Ekstraksi Relasi *Meronymy* dengan *Lexico-Syntactic Patterns*

Eunike Andriani Kardinata^{#1}, Nur Aini Rakhmawati^{*2}

[#]Departemen Sistem Informasi, Institut Sains dan Teknologi Terpadu, Surabaya, Indonesia

¹eunike@istts.ac.id

^{*}Departemen Sistem Informasi, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

²nur.aini@is.its.ac.id

Abstrak— Ontologi terdiri atas konsep dan relasi yang masing-masing dapat diekstrak dengan berbagai macam metode. Salah satu metode yang dapat digunakan untuk ekstraksi relasi adalah metode berdasarkan *Lexico-Syntactic Patterns*. Secara sederhana, ekstraksi relasi dilakukan dengan mendapatkan sebuah pola yang menunjukkan sebuah relasi. Kemudian dilakukan percobaan untuk menguji apakah pola yang didapatkan mampu memprediksi relasi dengan tepat. Pada penelitian ini dilakukan percobaan untuk menguji pola relasi *meronymy* yang didapatkan dari dataset penelitian terdahulu. Evaluasi dilakukan dengan menggunakan nilai *recall* dan *precision*. Dari penelitian ini, ditemukan bahwa banyaknya (keragaman) variasi dalam sekumpulan pola yang menunjukkan suatu relasi dapat mempengaruhi kemampuan kumpulan pola tersebut untuk memprediksi relasi dengan tepat. Semakin banyak variasi pola dalam satu relasi, maka ketepatan prediksi cenderung menurun.

Kata kunci— ekstraksi, *Lexico-Syntactic Patterns*, ontologi, *meronymy*, relasi

I. PENDAHULUAN

Ontologi adalah sekumpulan konsep atau entitas dalam sebuah domain yang saling terhubung melalui relasi antar konsep. Dalam pembelajaran ontologi (*ontology learning*) terdapat dua proses, yaitu ekstraksi konsep dan ekstraksi relasi [1].

Ekstraksi konsep berkaitan dengan proses untuk menemukan kata atau frasa benda yang menunjukkan sebuah entitas, sedangkan ekstraksi relasi ialah proses untuk menemukan hubungan antara konsep.

Ekstraksi relasi bertujuan untuk membangun sebuah pengetahuan yang terstruktur dari teks yang tidak terstruktur. Terdapat beberapa kategorisasi tipe relasi yang biasanya didasarkan pada konteks di mana relasi tersebut ditemukan [2]. Sebagai contoh, beberapa tipe relasi yang berbeda didefinisikan menurut:

- ACE 2003: *role, part, at, social*
- Geographical: *geographical entity (GE), natural GE, inhabited GE*
- Freebase: *people, person, nationality*

- dan sebagainya.

Dalam ekstraksi relasi, sangat penting untuk mempertimbangkan konteks dalam menentukan relasi apa yang dicari. Selain itu, pada konteks yang berbeda, relasi dapat memiliki makna atau definisi yang berbeda. Dengan demikian, evaluasi hasil ekstraksi relasi menjadi aspek yang signifikan dalam pembangunan ontologi.

Terdapat beberapa metode ekstraksi relasi yang didefinisikan oleh [3]. Beberapa metode tersebut adalah sebagai berikut:

- Metode berdasarkan *Lexico-Syntactic Patterns*
- Metode berdasarkan *Head-Modifier Detection*
- Metode berdasarkan *Distributional Analysis*
- Metode berdasarkan *Distributional Inclusion*

Ekstraksi relasi berbasis *Lexico-Syntactic Patterns* menginferensikan relasi dari berbagai pola yang ada, yang ditemukan merujuk kepada jenis relasi tertentu. Sedangkan metode berdasarkan *Head-Modifier Detection*, menentukan relasi dengan melihat kepala dari sebuah kata dan imbuhan atau kata apa yang menempel pada kepala tersebut.

Metode *Distributional Analysis* mencari relasi berdasarkan kemiripan pada konteks di mana sebuah kata ditemukan. Kemiripan ini bisa berdasarkan atribut (dari kata-kata yang muncul di sekitarnya) atau kemiripan relasi (dari pasangan kata yang menyertai). Lalu metode *Distributional Inclusion* menemukan relasi berdasarkan hirarki semantik. Ini berarti bahwa dalam satu hirarki, maka kata yang pangkatnya lebih tinggi akan menurunkan makna umumnya kepada kata di bawahnya.

Basis penelitian ini adalah ekstraksi relasi berdasarkan *Lexico-Syntactic Patterns* yang akan dijelaskan lebih lanjut pada Bagian II. Selanjutnya akan dijabarkan pula dataset dan metode yang digunakan pada Bagian III. Terakhir, akan diberikan hasil dari percobaan yang sudah dilakukan beserta diskusi mengenai hasil tersebut pada Bagian IV dan kesimpulan dari penelitian ini pada Bagian V.

II. PENELITIAN TERKAIT

Ekstraksi relasi berdasarkan *Lexico-Syntactic Patterns* memiliki pemahaman bahwa hubungan semantik antar

konsep atau entitas dapat diinferensikan dari pola yang ditemukan sekalipun istilah yang bersangkutan atau entitas yang berhubungan tersebut belum pernah dijumpai. Berikut contohnya pada Gambar 1.

Gambar 1. Pola dalam kalimat

Kalimat tersebut mempunyai pola frasa benda (NP—*noun phrase*) diikuti dengan kata “seperti” lalu frasa benda (NP) lainnya. Dari kalimat tersebut, dapat diduga bahwa “beluga” adalah sejenis “paus” sekalipun kata “beluga” tidak pernah ditemukan sebelumnya. Sebuah kalimat yang mengandung kata “seperti” bisa menunjukkan bahwa kalimat tersebut digunakan untuk memberikan contoh.

Ada berbagai metode untuk mendapatkan relasi berdasarkan *Lexico-Syntactic Patterns*. Beberapa metode telah dipublikasikan oleh peneliti berikut:

- Hearst [4] [5]
- Pantel dan Pennacchiotti [6]
- Ponzetto dan Strube [7]
- Cederberg dan Widdows [8]

Pada penelitian ini, secara spesifik akan dibahas mengenai relasi *meronymy* atau *part-whole*. Relasi *meronymy* adalah relasi antar komponen penyusun dengan kesatuan sebuah entitas. Relasi ini berbeda dengan relasi *hyponymy*, yaitu relasi antara entitas dengan subsetnya. Sebagai contoh, meronim dari “pohon” adalah “daun” atau “batang”; sedangkan hiponim dari “pohon” adalah “pohon apel” atau “pohon jeruk”.

Relasi *meronymy* memiliki beberapa sub-tipe. Salah satu kategori yang dapat digunakan adalah berdasarkan penelitian terdahulu oleh [9]. Berikut adalah sub-tipe tersebut dan definisinya:

- *Component-Of* menunjukkan relasi antara integral dan komponen fungsionalnya, contohnya: jari dan tangan.
- *Member-Of* menunjukkan relasi antara objek fisik atau peran dengan sebuah agregasi (tim, organisasi, dan sebagainya), contohnya: pemain dan tim.
- *Portion-Of* atau *Sub-Quantity-Of* menunjukkan relasi antara jumlah materi atau unit, contohnya: oksigen dan air.
- *Stuff-Of* atau *Substance-Of* atau *Constituted-Of* menunjukkan relasi antara objek fisik dan sejumlah materi, contohnya: besi dan mobil.
- *Located-In* menunjukkan relasi antar konsep dengan region dua dimensi, contohnya: Jakarta dan Indonesia.
- *Contained-In* menunjukkan relasi antar konsep dengan region tiga dimensi, contohnya: chip dan prosesor.
- *Phase-Of* atau *Involved-In* atau *Feature-Activity* menunjukkan relasi antara sebuah fase dengan sebuah proses, contohnya: mengunyah dan makan.
- *Participates-In* menunjukkan relasi antara konsep dan proses, contohnya: enzim dan reaksi.

Evaluasi akurasi relasi yang ditemukan dapat dilakukan dengan berbagai cara. Cara yang umum adalah dengan

melakukan pengecekan secara manual seperti pada penelitian [10]. Evaluasi juga dapat dilakukan dengan menggunakan pengukuran seperti *precision*, *recall*, dan sebagainya, seperti yang dijabarkan oleh [11].

III. METODE PENELITIAN

Terdapat beberapa tahapan dalam penelitian ini, yang mana tahapan tersebut dirancang dari adopsi pengaturan eksperimen yang dilakukan oleh [12]. Secara garis besar, peneliti mencari pola yang menunjukkan sebuah relasi dari dataset yang ada, kemudian peneliti memprediksi relasi dari dataset lain berdasarkan pola yang sudah ditemukan.

Dataset yang digunakan bersumber dari penelitian yang dilakukan oleh [13]. Pada penelitian tersebut, data yang digunakan dihimpun dari berbagai sumber dan diolah sehingga sudah diidentifikasi konsep dan jenis relasi *meronymy* terkait.

Terdapat sekitar 5700 baris dalam dataset yang digunakan. Setiap baris data memiliki format sebagai berikut:

- 1stEntity: konsep pertama
- Pattern: frasa atau kata yang menunjukkan relasi
- 2ndEntity: konsep kedua
- 1stEntity normalized: normalisasi konsep pertama
- Pattern normalized: normalisasi frasa atau kata yang menunjukkan relasi
- 2ndEntity normalized: normalisasi konsep kedua
- Kalimat sumber
- URL sumber artikel atau paragraf
- Subtipe relasi *part-whole*
- Urutan yang menunjukkan hubungan 1stEntity dan 2ndEntity: Part-Whole atau Whole-Part

Pada Gambar 2 di halaman berikut ialah contoh satu baris data sesuai dengan format yang dijelaskan.

Percobaan dilakukan dengan pengaturan berikut: sebanyak 4000 baris data akan digunakan untuk *training* sementara sisanya digunakan untuk *testing*. Pada halaman berikut, Gambar 3 adalah tahapan dalam penelitian ini. Selanjutnya akan dijabarkan masing-masing tahapan pada penelitian ini.

A. Ekstraksi Pola

Langkah pertama yang dilakukan ialah mengambil semua pola (*pattern*) yang menunjukkan sebuah relasi. Field yang diambil adalah *pattern normalized* karena pada field ini, semua kata dikembalikan ke dalam bentuk dasarnya sehingga meminimalkan duplikasi kata atau frasa yang sebenarnya merujuk pada hal yang sama.

Pada tahapan ini, data *training* yang digunakan murni berasal dari penelitian sebelumnya [13]. Sesuai dengan relasi yang didefinisikan pada [9], didapatkan semua kata atau frasa yang mengindikasikan adanya sebuah relasi antar entitas di dalam kalimat, beserta dengan jenis relasi yang diindikasikan tersebut.

B. Pengelompokan Pola

Berdasarkan jenis relasinya, semua kata atau frasa yang ditemukan dikelompokkan menjadi satu. Dengan demikian

telah didapatkan sekumpulan kata dan frasa yang pernah digunakan untuk menunjukkan suatu jenis relasi.

1ST ENTITY	the committee
PATTERN	is made up of
2ND ENTITY	eight members
1ST ENTITY NORMALIZED	the committee
PATTERN NORMALIZED	be make up of
2ND ENTITY NOMALIZED	eight member
KALIMAT SUMBER	the committee is made up of eight members, appointed by the board of voluntary planning
URL SUMBER	http://www.dominionpaper.ca/articles/2147
SUB-TIPE RELASI	Member-Of
URUTAN	Whole-Part

Gambar 2. Contoh baris data

Gambar 3. Tahapan penelitian

C. Penghapusan Duplikat Pola

Ada kalanya, suatu pola yang sama digunakan berulang kali dalam beberapa kalimat. Agar pemrosesan data pada tahap berikutnya tidak terlalu berat dan agar didapatkan daftar pola yang tepat, maka dilakukan pengecekan untuk menghilangkan duplikat kata atau frasa.

D. Deteksi Relasi

Lalu setelah didapatkan sekumpulan pola pada masing-masing kategori yang menunjukkan suatu jenis relasi, dilakukan percobaan untuk memprediksi jenis relasi dari dataset yang ditentukan. Dalam hal ini, digunakan data *testing* yang bersumber dari penelitian yang sama [13], namun diasumsikan jenis relasinya belum diketahui.

Deteksi relasi dilakukan dengan cara membandingkan apakah pola yang ditemukan pada kalimat juga ditemukan pada daftar pola yang disimpan menurut kategori jenis relasi. Jika pola ditemukan pada daftar, maka dapat dikatakan bahwa jenis relasi pada kalimat bersangkutan adalah sama dengan jenis relasi pada daftar tersebut.

E. Evaluasi Recall dan Precision

Evaluasi hasil percobaan dilakukan menggunakan perhitungan *precision* dan *recall* untuk masing-masing sub-tipe relasi. Jenis relasi yang ditemukan pada data *testing* berdasarkan hasil percobaan dibandingkan dengan jenis relasi sesuai label awal.

IV. HASIL DAN PEMBAHASAN

Dari 4000 baris data yang diproses, pada awalnya ditemukan pola untuk masing-masing sub-tipe relasi dengan jumlah seperti pada Gambar 4:

Initial length for Component-Of	: 438
Initial length for Contained-In	: 191
Initial length for Located-In	: 386
Initial length for Member-Of	: 911
Initial length for Participates-In	: 616
Initial length for Phase-Of	: 349
Initial length for Portion-Of	: 373
Initial length for Stuff-Of	: 744

Gambar 4 Jumlah pola awal

Setelah duplikasi pola dihilangkan, jumlah sisa pola yang dimasukkan ke dalam daftar pola dapat dilihat pada Gambar 5:

Final length for Component-Of	: 172
Final length for Contained-In	: 24
Final length for Located-In	: 19
Final length for Member-Of	: 217
Final length for Participates-In	: 292
Final length for Phase-Of	: 207
Final length for Portion-Of	: 125
Final length for Stuff-Of	: 375

Gambar 5. Jumlah pola akhir

Berdasarkan hasil tersebut, ditemukan duplikasi pola rata-rata mencapai setengah dari jumlah awal pola yang ditemukan; dengan pengecualian sub-tipe relasi *Contained-In* dan *Located-In* yang mana jumlah pola akhirnya sangat sedikit.

Duplikasi pola pada satu sub-tipe relasi memiliki jumlah yang berbeda dengan sub-tipe relasi lainnya karena variasi pola yang dapat digunakan untuk menunjukkan relasi pun memiliki keterbatasan yang berbeda. Demikian dapat disimpulkan bahwa variasi pola untuk relasi *Located-In* dan *Contained-In* jauh lebih sedikit dibandingkan dengan relasi lain, seperti *Participates-In* dan *Phase-Of*.

Selanjutnya, percobaan dilakukan menggunakan sisa baris data untuk memprediksi sub-tipe relasi. Gambar 6 menunjukkan jumlah data dan perhitungan untuk *recall* dan *precision* pada masing-masing sub-tipe relasi.

True positive for Component-Of	: 154
False negative for Component-Of	: 52
False positive for Component-Of	: 722
Recall for Component-Of	: 0.748
Precision for Component-Of	: 0.176
True positive for Contained-In	: 74
False negative for Contained-In	: 7
False positive for Contained-In	: 366
Recall for Contained-In	: 0.914
Precision for Contained-In	: 0.168
True positive for Located-In	: 146
False negative for Located-In	: 3
False positive for Located-In	: 238
Recall for Located-In	: 0.980
Precision for Located-In	: 0.380
True positive for Member-Of	: 297
False negative for Member-Of	: 65
False positive for Member-Of	: 362
Recall for Member-Of	: 0.820
Precision for Member-Of	: 0.451
True positive for Participates-In	: 178
False negative for Participates-In	: 79
False positive for Participates-In	: 421
Recall for Participates-In	: 0.693
Precision for Participates-In	: 0.297
True positive for Phase-Of	: 100
False negative for Phase-Of	: 49
False positive for Phase-Of	: 489
Recall for Phase-Of	: 0.671
Precision for Phase-Of	: 0.170
True positive for Portion-Of	: 144
False negative for Portion-Of	: 39
False positive for Portion-Of	: 448
Recall for Portion-Of	: 0.787
Precision for Portion-Of	: 0.243
True positive for Stuff-Of	: 206
False negative for Stuff-Of	: 133
False positive for Stuff-Of	: 600
Recall for Stuff-Of	: 0.608
Precision for Stuff-Of	: 0.256

Gambar 6. Hasil percobaan

Secara umum, hasil percobaan menunjukkan *recall* di atas 60 persen, sementara *precision* yang didapatkan semuanya di bawah 50 persen. *Recall* yang paling rendah adalah untuk relasi *Stuff-Of*, sementara yang tertinggi adalah untuk relasi *Located-In*. Lalu *precision* yang terendah adalah untuk relasi *Phase-Of*, dan yang tertinggi adalah untuk relasi *Member-Of*.

Jika persentase duplikasi, nilai *precision*, dan nilai *recall* diurutkan dari yang tertinggi berdasarkan sub-tipe relasi, maka didapatkan hasil berikut seperti pada Tabel 1.

TABEL 1
URUTAN BERDASARKAN SUB-TIPE RELASI

Sub-Tipe Relasi	Duplikasi	Recall	Precision
Component-Of	5	5	6
Contained-In	2	2	8
Located-In	1	1	2
Member-Of	3	3	1
Participates-In	6	6	3
Phase-Of	8	7	7
Portion-Of	4	4	5
Stuff-Of	7	8	4

Pada Tabel 1, ditemukan bahwa jenis relasi yang polanya memiliki banyak duplikasi—dengan kata lain, variasi polanya sedikit, menghasilkan nilai *recall* yang tinggi. Hal ini mungkin terjadi karena variasi pola yang sedikit menyebabkan prediksi relasi yang benar lebih mudah. Lain halnya jika ada banyak variasi pola dalam satu sub-tipe. Banyaknya variasi pola dapat meningkatkan kemungkinan terjadinya *overlap* dengan pola pada sub-tipe relasi lain. Dengan kata lain, satu pola yang sama bisa ada pada beberapa sub-tipe relasi.

Sementara itu untuk nilai *precision*, hasil yang didapat cenderung acak. Terdapat sub-tipe relasi yang duplikasinya banyak, namun memiliki *precision* yang tinggi dan sebaliknya. Secara umum, *precision* yang rendah ini dapat disebabkan oleh banyaknya *overlap* antara pola pada sub-tipe relasi yang berbeda. Demikian, ada banyak relasi yang diprediksi sebagai salah satu sub-tipe, namun kenyataannya bukan termasuk ke dalam sub-tipe tersebut.

V. KESIMPULAN

Pada penelitian ini, telah dilakukan percobaan sederhana untuk mengekstrak relasi berdasarkan *Lexico-Syntactic Patterns*. Relasi yang diekstrak adalah relasi *meronymy* dengan evaluasi memakai nilai *recall* dan *precision*. Disimpulkan bahwa relasi yang memiliki banyak variasi pola lebih mampu memprediksi dengan benar sub-tipe relasi tertentu. Sedangkan untuk hasil prediksi yang benar sesuai sub-tipenya tidak dapat diamati hanya dari variasi pola. Diperlukan evaluasi lain, seperti seberapa besar *overlap* pola antar sub-tipe relasi.

REFERENSI

- [1] K. Lakel and F. Bendella, "Dynamic Evaluation of Ontologies," *Procedia Computer Science*, vol. 73, pp. 16-23, 2015.
- [2] L. Zettlemoyer, "Relation Extraction," Department of Computer Science & Engineering, University of Washington, 2013.

- [3] R. Granada, R. Vieira, C. Trojahn and N. Aussenac-Gilles, "Evaluating the Complementarity of Taxonomic Relation Extraction Methods Across Different Languages," *arXiv*, 8 November 2018.
- [4] M. A. Hearst, "Automated discovery of wordnet relations," in *WordNet: An electronic lexical database and some of its applications*, 1998, p. 131–153.
- [5] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *Proceedings of the 14th conference on Computational linguistics - Volume 2, COLING '92*, Stroudsburg, PA, USA, 1992.
- [6] P. Pantel and M. Pennacchiotti, "Espresso: Leveraging generic patterns for automatically harvesting semantic relations," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, ACL-44*, Stroudsburg, PA., 2006.
- [7] S. P. Ponzetto and M. Strube, "Taxonomy induction based on a collaboratively built knowledge repository," *Artificial Intelligence*, vol. 175, no. 9, pp. 1737-1756, June 2011.
- [8] S. Cederberg and D. Widdows, "Using Isa and noun coordination information to improve the precision and recall of automatic hyponymy extraction," in *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4, CoNLL-2003*, Stroudsburg, PA, USA, 2003.
- [9] V.-T. Phi and Y. Matsumoto, "Integrating Word Embedding Offsets into the Espresso System for Part-Whole Relation Extraction," in *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers*, Seoul, South Korea, 2016.
- [10] C. Klaussner and D. Zhekova, "Lexico-Syntactic Patterns for Automatic Ontology Building," in *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, Hissar, Bulgaria, 2011.
- [11] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37-63, 2011.
- [12] W. R. v. Hage and G. Schreiber, "A Method for Learning Part-Whole Relations," in *ISWC 2006: 5th International Semantic Web Conference*, Athens, GA, USA, 2006.
- [13] V.-T. Phi, J. Santoso, M. Shimbo and Y. Matsumoto, "Ranking-Based Automatic Seed Selection and Noise Reduction for Weakly Supervised Relation Extraction," in *56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, 2018.