JURNAL TEKNOLOGI TECHNOSCIENTIA ISSN: 1979-8415
Vol. 7 No. 2 Februari 2015

EKSTRAKSI INFORMASI KONTEN WEB MENGGUNAKAN PENDEKATAN BERBASIS ONTOLOGI

Erma Susanti[1]

[1]Jurusan Teknik Informatika, Institut Sains & Teknologi AKPRIND Yogyakarta

Masuk: 19 September 2014, revisi masuk: 8 Januari 2015, diterima: 19 Januari 2015

ABSTRACT

Most of information on the Internet is transmitted through unstructured information via websites. Information in the web pages usually contains main content, advertisement, navigation and other additional information. The amount of information makes it difficult to obtain the core information, value and relevant knowledge inthe form of structured information, like databases. Information extraction is a process to converting unstructured information into structured information. Applying information extraction method using ontology called Ontology-Based Information Extraction (OBIE) aims to provide semantic content for semantic web and perform extracting process using semi-automatic method (a method to minimize human intevention in extracting process). A case study to apply information extraction used "LonelyPlanet" dataset.

Keywords: *information extraction, Ontology-Based Information Extraction, OBIE ontology*

INTISARI

Sebagian besar informasi yang ada di Internet ditransmisikan secara tidak terstruktur melalui website. Informasi pada halaman web biasanya memuat konten utama, iklan, navigasi, dan informasi tambahan lainnya. Banyaknya informasi tersebut berakibat pada sulitnya untuk mendapatkan informasi inti, nilai dan pengetahuan yang relevan dalam bentuk informasi terstruktur, seperti basis data. Ekstraksi informasi merupakan suatu proses untuk mengubah informasi tidak terstruktur menjadi informasi terstruktur. Penggunaan metode ekstraksi informasi menggunakan ontologi disebut *Ontology-Based Information Extraction* (OBIE) bertujuan untuk menyediakan konten semantik untuk web semantik dan dapat melakukan proses ekstraksi secara *semi-automatic*(suatu metode yang dapat meminimalkan keterlibatan manusia dalam proses ekstraksi). Studi kasus untuk penerapan ekstraksi informasi ini dilakukan menggunakan *dataset* "LonelyPlanet".

Kata Kunci: ekstraksi informasi, *Ontology-Based Information Extraction,* OBIE, ontologi

PENDAHULUAN

Ekstraksi informasi merupakan suatu proses untuk mengubah informasi tidak terstruktur menjadi informasi terstruktur. Contoh informasi tidak terstruktur adalah informasi yang ada pada halaman web. Artikel-artikel yang dimuat pada suatu website sebagian besar berupa informasi tidak terstruktur, karena biasanya memuat informasi utama atau konten utama, iklan, navigasi, dan informasi tambahan lainnya. Banyaknya informasi tersebut berakibat pada sulitnya untuk mendapatkan inti informasi utama, sulit menemukan nilai dan pengetahuan yang relevan dalam bentuk informasi terstruktur, seperti bentuk basis data. Mekanisme untuk mengekstraksi sekumpulan teks untuk mendapatkan fakta-fakta dalam bentuk *events* (kejadian), entitas, dan *relationship* (keterhubungan) dalam bentuk informasi terstruktur sebagai masukan untuk basis data atau ontologi disebut sebagai ekstraksi informasi (Piskorski dan Yangarber, 2013).

Pendekatan ekstraksi informasi awal dijelaskan oleh Appelt (1999) dibagi menjadi dua pendekatan yaitu *knowledge engineering* dan *machine learning*. Pendekatan *knowledge engineering* merupa-

kan suatu pendekatan ekstraksi yang dilakukan oleh seorang *knowledge engineer*(ahli/pakar) dimana proses ekstraksi dilakukan secara manual. Aturan *grammar rules* (tata bahasa) ditulis dengan tangan oleh ahli atau pakar dalam suatu domain aplikasi. Proses manual ini akan memakan waktu yang cukup lama. Kemudian mulai dikembangkan pendekatan *machine learning*yang merupakan pendekatan ekstraksi yang dilakukan dengan membentuk *rules* (*aturan*) secara otomatis dengan melakukan proses *training*(pelatihan) terhadap data yang ada.

Banyaknya sumber data yang akan diekstraksi juga akan menyulitkan manusia dalam melakukan proses ekstraksi secara manual. Ketersediaan sistem yang dapat mengekstraksi informasi dari teks secara otomatis, akan sangat membantu dalam proses ekstraksi informasi. Otomatisasi ekstraksi dengan *machine learning* memiliki kelebihan pada proses pengembangan yang tidak memerlukan banyak waktu, namun belum dapat mengatasi masalah penyediaan konten semantik untuk web semantik (Labsky, 2008).

Penggunaan ontologi dapat dilakukan untuk mengatasi permasalah tersebut. Pendekatan ekstraksi informasi yang menggunakan ontologi disebut sebagai *Ontology-Based Information Extraction* (OBIE) (Wimalasuriya dan Dou, 2009). Pendekatan ekstraksi dengan ontologi untuk mengekstraksi *instance* dari teks juga pernah dilakukan oleh Cimiano dkk. (2005) dan Wimalasuriya dan Dou (2009). Cimiano dkk. (2005) menggunakan hierarki konsep dari teks menggunakan teknik *clustering* dengan menggunakan pendekatan *Formal Concept Analysis* (FCA) untuk secara otomatis membentuk ontologi berdasarkan pada ekstraksi pengetahuan dari teks. Pendekatan FCA tersebut dievaluasi dengan membandingkan hasil hierarki konsep dari taksonomi yang dibangun secara manual pada domain *tourism* dan *finance*.

Penelitian lain yang menggunakan ontologi untuk ekstraksi informasi pernah dilakukan oleh Wimalasuriya dan Dou (2009). Teknik yang digunakan berbeda dari penelitian sebelumnya yaitu dengan kombinasi *extraction rules* dan klasifikasi. Domain yang digunakan adalah tentang *terroist attack*.

Ekstraksi informasi menggunakan pendekatan berbasis ontologi (OBIE) akan dibahas pada penelitian ini menggunakan teknik yang berbeda dari sebelumnya.Pada penelitian ini akan dibahas bagaimana melakukan ekstraksi informasi pada artikel dari sumber web. Studi kasus yang digunakan adalah ekstraksi data *tourism* dari penelitian Cimiano dkk. (2005). Penelitian ini menggunakan ontologi untuk ekstraksi informasi dengan mengkombinasikan metode *extraction rules* dan daftar *gazetteer* (*instance* dari suatu kelas). Tujuannya adalah untuk dapat menyediakan konten semantik untuk web semantik dan dapat melakukan ekstraksi secara *semiautomatic*(suatu metode yang dapat meminimalkan keterlibatan manusia dalam proses ekstraksi).

METODE

Penelitian dimulai dengan memilih teks korpus dan domain ontologi yang relevan dengan jenis informasi yang akan diekstraksi. Teks korpus merupakan kumpulan teks dari koleksi halaman web yang akan dijadikan sebagai masukan pada proses ekstraksi. Sistem OBIE yang digunakan di sini dikembangkan dengan menggunakan metode *extraction rules*(aturan-aturan ekstraksi) dan *gazetteer list*(daftar *instance*).

Langkah awal dilakukan proses *training* untuk mengekstraksi informasi berkaitan dengan konsep domain studi kasus. Proses *training* diperlukan untuk pembentukan *extraction rules*. Implementasi akan menggunakan arsitektur GATE (*General Architecture for Text Engineering*) (Cunningham, 2002). *Extraction rules* ditulis ke dalam format *Java Annotation Pattern Engine* (JAPE) dan diinterpretasikan dalam GATE.

Studi kasus yang dipilih pada penelitian ini adalah ekstraksi informasi dari sumber teks pada halaman web (*webpages*) pada suatu domain spesifik. Ontologi berkaitan dengan konsep dalam suatu domain spesifik juga disediakan

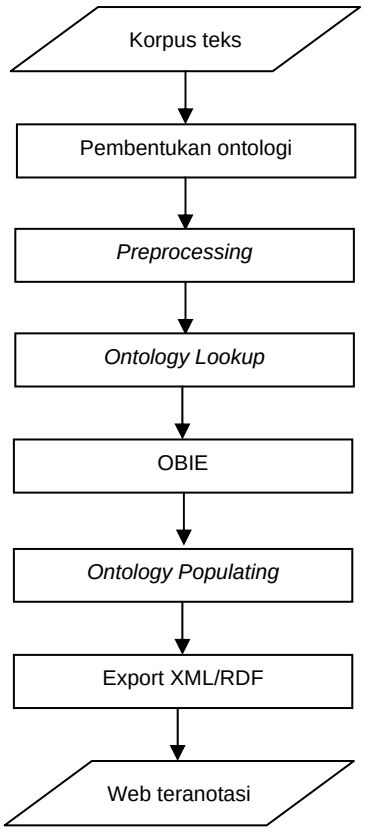
diawal untuk mendukung tujuan penelitian. Data untuk *benchmark* evaluasi menggunakan *dataset* "Lonely Planet" yang merupakan *dataset* dalam domain *tourism* (Cimiano dkk., 2005).

Sebelum memulai proses ekstraksi maka akan dilakukan *preprocessing* (pra-pemrosesan) dengan melakukan proses parsing teks. Parsing teks terdiri dari proses pemotongan kalimat (*Sentence Splitter*), pemotongan kata (*Tokenizer*), pemberian sintaksis informasi atau *Part of Speech* (POS) *Tagger* dan pemotongan frasa kata benda (*NP Chunker*). Tahapan awal ini bertujuan untuk mempersiapkan teks menjadi data yang dapat diproses sebagai masukan pada sistem ekstraksi informasi. Proses parsing dilakukan pada semua dokumen teks untuk mengidentifikasi semua kata benda (*noun*) atau frasa kata benda (*noun phrase*) dan konteksnya.

Proses pengembangan sistem ekstraksi informasi dengan menggunakan ontologi sebagai panduan digunakan untuk melakukan proses anotasi secara semantik dengan memberkan *link* pada entitas dalam teks menuju deskripsi semantik. Diagram alir untuk proses pengembangan sistem ekstraksi infomasi dapat dilihat pada Gambar 1.

Diagram alir proses pengembangan sistem ekstraksi dilakukan sebagai *pipeline* proses. Prosesnya dilakukan secara berurutan. Selesai proses yang satu, baru dapat beralih ke proses berikutnya. Pertama korpus teks masukan diproses ke dalam modul *preprocessing* untuk menghasilkan Ontologi akan digunakan untuk dihubungkan dengan teks melalui mekanisme anotasi semantik. Pemodelan ontologi untuk pembuatan API (*Application Programming Interface*) berdasarkan formalisasi OWL (*Ontology Web Language*), sedangkan untuk implementasi ontologi akan memanfaatkan OWLIM. Pengaturan untuk ontologi akan menggunakan GATE, dimana di dalamnya sudah tersedia fasilitas untuk editor ontologi, pembentukan daftar istilah (*OntoRoot Gazetteer*) dan sudah mendukung pembuatan *rules* dengan JAPE (*Java Annotatin Pattern Engine*). Fungsionalitas untuk pembentukan konsep baru dan *instances*, pendefinisian *properties* baru dan nilai *property* dan penghapusan dapat dilakukan melalui ontologi editor dalam GATE.Ontologi yang akan digunakan sebagai panduan proses ekstraksi di sini menggunakan ontologi yang sudah disediakan pada *dataset*. Taksonomi dari ontologi *e-tourism* bernama GETESS pada *dataset* ditunjukkan pada Gambar 2.

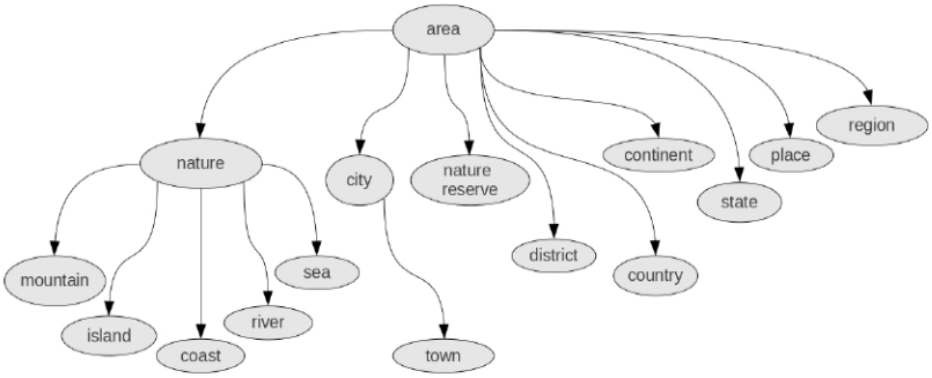
Format masukan yang dapat diterima oleh proses *Ontology lookup*. Hasil dari proses tersebut, kemudian diolah dalam modul OBIE untuk dapat dilakukan proses populasi ontologi. Selanjutnya hasil proses ekstraksi di-*export* ke dalam format XML/RDF. Keluaran dari proses tersebut berupa web teranotasi dan XML/RDF.



Gambar 1. Diagram alir proses pengembangan sistem ekstraksi informasi

Detail untuk sistem OBIE dijelaskan menggunakan arsitektur GATE. Komponen arsitektur GATE akan menggunakan arsitektur *Ontology-based Ga-*

zetteer (Cunningham dkk., 2011) seperti terlihat pada Gambar 3.



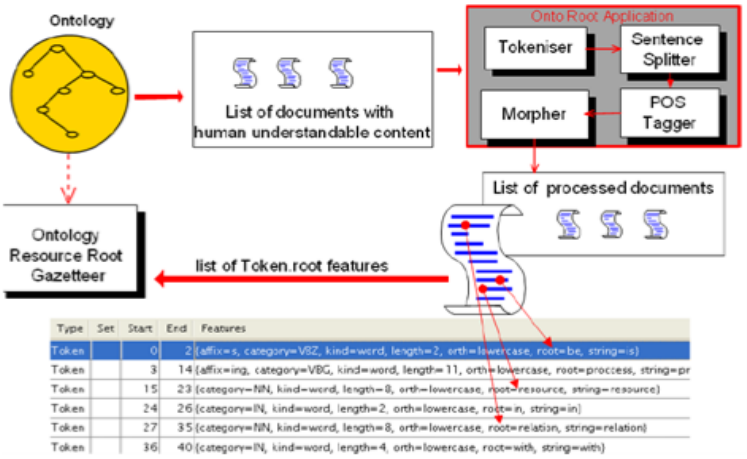
Gambar 2. Taksonomi *gold standard* untuk *dataset* "LonelyPlanet"

GATE akan dikoneksikan dengan OWLIM untuk membaca domain ontologi dan menyimpan hasil. Domain ontologi dimuat ke dalam GATE, kemudian teks akan dianotasi sesuai dengan kemunculan *instance* dalam teks dan nilai *property*. Komponen ekstraksi informasi akan menggunakan komponen bahasa (*Language Resource*) yang terdiri dari ontologi panduan dan korpus dokumen untuk ekstraksi teks. Korpus teks akan diproses sebagai masukan dalam aplikasi *OntoRoot*. Komponen pemrosesan (*Processing Resource*) menggunakan *OntoRoot* yang terdiri dari komponen *Tokenizer, Sentence Splitter, POS Tagger,* dan *Morpher.*

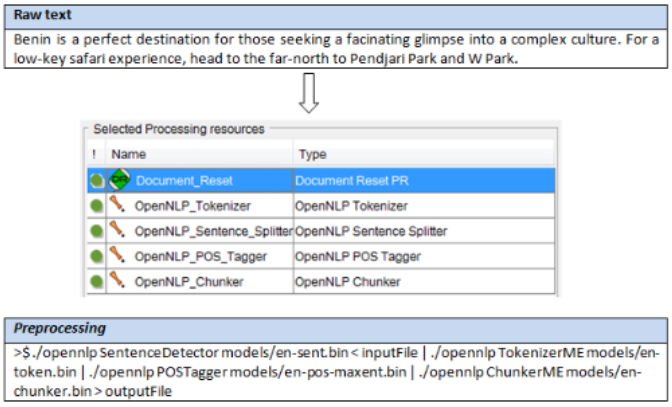
Hasil pemrosesan berupa anotasi dengan fitur sesuai dengan *gazetteer* pada ontologi. Setiap hasil *link* anotasi dari teks akan dipopulasi ke dalam domain ontologi yang digunakan. Hasil populasi adalah informasi terekstraksi yang akan di-*export* ke dalam bentuk XML/RDF.

PEMBAHASAN

Implementasi *preprocessing* melibatkan modul untuk ekstraksi informasi secara sintaksis dan leksikal dari teks, terdiri dari *Sentence Splitter*, *Tokenizer*, *POS Tagger*, dan *Chunker*. Implementasi *preprocessing* dijalankan sebagai komponen pemrosesan menggunakan *tools* Apache OpenNLP (Gambar 4).



Gambar 3. Ontology Based Gazetter (Cunningham dkk., 2011)



Gambar 4. Implementasi *preprocessing* dengan OpenNLP

Proses pembentukan ontologi diimplementasikan menggunakan komponen ontologi yang terintegrasi dalam GATE. Proses inisialisasi dilakukan dengan memuat *OWLIM Ontology* baru ke dalam aplikasi sebagai komponen *Language Resource* (LR).

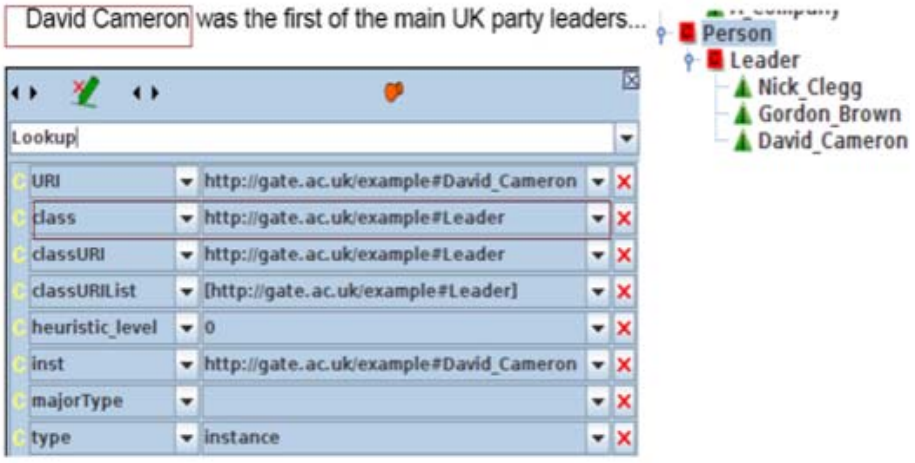
Implementasi *Ontology Lookup* dijalankan sebagai *pipeline*dari *Processing Resource* (PR). Implementasi *Ontology Lookup* terdiri dai komponen antara lain:

Morpher, pemrosesan sumber digunakan untuk mendapatkan lemma dari token (kata).

Gazetteer, mengklasifikasikan tipe kata yang ada seperti *cities*, *organization*, *countries*, *dates*, dan lain-lain.

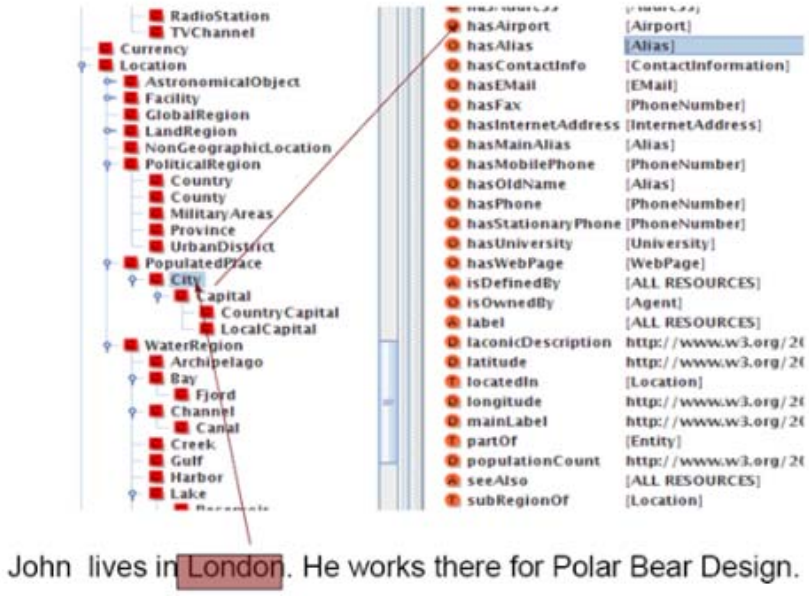
Lookup, menggunakan lemma dari token untuk melakukan *pattern matching*(pencocokan pola) untuk mendapatkan entitas nama.

Contoh hasil implementasi *Lookup* sesuai dengan ontologi panduan untuk ekstraksi dapat dilihat pada Gambar 5.



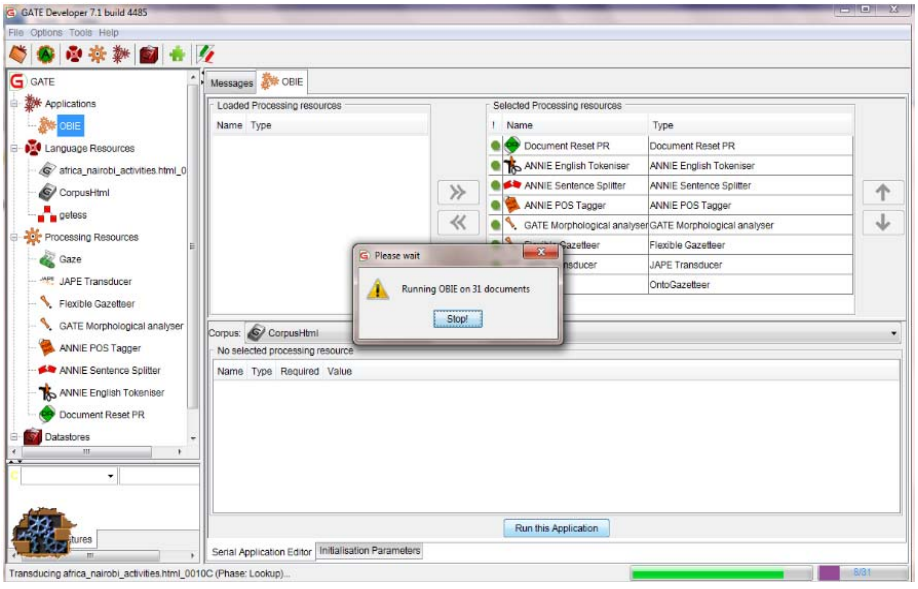
Gambar 5. Hasil pencocokan rules

Proses populasi ontologi merupakan proses untuk mendapatkan *instance* teks ke dalam ontologi, kemudian menghubungkannya *link* dari *instance* dari konsep dalam ontologi. Gambar 6. menunjukkan contoh hasil anotasi “London” dipopulasi sebagai *instance* ontologi pada *class* “City”.



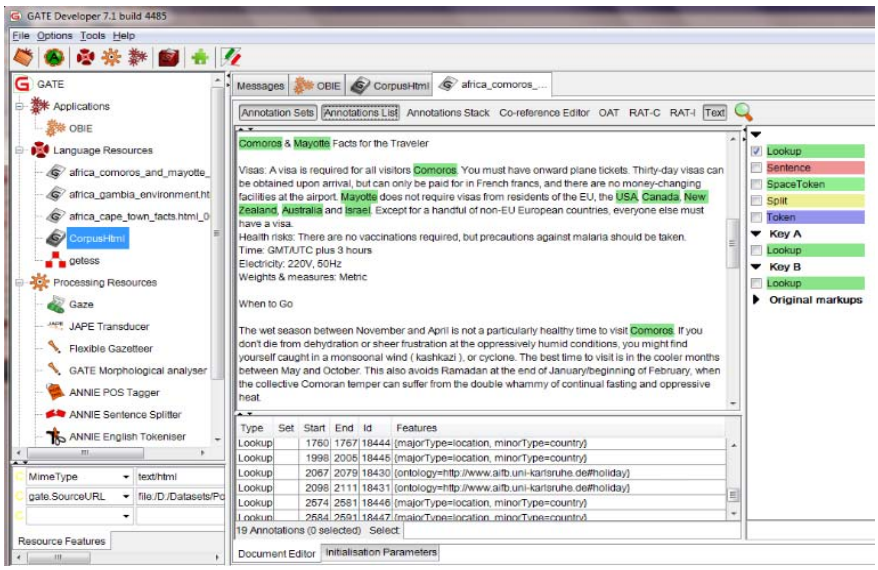
Gambar 6. Populasi ontologi

Hasil implementasi sistem OBIE dijalankan sebagai komponen pemrosesan dalam GATE. Komponen pemrosesan disimpan sebagai satu sumber *resource* aplikasi berekstensi "gapp" atau "xgapp". Implementasi sistem OBIE dijalankan sebagai *pipeline* komponen pemrosesan seperti terlihat pada Gambar 7. Ekstraksi korpus dilakukan dengan menjalankan aplikasi.



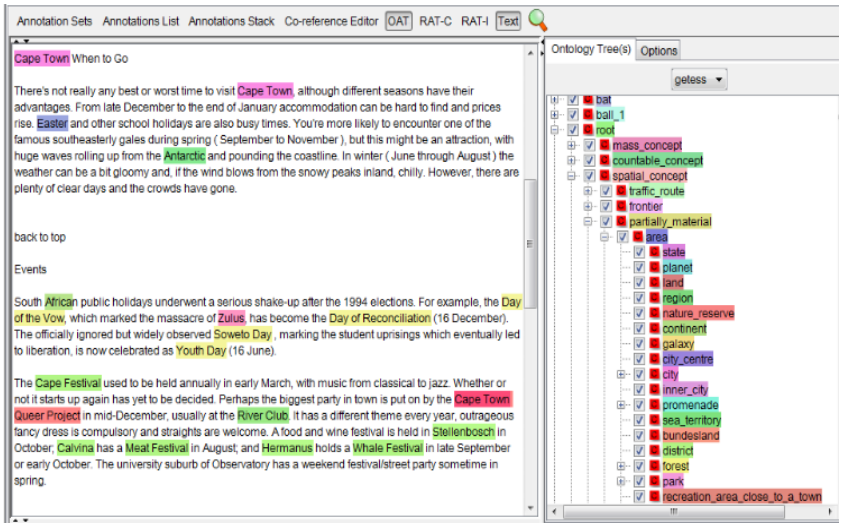
Gambar 7. Menjalankan aplikasi OBIE

Hasil proses *ontology lookup* pada korpus ditunjukkan pada Gambar 8. Keterangan tipe anotasi dapat dilihat pada bagian tengah bawah pada jendela aplikasi. Hasil anotasi menggunakan OAT (*Ontology Annota-tion Tools*) dapat dilihat pada Gambar 9.



Gambar 8. Hasil Lookup

Kata/frasa pada teks di bagian tengah jendela aplikasi diberi pelabelan berdasarkan ontologi pada bagian kanan jendela aplikasi. Keterangan tipe dan fitur kata/frasa yang menghubungkan ke kelas dapat dilihat pada Gambar 10.

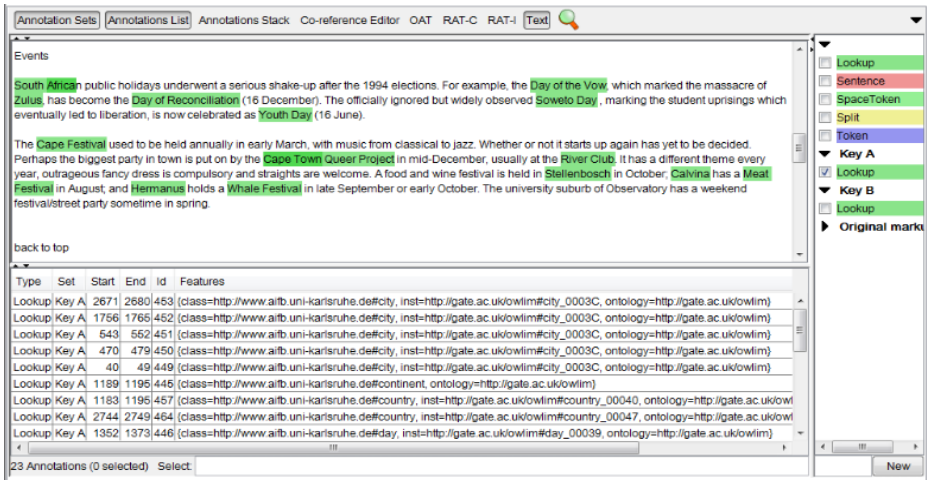


Gambar 9. Hasil anotasi menggunakan OAT

Setelah proses ekstraksi selesai, selanjutnya dilakukan evaluasi pengujian untuk mengidentifikasi *named entity* (entitas bernama) pada sistem OBIE pada penelitian ini yang diujikan menggunakan 30 *datatesting* dari *dataset* "LonelyPlanet". Evaluasi pengujian performa menggunakan *Corpus Benchmark Tool* (CBT) pada GATE. Hasil pengujian dengan menggunakan CBT menunjukkan hasil Presisi 73%, Recall 62% dan F-Measure 67%.

Hasil evaluasi ini dengan menggunakan *dataset* yang sama dengan penelitian Cimiano dkk. (2005) menunjukkan peningkatan kinerja F-Measure, dimana pada penelitian Cimiano dkk. (2005)nilai Presisi 29,33%,

Recall 65,49% dan *F-Measure* 40,52%. Nilai *F-Measure* dipilih karena merupakan nilai rata-rata harmonik dari Presisi dan Recall. Semakin tinggi nilai *F-Measure* di sini menunjukkan kinerja sistem yang cukup baik.



Gambar 10. *Link* hasil anotasi ke *instance* dari ontologi

KESIMPULAN

Ekstraksi informasi dari sumber web telah dilakukan dengan menjalankan proses ekstraksi dengan menggunakan panduan ontologi yang sudah ada. Proses ekstraksi ini disebut sebagai *Ontology-Based Information Extraction* (OBIE). Sistem ekstraksi yang dikembangkan menggunakan arsitektur GATE ini telah diuji coba menggunakan *dataset "LonelyPlanet"*. Hasil evaluasi pengujian yang telah dilakukan menunjukkan hasil peningkatan kinerja yang ditunjukkan dengan peningkatan hasil *F-Measure* bila dibandingkan dengan penelitian sebelumnya (Cimiano, 2005) dengan menggunakan *dataset* uji yang sama. Evaluasi menggunakan *Corpus Benchmark Tool* (CBT) pada GATE menunjukkan hasil Presisi 73%, *Recall* 62% dan *F-Measure* 67%.

Berdasarkan hasil penelitian tersebut dapat disimpulkan bahwa penggunaan ontologi untuk ekstrasi informasi merupakan pendekatan yang menjanjikan hasil performa yang dapat terus ditingkatkan untuk penelitian selanjutnya.

DAFTAR PUSTAKA

Appelt, D.E., 1999, Introduction to Information Extraction, *Ai Communication*, No. 3, Vol. 12, hal. 161-172.

Cimiano, P., Hotho, A. dan Staab, S., 2005, Learning concept hierarchies from text corpora using formal concept analysis, *Journal of Artificial Intelligence Research*, Vol. 24, hal. 305-339.

Cunningham, 2002, GATE A General Architectue of Text Engineering, *Computers and the Humanities*, No. 2, Vol. 36, hal. 223-254.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Robert, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M.A., Saggion, H., Petrak, J., Li, Y. dan Peters, W., 2011, *Text Processing with GATE (Version 6)*, Department of Computer Science University of Sheffield.

Labsky, M., 2008, Information Extraction from Website using Extraction Ontologies, *Disertasi*, University of Economic Prague.

Piskorski, J. dan Yangarber, R., 2013, *Information Extraction: Past,*

Present and Future, In: T. Poibeau *et al.*, eds., *Multi-source, Multilingual Information Extraction and Summarization,* Springer-Verlag, Berlin.

Wimalasuriya, D.C. dan Dou, D., 2009, Ontology-Based Information Extraction: An Introduction and a Survey of Current Approach, *Journal of Information Science*, No. X, Vol. XX, hal. 1-20.