115, 2003.