IMPLEMENTASI INFORMATION EXTRACTION PADA DOMAIN BUKU DENGAN METODE SUPERVISED LEARNING OF EXTRACTION PATTERNS AND RULES DENGAN HTML TEXT PROCESSING

Ummi Hamidah[1], Ade Romadhony[2], Retno Novi Dayawati[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Seiring cepatnya pertambahan data pada internet, internet kini dimanfaatkan menjadi sumber data bagi berbagai keperluan. Automatic Cataloging (ACat) adalah sistem IE yang digunakan untuk otomasi proses pengkatalogkan buku dengan input data dari internet yang berupa halaman offline html.

Dengan menggunakan rule yang dibentuk dari learning corpus menggunakan natural language tools, informasi tentang buku dapat diambil dari suatu halaman html. Nilai precision dan recall dari penggunaan rule hasil learning dipengaruhi oleh nilai maksimum dan minimum slot filler length serta penghilangan uncoupled tag.

Motode yang digunakan adalah Supervised Learning of Extraction Patterns and Rules di mana learning corpus perlu dibuat sesuai dengan domain yang diharapkan, dalam kasus ini merupakan domain buku. Sistem IE dibuat bekerja sebagai tagger yang berfungsi memberikan tag pada informasi relevant yang akan diekstrak.

Kata Kunci : tagger, information extraction, supervised learning, rule, natural language, POS tagger

Abstract

As fast as the adding of information in internet, nowadays internet is used for data resource for many purposes. Automatic Cataloging (ACat) is IE system for automaton process in book cataloging using html offline page as the input data.

Using the rule that made from learning corpus using natural language tools, book information can be found from a html page. Precision and recall values in the tagging process using the rule is depend on the value of minimum and maximum slot filler length and the uncoupled tag removal.

Supervised Learning of Extraction Patterns and Rules is the method in used, where learning corpus is needed to be made based on the domain, in this case in book domain. IE system is made to be a tagger that for tagging the relevant information that will be extracted.

Keywords : tagger, information extraction, supervised learning, rule, natural language, POS tagger

1. Pendahuluan

1.1 Latar belakang

Ketika perpustakaan mendapatkan buku baru (belum ada dalam *database*) dan perlu dimasukkan dalam katalog, cara yang paling mudah untuk melakukannya adalah dengan memasukkan data secara manual. Jika pustakawan ingin mempercepat penyusunan katalog tanpa input secara manual, diperlukan sumber informasi selain buku itu sendiri. Banyak terdapat informasi mengenai buku yang tidak terstruktur, semi-struktur [14] dan terstruktur tersebar di Internet. Salah satu cara untuk mendapatkan informasi adalah dengan mempergunakan sistem *information extraction*. *Information extraction* (IE) adalah sebuah teknologi *Natural Language Processing* (NLP), yang berfungsi untuk memproses data tidak terstruktur, teks bahasa natural, untuk menemukan potongan informasi yang spesifik, atau fakta, di dalam teks, dan menggunakannya untuk mengisi *database* [15]. Sistem *Information extraction* (IE) umumnya berfokus pada domain atau topik yang spesifik, dan hanya mencari informasi yang relevan dengan minat pengguna [4-6][8-10] [12-14].

Dalam kasus mencari informasi untuk penyusunan katalog buku, pustakawan dapat mencari di Internet. Dengan mempergunakan ISBN buku, informasi yang diperlukan dapat diperoleh dari mesin pencari karena ISBN berisi informasi tentang grup, penerbit dan judul buku. Informasi lain yang bisa didapat adalah review buku yang berisikan informasi yang dapat dipergunakan untuk mengisi form pengisian untuk penyusunan katalog. Hasil pencarian akan digunakan sebagai sumber utama bagi sistem *Information extraction* (IE). *Supervised Learning of Extraction Patterns and Rules*, salah satu dari metode *learning-based* dalam IE akan diimplementasikan di sini. Tugas Akhir ini akan mempergunakan pendekatan yang sama dengan sistem (LP)2 (*Learning Pattern by Learning Processing*) dimana penandaan teks digunakan [4-5]. Input dokumen yang akan diproses adalah dokumen HTML, dan akan dilakukan *preprocessing* untuk memperoleh *main content*. *Main content* adalah bagian dokumen yang mengandung artikel utama. *Preprocessing* perlu dilakukan supaya input tidak mengandung informasi yang tidak terkait yang akan mempengaruhi terhadap hasil *learning*. *Preprocessing* akan dilakukan dengan cara analisis terhadap struktur *Document Object Model* (DOM) *tree*. Performansi sistem IE diukur dengan *recall* (persentase jawaban yang benar yang didapat oleh sistem), *precision rate* (persentase jawaban yang benar dari sistem) dan *F-score* [3-6][8-10][13-15]. Data yang telah terekstrak oleh sistem IE akan muncul di dalam form *interface*, beberapa *field*-nya dapat di-*edit* seperti edisi pencetakan buku. *Interface* muncul hanya untuk validasi dari keseluruhan sistem yang dilakukan oleh manusia.

Berdasarkan hasil percobaan yang dilakukan pada [5] dalam *the Message Understanding Conference* (MUC), (LP)2 (*Learning Pattern by Learning Processing*) memiliki performansi yang lebih baik dibandingkan algoritma lain yang menggunakan informasi NLP seperti Rapier dan Whisk (akurasi lebih dari 9% dan 20%). Diharapkan dengan mempergunakan sistem IE, informasi yang dibutuhkan dalam penyusunan katalog dapat terekstrak dengan baik dari hasil pencarian di Internet.

1.2 Perumusan masalah

Rumusan masalah yang muncul pada Tugas Akhir ini adalah:

1. Bagaimana melakukan preprocessing berdasar analisis DOM tree untuk input dokumen ?
2. Bagaimana implementasi metode *Supervised Learning of Extraction Patterns and Rules* untuk mengekstrak informasi pada domain buku?
3. Bagaimana performansi sistem, dihitung berdasarkan persentase *recall* dan *precision rate* pada metode *Supervised Learning of Extraction Patterns and Rules* untuk mengekstraksi informasi pada domain buku?
4. Faktor apa yang mempengaruhi *recall* dan *precision* pada metode *Supervised Learning of Extraction Patterns and Rules* untuk mengekstraksi informasi pada domain buku?

Batasan Masalah

Batasan masalah dari Tugas Akhir ini adalah:

1. buku memiliki kode ISBN yang valid (satu ISBN untuk satu judul buku, grup ISBN tidak tercangkup pada sistem),
2. nomer ISBN diperoleh dengan input manual maupun menggunakan *scanner* ISBN *barcode*,
3. lingkungan implementasi sistem memiliki koneksi Internet untuk mendapatkan input dari sistem *Information extraction* (IE),
4. sistem menggunakan mesin pencari yang sudah ada,
5. input dari sistem *Information extraction* (IE) berasal dari hasil pencarian dengan mesin pencari (halaman *offline* HTML), berasal dari situs gramedia.com

1.3 Tujuan

Tujuan dari Tugas Akhir ini adalah:

1. Mengimplementasikan metode *Supervised Learning of Extraction Patterns and Rules* untuk *Information extraction* (IE) pada domain buku.
2. Mengukur performansi sistem dengan menghitung nilai *recall* (persentase jawaban benar yang didapat sistem) dan *precision rate* (persentase jawaban sistem adalah benar).
3. Menganalisis faktor-faktor yang mempengaruhi *recall* dan *precision* pada metode *Supervised Learning of Extraction Patterns and Rules* pada domain buku.

Hipotesa

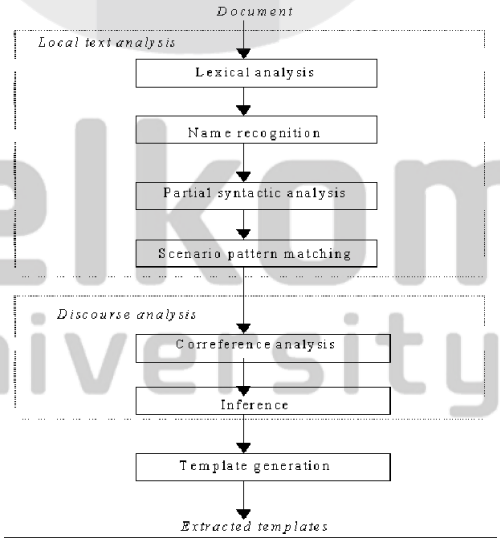
Mempergunakan algorithma $(LP)^2$ (*Learning Pattern by Learning Processing*), diharapkan hasil dari penelitian pada penyusunan katalog buku mempergunakan metode *Supervised Learning of Extraction Patterns and Rules* akan menghasilkan *recall* dan *precision* di atas 60%.

1.4 Metodologi penyelesaian masalah

Penyelesaian masalah dalam Tugas Akhir ini adalah:

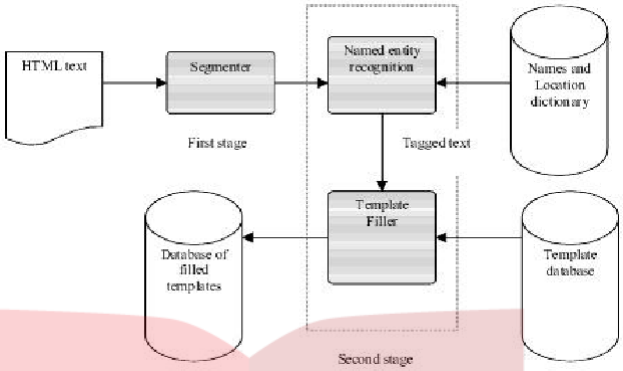
1. Identifikasi masalah
   Pada tahap ini, dilakukan identifikasi terhadap permasalahan yang akan diangkat serta pencarian solusinya.

2. Studi Literatur
   Pada tahap ini, dilakukan penacarian referensi dan sumber mengenai *Information extraction*. Terdapat empat metode *learning-based Information extraction* [9]:
   a. *Supervised Learning of Extraction Patterns and Rules*
      Dibandingkan menulis *patterns* dan *rules* secara manual, rekayasa pengetahuan dapat dikurangi menjadi pemberian penjelasan secara manual terhadap teks data *training*.
   b. *Supervised Learning of Sequential Classifier Models*
      Dibandingkan menggunakan pola eksplisit atau aturan untuk mengekstrak informasi, sebuah *classifier* mesin dilatih untuk memindai teks secara berurutan dari kiri ke kanan dan melabeli setiap kata sebagai ekstraksi atau non-ekstraksi.
   c. *Weakly Supervised and Unsupervised Approach*
      Teknik *supervised learning* secara substansial mengurangi upaya manual yang diperlukan dalam pembuatan sistem IE pada domain yang baru. Namun, pemberian penjelasan pada teks untuk *learning* masih membutuhkan waktu yang banyak dan pemberian penjelasan pada dokumen untuk *information extraction* bisa menjadi komplek. Karena *sistem* IE adalah domain spesifik, *corpora* yang telah diberi penjelasan tidak dapat digunakan pada domain yang baru.
   d. *Discourse-oriented Approaches to IE*
      Sistem IE memandang dengan lebih global dalam proses ekstraksi.
3. Pengumpulan data
   Pada tahap ini, dilakukan pengumpulan data dan pemberian penjelasan pada data tersebut untuk digunakan dengan tujuan pembelajaran dan validasi.
4. Disain Pembangunan Sistem
   Arsitektur sistem *Information extraction* digambarkan pada [2] ditunjukkan pada diagram di bawah.



Gambar 1-1: Arsitektur Sistem Information extraction

Diagram data alir proses *information extraction* digambarkan pada [2] ditunjukkan pada diagram di bawah.



Gambar 1-2: Diagram Data Alir Proses Information extraction

5. Implementasi Sistem
   Dalam kasus ini, ini akan digunakan pada Otomasi Penyusunan Katalog.
6. Skenario Pengujian
   Sistem diuji dengan buku (dicetak dalam bahasa Indonesia), yaitu sebanyak 10 buku (diambil acak tanpa mempertimbangkan jenis maupun kategori buku).
7. Analisis Hasil Implementasi
   Setelah sistem diimplementasi, hasil pengujian akan dianalisis *recall* (persentase jawaban benar yang didapat sistem) dan *precision rate* (persentase jawaban sistem adalah benar) dari performansi sistem dan faktor yang mempengaruhi nilai *recall* and *precision*.
8. Penyusunan Laporan dan Penyimpulan Hasil Analisis
   Mendokumentasikan penelitian serta hasil analisis agar dapat dimanfaatkan pada penelitian terkait selanjutnya.

1.5 Sistematika Penulisan

Sistematika penulisan laporan Tugas Akhir ini terdiri dari 5 bab, yaitu :

Bab I : Pendahuluan
Pendahuluan berisi latar belakang, rumusan masalah, batasan masalah, tujuan, hipotesa, metode penyelesaian masalah dan sistematika penulisan.

Bab II : Dasar Teori
Dasar teori berisi penjelasan singkat mengenai konsep-konsep yang mendukung dikembangkannya sistem ini.

Bab III : Analisis Perancangan dan Implementasi Sistem
Bab analisis perancangan sistem berisi rancangan sistem yang dibangun dan implementasi yang dilakukan.

Bab IV : Pengujian dan Analisis
Bab pengujian dan analisis berisi scenario pengujian dan hasil pengujian terhadap sistem yang sudah diimplementasikan serta analisis terhadap hasil pengujian.

Bab V : Kesimpulan dan Saran
Bab ini berisi kesimpulan yang didapat dari analisis yang dilakukan dan saran untuk penelitian serupa selanjutnya.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Dari hasil *learning* maupun pengujian yang dilakukan oleh sistem, dapat disimpulkan beberapa hal penting diantaranya:

1. Pemanfaatan nilai minimum dan maksimum *slot filler length* mempengaruhi hasil learning untuk menghasilkan *correction rule*.
2. Pemanfaatan nilai minimum dan maksimum *slot filler length* dalam proses *tagging* dapat meningkatkan nilai *recall* dan *precision*.
3. Nilai rata-rata *precision* dan *recall* pada *deployment testing* memenuhi hipotesis (melebihi nilai 60%) sehingga memiliki hasil yang baik.

5.2 Saran

Untuk mendapatkan hasil hasil yang lebih baik, disarankan untuk melakukan beberapa hal berikut:

1. Melakukan pengecekan serta proses *tagging* ulang sebelum dilakukan penghilangan *uncoupled* tag untuk mendapatkan nilai *recall* yang lebih tinggi.
2. Melakukan modifikasi pada algoritma *tagging* seperti dengan penambahan bobot tertentu terhadap nilai lexitem suatu rule (agar dapat dipergunakan terlebih dahulu maupun tidak), sehingga menjadikan implementasi menjadi lebih cepat dan akurat.
3. Mempergunakan NLP tool yang dapat memberikan nilai lebih terhadap properti lexitem dari rule maupun instance seperti *stammer* dan *entity recognation*, sehingga dihasilkan nilai generalisasi yang lebih beragam dan berbobot.

Referensi

[1] Alpaydin, Athem (2010). Introduction to Machine *learning*. The MIT Press, London.

[2] Bia, A., Munoz, R., (2002). *Information Exctraction to feed Digital Library Databases*. Universidad de Alicante, apartado de correos 99, E-03080, Espana. www.dlsi.ua.es/~abia/articles/sepln00.pdf

[3] Ciravegna, F., (2000). *Learning* to Tag for *Information extraction* from Text. In *Processing of the ECAI-2000 Workshop on Machine Learning for Information extraction*, F. Ciravegna, R. Basili, R. Gaizauskas (Eds.), Berlin.

[4] Ciravegna, F., (2001). Adaptive *Information extraction* from Text by *Rule* Induction and Generalisation. In *Preceedings of 17th Internation Joint Conference on Artificial Intelligence* (IJCAI). Seattle.

[5] Ciravegna, F., (2001). (LP)2, an Adaptive Algorithm for *Information extraction* for Web-related Texts. In *Proceedings of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining*.

[6] Ciravegna, F., (2003). *LearningPinocchio: Adaptive Information extraction for Real World Applications. Natural Language Engineering 1*. Cambridge University Press, United Kingdom.

[7] DOM Working Group (1998). *Document Object Model (DOM) Level 1 Specification*. World Wide Web Consortium. http://www.w3.org/TR/1998/REC-DOM-Level-1-19981001/DOM.pdf

[8] Feldman, R., Sanger, J., (2007). *The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, New York.

[9] Freitag, D (1998). *Information extraction* from HTML: Application of a general Machine *Learning* Approach. In *Proceeding of the Fifteen National Conference on Artificial Intelligence* (AAAI-98).

[10] Indurkhya, N., Fred J. Damerau(2010). *Handbook of Natural Language Processing*. Taylor and Francis.

[11] Mitchell, Tom M (1997). Machine *Learning*, McGRAW-HILL International Editions, Singapore.

[12] Nedellec, C (20xx). Machine *Learning* for *Information extraction*. *Literature review on the application of learning to research information*. http://caderige.imag.fr/Articles/Machine-*learning*-IE.pdf

[13] Riloff, E (1993). Automatically Constructing a Dictionary for *Information extraction* Tasks. In *Processing of the Eleventh National Conference on Artificial Intelligence*, pp. 811-816. AAAI Press / MIT Press.

[14] Soderland, S (1999). *Learning Information extraction Rules* for Semi-structured and Free Text. In *Machine Learning Journal*. Kluwer Academic Publisher, Boston.

[15] Yangarber, R (2001). *Scenario Customization for Information extraction*. Ph.D. Dissertation, New York University.

[16] Wicaksono, Alfan Farizki dan Purwarianti, Ayu (2010). HMM Based POS Tagger for Bahasa Indonesia. *On Proceedings of 4th International MALINDO (Malay - Indonesian Language) Workshop*.