WordNet Sebagai Basis Pengetahuan Untuk Mengatasi Polisemi Kata-Kata

Sherly Christina
Program Studi Teknik Informatika, Fakultas Teknik, Universitas Palangkaraya
Email: sherly.christina.upr@gmail.com

Abstract

Information system that accepts data in the form of natural languages often get problems such as polysemy in words. Polysemy affects performance of the system to provide accurate information. Therefore, it's needed a knowledge base that can be used to differentiate the meaning of words and find the relationship between words.

WordNet is a lexical database that is suggested by several studies to address the polysemy in words. WordNet contains the set of meanings of words and organized in semantic relations. This paper will study some researches that use Wordnet as a knowledge base to solve some ambiguity in polysemy.

Key Words: polysemy, wordnet, semantic

1. Pendahuluan

Sistem temu kembali informasi sangat rentan dengan masalah polisemi kata-kata. Polisemi kata adalah sifat bawaan dari bahasa alami. Polisemi dapat menjadi penghalang bagi kinerja sistem yang menerima input data berupa bahasa alami. Polisemi menyebabkan suatu kata memiliki beberapa makna, sehingga polisemi cenderung menimbulkan kerancuan dalam memahami makna suatu kata. Kerancuan dalam Kamus Besar Bahasa Indonesia berasal dari kata rancu yang artinya: tidak teratur, campur aduk, kacau. Sedangkan dalam bahasa Inggris kata rancu disebut dengan *ambiguous, ambiguity* (*http://glosbe.com/id/en/rancu*).

Rodd *et. al* (2004), menyatakan dalam interpretasi yang berbeda banyak kata dalam bahasa Inggris yang rancu, kata bisa memiliki makna yang berbeda. Polisemi sering terjadi karena banyaknya cara dalam mendeskripsikan suatu konsep. Contohnya dalam bahasa Inggris kata *bark* bisa berarti bagian dari sebuah pohon, atau gonggongan anjing. Kedua makna tersebut tidak berhubungan tetapi memiliki bentuk penulisan dan pengucapan yang sama.

Sistem temu kembali informasi harus dapat memilih makna yang tepat dari suatu konsep. Oleh karena itu muncul kebutuhan terhadap basis pengetahuan yang dapat digunakan untuk membedakan makna kata dan menemukan hubungan semantik antara kata. Beberapa penelitian telah mencoba menggunakan basis data Leksikal berbahasa Inggris bernama WordNet sebagai basis pengetahuan untuk mengatasi polisemi kata.

Studi literatur ini akan menjelaskan definisi WordNet, kemudian akan dibahas beberapa penelitian yang menggunakan WordNet sebagai basis pengetahuan untuk mengatasi polisemi kata. Pada akhirnya akan disimpulkan kelebihan dan kelemahan penggunaan WordNet sebagai basis pengetahuan.

2. Pembahasan

2.1. Definisi WordNet

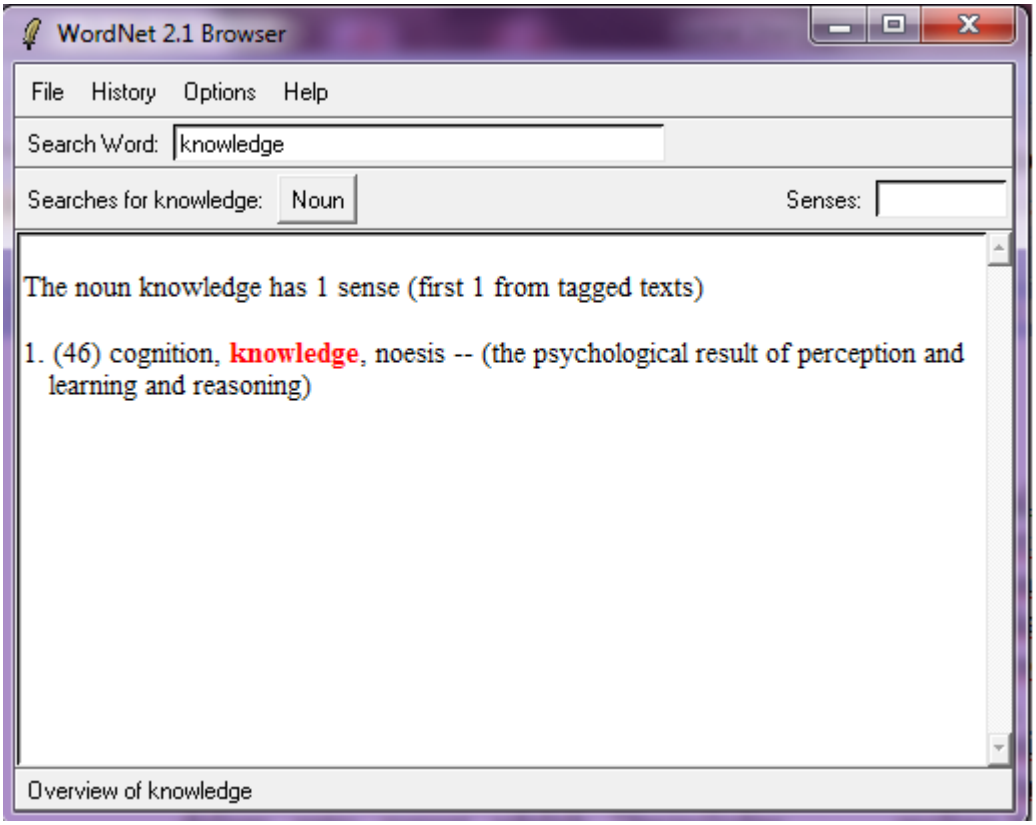
Pada WordNet, kata benda (*Noun*), kata kerja (*verb*) dan kata sifat (*adverb*) diorganisasikan ke dalam *synonym sets* (*synset*). Tiap *synset* mewakili konsep

leksikal dasar. Contoh kata yang berada dalam satu *synset* adalah "*knowledge, cognition, noesis*" yang berarti " *the psychological result of perception and learning and reasoning*". Pada Gambar 1 ditunjukkan browser Wordnet 2.1.

WordNet dikelola berdasarkan hubungan semantik. Hubungan semantik adalah hubungan antara makna yang diwakili dengan *synset* maka dapat dikatakan bahwa antar *synset* terjalin hubungan semantik. Hubungan yang terbentuk antara lain sinonim, antonim, hiponim, dan meronim Miller(1995).

Hal yang paling penting dalam WordNet adalah kemiripan makna karena penentuan hubungan antar bentuk kata merupakan syarat untuk merepresentasikan makna dalam matriks leksikal. Dua buah kata dianggap bersinonim jika penggantian kata tersebut oleh kata yang lain tidak akan merubah makna kalimat. Jarang sekali ditemukan kata yang bersinonim dan dapat digunakan dalam konteks yang sama. Kata "mobil" dan "truk" memiliki makna yang serupa tetapi keduanya memiliki perbedaan konteks penggunaan.

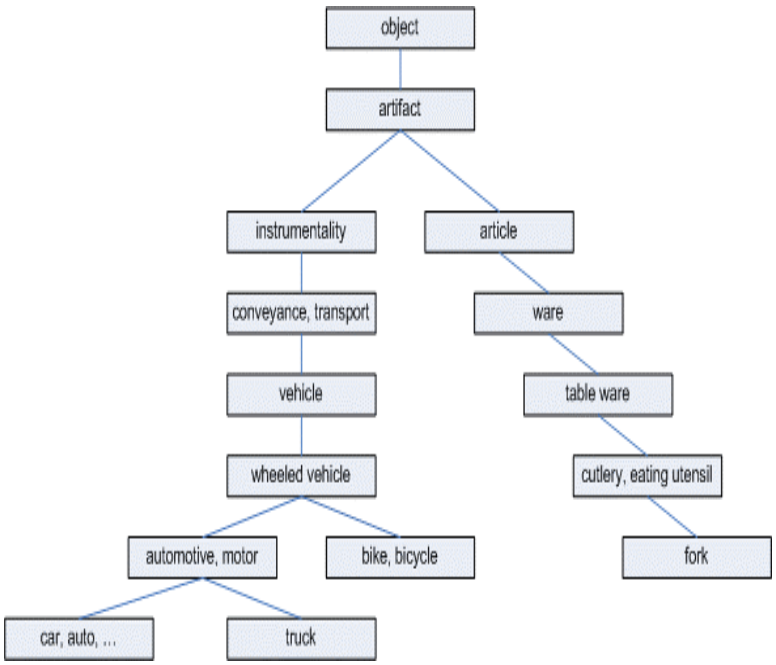
Relasi Semantik pada WordNet ditunjukkan oleh Tabel 1 Kolom *Semantic Relation* menunjukkan relasi semantik antar kata, sedangkan kolom *Syntactic Category* menunjukkan tipe *part of speech* untuk tiap kata. WordNet memiliki empat tipe *part of speech* (POS) yaitu *noun (N), verb (V), adjective (Aj),* dan *adverb (Av).*



Gambar 1. Browser wordnet 2.1

Tabel 1. Relasi semantik antar synset di wordnet Miller(1995)

Semantic Relation	**Syntactic Category**	**Examples**
Synonymy (similar)	N, V, Aj, Av	pipe, tube, rise, ascend sad, unhappy,rapidly, speedily
Antonymy (opposite)	Aj, Av, (N, V)	wet, dry powerful, powerless friendly, unfriendly, rapidly, slowly
Hyponymy (subordinate)	N	sugar maple, maple,maple, tree, tree, plant
Meronymy (part)	N	brim, hat gin, martini ship, fleet
Troponomy (manner)	V	march, walk,whisper, speak
Entailment	V	drive, ride divorce, marry
Note: N= Nouns, Aj = Adjectives, V = Verbs, Av = Adverbs		



Gambar 2. Contoh hirarki taksonomi hiponim pada wordnet (Dao, 2006)

POS *noun* dan v*erb* dalam WordNet tersusun dalam bentuk hirarki berdasarkan pada relasi hipernim/hiponim antara *synset*. Contoh hirarki hiponim pada WordNet dapat dilihat pada Gambar 2.

2.2 Penggunaan WordNet sebagai Basis Pengetahuan

Pederson *et. al* (2005) menggunakan WordNet sebagai basis pengetahuan untuk proses diambiguasi kata. Disambiguasi adalah proses menemukan makna dari suatu kata yang digunakan dalam sebuah kalimat. Proses disambiguasi mengadaptasi algoritma yang diusulkan oleh Michael Lesk

dengan menggunakan WordNet sebagai perangkat utama. Berikut adalah langkah-langkah disambiguasi yang digunakan oleh Pederson *et. al* (2005).

1. Mengakses kamus WordNet yang maknanya tersusun dalam hirarki.
2. Menerapkan mekanisme skoring untuk mengukur glosarium yang tumpang-tindih, dengan cara:
    1. Memilih konteks: Jika kalimat terdiri atas sejumlah *n* kata, maka didefinisikan *k* konteks disekitar kata target (kata target adalah kata yang mengalami disambiguasi) sebagai urutan. Jika *k* adalah empat, akan ada dua kata ke kiri dari kata target dan dua kata ke kanan dari kata target.
    2. Untuk setiap kata dalam konteks yang dipilih, dicari dan dibuat daftar semua kemungkinan makna pada kedua POS kata benda dan kata kerja.
    3. Untuk setiap makna kata (*WordSense*), dibuat daftar hubungan seperti berikut.
        - glossarium/definisi kata yg disediakan oleh WordNet
        - glossarium dari *synsets* yang terhubung melalui hubungan hipernim. Jika ada lebih dari satu hipernim untuk sebuah *sense* kata, maka glossarium untuk tiap hipernim dirangkai menjadi glosarium string tunggal.
        - glossarium dari *synsets* yang terhubung melalui hubungan hiponim
        - glossarium dari *synsets* yang terhubung melalui meronim
        - glossarium dari *synsets* yang terhubung melalui troponim
    4. Kombinasikan semua pasangan glosarium dan hitung keterkaitan dengan mencari overlapping. Skor keseluruhan adalah jumlah skor dari tiap pasangan relasi.
    5. Setelah skor setiap kombinasi diperoleh, maka diambil makna yang memiliki skor tertinggi untuk menjadi makna yang paling tepat untuk kata target dalam ruang konteks yang dipilih.

Basis data WordNet mengatur hirarki hubungan antara kata, hanya untuk kata pada POS yang sama. Sehingga tidak dapat dilakukan disambiguasi kata pada POS yang berbeda.

Penelitian berikutnya dilakukan oleh Dao dan Simpson (2006). Dao dan Simpson menghitung nilai kemiripan semantik antara kalimat-kalimat dengan menggunakan WordNet sebagai basis pengetahuan. Nilai kemiripan semantik adalah nilai eksak yang merefleksikan hubungan semantik antara makna dari dua kalimat yang mengalami polisemi.

Langkah-langkah untuk menghitung kemiripan semantik adalah sebagai berikut.

1. Tokenisasi.
   Tahap tokenisasi adalah tahap pemotongan *string* masukan berdasarkan tiap kata yang menyusunnya.
2. *Stemming*.
   Tahap *stemming* adalah tahap mencari akar kata dari tiap kata hasil tokenisasi.
3. Melakukan *speech tagging*.
   Tahap menentukan POS *noun, verb, adjective* dan *adverb*.
4. *Word Sense Disambiguation*.

Tahap disambiguasi adalah tahap mencari makna kata yang tepat.

5. Membuat *Semantic Similarity Relative Matrix R [m,n]* untuk tiap pasang *word sense*, di mana *R[i,j]* adalah kemiripan semantik antara *senses* yang paling cocok dari kata pada posisi *i* dari kalimat *X* dengan *senses* yang paling cocok dari kata pada posisi *j* dari kalimat *Y*. Jadi, *R[i,j]* adalah bobot koneksi tepi dari *i* ke *j*. Jika sebuah kata tidak ada pada kamus, maka sebagai gantinya menggunakan *Levenshtein Distance Similarity*. *Levenshtein Distance Similarity* mengukur jarak antara dua string dengan menghitung jumlah pengoperasian yang perlu dilakukan untuk mengubah string satu menjadi string dua. Pengoperasian yang dilakukan termasuk operasi operasi sisip, hapus dan substitusi.

6. Menghitung kemiripan semantik antara dua *word sense* menggunakan *Matching Average, Dice's Coefficient, Cosine Similarity, Jaccard Coefficient.* Jika waktu komputasi kritis maka digunakan rumus heuristik cepat untuk menghitung nilai kemiripan semantik.

Penggunaan WordNet sebagai basis pengetahuan untuk mengatasi polisemi kata oleh Dao dan Simpson dapat menghasilkan nilai kemiripan semantik yang lebih akurat dibandingkan cara sintatiks seperti *Levensthein Distance Similarity.* Namun karena konten WordNet yang terbatas dan koleksi makna dari kata-kata dalam WordNet yang bersifat umum, menyebabkan nilai kemiripan semantik tidak akurat untuk makna kata dalam konteks khusus.

Penelitian berikutnya dilakukan oleh Richardson et.al (1994). WordNet digunakan sebagai basis pengetahuan pada sebuah sistem temu kembali informasi. *Query* dalam bentuk bahasa alami menjadi masukan bagi sistem untuk menampilkan artikel dari 153.000 artikel. Sistem kemudian harus dapat menampilkan 1000 artikel pertama, yang paling mirip dengan kata-kata kunci query. Strategi yang digunakan adalah membandingkan setiap *term* pada query dengan indeks *term* dari suatu artikel dan mengumpulkan semua hasil perbadingan untuk menghitung skor yang paling relevan dari artikel terhadap *query.* Mekanisme yang digunakan adalah *The Information Based* dan *Conceptual Distance Semantic Similarity Estimator* yang diimplementasikan pada hirarki Wordnet.

Hasil experimen Richardson et. al menyimpulkan, penggunaan WordNet sebagai basis pengetahuan menghasilkan nilai kemiripan semantik yang lebih akurat. Walaupun ada kelemahan dalam perhitungan *conceptual distance* pada hirarki WordNet karena struktur kepadatan *link* antar konsep yang tidak teratur.

3. Kesimpulan

Beberapa penelitian tersebut telah merekomendasikan WordNet sebagai basis pengetahuan yang dapat memberi jalan keluar dari kerancuan karena polisemi kata. WordNet menyimpan makna-makna kata dan hubungan diantara kata-kata. Namun karena basis data WordNet bersifat umum, maka untuk mengatasi polisemi kata pada konteks yang lebih spesifik dibutuhkan tesaurus yang lebih relevan.

Pada penelitian yang lebih lanjut diusulkan pengembangan basis data WordNet dapat menghubungkan kata-kata pada POS yang berbeda. Dan diharapkan pengembangan WordNet berbahasa Indonesia dapat terus dikembangkan mengingat reliabilitas yang ditawarkan oleh WordNet dapat dimanfaatkan dalam sistem informasi berbahasa Indonesia.

Daftar Pustaka

1. Dao, T. N. & Simpson, T. (2006), "Measuring Similarity Between Sentences"
2. Miller, G. (1995), *WordNet: A Lexical Database for English*, Communication of The ACM, volume 38
3. Richardson R., Smeaton A. F., Murphy J., (1994), Using WordNet as a Knowledge Base for Measuring Semantic Similarity Between Words, Dublin City University, Ireland
4. Rodd J. M., Gaskel M. G., Wilson W. D. M., (2004), Modelling The Effects of Semantic Ambiguity in Word Recognition, *Cognitive Science*, *Elsevier*, No. 28, hal. 89-104
5. Pedersen T., Banerjee S., Patwardhan S., (2005), Maximize semantic relatedness to perform word sense disambiguation"
6. Pusat Bahasa Departemen Pendidikan Nasional, 2008, Kamus Besar Bahasa Indonesia, Jakarta
7. http://glosbe.com/id/en/rancu, diakses tanggal 11 juli 2012, pukul 19.01 WIB.