Pembelajaran Relasi Non-Taksonomi Dari Dokumen Web

Elang Cergas Pembrani
Teknik Informatika
Universitas Islam Indonesia
Yogyakarta, Indonesia
14523290@students.uii.ac.id

Abstract*—Ontologi memiliki peran yang sangat penting dalam sistem ilmu pengetahuan modern karena ontologi memberikan standar formal dalam penyusunan struktur basis pengetahuan serta dalam penggunaan kembali (reuse) dan berbagi pakainya (sharing). Pembangunan ontologi skala besar dimana terdapat sangat banyak data dan informasi, penentuan nama relasi yang ideal dilakukan secara otomatis tanpa keterlibatan manusia. Penentuan nama relasi secara otomatis inilah yang menjadi tantangan dalam pembangunan ontologi secara otomatis. Salah satu alternatif solusi untuk memecahkan masalah ini adalah dengan memanfaatkan data dan informasi berlimpah dari Web untuk menemukan kemungkinan nama relasi yang sesuai. Dengan memanfaatkan pengolahan bahasa alami, komputer dapat diprogram untuk mencari kemungkinan nama relasi yang sesuai dari kelas-kelas yang diberikan. Dalam penelitian ini, dibangun rancangan alur kerja dan program untuk melakukan pembelajaran relasi non-taksonomi (*Learning Non-Taxonomic Relationships of Ontologies* atau LNTRO) secara otomatis dengan bersumber data dari dokumen Web menggunakan mesin pencari. Untuk mencapai hal tersebut, penelitian dilakukan dimulai dengan studi pustaka untuk meneliti penyelesaian masalah LNTRO terdahulu, kemudian merancang alur kerja pembelajaran relasi non-taksonomi, dilanjutkan dengan implementasi program, dan terakhir pengujian alur kerja dan program yang dibangun. Ada dua metode ekstraksi kata yang diusulkan, yaitu metode MAANS (*Maximum of Absolute Average of Normalized Score*) dan metode MFLR (*Most Frequent Local Recommendation*). Proses penelitian tersebut menghasilkan suatu program yang dapat digunakan untuk melakukan pembelajaran relasi non-taksonomi secara otomatis. Pengujian dilakukan dengan dua data uji yang berbeda dengan empat parameter pembelajaran yang berbeda. Dari hasil pembelajaran dan pengujian, nama relasi non-taksonomi yang dihasilkan sudah sesuai dan relevan dengan penalaran peneliti. Hasil evaluasi kata menunjukkan nama relasi non-taksonomi juga cukup relevan.

Keywords— ontologi, relasi non-taksonomi, LNTRO, web, mesin pencari

I. PENDAHULUAN

Ontologi adalah spesifikasi eksplisit dan formal dari suatu konseptualisasi bidang keilmuan tertentu yang mendefinisikan kumpulan konsep dan relasinya untuk merepresentasikan pengetahuan pada bidang yang tersebut [1]. Ontologi memiliki peran yang sangat penting dalam sistem ilmu pengetahuan modern karena ontologi memberikan standar formal dalam penyusunan struktur basis pengetahuan serta dalam penggunaan kembali (reuse) dan berbagi pakainya (sharing) [2]. Karena peran ontologi yang sangat penting inilah maka banyak usaha yang dilakukan untuk dapat membangun ontologi berbantuan komputer, baik menggunakan metode pembelajaran otomatis yang tidak melibatkan peran manusia maupun metode pembelajaran semi-otomatis yang melibatkan peran manusia.

Untuk pembangunan ontologi skala besar dimana terdapat sangat banyak data dan informasi, penentuan nama relasi yang ideal dilakukan secara otomatis tanpa keterlibatan manusia. Penentuan nama relasi secara otomatis inilah yang menjadi tantangan dalam pembangunan ontologi secara otomatis. Penentuan nama relasi secara otomatis sulit dilakukan karena relasi menentukan konteks dari kelas-kelas atau konsep-konsep yang saling dihubungkan. Hal ini berlaku terutama untuk relasi yang bersifat non-taksonomis, karena dalam ilmu pengetahuan lebih banyak konsep yang terhubung secara non-taksonomis, dimana hubungan yang ada bukanlah hubungan dari umum ke khusus atau sebaliknya.

Salah satu alternatif solusi untuk memecahkan masalah ini adalah dengan memanfaatkan data dan informasi berlimpah dari Web untuk menemukan kemungkinan nama relasi yang sesuai. Dengan memanfaatkan pengolahan bahasa alami, komputer dapat diprogram untuk mencari kemungkinan nama relasi yang sesuai dari kelas-kelas yang diberikan. Namun bahasa yang digunakan sangat mempengaruhi masukan dan keluaran dari program. Meskipun program komputer yang dapat melakukan hal ini sudah banyak dikembangkan [3], namun program-program tersebut tidak dikembangkan untuk memproses data yang menggunakan Bahasa Indonesia. Selain itu, proses pengolahan bahasa alami yang dijalankan juga sangat dipengaruhi oleh bahasa yang diproses, sehingga bahasa yang berbeda pun perlu pendekatan pengolahan yang berbeda pula. Oleh karena itulah penelitian ini dilakukan, yaitu untuk menentukan nama relasi yang mungkin dari data kelas ontologi yang menggunakan Bahasa Indonesia secara otomatis. Kemungkinan nama relasi didapatkan dari dokumen Web yang jumlahnya melimpah, yang kemudian diproses secara otomatis

untuk menentukan satu nama relasi yang dapat menghubungkan kelas-kelas yang diberikan.

II. STUDI LITERATUR

Pembelajaran relasi non-taksonomi dalam bidang ontologi bukan hal baru. Sudah terdapat beberapa penelitian yang membahas pendekatan-pendekatan tertentu untuk dapat menentukan nama relasi non-taksonomi dari suatu pasangan data kelas ontologi. Pendekatan pembelajaran menggunakan dokumen web juga bukan hal baru, karena dokumen web merupakan sumber data dengan jumlah yang sangat besar.

Penelitian yang dilakukan oleh [4] membahas empat teknik state-of-the-art LNTRO. Solusi yang ditawarkan oleh keempat teknik tersebut dibahas dalam penelitian tersebut beserta kelebihan, kekurangan, dan batasan penggunaan teknik tersebut. Uraian ringkas mengenai pembahasan keempat teknik/metode LNTRO oleh [4] dijabarkan dalam bagian-bagian berikut.

A. LNTRO based on the Extraction of Association Rules

LNTRO based on the Extraction of Association Rules (LNTRO berdasarkan Ekstraksi Aturan Asosiasi) diambil dari penelitian yang dilakukan oleh [5]. Metode ini memiliki dua fase, yaitu "Identifikasi kemunculan relasi" dan "Mining asosiasi". Fase "Identifikasi kemunculan relasi" menerima korpus dan himpunan konsep/kelas ontologi, yang kemudian menghasilkan suatu tuple dalam bentuk < c1, v, c2 > dimana c1 dan c2 adalah kelas ontologi dan v adalah kata kerja. POS Tagging dilakukan untuk mengenali kata kerja. Hanya kalimat yang memenuhi kondisi berikut yang menghasilkan tuple: (1) kalimat yang memiliki tepat dua konsep/kelas dan satu kata kerja diantaranya, dan (2) kedua konsep tersebut berjarak maksimal N kata dimana N adalah parameter yang ditentukan oleh ahli. Setelah himpunan tuple terbentuk, selanjutnya fase "Mining asosiasi" dilakukan dengan menerapkan algoritma yang diberikan untuk memperbaiki hasil pada fase sebelumnya.

B. LNTRO based on the Extraction of Generalized Association Rules

LNTRO based on the Extraction of Generalized Association Rules (LNTRO berdasarkan Ekstraksi Aturan Asosiasi yang Digeneralisasi) diambil dari penelitian yang dilakukan oleh [6]. Metode ini hampir sama dengan metode LNTRO based on the Extraction of Association Rules yang diusulkan oleh [5], hanya saja perbedaannya terletak pada penggunaan algoritma generalized association rules untuk mengajukan kemungkinan tingkat hirarki relasi yang paling tepat dan dapat digunakan pada teks berbahasa Jerman. Metode ini memiliki dua fase, yaitu "Pemrosesan teks" dan "Mining asosiasi". Fase "Pemrosesan teks" bertujuan untuk mengekstrak pasangan konsep yang berhubungan dengan kandidat relasi. Untuk mencapai hal ini, digunakan teknik title heuristic dan sentence heuristic. Teknik pertama mendefinisikan bahwa sepasang konsep yang berhubungan perlu dibuat untuk setiap konsep dalam teks dengan setiap konsep pada judul teks. Teknik kedua menghasilkan suatu tuple untuk setiap pasangan konsep yang muncul dalam kalimat yang sama. Pada fase "Mining asosiasi", relasi yang berbentuk pasangan konsep dari fase sebelumnya diolah menggunakan algoritma Mining Generalized Association Rules untuk mengekstrak relasi non-taksonomi dalam bentuk aturan asosiasi.

C. LNTRO based on Queries on Web Search Engines

LNTRO based on Queries on Web Search Engines (LNTRO berdasarkan query pada mesin pencari Web) diambil dari penelitian yang dilakukan oleh [3]. Dalam penelitian tersebut, pembelajaran relasi non-taksonomi mengandalkan identifikasi kata kerja (verb) sebagai kandidat nama relasi yang akan dipilih. Selain itu, karena data yang digunakan adalah data berbahasa Inggris, maka kata kerja yang akan dipilih juga memiliki batasan-batasan tertentu, yaitu:

1) Hanya memilih kata kerja yang masuk dalam kategori kalimat present tense saja.

2) Frasa kata kerja yang melibatkan modifier berupa kata keterangan (adverb) atau lainnya juga tidak dipilih.

3) Frasa kata kerja yang terdiri dari beberapa kata kerja tidak dipilih karena dianggap sulit untuk mengenali kata kerja utama, dengan pengecualian "to be" yang sangat umum dalam bahasa inggris.

4) Kata kerja yang merepresentasikan relasi taksonomi (is/are, include, dan lain-lain) juga tidak dipilih, karena relasi taksonomi dianggap sebagai kasus yang berbeda.

Penelitian tersebut juga mengklasifikasikan kata kerja yang terpilih berdasarkan posisi katanya menjadi dua, yaitu kata kerja aktif dan kata kerja pasif. Kemudian dilakukan penyaringan lebih lanjut dengan hanya memilih kata kerja yang sesuai dan digunakan spesifik dalam domain ontologi yang diberikan. Hal ini dilakukan dengan menggunakan analisis statistik secara otomatis untuk menentukan tingkat relevansi antara domain dengan kata kerja yang dipilih. Penelitian tersebut juga memanfaatkan wordnet untuk menganalisis kata kerja yang dipilih lebih jauh, yang tidak diterapkan dalam penelitian ini karena kurangnya ketersediaan wordnet bahasa Indonesia.

D. LNTRO based on logistic regression

LNTRO based on logistic regression (LNTRO berdasarkan regresi logistik) diambil dari penelitian yang dilakukan oleh [7]. Metode ini tidak terikat oleh domain manapun dan mengekstrak relasi non-taksonomi dari korpus berbahasa Inggris. Metode ini menggunakan pembatas sintaksis dan pembatas leksikal. Pembatas sintaksis mengharuskan setiap frasa kata kerja untuk memenuhi pola berikut: (1) merupakan kata kerja, (2) merupakan kata kerja yang langsung diikuti oleh kata depan/preposisi, atau (3) merupakan kata kerja yang diikuti oleh kata benda, kata sifat, atau kata keterangan yang diakhiri dengan kata depan/preposisi. Pembatas sintaksis digunakan untuk mengurangi hasil ekstraksi yang tidak informatif atau tidak relevan. Namun karena pembatas sintaksis memungkinkan terpilihnya frasa yang terlalu spesifik, maka pembatas leksikal digunakan untuk membedakan antara kalimat yang memiliki relasi nyata dengan kalimat yang frasa relasinya terlalu spesifik. Metode ini memiliki tiga fase, yaitu "ekstraksi relasi", "ekstraksi argumen", dan "perbaikan/refinement". Fase "ekstraksi relasi" bertujuan untuk mengekstrak relasi dengan melibatkan pembatas sintaksis dan leksikal. Fase "ekstraksi argumen" bertujuan untuk memilih relasi argumen-argumen yang berbeda (sehingga relasi dianggap cukup umum digunakan). Kemudian

fase “perbaikan/refinement” menggunakan logistic regression classifier dan bertujuan untuk mengurutkan relasi berdasarkan probabilitas untuk merepresentasikan relasi valid.

III. RANCANGAN ALUR KERJA

Secara garis besar, tahapan alur kerja pembelajaran relasi non-taksonomi yang dirancang ada empat tahapan utama dan satu tahapan tambahan. Empat tahapan utama meliputi: pengembangan korpus (*corpus construction*), anotasi korpus (*corpus annotation*), ekstraksi kata (*term extraction*), dan pengembangan kata (*term refinement*). Sedangkan tahapan tambahannya adalah evaluasi kata (*term evaluation*). Tahap pengembangan korpus bertujuan untuk membangun suatu korpus yang akan dijadikan sebagai sumber data utama dalam menentukan kandidat nama relasi non-taksonomi. Tahap anotasi korpus bertujuan untuk mengolah dan menyaring korpus sehingga didapatkan kandidat nama relasi non-taksonomi. Tahap ekstraksi korpus bertujuan untuk melakukan penilaian (*scoring*) pada setiap kandidat nama relasi non-taksonomi untuk menentukan kandidat mana yang paling mungkin dipilih sebagai nama relasi non-taksonomi. Tahap pengembangan kata bertujuan untuk mengolah nama relasi non-taksonomi yang sudah dipilih sehingga lebih informatif (dalam kasus penelitian ini, nama relasi dicari bentuk kata kerja aktif dan pasifnya). Yang terakhir, tahap tambahan evaluasi kata bertujuan untuk membantu peneliti melakukan pengujian alur kerja yang diusulkan.

A. Pembangunan Korpus (Corpus Construction)

Tahap pertama pembelajaran relasi non-taksonomi dimulai dari pembangunan korpus. Secara garis besar, pembangunan korpus dilakukan dengan cara melakukan pencarian untuk setiap pasangan data melalui mesin pencari internet (search engine). Setiap hasil pencarian yang ditemukan akan mengandung judul dokumen dan ringkasan isi dokumen (snippet) yang diasumsikan mengandung kalimat yang relevan dengan pasangan data yang dicari. Oleh karena itu, sejumlah hasil pencarian tersebut disimpan dalam koleksi korpus sebagai referensi untuk mencari nama relasi non-taksonomi. Hasil pencarian dari setiap pasangan data membentuk korpus masing-masing untuk kemudian diproses pada tahap selanjutnya. Hasil dari tahap pembangunan korpus adalah himpunan yang berisi korpus untuk setiap pasangan data, dimana setiap korpus berisi kumpulan hasil pencarian dalam bentuk kalimat teks.

1) Mesin Pencari Web (Web Search Engine)

Mesin pencari adalah suatu layanan internet yang digunakan untuk mencari dokumen di internet. Dalam penelitian ini, mesin pencari digunakan sebagai alat utama untuk membangun korpus. Cara kerja mesin pencari cukup sederhana, yaitu memasukkan query/pertanyaan berupa kata kunci (keyword) yang ingin ditemukan. Mesin pencari akan memberikan hasil pencarian berupa daftar dokumen, yang terdiri dari judul, pranala (link) atau alamat dokumen, dan ringkasan isi dokumen yang dianggap relevan dengan kata kunci yang diberikan.

Hasil pencarian dari mesin pencari diolah menjadi bentuk korpus yang berisi kumpulan kalimat. Dalam penelitian ini, bagian dari hasil pencarian yang digunakan sebagai sumber korpus adalah judul dokumen dan ringkasan dokumen yang ada dalam setiap butir hasil pencarian.

Mesin pencari yang digunakan dalam penelitian ini yaitu:

a) Google (https://www.google.com)

b) Bing (https://www.bing.com)

c) DuckDuckGo (https://duckduckgo.com)

d) Yahoo (https://search.yahoo.com)

e) Ecosia (https://www.ecosia.org)

f) Dogpile (https://www.dogpile.com)

2) Pembangunan Kata Kunci (Keyword)

Mesin pencari umumnya melakukan pencarian dengan masukan suatu kata kunci (keyword) yang berbentuk kalimat dengan sintaks tertentu. Oleh karena itu, pencarian dimulai dengan mengubah setiap pasangan data ke dalam bentuk kata kunci (keyword) yang dapat dipahami oleh mesin pencari yang digunakan seperti terlihat pada Tabel 3.1. Perlu dipahami bahwa bentuk kata kunci dari pasangan data yang sama mungkin berbeda untuk setiap mesin pencari yang berbeda.

3) Proses Pencarian pada Mesin Pencari

Untuk setiap pasangan data, kata kunci yang dihasilkan kemudian diberikan pada mesin pencari untuk mendapatkan hasil pencarian. Untuk setiap butir hasil pencarian, dilakukan ekstraksi judul halaman dan ringkasan yang diberikan oleh mesin pencari, seperti terlihat pada Gambar 3.3. Dari proses ekstraksi ini, setiap pasangan data menghasilkan sekumpulan kalimat yang bersama-sama membentuk korpus untuk pasangan data tersebut.

4) Preprocessing

Setelah setiap pasangan data menghasilkan korpus, selanjutnya dilakukan preprocessing, yaitu proses pengolahan awal terhadap suatu teks untuk menghilangkan elemen-elemen yang dianggap tidak penting dan tidak sesuai konteks. Dalam tahap pembangunan korpus ini, preprocessing yang dilakukan yaitu:

a) Case Folding

b) Remove Number Token

c) Remove Punctuation Token

d) Remove Extra Whitespace

5) Ekstraksi Kalimat Relevan dalam Korpus

Setelah proses preprocessing selesai, maka proses selanjutnya adalah melakukan ekstraksi kalimat yang dianggap relevan dalam korpus, dengan cara hanya memilih kata-kata yang letaknya berada di antara kedua pasangan data yang bersangkutan. Aturan ekstraksi kalimat relevan adalah sebagai berikut:

a) Pilih rangkaian kata yang letak/posisinya berada diantara kedua pasangan data yang bersangkutan.

b) Apabila dalam satu kalimat terdapat lebih dari satu rangkaian kata relevan yang terdeteksi, maka akan dihasilkan kalimat baru untuk setiap rangkaian kata relevan tersebut.

B. Anotasi Korpus (Corpus Annotation)

Tahap kedua pembelajaran relasi non-taksonomi yaitu anotasi korpus. Pada penelitian ini, kata yang memiliki kemungkinan merupakan nama relasi yang dicari adalah kata-kata yang menghubungkan pasangan data kelas dalam sebuah kalimat. Dalam suatu ontologi, nama kelas biasanya berbentuk kata benda (noun) dan nama relasi biasanya berbentuk kata kerja (verb). Oleh karena itu, dalam penelitian ini diasumsikan bahwa nama relasi yang dicari adalah kata yang merupakan kata kerja. Penyaringan kata dilakukan dengan melihat bentuk dari setiap kata yang ada dalam korpus dan menyeleksi hanya kata-kata yang berbentuk kata kerja dan letak katanya berada di antara kedua pasangan data. Sistem harus dapat menentukan bentuk dari setiap kata dalam korpus, untuk itulah dilakukan anotasi korpus. Anotasi korpus adalah proses untuk menandai setiap kata dalam korpus dengan bentuk atau kelas kata yang tepat. Hasil dari tahap anotasi korpus adalah korpus yang sama dengan tahap sebelumnya, tetapi hanya berisi kata kerja saja.

1) Part-of-Speech Tagging

Dalam penelitian ini, POS Tagging dilakukan dengan menggunakan library POS Tagger dari Stanford (https://nlp.stanford.edu/software/tagger.shtml) yang merupakan implementasi dari konsep log-linear POS Tagger oleh [8]. Dengan model yang tepat, POS Tagger tersebut dapat mengenali jenis kata dalam bahasa yang diinginkan. Model yang baik dapat dihasilkan dengan cara melakukan pelatihan (model training) POS Tagger menggunakan korpus dimana setiap kata telah ditandai dengan jenis kata yang sesuai (POS Tag) secara manual.

Model POS Tagger untuk bahasa Indonesia tidak disediakan oleh Stanford, oleh karena itu dalam penelitian ini dibuat model tersendiri yang dilatih khusus untuk mengenali jenis kata dalam bahasa Indonesia. Model dilatih menggunakan korpus berbahasa Indonesia yang telah ditandai secara manual oleh [9].

2) Seleksi Kata Kerja

Setelah proses POS Tagging selesai, selanjutnya dilakukan penyaringan kata sehingga hanya kata kerja saja yang diseleksi. Aturan penyaringan kata untuk setiap kalimat yang ada dalam korpus adalah sebagai berikut:

a) Dalam satu kalimat, hanya diambil kata-kata yang merupakan kata kerja saja, karena asumsinya adalah bahwa nama relasi selalu berbentuk kata kerja.

b) Apabila dalam satu kalimat menghasilkan lebih dari satu kata kerja, maka kalimat tersebut dianggap tidak memenuhi kriteria dan semua kata kerja yang ada dalam kalimat tersebut tidak diseleksi (dibuang). Hal ini dilakukan karena kalimat-kalimat yang mengandung lebih dari satu kata kerja akan menimbulkan ambiguitas konteks, karena secara komputasi sulit menentukan kata kerja mana yang lebih tepat sebagai nama relasi.

C. Ekstraksi Kata (Term Extraction)

Tahap ketiga pembelajaran relasi non-taksonomi yaitu ekstraksi kata. Pada tahap ini, setiap kata yang telah ditandai telah disaring sehingga hanya kata kerja saja yang diambil. Kemudian setiap kata kerja yang muncul akan dihitung jumlah kemunculannya secara terpisah untuk tiap korpus, sehingga setiap korpus menghasilkan himpunan kata kerja masing-masing beserta jumlah kemunculan kata tersebut dalam korpus yang bersangkutan. Dari titik ini, diusulkan dua metode terpisah yang dilakukan untuk memilih kata yang tepat dijadikan sebagai nama relasi, yaitu metode Maksimum dari Rata-rata Absolut Skor Ternormalisasi (Maximum of Absolute Average of Normalized Score - MAANS) dan metode Frekuensi Terbesar dari Rekomendasi Lokal (Most Frequent Local Recommendation - MFLR). Kedua metode tersebut akan menghasilkan himpunan kata kerja keseluruhan dari seluruh korpus dimana setiap kata kerja akan memiliki nilai skor masing-masing. Meskipun kedua metode menerapkan pendekatan yang berbeda untuk memberikan nilai skor untuk masing-masing kata kerja, namun pada kedua metode tersebut kata kerja yang memiliki nilai skor paling tinggi adalah kata yang dipilih sebagai nama relasi.

1) Perhitungan Frekuensi Kemunculan Kata

Sebelum dilakukan proses pemilihan nama relasi, frekuensi kemunculan tiap kata kerja untuk setiap korpus dari setiap pasangan data harus dihitung terlebih dahulu. Setiap pasangan data dengan korpus masing-masing akan menghasilkan suatu daftar kata kerja dimana setiap kata kerja memiliki nilai masing-masing, yaitu jumlah frekuensi kemunculan kata kerja tersebut pada korpus yang bersangkutan.

2) Maximum of Absolute Average of Normalized Score (MAANS)

Metode Maximum of Absolute Average of Normalized Score (MAANS) menerapkan pendekatan secara statistik dimana setiap kata dalam setiap pasangan data dihitung bobotnya relatif terhadap banyak kata dalam pasangan data tersebut sebagai bentuk relevansi terhadap pasangan data. Kemudian setiap kata dihitung rata-rata nilai skornya dengan diagregasikan dengan kata yang sama pada pasangan data yang berbeda. Metode ini diusulkan untuk meningkatkan probabilitas kata yang lebih sering muncul dalam pasangan data yang berbeda, meskipun frekuensi kemunculan katanya tidak selalu tinggi pada pasangan data tertentu. Berikut adalah langkah-langkah penyelesaian untuk metode MAANS:

a) Buat suatu himpunan yang berisi semua kata kerja unik yang muncul dari hasil pencarian semua pasangan data.

b) Untuk setiap kata kerja yang ada dalam setiap pasangan data, beri nilai skor ternormalisasi berdasarkan frekuensi kemunculan kata kerja pada pasangan data tersebut.

c) Untuk setiap kata kerja yang ada dalam himpunan kata kerja unik dari semua data, jumlahkan skor dari kata kerja yang sama untuk setiap pasangan data. Apabila suatu kata kerja tidak muncul pada suatu pasangan data, maka skor dianggap 0.

d) Skor akhir setiap kata kerja merupakan rata-rata skor, maka dari itu setiap skor dari hasil penjumlahan dibagi dengan banyaknya pasangan data yang ada.

Kata kerja dengan rata-rata nilai skor tertinggi adalah kata yang dipilih sebagai nama relasi.

3) Most Frequent Local Recommendation (MFLR)

Metode Most Frequent Local Recommendation (MFLR) menerapkan pendekatan sederhana dimana setiap pasangan data

diharuskan menghasilkan keluaran berupa satu kata yang paling banyak muncul dalam pasangan data tersebut. Kemudian setelah semua pasangan data memiliki satu kata pilihan, pilih kata yang paling banyak dipilih oleh pasangan-pasangan data. Berikut adalah langkah-langkah penyelesaian untuk metode MFLR:

a) Untuk setiap pasangan data, pilih salah satu kata kerja yang frekuensi kemunculannya tertinggi dan masukkan ke dalam suatu himpunan pasangan baru dimana setiap pasangan adalah pasangan data dengan kata kerja terpilih.

b) Untuk setiap kata kerja yang muncul sebagai kata kerja terpilih suatu pasangan data, hitung banyak data yang memilih kata kerja tersebut sebagai kata kerja terpilih. Hasil pencacahan pemilihan tersebut menjadi skor akhir untuk setiap kata kerja.

Kata kerja dengan nilai skor tertinggi adalah kata yang dipilih sebagai nama relasi.

4) Penyaringan Nama Relasi Taksonomis

Karena nama relasi yang dicari adalah nama relasi non-taksonomi, maka nama-nama relasi yang digunakan untuk merepresentasikan nama relasi taksonomis harus disaring sehingga tidak mempengaruhi hasil akhir ekstraksi nama relasi. Nama-nama relasi taksonomis disimpan sebagai suatu himpunan stopwords. Dalam perhitungan skor total ekstraksi kata, apabila suatu kata terdeteksi sebagai anggota dari himpunan nama relasi taksonomis maka kata tersebut akan dibuang terlepas dari berapapun skor yang dihasilkan.

5) Validasi Kata Kerja Menggunakan Kamus

Sebelum dilakukan pemilihan kata kerja sebagai nama relasi non-taksonomi dari hasil perhitungan skor tahap sebelumnya, ada satu validasi yang harus dilakukan, yaitu memeriksa apakah kata kerja yang ada dalam tabel skor adalah kata kerja yang valid.

Dalam penelitian ini, proses klasifikasi yang diterapkan adalah mencari kata kerja yang ada dalam korpus, yang didapat dari penentuan jenis kata yang dilakukan menggunakan POS Tagger. Pemeriksaan kebenaran penentuan jenis kata dilakukan dengan cara pemeriksaan ulang pada kamus, yaitu Kamus Besar Bahasa Indonesia. Namun karena kamus yang dipakai merupakan layanan eksternal (diluar program), maka pemeriksaan semua kata satu per satu tidak mungkin dilakukan. Untuk mengurangi beban, maka dari itu dalam penelitian ini diputuskan bahwa hanya kata-kata yang terdeteksi sebagai kata kerja saja yang akan diperiksa menggunakan kamus. Pada tahap anotasi korpus, True Negative dan False Negative sudah dieliminasi dengan menyingkirkan kata-kata yang dideteksi oleh POS Tagger sebagai bukan kata kerja. Artinya tahap ini hanya diterapkan untuk menyingkirkan False Positives (FP), yaitu kata yang sebenarnya bukan kata kerja tetapi terdeteksi oleh POS Tagger sebagai kata kerja. Sehingga pada tahap akhir kata kerja yang terdaftar adalah kata kerja yang masuk dalam kategori True Positive (TP), yaitu kata kerja yang terdeteksi oleh POS Tagger sebagai kata kerja dan benar-benar merupakan kata kerja.

D. Pengembangan Kata

Tahap terakhir pembelajaran relasi non-taksonomi yaitu pengembangan kata. Pada tahap ini, daftar skor akhir yang sudah diolah kemudian diurutkan berdasarkan nilai skornya dari nilai terbesar ke terkecil (descending). Nama relasi non-taksonomi yang dipilih adalah kata kerja dengan skor tertinggi. Namun untuk memberikan pilihan tambahan bagi pengguna, maka diambil lima kata kerja dengan skor tertinggi yang kemudian dijabarkan bentuk katanya menjadi tiga bagian, yaitu kata dasarnya, bentuk aktifnya, dan bentuk pasifnya.

1) Deteksi Kata Kerja Aktif dan Pasif

Pendeteksian kata kerja aktif dan pasif dilakukan terhadap lima kata kerja dengan skor tertinggi. Program pendeteksi kata kerja aktif dan pasif dikembangkan sendiri oleh peneliti dengan berbasis aturan (rule-based). Aturan-aturan dalam program tersebut dibuat berdasarkan aturan linguistik bahasa Indonesia yang merujuk ke buku oleh [10]. Aturan-aturan linguistik yang digunakan meliputi pemenggalan suku kata dan aturan bentuk kata (imbuhan). Aturan-aturan tersebut kemudian diimplementasikan dalam bentuk regular expression dan percabangan program. Dalam pendeteksian juga dilakukan stemming untuk mendapatkan kata dasar. Hasil inferensi kata dasar dan perubahan imbuhan kata kerja aktif dan pasif kemudian diperiksa ulang ke kamus untuk menentukan apakah hasil inferensi sudah benar. Meski demikian, program pendeteksi yang dibuat belum sempurna dan masih banyak ruang untuk pengembangan lebih lanjut.

2) Pemilihan Nama Relasi Non-Taksonomi

Setelah semua proses telah dilalui, pemilihan nama relasi non-taksonomi cukup mudah, yaitu memilih kata kerja dengan skor tertinggi sebagai nama relasi non-taksonomi. Karena ada dua metode penilaian skor, maka program akan memilih salah satu metode secara acak, kemudian mengambil kata kerja dengan skor tertinggi pada metode tersebut.

E. Evaluasi Kata (Term Evaluation)

Tambahan proses untuk pembelajaran relasi non-taksonomi yaitu evaluasi kata. Pada proses ini, beberapa kata yang memiliki nilai skor tertinggi akan diambil untuk dilakukan pencarian ulang melalui mesin pencari (search engine) bersama dengan pasangan data yang ada di awal. Tahap ini dilakukan dengan tujuan untuk menilai apakah kata dengan nilai skor tertinggi merupakan nama relasi yang tepat apabila dikombinasikan dengan pasangan-pasangan data yang dimiliki. Penilaian dilakukan dengan cara sederhana, yaitu melihat jumlah hit sampel pasangan data dan kata kerja yang dievaluasi pada hasil pencarian mesin pencari. Dari jumlah hasil pencarian untuk masing-masing kata, pengguna bisa menilai apakah kata yang dipilih oleh sistem merupakan kata yang tepat.

1) Sampling Data untuk Evaluasi

Untuk mengurangi beban evaluasi dalam kasus dimana pasangan data kelas yang digunakan berjumlah banyak, maka untuk tahap evaluasi dilakukan sampling data untuk mengambil beberapa pasangan data kelas secara acak. Dalam penelitian ini, jumlah sampel yang digunakan ada sebanyak 10 data sampel (atau sesuai jumlah pasangan data jika jumlahnya kurang dari 10).

2) Seleksi Kata Berdasarkan Skor Tertinggi

Setiap pasangan data sampel akan diujikan dengan setiap kata kerja yang akan dievaluasi. Kata-kata kerja yang dievaluasi

adalah beberapa kata kerja yang skor akhirnya paling tinggi, yaitu lima (5) kata kerja dengan skor tertinggi. Karena setiap metode (MAANS dan MFLR) mempunyai kemungkinan untuk menghasilkan lima (5) kata kerja yang berbeda, maka masing-masing metode dievaluasi secara independen.

3) Proses Pencarian Jumlah Hit pada Mesin Pencari

Kemudian untuk setiap pasangan data sampel dan kata kerja dilakukan pencarian melalui mesin pencari untuk menemukan jumlah hit total pencarian.

Setelah setiap data sampel menghasilkan jumlah hit masing-masing berdasarkan mesin pencari, selanjutnya dilakukan pengelompokan jumlah hit dari setiap data sampel yang dievaluasi berdasarkan mesin pencari.

Langkah terakhir yaitu dilakukan penjumlahan jumlah hit dari semua pasangan data untuk masing-masing kata kerja yang dikelompokkan berdasarkan mesin pencari. Hasil penjumlahan untuk setiap mesin pencari ditambah dengan hasil penjumlahan total dari semua mesin pencari merupakan hasil akhir proses evaluasi.

IV. PENGUJIAN

A. Data Uji

Untuk menguji hasil pembelajaran relasi non-taksonomi, peneliti menggunakan dua jenis pasangan data kelas yang berbeda. Kedua data tersebut kemudian dimasukkan ke dalam program sebagai masukan pembelajaran. Berikut adalah deskripsi dari kedua data yang diujikan:

1) Data [herbal_penyakit]

Data pertama merupakan pasangan data tanaman herbal dengan penyakit tertentu. Data ini masuk ke dalam domain kesehatan dan farmasi. Menurut penalaran peneliti, seharusnya nama relasi yang dihasilkan oleh data ini adalah salah satu atau merupakan sinonim dari kata berikut: mengobati, menyembuhkan, mengatasi

2) Data [tempat_wisata]

Data kedua merupakan pasangan data tempat wisata dengan provinsi. Data ini masuk ke dalam domain pariwisata. Menurut penalaran peneliti, seharusnya nama relasi yang dihasilkan oleh data ini adalah salah satu atau merupakan sinonim dari kata berikut: terletak, berada

B. Parameter Pembelajaran

Program yang dikembangkan memberikan beberapa pilihan sebagai parameter untuk mengatur pembelajaran relasi non-taksonomi. Mesin pencari yang mendukung proses evaluasi kata secara otomatis yang digunakan dalam penelitian ini hanya tiga, yaitu Google, Yahoo, dan Ecosia. Untuk ketiga mesin pencari tersebut, masing-masing dibuatkan parameter pembelajaran untuk menguji keakuratan penggunaan mesin pencari tersebut. Kemudian ada satu parameter tambahan yang menguji pembelajaran yang menggunakan semua jenis mesin pencari. Untuk itu, ada empat jenis parameter pembelajaran yang diterapkan untuk setiap data, yaitu tiga parameter pertama untuk menguji berdasarkan mesin pencari dan parameter terakhir yang menguji pembelajaran yang mengerahkan semua mesin pencari secara penuh. Dengan dua jenis data, maka total pengujian dilakukan sebanyak delapan kali pembelajaran. Kedua metode ekstraksi data digunakan dalam semua parameter pembelajaran, karena kedua metode tidak mempengaruhi satu sama lain.

C. Pembahasan Pengujian

Dari delapan hasil pengujian yang telah dipaparkan, ringkasan hasil ekstraksi kata untuk setiap data uji beserta hasil yang diharapkan dapat diamati pada Tabel 5.10. Berdasarkan tabel tersebut, hasil ekstraksi kata yang dipilih oleh setiap metode ekstraksi kata (MAANS dan MFLR) dengan beberapa parameter pembelajaran yang berbeda menghasilkan kata yang seragam, yaitu kata "mengobati" untuk data [herbal_penyakit] dan kata "terletak" untuk data [tempat_wisata].

1) Analisis Akurasi Hasil Pembelajaran

a) Analisis Akurasi pada Data [herbal_penyakit]

- Untuk data [herbal_penyakit], hasil evaluasi untuk kata terpilih "mengobati" menempati urutan pertama pada mesin pencari Google.
- Pada mesin pencari Yahoo, kata terpilih "mengobati" menempati urutan kedua setelah kata "berkhasiat".
- Pada mesin pencari Ecosia, kata terpilih "mengobati" menempati urutan pertama.
- Pada pengujian dimana ketiga mesin pencari (Google, Yahoo, Ecosia) digunakan, kata terpilih "mengobati" menempati urutan kedua setelah kata "menurunkan".

Dari beberapa hasil analisis diatas, peneliti menyimpulkan bahwa kata terpilih "mengobati" adalah kata yang tepat sebagai hasil ekstraksi data [herbal_penyakit], karena kata "mengobati" merupakan salah satu kata yang diasumsikan sebagai kandidat nama relasi oleh penalaran peneliti dan hasil evaluasi kata juga menunjukkan bahwa kata terpilih "mengobati" konsisten muncul sebagai kata kerja yang menghasilkan hasil pencarian dengan urutan teratas. Meski demikian, dengan adanya kasus dimana salah satu sampel pasangan data memiliki jumlah hasil pencarian jauh lebih tinggi dibandingkan dengan sampel data lain, maka hasil evaluasi relatif kurang bisa diandalkan. Peneliti berasumsi ukuran sampel relatif kurang besar sehingga masih memungkinkan terjadinya dominasi sampel tertentu.

b) Analisis Akurasi pada Data [tempat_wisata]

- Untuk data [tempat_wisata], hasil evaluasi untuk kata terpilih "terletak" menempati urutan pertama pada mesin pencari Google.
- Pada mesin pencari Yahoo, kata terpilih "terletak" menempati urutan pertama.
- Pada mesin pencari Ecosia, kata terpilih "terletak" menempati urutan kedua setelah kata "memiliki".
- Pada pengujian dimana ketiga mesin pencari (Google, Yahoo, Ecosia) digunakan, kata terpilih "terletak" menempati urutan kedua setelah kata "berada".

Dari beberapa hasil analisis diatas, peneliti menyimpulkan bahwa kata terpilih "terletak" adalah kata yang tepat sebagai hasil ekstraksi data [tempat_wisata], karena kata "terletak" merupakan salah satu kata yang diasumsikan sebagai kandidat

nama relasi oleh penalaran peneliti dan hasil evaluasi kata juga menunjukkan bahwa kata terpilih "terletak" konsisten muncul sebagai kata kerja yang menghasilkan hasil pencarian dengan urutan teratas. Meski demikian, akurasi dan kemampuan setiap mesin pencari berbeda-beda sehingga hasil evaluasi secara keseluruhan relatif kurang bisa diandalkan. Peneliti berasumsi masih ada ruang untuk peningkatan dalam proses ekstraksi untuk memilih kata kerja yang tepat sehingga dapat memilih kata-kata kerja yang lebih relevan.

2) Analisis Metode Ekstraksi

Dari analisis diatas, peneliti menyimpulkan bahwa metode MAANS maupun metode MFLR memiliki akurasi yang hampir sama dalam menentukan kata kerja terpilih, karena urutan pertama dan keduanya hampir selalu konsisten. Namun tentu saja urutan kata kerja juga dipengaruhi oleh mesin pencari yang digunakan (korpus), tidak hanya metode ekstraksi semata. Secara keseluruhan, peneliti lebih memilih metode MAANS dibandingkan metode MFLR karena metode MAANS lebih konsisten dalam menentukan urutan rangking kata kerja.

V. SIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, maka dapat diambil beberapa kesimpulan sebagai berikut:

- Penelitian ini telah menghasilkan suatu program yang dapat digunakan untuk melakukan pembelajaran relasi non-taksonomi (LNTRO) dengan data bersumber dari dokumen Web.
- Dari hasil pengujian dengan dua jenis data uji yang berbeda, dapat disimpulkan bahwa rancangan alur kerja yang diusulkan untuk melakukan pembelajaran relasi non-taksonomi berhasil menghasilkan nama relasi non-taksonomi yang sesuai dan relevan dengan data yang diberikan.
- Dari hasil evaluasi kata dengan dua jenis data uji yang berbeda dan empat parameter pembelajaran yang berbeda, disimpulkan bahwa hasil evaluasi tidak selalu menghasilkan jumlah hasil pencarian yang akurat dan sesuai dengan nama relasi non-taksonomi terpilih karena berbagai faktor, seperti faktor sampling pasangan data dan faktor kemampuan setiap mesin pencari yang berbeda-beda.
- Metode ekstraksi kata yang diusulkan, yaitu MAANS (Maximum of Absolute Average of Normalized Score) dan MFLR (Most Frequent Local Recommendation) menghasilkan nama relasi non-taksonomi teratas yang sama, sehingga kedua metode tersebut dianggap berhasil. Namun untuk masalah konsistensi urutan rangking metode MAANS lebih konsisten dibandingkan dengan metode MFLR.

DAFTAR PUSTAKA

[1] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing?," *Int. J. Hum.-Comput. Stud.*, vol. 43, no. 5–6, hlm. 907–928, Nov 1995.

[2] V. Alexiev, M. Breu, dan J. de Bruijn, *Information integration with ontologies : experiences from an industrial showcase*. John Wiley & Sons, 2005.

[3] D. Sánchez dan A. Moreno, "Learning non-taxonomic relationships from web documents for domain ontology construction," *Data Knowl. Eng.*, vol. 64, no. 3, hlm. 600–623, Mar 2008.

[4] I. Serra, R. Girardi, dan P. Novais, "The Problem of Learning Non-taxonomic Relationships of Ontologies from Text," dalam *Distributed Computing and Artificial Intelligence*, vol. 151, S. Omatu, J. F. De Paz Santana, S. R. González, J. M. Molina, A. M. Bernardos, dan J. M. C. Rodríguez, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, hlm. 485–492.

[5] J. Villaverde, A. Persson, D. Godoy, dan A. Amandi, "Supporting the discovery and labeling of non-taxonomic relationships in ontology learning," *Expert Syst. Appl.*, vol. 36, no. 7, hlm. 10288–10294, Sep 2009.

[6] A. Maedche dan S. Staab, "Mining Non-Taxonomic Conceptual Relations from Text," hlm. 13.

[7] A. Fader, S. Soderland, dan O. Etzioni, "Identifying relations for open information extraction," dalam *Proceedings of the conference on empirical methods in natural language processing*, 2011, hlm. 1535–1545.

[8] K. Toutanova, D. Klein, C. D. Manning, dan Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," dalam *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, Edmonton, Canada, 2003, vol. 1, hlm. 173–180.

[9] A. Dinakaramani, F. Rashel, A. Luthfi, dan R. Manurung, "Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus," 2014, hlm. 66–69.

[10] Sugihastuti dan S. Saudah, *Buku Ajar Bahasa Indonesia Akademik*. Yogyakarta: Pustaka Pelajar, 2016.