

Pendekatan *Schema Matching* dalam Bahasa Indonesia

I Wayan Simri Wicaksana*, Reza A. Hakim[†]

*Universitas Gunadarma

[†]PT. Radiant Centra Nusa

E-mail: iwayan@staff.gunadarma.ac.id, reza.aulia@gmail.com

Schema matching adalah problem dasar dari aplikasi database, *semantic web*, *web services* pada e-business, query berbasis semantik dan sebagainya. Pada implementasi *matching* saat ini sebagian besar dilakukan secara manual. Disisi lain, beberapa riset telah merancang beberapa teknik untuk mendukung otomatisasi *schema matching* pada beberapa domain.

Beberapa pendekatan [4], [5] yang telah ada saat ini untuk melakukan *schema matching*, dengan pendekatan seperti pada *schema level*, *instance level*, *language based* dan sebagainya. Pada paper ini kami akan menguji *language based* untuk kasus dalam sumber yang berbahasa Indonesia.

Kami akan melakukan pendekatan terminologi *schema matching* untuk menguji kasus pada sumber yang berbahasa Indonesia. Alasan kami, bahwa sebagian besar pendekatan, terutama pada *language based* adalah berbasiskan bahasa Inggris. Proses pendekatan kami akan memanfaatkan *on-line* kamus yang dikombinasikan dengan WordNet dengan metode *path length* [6] sebagai input untuk *language based*. Hasil *matching* dari pendekatan kami akan dievaluasi dengan membandingkan terhadap hasil *matching* secara manual. Perbandingan ini akan memberikan informasi apakah pendekatan ini layak atau tidak untuk diterapkan.

Pendekatan kami dapat memberikan kontribusi untuk dikembangkan sebagai model untuk pemetaan pada interoperabilitas pada sumber informasi berbahasa Indonesia yang semakin dinamis, *semi structure* dan terbuka di Internet baik untuk aplikasi sebagai *semantic web* ataupun *web services*.

Kata Kunci : *interoperabilitas*, *mapping*, *schema matching*, *semantic similarity*, *semantic web*, *WordNet*.

1. Pendahuluan

Interoperabilitas informasi pada masa Internet tidak saja memberikan dampak positif, tetapi juga mengantar kepada berbagai masalah baru. Salah satu permasalahan baru adalah pada keragaman pada sintatik, skematik dan semantik. Permasalahan ini juga terjadi pada berbagai sumber informasi berbahasa Indonesia.

Penelitian hubungan semantik telah merupakan bagian dari intelegensi buatan dan psikologi untuk dekade terakhir ini. Sebagian besar relasi semantik adalah berhubungan dengan *natural language processing* (NLP) yang mengacu ke pada *thesaurus* Roget. Beberapa saat kemudian WordNet menjadi hal yang menarik perhatian para peneliti untuk melihat kesamaan semantik. Hal ini disebabkan adalah

WordNet tersedia secara free, cukup besar dan dirancang untuk digunakan pada komputer.

Sebagai sifat dasar dari manusia akan selalu mencoba mencari relasi antara dua konsep. Seperti contoh, semua hampir setuju kalau dikatakan *automotive* memiliki makna *car* dan *tire* memiliki hubungan dengan *car*, tetapi *tree* tidak. Tetapi untuk memberikan sebuah nilai kuantitatif untuk menunjukkan tingkat relasi dua konsep adalah sangat sulit. Sebagai ilustrasi lain, dua konsep dapat memiliki relasi, karena sebuah konsep lebih umum dari konsep lainnya (contoh *car* adalah jenis dari *vehicle*) atau juga karena merupakan bagian dari yang lain (contoh *tire* adalah bagian dari *car*).

Jika melihat hal di atas maka semua pendekatan adalah dimodelkan untuk bahasa Inggris. Pada paper ini kami akan menguji

penggunaan dalam bahasa Indonesia dengan menggunakan kamus on-line. Alasannya adalah pertama, dapat mengetahui apakah cukup memadai dengan kamus atau perlu membuat WordNet versi Indonesia. Sisi lainnya dengan kamus on-line, bisa dikembangkan untuk pemanfaatan pada web services pada pencarian dan pertukaran informasi dan service.

Paper ini akan memiliki 4 bagian, bagian pertama adalah pendahuluan. Bagian ke-dua merupakan uraian singkat tentang WordNet. Evaluasi dari pendekatan ini akan dibahas pada bagian 3. Yang terakhir bagian 4 merupakan kesimpulan.

2. WordNet

WordNet adalah sebuah database network semantik untuk bahasa Inggris yang dikembangkan di Princeton University (<http://wordnet.princeton.edu/>). Beberapa versi dalam bahasa lain juga telah dikembangkan seperti EuroNet. WordNet merupakan database lexical yang memiliki arti unik dari sebuah kata yang dipresentasikan dalam *synonym set* atau *synset*. Setiap *synset* memiliki sebuah glosari yang mendefinisikan konsep yang direpresentasikannya. Sebagai contoh kata *car*, *auto*, *automobile* dan *motorcar* memiliki satu *synset* dengan glosari sebagai berikut sebuah kendaraan beroda empat yang dikerakkan oleh mesin.

Synset dihubungkan dengan berbagai bentuk relasi seperti *hypernym* (adalah jenis dari), *meronymy* (adalah bagian dari), *antonymy* (adalah lawan dari) dan sebagainya.

Jika sebuah kata benda A dihubungkan dengan kata benda B dengan 'jenis dari', maka B adalah *hypernym* dari A atau A adalah *hyponym* dari B. Sebagai contoh *car* adalah *hypernym hatchback*, atau *hatchback* adalah *hyponym* dari *car*.

Metode kesamaan semantik perhitungan pada WordNet dibagi dalam dua kelompok besar pendekatan [6], yaitu *path length* dan *information content*. *Path length* secara sederhana menghitung jumlah node atau relasi yang menghubungkan antar node dalam taksonomi. Jarak yang lebih pendek antara dua konsep, berarti memiliki kesamaan lebih tinggi. *Path length* memberikan keuntungan dengan tidak bergantung pada statistik *corpus* dan tidak terpengaruh dengan penyebaran kata. Tetapi

memiliki kelemahan dalam taksonomi yang memiliki jarak yang *uniform*/sama. Beberapa contoh pendekatan dengan *path length* adalah Leacock-Chodorow, Resnik, Wu-Palmer. Pada paper ini, pendekatan Wu Palmer adalah model perhitungan yang digunakan untuk menguji, persamaanya seperti pada formula 1

Wu-Palmer:

$$sim_{WUP} = \max \left[\frac{2 \times depth(LCS(a, b))}{length(a, b) + 2 \times depth(LCS(a, b))} \right] \quad (1)$$

dimana $length(a, b)$ adalah jumlah panjang path antara a dan b; $depth(LCS(a, b))$ adalah jumlah panjang path dari konsep umum dari a dan b ke root.

Information content sebuah node adalah $-\log$ dari jumlah semua kemungkinan (dihitung berdasarkan frekuensi *corpus*) dari semua akata yang memiliki *synset*. Dengan kata lain jika $p(x)$ adalah probability dari sebuah instace dari x , maka *information content* dari x adalah $-\log p(x)$.

3. Evaluasi

Tujuan dari percobaan adalah untuk menguji untk semantic matching dengan menggunakan WordNet tanpa perlu membuat WordNet versi bahasa Indonesia tapi dapat menggunakan kamus. Langkah yang akan dilakukan adalah:

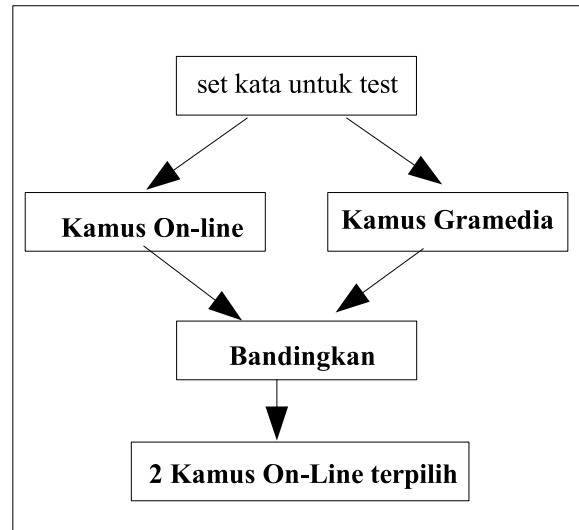
- Memilih kamus on-line yang akan digunakan
- Hasil terjemahan ke bahasa Inggris dari kamus on-line akan merupakan input untuk WordNet. Pengujian akan berdasarkan perbandingan terhadap evaluasi dari ekspert dengan melihat faktor Recall, Precision dan F-measure.

3.1. Persiapan Evaluasi

Beberapa hal yang dipersiapkan untuk pengujian adalah :

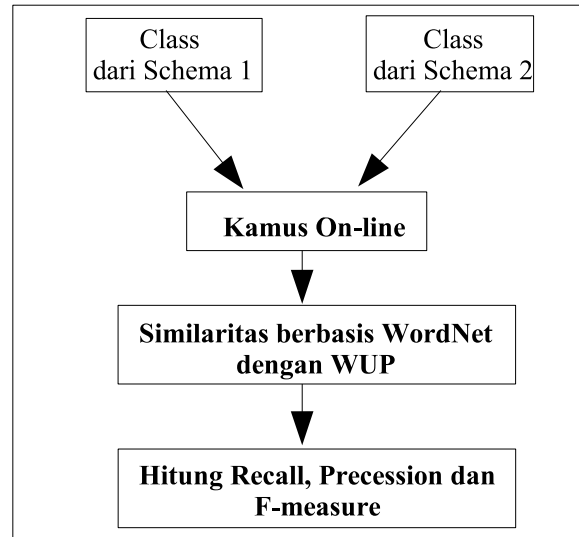
- Memilih kamus on-line yang akan digunakan. Pemilihan kamus diawalnya akan mengacu kepada hasil riset yang telah dilakukan. Ternyata tidak ditemukan referensi yang menguji kamus on-line Indonesia-Inggris untuk digunakan sebagai acuan. Untuk itu akan dilakukan langkah sebagai berikut dalam pemilihan kamus on-line seperti pada gambar 1 :

- Mencari ketersediaan kamus on-line dari mesin pencari google.
- Uji dari halaman pertama hasil searching. Pengujian akan dilakukan dengan memasukkan set kata sederhana berjumlah 30 kata dan hasil translasi kamus on-line akan dibandingkan dengan kamus Indonesia ke Inggris oleh John M Echols dan Hassan Shadily dari penerbit Gramedia. Dari hasil uji ini kami akan memilih dua kamus on-line untuk digunakan.



Gambar. 1. Proses Pemilihan Kamus On-line

- Mencari atau mengembangkan tool untuk menghitung similaritas berdasarkan pendekatan di atas. Pada percobaan ini digunakan tool on-line yang telah tersedia, untuk WordNet digunakan <http://marimba.d.umn.edu/cgi-bin/similarity.cgi>.
- Perhitungan evaluasi similariti untuk bahasa Indonesia akan mengikuti proses seperti gambar 2 sebagai berikut:



Gambar. 2. Proses Evaluasi Schema Matching Bahasa Indonesia

- Menhitung semua kombinasi konsep antara dua sumber dalam satu domain berdasarkan Wu-Palmer.
- Memfilter hasil perhitungan dengan memberikan nilai threshold, tujuannya adalah untuk mempermudah dalam analisis. Penentuan nilai threshold dilakukan dengan cara try-error untuk mendapatkan nilai optimal, dimulai dengan initial nilai dari 0.7 ke 1,0. Ini disebut dengan tabel hasil perhitungan (ζ).
- Membuat tabel perhitungan dari expert atau disebut tabel referensi, ini disebut β
- Membandingkan hasil perhitungan terhadap referensi ini adalah Δ
- Kemudian menghitung

$$\text{nilai Recall } (Recall = (\Delta/\beta)), \quad \text{Precession } (Precession = (\Delta/\zeta)) \text{ dan } F\text{-Measure } (Fmeasure = 2/((1/Recall) + (1/Precession))).$$

- Hasil perhitungan akan ditampilkan pada tabel untuk dievaluasi apakah pendekatan ini memadai.

3.2. Hasil dan Diskusi

Hasil evaluasi adalah sebagai berikut

- Terjemahan ke bahasa Inggris dari dua domain, tabel 1, dan 2. Kamus on-line yang digunakan adalah

(i) <http://www.kamus-online.com/index.php?lang=en> dan (ii) <http://www.seasite.niu.edu/Indonesian/TataBahasa/dictionary/Default.htm>. Untuk domain mobil mengambil dari <http://www.mobilbabe.com/> dan <http://www.autocybercenter.com/>. Untuk domain komputer mengambil dari <http://www.bhinneka.com/> dan <http://www.belikomputer.com/>.

- Perhitungan dengan menggunakan WordNet dengan formula Wu-Palmer. Pengolahan WordNet hanya menggunakan output dari kamus on-line 1, dikarenakan output kamus on-line 2 kurang memadai. Hasil perhitungan WordNet dapat dilihat pada tabel 3 dan 4.
- Hasil dari tabel WordNet dibandingkan dengan hasil ekspert untuk menghitung Recall, Precision dan F-measure, dapat dilihat pada tabel 5.

Tabel 1

TERJEMAHAN KAMUS ON-LINE DOMAIN MOBIL

| Sumber | Class | Kamus 1 | Kamus 2 |
|------------------|-----------|-------------|-----------|
| mobil-babe | tahun | year | year |
| | merk | tademark | NA |
| | deskripsi | description | NA |
| | harga | price | price |
| | kilometer | kilometer | kilometer |
| autocyber-center | kota | city | city |
| | foto | photo | picture |
| | merk | tademark | NA |
| | tahun | year | year |
| autocyber-center | warna | color | color |
| | lokasi | location | NA |
| | harga | price | price |

Tabel 2

TERJEMAHAN KAMUS ON-LINE DOMAIN KOMPUTER

| Sumber | Class | Kamus 1 | Kamus 2 |
|---------------|-----------|-------------|---------|
| bhinneka | deskripsi | description | NA |
| | mata uang | currency | NA |
| | harga | price | price |
| beli-komputer | produk | product | product |
| | harga | price | price |

Dari tabel-tabel di atas, maka didapatkan hasil sebagai berikut :

Tabel 3

SIMILARITAS WORDNET-WUP DI DOMAIN MOBIL

| | trd | yr | clr | loc | pr |
|--------------------|------|------|------|------|------|
| trademark | 1.00 | 0.50 | 0.75 | 0.47 | 0.75 |
| description | 0.59 | 0.50 | 0.92 | 0.78 | 0.71 |
| price | 0.75 | 0.50 | 0.75 | 0.47 | 1.00 |
| kilometer | 0.47 | 0.59 | 0.50 | 0.44 | 0.47 |
| city | 0.47 | 0.82 | 0.63 | 0.71 | 0.47 |
| photo | 0.22 | 0.22 | 0.38 | 0.53 | 0.35 |
| trd=trademark | | | | | |
| yr=year | | | | | |
| clr=color | | | | | |
| loc=location | | | | | |
| pr=price | | | | | |

Tabel 4

SIMILARITAS WORDNET-WUP DI DOMAIN KOMPUTER

| | product | price |
|--------------------|---------|-------|
| description | 0.78 | 0.71 |
| currency | 0.47 | 0.71 |
| price | 0.50 | 1.00 |

- Kamus on-line belum memiliki kelengkapan kata dan penjelasan tambahan yang memadai. Dari 10 kamus on-line yang kami ambil dari mesin pencari google, ternyata hanya dua yang terpilih cukup memadai. Dari dua itupun, akhirnya pada evaluasi kami hanya kamus pertama yang dapat kami gunakan.
- Melihat hasil akhir analisis pada nilai Recall, Precision dan F-measure dari dua domain mobil dan komputer yang kemudian dibandingkan eksperimen dari [1] pada bahasa Inggris. Hasil schema matching pada terminologi level/label matching dengan menggunakan kamus on-line dan WordNet v2.1 (berbahasa Inggris) adalah memiliki nilai di atas nilai minimum dari referensi. Dengan kata lain pendekatan ini dapat digunakan untuk schema matching dan aplikasi terkait lainnya.

4. Penutup

Pada paper ini telah dilakukan evaluasi schema matching dalam bahasa Indonesia dengan memanfaatkan kamus on-line dan

Tabel 5
RECALL, PRECESSION DAN F-MEASURE

| | |
|--------------------------|------------|
| Domain Mobil | |
| threshold | 0.91 |
| Calculation | 3 |
| Expert | 4 |
| Match Calculation-Expert | 3 |
| | |
| Recall | 75% |
| Precession | 100% |
| F-Measure | 86% |
| Domain Komputer | |
| threshold | 0.91 |
| Calculation | 1 |
| Expert | 2 |
| Match Calculation-Expert | 1 |
| | |
| Recall | 50% |
| Precession | 100% |
| F-Measure | 67% |

WordNet dalam bahasa Inggris. Hasil yang didapat dari evaluasi adalah bahwa pendekatan ini memadai, dikarenakan mendapatkan hasil Recall, Precession dan F-measure di atas nilai minimum dari referensi.

Kontribusi dari paper ini adalah hasil evaluasi dapat dimanfaatkan untuk pengembangan aplikasi semantic web, perawatan ontology, pemetaan konsep dalam menghadapi keragaman semantik pada sumber informasi berbahasa Indonesia.

Rencana kedepan kami akan mengujikan dengan membuat WordNet mini versi Indonesia untuk dibandingkan dengan pendekatan pada paper yang kami buat ini. Untuk mengetahui pendekatan mana yang lebih baik untuk perhitungan kesamaan semantik, walaupun secara hipotesa teoritis WordNet versi Indonesia akan memiliki hasil lebih baik dibandingkan dengan menggunakan kamus dan WordNet versi bahasa Inggris.

DAFTAR PUSTAKA

- [1] Satanjeev Banerjee and Ted Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In Georg Gottlob and Toby Walsh, editors, *IJCAI*, pages 805–810. Morgan Kaufmann, 2003.
- [2] Thomas Hofmann. Probabilistic latent semantic analysis. In Kathryn B. Laskey and Henri Prade, editors, *UAI*, pages 289–296. Morgan Kaufmann, 1999.

- [3] Dekang Lin. An information-theoretic definition of similarity. In Jude W. Shavlik, editor, *ICML*, pages 296–304. Morgan Kaufmann, 1998.
- [4] Erhard Rahm and Philip A. Bernstein. A Survey of Approaches to Automatic Schema Matching. *The VLDB Journal*, 10:334–250, 2001.
- [5] Pavel Shvaiko and Jerome Euzenat. A Survey of Schema-based Matching Approaches. *J. Data Semantics IV*, pages 146–171, 2005.
- [6] Martin Warin. Using WordNet and Semantic Similarity to Disambiguate an Ontology. http://ling16.ling.su.se:8080/PubDB/doc_repository/warin2004usingwordnet.pdf, 2004.