IJCCS, Vol.9, No.2, July 2015, pp. 111~120
ISSN: 1978-1520

Ekstraksi Informasi Halaman Web Menggunakan Pendekatan *Bootstrapping* pada *Ontology-Based Information Extraction*

Erma Susanti*[1], Khabib Mustofa[2]
[1]Program Studi S2/S3 Ilmu Komputer, FMIPA UGM, Yogyakarta
[2]Jurusan Ilmu Komputer dan Elektronika, FMIPA UGM, Yogyakarta
e-mail: ***[1]erma.susan@gmail.com,** [2]khabib@ugm.ac.id

Abstrak

Ekstraksi informasi merupakan suatu bidang ilmu untuk pengolahan bahasa alami, dengan cara mengubah teks tidak terstruktur menjadi informasi dalam bentuk terstruktur. Berbagai jenis informasi di Internet ditransmisikan secara tidak terstruktur melalui website, menyebabkan munculnya kebutuhan akan suatu teknologi untuk menganalisa teks dan menemukan pengetahuan yang relevan dalam bentuk informasi terstruktur. Contoh informasi tidak terstruktur adalah informasi utama yang ada pada konten halaman web. Bermacam pendekatan untuk ekstraksi informasi telah dikembangkan oleh berbagai peneliti, baik menggunakan metode manual atau otomatis, namun masih perlu ditingkatkan kinerjanya terkait akurasi dan kecepatan ekstraksi. Pada penelitian ini diusulkan suatu penerapan pendekatan ekstraksi informasi dengan mengkombinasikan pendekatan bootstrapping dengan Ontology-based Information Extraction (OBIE). Pendekatan bootstrapping dengan menggunakan sedikit contoh data berlabel, digunakan untuk memimalkan keterlibatan manusia dalam proses ekstraksi informasi, sedangkan penggunakan panduan ontologi untuk mengekstraksi classes (kelas), properties dan instance digunakan untuk menyediakan konten semantik untuk web semantik. Pengkombinasian kedua pendekatan tersebut diharapkan dapat meningkatan kecepatan proses ekstraksi dan akurasi hasil ekstraksi. Studi kasus untuk penerapan sistem ekstraksi informasi menggunakan dataset "LonelyPlanet".

***Kata kunci**—Ekstraksi informasi, ontologi, bootstrapping, Ontology-Based Information Extraction, OBIE, kinerja*

Abstract

Information extraction is a field study of natural language processing by converting unstructured text into structured information. Several types of information on the Internet is transmitted through unstructured information via websites, led to emergence of the need a technology to analyze text and found relevant knowledge into structured information. For example of unstructured information is existing main information on the content of web pages. Various approaches for information extraction have been developed by many researchers, either using manual or automatic method, but still need to be improved performance related accuracy and speed of extraction. This research proposed an approach of information extraction that combines bootstrapping approach with Ontology-Based Information Extraction (OBIE). Bootstrapping approach using small seed of labelled data, is used to minimize human intervention on information extraction process, while the use of guide ontology for extracting classes, properties and instances, using for provide semantic content for semantic web. Combining both approaches expected to increase speed of extraction process and accuracy of extraction results. Case study to apply information extraction system using "LonelyPlanet" datasets.

***Keywords—** Information extraction, ontology, bootstrapping, Ontology-Based Information Extraction, OBIE, performance*

Received July 23th,2014; Revised January 8th, 2015; Accepted July 10th, 2015

1. PENDAHULUAN

Pertumbuhan jumlah web di Internet berdasarkan *survey* dari [1] menunjukkan peningkatan pesat dari 18 juta pada tahun 2000 menjadi 716 juta web pada tahun 2013. Berbagai jenis informasi di Internet ditransmisikan secara tidak terstuktur menyebabkan munculnya kebutuhan akan suatu teknologi ekstraksi informasi untuk menganalisa teks dan menemukan pengetahuan yang relevan dalam bentuk informasi terstruktur. Ekstraksi informasi bertujuan untuk mengekstraksi sekumpulan data teks untuk mendapatkan "fakta-fakta berkaitan dengan kejadian (*events*), entitas, atau keterhubungan (*relationship*)" dalam bentuk informasi terstruktur sebagai masukan untuk basis data [2].

Pendekatan yang dapat digunakan untuk membangun sistem ekstraksi informasi menurut [3] dibagi menjadi dua, yaitu *knowledge engineering* dan *automatic training*. Pendekatan *knowledge engineering* atau *rule-based* menggunakan komponen berupa *grammar*/*rules* yang ditulis secara manual oleh *knowledge engineer* (pakar). Sedangkan, pada pendekatan *automatic training* atau *machine learning*, pembentukan *rules* dilakukan secara otomatis dengan mempelajari dari data latih yang ada. Pendekatan *automatic training* secara umum diaplikasikan menggunakan algoritma klasifikasi, contohnya menggunakan Support Vector Machine (SVM) atau Conditional Random Field (CRF) [4].

Kelebihan dari pendekatan *knowledge engineering* menurut [3] adalah pada kinerja yang baik karena *rules* dituliskan secara manual oleh pakar sehingga hasilnya lebih akurat dan tidak diperlukan data latih. Kekurangannya terletak pada proses pengembangannya yang sangat menyulitkan karena memerlukan banyak tenaga, sulit untuk dilakukan perubahan spesifikasi setelah sistem jadi, dan juga perlu tersedianya orang yang ahli dalam membuat *rules*. Selain itu, *rules* dapat berjalan lambat dan proses pengembangannya memakan waktu yang lama [5]. Ketersediaan sistem yang dapat mengekstraksi informasi secara otomatis akan sangat membantu proses ekstraksi informasi. Otomatisasi ekstraksi informasi dapat dilakukan dengan menggunakan pendekatan *machine learning*. Proses pengembangannya tidak memerlukan banyak waktu, namun diperlukan ketersediaan data latih yang besar [4]. Permasalahan lainnya adalah terkait dengan penyediaan konten semantik untuk web semantik [6].

Penggunaan ontologi sebagai panduan untuk ekstraksi informasi disebut sebagai *Ontology-Based Information Extraction* (OBIE), dikembangkan oleh [6] merupakan pendekatan ekstraksi informasi yang memanfaatkan ontologi sebagai panduan untuk mengekstraksi informasi dari dokumen. Permasalahan pada pendekatan tersebut adalah perlunya suatu semantik leksikon (kamus kata atau kosa kata) untuk domain spesifik data yang diekstraksi. Oleh karena itu, pendekatan *bootstrapping* menggunakan sedikit contoh data berlabel, diusulkan untuk mengatasi permasalahan tersebut. Penggunaan pendekatan *bootstrapping* dilakukan untuk meminimalkan campur tangan manusia dalam pembentukan pengetahuan [7]. Pada penelitian ini diusulkan pendekatan *bootstrapping* dan *Ontology-Based Information Extraction* untuk mengekstraksi informasi dari sumber halaman web. Pengkombinasian kedua pendekatan tersebut diharapkan dapat meningkatkan akurasi hasil ekstraksi dan kecepatan proses ekstraksi informasi.

2. METODE PENELITIAN

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

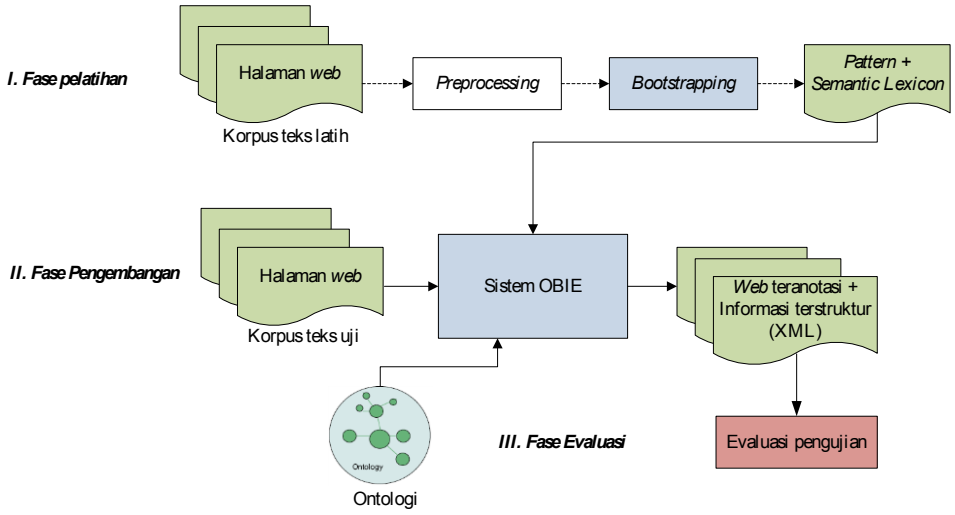
2.1 Deskripsi Sistem

Sistem diawali dengan proses pelatihan menggunakan pendekatan *bootstrapping* untuk mengekstraksi informasi berkaitan dengan konsep. Proses pelatihan ini diperlukan untuk pembentukan *extraction rules* dan daftar kamus. *Extraction rules* dan daftar kamus akan digunakan sebagai komponen untuk sistem OBIE. Selanjutnya sistem OBIE diimplementasikan menggunakan *framework* GATE (*General Architecture for Text Engineering*) [8]. *Extraction rules* ditulis ke dalam format yang diketahui sebagai *Java Annotation Pattern Engine* (JAPE) sebagai komponen GATE. Penggunaan studi kasus digunakan untuk menunjukkan penggunaan dari sistem ekstraksi informasi yang diteliti dan untuk mengetahui hasil evaluasi dari penerapan pendekatan yang digunakan untuk ekstraksi informasi. Setelah proses ekstraksi selesai, dilakukan pengujian evaluasi untuk mengetahui kinerja sistem, terkait nilai *precision*, *recall*, *F-measure* dan kecepatan waktu ekstraksi. Selanjutnya berdasarkan hasil evaluasi yang diperoleh, dilakukan perbandingan kinerja ekstraksi informasi dengan penelitian terdahulu.

2. 2 Arsitektur Sistem

Arsitektur sistem ekstraksi informasi pada penelitian ini merupakan adaptasi dari arsitektur umum sistem ekstraksi informasi dari [9] dan arsitektur umum sistem OBIE dari [6]. Masukan sistem ekstraksi informasi berupa teks bahasa alami (*natural language text*) tidak terstruktur dari sumber teks pada halaman *web* (teks HTML). Proses ekstraksi informasi sesuai dengan konsep OBIE akan melibatkan ontologi sebagai panduan ekstraksi dan menghasilkan keluaran berupa informasi terekstraksi yang direpresentasikan dalam bentuk XML dan teks teranotasi. Proses ekstraksi yang dipandu oleh ontologi akan mengekstraksi sesuatu (*things*) seperti kelas (*classes*), *properties*, dan *instance*. Selanjutnya dilakukan pengujian evaluasi kinerja dari sistem ekstraksi informasi menggunakan parameter pengukuran *precision*, *recall*, *F-measure*, dan kecepatan.

Arsitektur sistem ekstraksi informasi pada penelitian ini dibagi menjadi tiga fase yaitu fase pelatihan, fase pengembangan dan fase evaluasi. Pada fase pelatihan, sistem mengidentifikasi pola (*patern*) dan daftar kamus (disebut dengan semantik leksikon), yang dipelajari menggunakan pendekatan *bootstrapping*. Sebelumnya, korpus harus melewati tahapan *preprocessing* sebelum proses pelatihan. Tujuan dari fase pelatihan adalah untuk menghasilkan *pattern* dan semantik leksikon. Pada fase pengembangan merupakan fase untuk mengidentifikasi dan mengklasifikasi informasi yang relevan pada kumpulan teks baru. Teks yang digunakan tersebut tidak termasuk dalam korpus pada proses pelatihan. *Patern* di-*generate* untuk mendapatkan *extraction rules*. Pada fase pengembangan, teks masukan dilewatkan ke sistem OBIE untuk menghasilkan suatu keluaran. Fase terakhir adalah fase evaluasi atau pengujian. Arsitektur sistem OBIE pada penelitian ini dapat dilihat pada Gambar 1.



Gambar 1 Arsitektur sistem ekstraksi informasi

2.2.1 Preprocessing

Tahap awal dari ekstraksi informasi adalah dengan melakukan *preprocessing* pada masukan teks yang bertujuan untuk mempersiapkan teks menjadi data yang dapat diproses sebagai masukan sistem ekstraksi informasi. Proses *parsing* dilakukan pada semua dokumen teks untuk mengidentifikasi semua kata benda (*noun*) atau frasa kata benda (noun phrase/NP) dan konteksnya.

Proses *parsing* dalam *preprocessing* terdiri dari deteksi kalimat (*sentence detection*), pemotongah kalimat menjadi *token* atau kata (tokenization), pemberian sintaksis informasi (*POS tagging*), dan pemotongan frasa (*NP chunker*). Setelah *parsing* teks selesai, selanjutnya dilakukan proses *indexing* dan *filtering*. Proses *indexing* kalimat dilakukan dengan mendeteksi kata/frasa NP dalam kalimat, *token* di sebelah kiri NP dan *token* di sebelah kanan NP. Setelah proses *indexing* selasai, maka keluaran (disebut *document-set*) disimpan ke dalam basis data untuk diproses dalam proses *bootstrapping*. Sebelum diproses dalam algoritma *bootstrapping*, *document-set* perlu di-*filtering* dengan membuang kata/frasa NP yang tidak akan menghasilkan ekstraksi. Contoh *document-set* yang perlu untuk dilakukan *filtering* yaitu kata benda yang mengandung angka, karena angka tidak akan mengekstraksi entitas nama. Proses *indexing* terhadap *token* dari *window* kiri dan kanan NP memanfaatkan *query* dalam *Lucene Based Search* (terintegrasi dalam GATE) untuk mendapatkan konteks yang dapat digunakan sebagai masukan dalam *bootstrapping*.

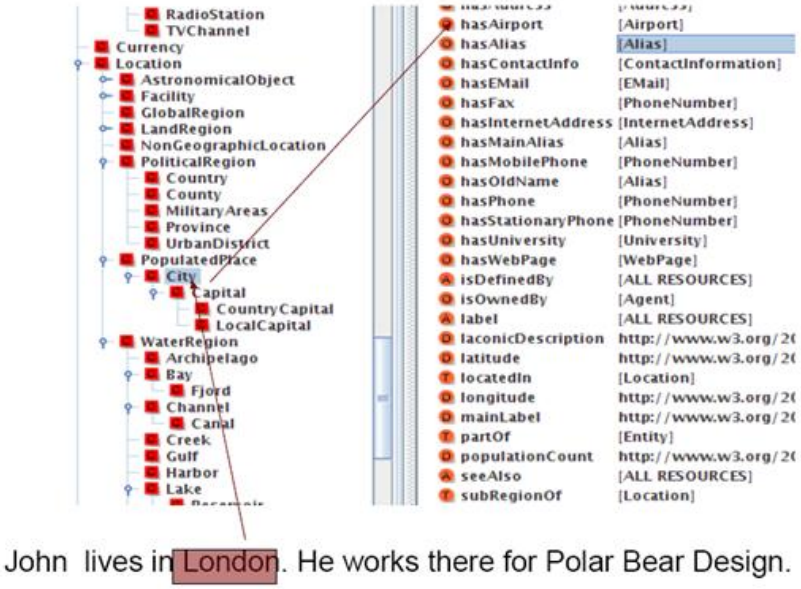
2. 2.2 Bootstrapping

Tahapan ekstraksi menggunakan algoritma *bootstrapping* langkah-langkahnya antara lain:

1. Mulai dengan sejumlah kecil *seed* (contoh) yang diambil dari *instance* pada ontologi panduan.
2. Masukkan jumlah iterasi yang akan dilakukan.
3. Cocokkan *seed* dengan *document-set* dari korpus teks yang telah di-*preprocessing* untuk mendapatkan kemunculan *term* (*occurrence*) dan *pattern extraction*.
4. Hitung skor untuk *pattern extraction* berdasarkan kemunculan *occurrence* dan *document-set*.
5. Aplikasikan *pattern extraction* dengan skor terbaik pada seluruh *document-set* dan catat hasil ekstraksi sebagai anggota kandidat leksikon.
6. Hitung skor untuk kandidat leksikon berdasar pada jumlah *pattern* yang berhasil melakukan ekstraksi.
7. Tambahkan kandidat *seed* baru terbaik ke dalam leksikon dan gunakan sebagai *seed* pada iterasi berikutnya.
8. Kembali ke langkah tiga. Berhenti jika proses sudah tidak dapat menghasilkan *pattern* dan *seed* baru, atau berhenti jika iterasi telah mencapai maksimum.

2.2.3 Ontology-Based Information Extraction

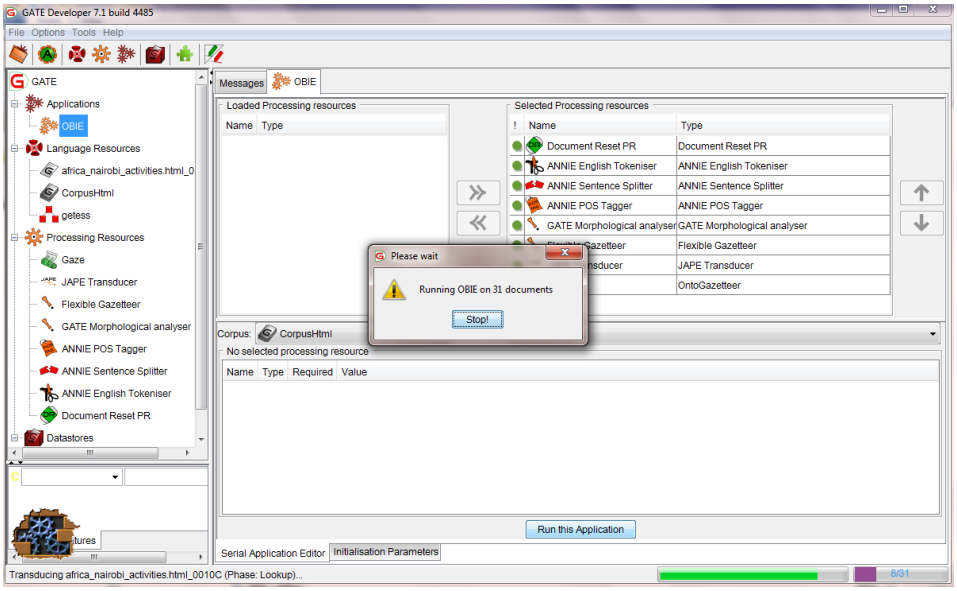
Rancangan diagram alir proses pengembangan sistem dijalankan sebagai *pipeline* proses. Prosesnya dilakukan secara berurutan. Selesai proses yang satu, baru dapat beralih ke proses berikutnya. Pertama korpus teks masukan diproses ke dalam modul *preprocessing* untuk menghasilkan format masukan yang dapat diterima oleh proses *Ontology Lookup*. Hasil dari proses tersebut kemudian diolah dalam modul OBIE untuk dapat dilakukan proses populasi ontologi. Proses populasi pada ekstraksi informasi dengan panduan ontologi digambarkan sebagai proses untuk mendapatkan dan menghubungkan entitas yang ada dalam teks dengan *classes, properties, atau instance* pada ontologi (lihat Gambar 2).



Gambar 2 Proses populasi ontologi

Detail rancangan sistem OBIE akan menggunakan Ontology-based Gazetteer pada GATE. GATE akan dikoneksikan dengan OWLIM untuk membaca domain ontologi dan menyimpan hasil. Domain ontologi dimuat ke dalam GATE, teks dianotasi sesuai dengan kemunculan *instance* dalam teks dan nilai *property*. Komponen ekstraksi informasi akan menggunakan komponen bahasa (*Language Resource*/LR) yang terdiri dari ontologi panduan dan korpus dokumen untuk ekstraksi teks. Korpus teks berupa halaman web akan diproses sebagai masukan dalam aplikasi *OntoRoot*.

Komponen pemrosesan (*Processing Resource*/PR) pada sistem OBIE ditunjukkan pada Gambar 3 terdiri dari komponen *Tokenizer*, *Sentence Splitter*, *POS Tagger*, dan *Morpher*. Hasil pemrosesan berupa anotasi dengan fitur sesuai dengan *gazetteer* pada ontologi. Selanjutnya hasil proses ekstraksi di-*export* kedalam format XML. Keluaran dari proses tersebut adalah berupa *web* teranotasi dan XML.



Gambar 3 Komponen pemrosesan pada sistem OBIE

2.2.4 Pengukuran Kinerja

Pengukuran evaluasi kinerja ekstraksi informasi didefinisikan sebagai pengukuran kinerja sistem ekstraksi informasi dibandingkan dengan *human-annotated gold standard*. Pengukuran evaluasi kinerja sistem ekstraksi informasi dapat menggunakan model pengukuran kuantitatif. Pengukuran kuantitatif ini terkait dengan perhitungan hasil keluaran yang diberikan oleh sistem dibandingkan dengan hasil yang seharusnya dikeluarkan oleh sistem. Evaluasi pengukuran hasil ekstraksi informasi disini menggunakan perbandingan hasil ekstraksi informasi dengan evaluasi *gold standard*.

Model pengukuran evaluasi ekstraksi informasi mengadopsi model pengukuran untuk tugas klasifikasi teks yaitu menggunakan *precision* dan *recall*. Perbandingan dua pengklasifikasi (*classifiers*) menggunakan nilai pengukuran tunggal *F-measure*, yang merupakan kombinasi nilai *precision* dan *recall* [9]. Parameter-parameter tersebut digunakan untuk mengetahui kinerja tingkat keberhasilan sistem ekstraksi informasi yang dikembangkan. Persamaan untuk hitung *precision (P), recall (R), F-measure* dan kecepatan (*Kec*) dapat dilihat pada persamaan (1), (2), (3), dan (4).

$$\boldsymbol{P} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \quad (1)$$

$$\boldsymbol{R} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \quad (2)$$

$$F - measure = \frac{2PR}{P + R} \quad (3)$$

$$Kec = \frac{JumlahDok}{Waktu} \quad (4)$$

Dimana:

TP = jumlah dokumen relevan yang ditemukan (*true positive*).
FP = jumlah dokumen tidak relevan yang ditemukan (*false positive*).
FN = jumlah dokumen tidak relevan dan tidak ditemukan (*false negative*).
JumlahDok = jumlah dokumen ekstraksi.

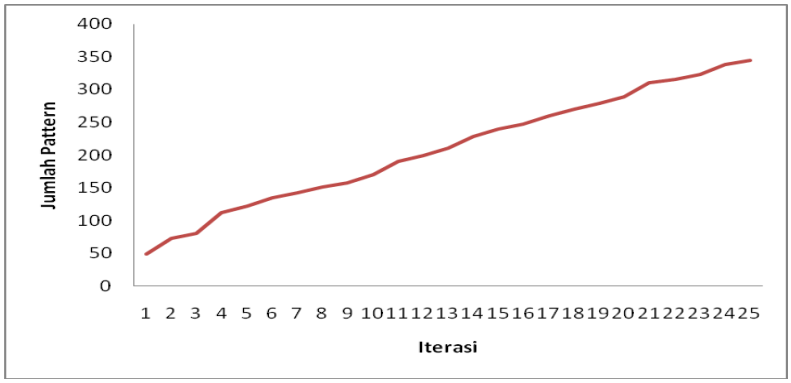
2.2.5 Evaluasi Pengujian

Korpus teks untuk pelatihan dan pengujian menggunakan *dataset* "LonelyPlanet" dari penelitian [8]. Dataset ini juga telah digunakan sebagai *benchmark* pada penelitian [10]. Motivasi pemilihan dataset tersebut sebagai *benchmark* karena penggunaan *dataset* secara umum digunakan pada penelitian ekstraksi informasi sehingga hasil dari metode yang digunakan dapat secara langsung dibandingkan dengan hasil dari metode lain. *Dataset* "LonelyPlanet" (LP) dijelaskan dalam penelitian [11] merupakan korpus dalam domain *e-tourism*, berisi kumpulan teks bahasa alami (*natural language text*) berbahasa Inggris yang terdiri dari 1801 file html yang dikoleksi dari situs http://www.lonelyplanet.com. *Dataset* LP berisi kumpulan halaman web tentang deskripsi tujuan wisata dari berbagai negara di dunia. Korpus tesebut dipilih karena artikel di dalamnya memuat informasi yang terkait dengan target informasi yang akan diekstraksi.

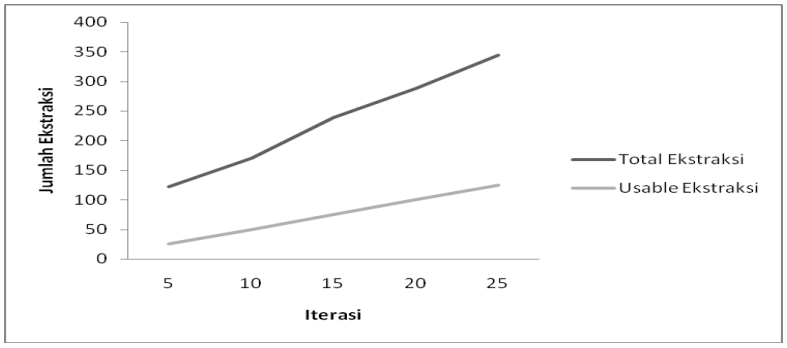
3. HASIL DAN PEMBAHASAN

Hasil dari fase pelatihan mengunakan algoritma *bootstrapping* telah dilakukan sebanyak 25 iterasi menggunakan data latih 1771 file teks (HTML) dan 5 *seed* (contoh *term*) awal menghasilkan kemunculan *term* (*occurrence*), *pattern*, dan *term* baru. Hasil yang diperoleh seperti terlihat pada Gambar 4 menunjukkan kecenderungan bahwa jumlah ekstraksi mengalami

peningkatan pada setiap penambahan iterasi. Sedangkan, pada Gambar 5 menunjukkan bahwa dari hasil *term* keseluruhan yang dapat diekstraksi hanya sebagian kecil *term* saja yang dapat digunakan.

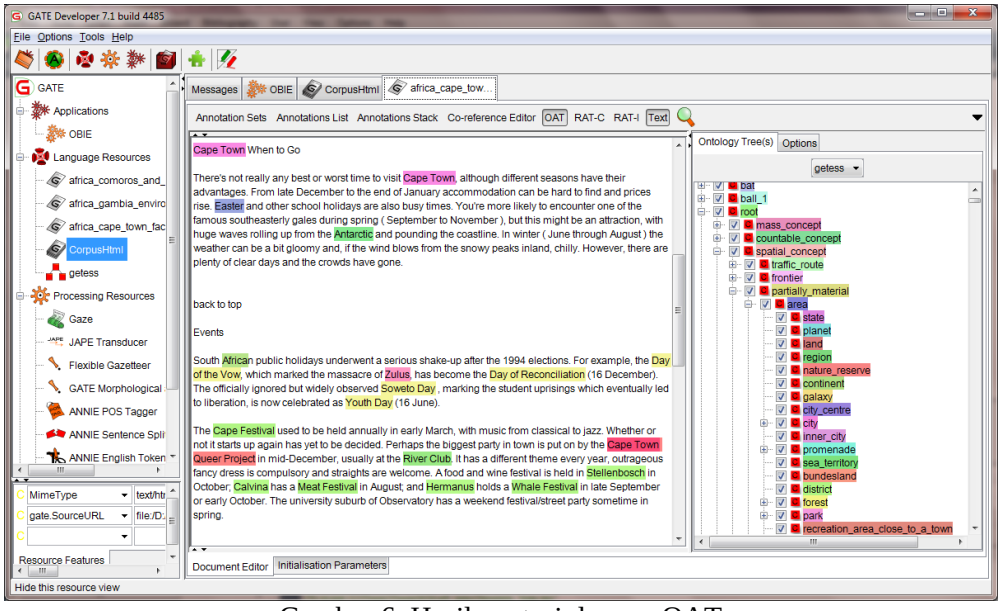


Gambar 4 Grafik pengaruh jumlah iterasi terhadap jumlah *pattern*



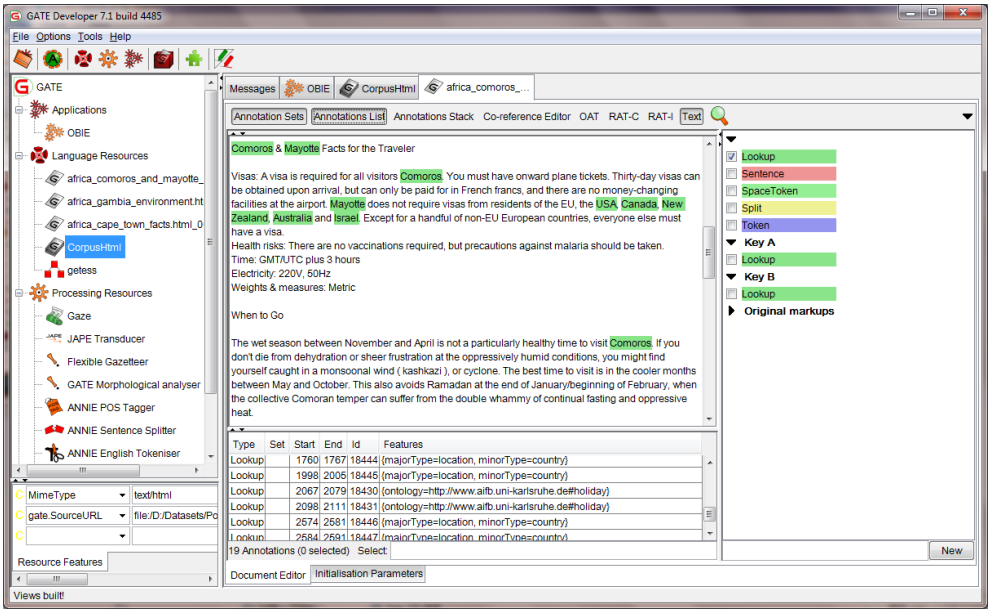
Gambar 5 Grafik pengaruh jumlah iterasi terhadap jumlah ekstraksi dan penggunaan ekstraksi

Hasil dari fase pengembangan sistem OBIE berupa teks yang telah dianotasi berdasarkan *classes*, *properties*, *instance* pada ontologi panduan. Hasil anotasi dengan OAT (*Ontology Annotation Tools*) GATE dapat dilihat pada Gambar 6.



Gambar 6 Hasil anotasi dengan OAT

Hasil proses populasi ontology dengan *Ontology Lookup* pada korpus ditunjukkan pada Gambar 7. Keterangan tipe anotasi dapat dilihat pada bagian tengah bawah pada jendela aplikasi.



Gambar 7 Hasil populasi ontologi

Hasil pengujian ekstraksi informasi untuk mendapatkan entitas baru dari teks (*Named Entity Recognition* / NER) menggunakan 30 *dataset* uji dilakukan dengan membandingkan dengan hasil anotasi *gold standard* dari *dataset* yang ada. Hasil rata-rata pengujian sistem OBIE dan perbandingan dengan hasil dari penelitian lain disajikan pada Tabel 1. Sedangkan perbandingan kecepatan proses sistem OBIE dengan sistem ekstraksi lain yang tidak menggunakan ontologi dapat dilihat pada Tabel 2.

Tabel 1 Perbandingan beberapa sistem ekstraksi informasi

Evaluasi Kinerja (%)	Penelitian [11]	Penelitian [10]	Penelitian ini (OBIE)
Precision	29,33	67,37	73
Recall	65,49	85,13	62
F-measure	40,52	75,21	67

Perbandingan dengan penelitian [10] memperlihatkan bahwa nilai *precision* pada penelitian ini lebih tinggi dari pada nilai *recall*, sebaliknya pada penelitian [10] nilai *recall*-nya lebih tinggi daripada nilai *precision*. Hasil kinerja secara keseluruhan dapat dilihat pada nilai *F-measure*. Hasil perbandingan menunjukkan bahwa sistem OBIE memiliki kecenderungan yang meningkat bila dibandingkan dengan penelitian [11], namun cenderung menurun bila dibandingkan dengan penelitian [10].

Tabel 2 Perbandingan kecepatan proses pada beberapa sistem ekstraksi informasi

Sistem Ekstraksi Informasi	Kecepatan rata-rata (dokumen/detik)
OBIE	0,61433
ANNIE	0,608845
LingPipe IE Systems	1,049218333

Kecepatan rata-rata sistem OBIE menunjukkan waktu pemrosesan yang tidak jauh berbeda bila dibandingkan dengan sistem ANNIE. Sedangkan bila dibandingkan dengan LingPipe IE Systems waktu prosesnya cenderung lebih cepat.

4. KESIMPULAN

Berdasarkan hasil uji coba yang telah dilakukan, dapat disimpulkan bahwa:

1. Hasil uji coba ekstraksi informasi ekstraksi informasi menggunakan dataset "LonelyPlanet" menunjukkan kinerja yang cenderung meningkat jika dibandingkan dengan penelitian [11]. Hasil evaluasi pengujian sistem didapatkan precision 73%, recall 62% dan F-measure 67%.
2. Hasil uji terhadap lama waktu yang diperlukan dalam pemrosesan 30 dokumen teks untuk sistem OBIE menunjukkan kecenderungan waktu proses yang hampir sama dengan sistem ANNIE. Sedangkan waktu proses sistem OBIE terhadap sistem LingPipe memiliki kecenderungan dapat memproses dokumen lebih cepat.
3. Proses iterasi pada algoritma *bootstrapping* untuk ekstraksi *pattern* menunjukkan bahwa jumlah *extraction pattern* yang didapatkan berbanding lurus dengan jumlah iterasi.
4. Jumlah *term* hasil ekstraksi yang dapat digunakan pada setiap iterasi *bootstrapping* menunjukkan hanya sebagian kecil saja yang dapat digunakan dari keseluruhan total *term* yang dapat diekstraksi.

5. SARAN

Sistem ekstraksi informasi pada penelitian ini masih perlu ditingkatkan lagi kinerja terkait nilai *precision* dan *recall*-nya, sehingga pemrosesan ekstraksi informasi menjadi lebih efektif. Penemuan teknik ekstraksi informasi baru yang dapat diintegrasikan dengan sistem OBIE dapat berperan untuk proses ini. Hal ini karena sistem ekstraksi informasi menggunakan panduan ontologi merupakan pendekatan ekstraksi informasi yang relevan dengan kondisi saat ini, dapat diintegrasikan dengan web semantik untuk menyediakan konten semantik untuk web semantik.

DAFTAR PUSTAKA

[1] Netcraft, 2013, August 2013 Web Server Survey, http://news.netcraft.com/archieves/2013/08/09/august-2013-web-server-survey.html, diakses tgl 1 Desember 2013.

[2] Piskorski, J., dan Yangarber, R., 2013, *Information Extraction: Past, Present and Future*, Poibeau, T. (ed): *Multi-source, Multilingual Information Extraction and Summarization*, Ch. 2, Springer-Verlag, Berlin.

[3] Appelt, D. E, 1999, Introduction to Information Extraction. *AI Communications*, No.3, Vol.12, 161-172.

[4] Labsky, M., 2008, Information Extraction from Websites using Extraction Ontologies, *Tesis*, University of Economic Prague.

[5] Maynard, D., Li, Y., dan Peters, W., 2007, *NLP Techniques for Term Extraction and Ontology Population*, Buitelaar, P. dan Cimiano, P. (ed): *Bridging the Gap between Text and Knowledge-Selected Contributions to Ontology Learning and Population from Text*, IOS Press.

[6] Wimalasuriya, D. C, and Dou, D. , 2009, Ontology-Based Information Extraction: An Introduction and a Survey of Current Approach, *Journal of Information Science*, vol XX, hal 1-20.

[7] Huang, R, and Riloff, E, 2012, Bootstrapped Training of Event Extraction Classifiers, *Proceeding of 13th Conference of the European Chapter of the Association for Computational Linguistics*.

[8] Cunningham, H., 2002, GATE, A General Architecture of Text Engineering, *Computers and the Humanities*, vol 36, hal 223-254.

[9] Moens, M., 2006, *Information Extraction: Algorithms and Prospects in a Retrieval*, Springer, Netherland.

[10] Rios-Alvarado, A. B, Lopez-Arevalo, I., and Sosa-Sosa, V. J., 2013, Learning Concept Hierarchies from Textual Resources for Ontologies Construction, *Expert Systems with Applications*, vol 40, hal 5907-5915.

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