BAB 1
PENDAHULUAN

Bab ini membahas mengenai latar belakang penelitian, perumusan masalah, tujuan dan manfaat dilakukannya penelitian, ruang lingkup penelitan, dan terakhir metodologi yang digunakan.

1.1 Latar Belakang

Berbagai penelitian yang berkaitan dengan *Information Retrieval*dan *Natural Language Processing* mulai muncul menggunakan Bahasa Indonesia. Informasi mengenai kata dan artinya sangat dibutuhkan. Sayangnya, *resource* yang dimiliki berbasis Bahasa Indonesia masih sangat terbatas. Informasi tersebut umumnya disimpan dalam kamus digital seperti WordNet. Penelitian ini dilaksanakan untuk mengembangkan *resource* dalam Bahasa Indonesia.

WordNet adalah kamus digital yang dapat digunakan untuk menunjang penelitian di bidang *Information Retrieval*dan *Natural Language Processing*. WordNet yang paling sering digunakan dalam berbagai penelitian adalah WordNet Princeton berbahasa Inggris yang dibentuk secara manual oleh ahli linguistik. Setiap *entry* pada WordNet disimpan dalam bentuk set sinonim atau biasa disebut *synset* dan arti dari *synset* tersebut atau biasa disebut *sense*. Informasi lain yang disimpan dalam suatu synset adalah relasi antar synset. Relasi semantik yang tersimpan dalam WordNet adalah *synonym*, *antonym*, *hypernym*, *hyponym*, *holonym*, *meronym*, *troponym*, dan *entailment*.

Beberapa penelitian telah dilakukan untuk membangun WordNet Bahasa Indonesia. WordNet Bahasa Indonesia yang telah ada adalah Indonesian WordNet (IWN) yang dibuat oleh Fakultas Ilmu Komputer Universitas Indonesia (Fasilkom UI) dan WordNet Bahasa yang dibuat oleh Nanyang Technology University (NTU). Salah satu kelemahan kedua WordNet tersebut adalah ukurannya yang masih sangat terbatas. Selain itu, informasi mengenai relasi kata juga belum dapat tersimpan secara baik. Kedua WordNet memetakan *synset* Bahasa Indonesia ke WordNet Princeton dan memanfaatkan relasi yang di dalamnya. Mengetahui hal tersebut, dilakukan penelitian yang dapat mengekstrak relasi antar kata dengan hanya menggunakan korpus Bahasa Indonesia. Proses ekstraksi berjalan secara cepat dan data yang dihasilkan berjumlah besar. Korpus yang dihasilkan diharapkan dapat berguna

untuk penelitian selanjutnya.

Relasi kata adalah salah satu hal penting yang perlu diketahui jika ingin mengetahui hubungan antar kata secara semantik. Informasi yang berkaitan dengan semantik atau arti kata sulit diperoleh tanpa adanya pengetahuan sebelumnya. Kata-kata yang mirip secara leksikal belum tentu berelasi secara semantik. Sementara kata-kata yang tidak memiliki kesamaan secara leksikal bisa memiliki arti yang mirip atau berhubungan secara semantik. Korpus relasi yang dibuat diharapkan dapat membantu memperoleh informasi tersebut. Selain itu, pengetahuan mengenai relasi kata dapat dimaanfatkan dalam berbagai penelitian lain seperti *question answering* (Ravichandran dan Hovy, 2002), *information extraction*, dan *anaphora resolution*.

Melihat adanya kebutuhan akan korpus relasi kata, dilakukanlah penelitian *word relation extraction*. Penelitian ini berusaha mengekstrak kata berdasarkan relasi tertentu dari suatu dokumen sehingga dihasilkan korpus relasi kata. Penelitian kali ini akan fokus pada relasi kata *hypernym-hyponym*. Keduanya menyatakan relasi antara kata yang lebih umum (*hypernym*) dengan kata yang lebih khusus (*hyponym*). Metode yang digunakan adalah pattern matching dengan memanfaatkan korpus Wikipedia Bahasa Indonesia. Wikipedia memuat banyak kata dari berbagai domain sehingga dapat dimanfaatkan untuk membuat pattern yang general serta menghasilkan korpus relasi berukuran besar.

1.2 Perumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, pertanyaan yang menjadi rumusan penelitian adalah sebagai berikut.

1. Bagaimana cara membangun korpus relasi secara cepat dan berkualitas baik secara otomatis?

2. Apakah metode *pattern extraction* dan *matching* baik dilakukan untuk ekstraksi relasi kata Bahasa Indonesia?

3. Bagaimana cara mengevaluasi *pattern* dan korpus relasi kata dari hasil eksperimen?

1.3 Tujuan dan Manfaat Penelitian

Tujuan dari penelitian *word relation extraction* ini adalah membangun korpus relasi kata Bahasa Indonesia berukuran besar dan berkualitas baik secara otomatis. Selain

itu, ingin diketahui pula apakah metode *pattern extraction* dan *matching* baik digunakan untuk mengekstrak relasi kata Bahasa Indonesia. Diharapkan korpus relasi kata yang dihasilkan dapat menunjang berbagai penelitian selanjutnya.

Penelitian ini juga diharapkan dapat memotivasi adanya penelitian selanjutnya di bidang *Language Resource Development*, terutama pembangunan WordNet Bahasa Indonesia. Penelitian mengenai ekstraksi relasi kata berikutnya dengan berbagai metode lain diharapkan terus dilaksanakan sehingga Bahasa Indonesia memiliki *resource* yang semakin baik.

1.4 Ruang Lingkup Penelitian

Penelitian ini hanya fokus pada pembuatan korpus pasangan kata dengan relasi *hypernym* dan *hyponym* Bahasa Indonesia. Kelas kata yang menjadi fokusi penelitan adalah kata benda (*noun*). Pengembangan korpus dilakukan secara *semi-supervised* dengan metode *pattern matching*. *Pattern* yang dibuat masih terbatas hanya merupakan *pattern* leksikal. Evaluasi akan dilakukan pada *pattern* dan korpus pasangan kata yang dihasilkan. Proses evaluasi korpus pasangan kata dilakukan menggunakan teknik *random sampling*. Data yang digunakan untuk pembuatan *pattern* maupun untuk ekstraksi korpus pasangan kata baru adalah Wikipedia Bahasa Indonesia. Pemilihan relasi *hypernym-hyponym* dalam penelitian karena memiliki berbagai manfaat. Relasi tersebut dapat mengidentifikasi *Proper Noun* baru yang belum diketahui maknanya.

1.5 Tahapan Penelitian

Proses penelitian dilakukan dalam beberapa tahapan sebagai berikut.

1. Studi Literatur
   Pada tahap ini, dilakukan pembelajaran mengenai penelitian-penelitian sebelumnya yang telah dilakukan di bidang *word relation extraction* sehingga diketahui langkah yang perlu diambil selanjutnya.

2. Perumusan Masalah
   Perumusan masalah dilakukan untuk mendefinisikan masalah yang ingin diselesaikan, tujuan penelitian, dan hasil yang diharapkan sehingga proses penelitian dapat berjalan dengan baik.

3. Perancangan Penelitian
   Setelah diketahui hasil yang ingin dicapai, dirancang tahap-tahap eksperimen

secara terstruktur. Hal-hal yang diperhatikan mulai dari pengumpulan korpus awal (*seed*), *pre-processing* dokumen, perancangan implementasi *pattern extraction matching*, hingga proses evaluasi.

4. Implementasi
   Implementasi dilaksanakan sesuai dengan rancangan penelitian untuk menjawab rumusan masalah. Segala hasil yang ditemukan digunakan untuk terus memperbaiki metode dan teknik penelitan sehingga didapatkan hasil yang semakin baik.

5. Analisis dan Kesimpulan
   Tahap terakhir dari penelitian ini adalah menganalisis korpus pasangan kata relasi yang dihasilkan. Pertanyaan dari perumusan masalah dijawab, kemudian ditarik kesimpulan.

BAB 2
TINJAUAN PUSTAKA

Pada bab ini, dijelaskan mengenai studi literatur yang dilakukan. Studi literatur yang dilakukan digunakan sebagai dasar konsep dan teknik penelitian. Dipaparkan pula berbagai istilah dan metode yang digunakan dalam penelitian.

2.1 WordNet

WordNet adalah kamus leksikal yang tersimpan secara digital dan digunakan untuk berbagai keperluan komputasi (Miller, 1995). Pembuatan WordNet dilatarbelakangi keperluan mendapatkan *sense* atau arti semantik suatu kata. Informasi tersebut perlu disimpan dan dapat dibaca oleh mesin. WordNet pertama dibuat oleh Miller (1995) berbasis Bahasa Inggris dan sekarang dikenal dengan nama Princeton WordNet (PWN). WordNet menyimpan informasi dalam bentuk database dimana setiap entry-nya adalah pasangan *synset* dan arti semantiknya (*sense*). Set sinonim (*synset*) adalah himpunan kata yang memiliki arti yang sama atau saling berelasi *synonym*.

WordNet mengandung beberapa kelas kata seperti kata benda (*noun*), kata kerja (*verb*), kata sifat (*adjective*), dan kata keterangan (*adverb*). WordNet juga menyimpan informasi mengenai relasi semantik antar *synset*. Relasi yang disimpan adalah *synonymy*, *antonymy*, *hyponymy*, *hypernym*, *meronymy*, *holonymy*, *troponymy*, dan *entailment*.

2.1.1 Indonesian WordNet

Penelitian mengenai WordNet Bahasa Indonesia pernah dilakukan sebelumnya oleh Desmond Darma Putra (2008) dan Margaretha dan Manurung (2008). Indonesian WordNet (IWN) dibangun menggunakan metode mapping antara WordNet yang sudah ada ke dalam Bahasa Indonesia (Desmond Darma Putra, 2008). WordNet yang digunakan sebagai dasarnya adalah Princeton WordNet. Synset dalam PWN akan dipetakan ke dalam entry Kamus Besar Bahasa Indonesia (KBBI), sehingga menghasilkan hasil yang berkualitas baik secara cepat dan mudah. Penelitian tersebut menghasilkan 1441 synset dan 3074 sense. Relasi semantik antar synset diperoleh dengan memetakan IWN synset dengan PWN synset, sehingga relasi yang dimiliki dalam PWN dapat diturunkan.

2.1.2 WordNet Bahasa

Pengembangan WordNet untuk Bahasa Indonesia juga dilakukan oleh Nanyang Technology University (NTU) sejak tahun 2011 dan diberi nama WordNet Bahasa (Nurril Hirfana Mohamed Noor, 2011). WordNet ini telah diintegrasi dengan salah satu *tools* NLP berbasis Python yaitu nltk sehingga dapat dengan mudah digunakan dalam komputasi. WordNet Bahasa juga memanfaatkan PWN untuk mendapatkan relasi semantik antar *synset*. Pada penelitian ini, dimanfaatkan *tools* tersebut untuk mendapatkan *seed* relasi semantik antar kata dalam Bahasa Indonesia.

2.1.3 Kelemahan WordNet Bahasa Indonesia

Hingga saat ini, relasi semantik antar kata yang dimiliki oleh WordNet Bahasa Indonesia merupakan hasil turunan dari relasi semantik WordNet Princeton. Sayangnya, pemanfaatan PWN untuk mendapatkan relasi semantik kata Bahasa Indonesia menemui beberapa hambatan. WordNet Bahasa Indonesia menjadi sangat tergantung dengan struktur PWN untuk mendapatkan relas semantik suatu *synset*. Selain itu, beberapa *synset* Bahasa Indonesia juga tidak dapat dipetakan secara tepat ke *synset* PWN, beberapa kata kehilangan arti atau mendapat arti yang kurang tepat. Hal ini menyebabkan pasangan kata relasi semantik Bahasa Indonesia yang dihasilkan terlihat kurang baik. Untuk itu, dicetuskanlah penelitian untuk mengekstrak relasi semantik dalam Bahasa Indonesia secara mandiri.

2.2 Relasi

Relasi menggambarkan hubungan atau koneksi yang dimiliki oleh suatu hal dengan yang lain (KBBI). Dalam bidang matematika, relasi memetakan suatu anggota dari himpunan satu ke himpunan lain sesuai dengan hubungan yang didefinisikan. Dalam penelitian ini, relasi yang diperhatikan adalah relasi semantik antar kata. Domainnya adalah kata-kata yang tergabung dalam kelas kata benda (*noun*).

Satu relasi dapat terdiri dari beberapa entitas dan dituliskan dalam bentuk tuple $t = (e_1, e_2, ..., e_n)$ dimana e_i adalah suatu entitas yang memiliki relasi r dalam dokumen D (Bach dan Badaskar, 2007). Relasi sinonim dapat ditulis dalam notasi tersebut. Banyak pula relasi yang hanya menghubungkan antar dua entitas (relasi biner), seperti *terletak-di(Universitas Indonesia, Depok)* atau *ditulis-oleh(Habis Gelap Terbitlah Terang, RA Kartini)*.

2.2.1 Relasi Semantik

Semantik adalah arti (*sense*) dari suatu kata. Umumnya informasi tersebut disimpan dalam kamus bahasa. Relasi semantik adalah hubungan yang dimiliki antar kata berdasarkan arti atau makna dari kata tersebut. Beberapa relasi semantik adalah sebagai berikut (Miller, 1995).

- *Synonymy* adalah relasi antar kata dimana dua kata yang berbeda memiliki arti yang sama. Semua kelas kata dapat memiliki relasi *synonymy*. Dalam WordNet, relasi ini direpresentasikan dalam bentuk *synset* dan bersifat simetris. Sebagai contoh 'makan', 'melahap', dan 'menyantap' memiliki makna yang sama.

- *Antonymy* adalah yang menggambarkan arti yang saling berkebalikan antar kata. Umumnya relasi ini digunakan pada kelas kata sifat (*adverb*) dan kata keterangan (*adjective*). Sama seperti synonymy, relasi ini memiliki sifat simetris. Sebagai contoh kata 'tinggi' memiliki makna yang berkebalikan dengan kata 'pendek'.

- *Hyponymy* adalah relasi yang menyatakan hubungan kata yang lebih khusus. Sementara untuk kata yang lebih umum dikenal dengan relasi *hypernym*. Kedua relasi ini diperuntukan kelas kata benda (*noun*) dan umumnya satu kata memiliki hanya satu *hypernym*. Kedua relasi ini bersifat transitif, sehingga dapat digambarkan dalam bentuk hirarki. Sebagai contoh kucing, ikan, kelinci (*hyponymy*) adalah binatang (*hypernym*). Binatang adalah *hyponym* dari makhluk hidup. Sehingga dapat dikatakan pula bahwa kucing, ikan, kelinci (*hyponym*) adalah makhluk hidup (*hypernym*).

- *Meronymy* dan *holonym* adalah relasi yang menyatakan hubungan bagian satu dengan yang lain, dimana *meronym* menyatakan sub-bagian dan *holonym* menyatakan bagian yang lebih besar. Seperti relasi hyponym-hypernym, relasi *meronym-holonym* bersifat transitif dan dapat digambarkan dalam bentuk hirarki. Dalam WordNet, relasi ini dibagi ke dalam tiga bagian yaitu *part-meronym*, *member-meronym*, dan *substance-meronym*. Sebagai contoh sebuah sel (*holonym*) memiliki nukleus, ribosom, mitokondria (*meronym*).

- *Troponymy* adalah relasi seperti *hyponymy-hypernymy* yang khusus untuk kelas kata kerja (*verb*). Dalam Bahasa Inggris, contoh kata yang memiliki relasi ini adalah 'stroll' dan 'walk'.

DAFTAR REFERENSI

Bach, N. dan Badaskar, S. (2007). A review of relation extraction. *Literature review for Language and Statistics II*.

Desmond Darma Putra, Abdul Arfan, R. M. (2008). Building an indonesian wordnet.

Margaretha, E. dan Manurung, R. (2008). Comparing the value of latent semantic analysis on two english-to-indonesian lexical mapping tasks.

Miller, G. A. (1995). Wordnet: A lexical database for english.

Nurril Hirfana Mohamed Noor, Suerya Sapuan, F. B. (2011). Creating the open wordnet bahasa.

Ravichandran, D. dan Hovy, E. (2002). Learning surface text patterns for a question answering system.