PENGHILANGAN AMBIGUITAS MAKNA KATA DALAM KALIMAT BERBAHASA INDONESIA DENGAN MENGGUNAKAN PARSER, WORDNET DAN ALGORITMA LESK
WORD SENSE DISAMBIGUATION IN INDONESIAN SENTENCE USE PARSER, WORDNET AND LESK ALGORITHM

Regina Malvinasrani Gitasari[1], -[2]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Tugas akhir ini bertujuan untuk membuat perangkat lunak yang dapat menghilangkan ambiguitas makna kata dalam kalimat berbahasa Indonesia. Latar belakang pembuatan adalah bahwa bahasa Indonesia sebagai bahasa alami memiliki kata yang bermakna lebih dari satu, sesuai konteks kalimat. Kata yang bermakna lebih dari satu tersebut, berpotensi menyebabkan keragu-raguan atau ambigu. Menghilangkan ambiguitas makna kata atau disebut juga word sense disambiguation dilakukan dengan memilih makna yang tepat dari kata ambigu dalam kalimat. Dalam tugas akhir ini, kata ambigu yang digunakan adalah homograf.

Pemilihan makna dilakukan dengan menggunakan teknik pennggabungan parser, wordnet, dan algoritma Lesk. Parser yang digunakan adalah PC-PATR, yang merupakan tools implementasi aturan sintak kalimat bahasa Indonesia, hasil penelitian [10].

Berdasarkan hasil pencarian terhadap berbagai macam artikel, maka kriteria pengujian terhadap kalimat dibagi menjadi 7 tipe, yaitu berdasarkan jumlah homograf dan jenis kelas kata yang dimiliki. 7 tipe kalimat tersebut dan hasil pengujiannya, yaitu: tipe 1 (kalimat yang memiliki 1 homograf, kelas kata berbeda jenis (6 kalimat)), berhasil menghilangkan ambiguitas makna sebanyak 5 kalimat; tipe 2 (kalimat yang memiliki 2 homograf sama, kelas kata berbeda jenis untuk setiap homograf (2 kalimat)), berhasil menghilangkan ambiguitas makna sebanyak 2 kalimat, tipe 3 (kalimat yang memiliki 2 homograf yang berbeda, kelas kata berbeda jenis untuk setiap homograf (3 kalimat)), berhasil menghilangkan ambiguitas makna sebanyak 2 kalimat; tipe 4 (kalimat yang memiliki 2 buah homograf yang berbeda, kelas kata berbeda jenis dan sama jenis(3 kalimat)), berhasil menghilangkan ambiguitas makna sebanyak 3 kalimat; tipe 5 (kalimat yang memiliki 1 buah homograf, kelas kata yang sama jenis(6 kalimat)), berhasil menghilangkan ambiguitas makna sebanyak 4 kalimat; tipe 6 (kalimat yang memiliki 2 homograf, kelas kata yang sama jenis (3 kalimat)), berhasil menghilangkan ambiguitas makna sebanyak 1 kalimat; tipe 7 (kalimat yang memiliki 2 homograf berbeda, kelas kata yang sama jenis(3 kalimat)), berhasil menghilangkan ambiguitas makna sebanyak 3 kalimat.

Hasil pengujian tersebut, sangat bergantung kepada kelengkapan basis data dan hasil penguraian kalimat.

Kata Kunci : word sense disambiguation; parser; wordnet;. algoritma Lesk.

Abstract

The purpose of this end task is making software of word sense disambiguation in the Indonesian sentences.

The production background is Indonesian language has word which more than one meaning, congruent with the sentence context. The word, which more than one meaning, can cause ambiguous or hesitancy. Word sense disambiguation do it by choosing appropriate meaning of ambiguous word in sentence. In this end task, ambiguous word used is homograf.

Sense elections do it by using integration parser, wordnet, and Lesk algorithm techniques. Parser which used is PC-PATR, which tools implementation of syntax regulation of Indonesian sentence, based on research [10].

Based on search to many articles, then type of sentence test divided into 7 type, which every type based on number of homograf and class type. Those 7 sentence type and test results are : type 1 (sentence which has 1 homograf, different word class type (6 sentence)), successful to word sense disambiguation 5 sentences; type 2 (sentence which has 2 same homograf, different word class type for every homograf (2 sentences)), successful to word sense disambiguation 2 sentences; type 3 (sentence which has 2 homograf different, different word class type for every homograf (3sentences)), successful to word sense disambiguation 2 sentences; type 4 (sentence which has 2 homograf different, different word class type and same word class type (3 sentences)), successful to word sense disambiguation 3 sentences; type 5 (sentence which has 1 homograf same word class type (6 sentences)), successful to word sense disambiguation 4 sentences; type 6 (sentence which has 2 same homograf , same word class type (3 sentences)), successful to word sense disambiguation 1 sentence; type 7 (sentence which has 2 different homograf, , same word class type and have different meaning (3 sentences)), successful to word sense disambiguation 3 sentence. Those testing result, depend on the data base completeness and sentence decomposition result.

Keywords : word sense disambiguation; parser; wordnet; Lesk algorithm.

1. Pendahuluan

1.1 Latar belakang

Setiap bahasa alami memiliki kata yang dapat bermakna lebih dari satu, sesuai dengan konteks kalimat yang menyertainya. Kata bermakna lebih dari satu tersebut, dapat berpotensi menyebabkan keragu-raguan atau ambigu. Usaha untuk memilih makna dari kata tersebut berdasarkan konteks kalimat disebut *word sense disambiguation*. Bahasa Indonesia sebagai salah satu bahasa alami, memiliki kata yang bermakna ambigu dan memiliki struktur kalimat yang khas. Hal tersebut yang menarik minat penulis untuk membuat aplikasi yang dapat menghilangkan ambiguitas makna kata dalam kalimat berbahasa Indonesia.

Ada beberapa pendekatan untuk menghilangkan ambiguitas makna kata dalam kalimat berbahasa Inggris, yaitu *supervised learning*, dan u*nsupervised*. Pendekatan *supervised learning* menggunakan data latih yang mengandung kumpulan besar contoh kalimat yang mengandung kata yang bermakna ambigu, setiap kata tersebut ditandai oleh manusia disertai makna dimana kata tersebut digunakan. Kemudian sekumpulan aturan secara otomatis belajar dari data latih tersebut. Sebagai contoh jika kata *dog* dan *bark* kedua-duanya ada dalam kalimat sedangakan kata *tree* tidak ada dalam kalimat tersebut, maka kata bark dapat diartikan sebagai lolongan anjing. Metode ini memiliki kelemahan yaitu tidak adanya sekumpulan aturan yang dapat melakukan penentuan makna untuk seluruh kata yang bermakna ambigu dan ketergantungan terhadap penandaan yang dilakukan oleh manusia kepada data latih, sehingga penentuan makna tidak dapat dilkukan terhadap kata yang tidak ditandai.
Pendekatan *unsupervised* menggunakan sumber informasi lain sebagai pengganti penandaan terhadap kata yang bermakna ambigu, yaitu kamus. Pendekatan ini diterapkan oleh algoritma Lesk. Algoritma ini berdasarkan intuisi bahwa kata yang bermakna ambigu yang terdapat bersamaan dalam kalimat, digunakan untuk merujuk topik yang sama dan makna yang berhubungan dengan topik tersebut didefinisikan di dalam kamus dengan menggunakan kata yang sama. Algoritma ini cocok untuk kalimat yang pendek, sedangkan untuk kalimat yang lebih panjang membutuhkan relasi antar kata dalam kalimat. Oleh karena itu maka algoritma Lesk diimplementasikan kepada semantik peristilahan yang terdapat dalam basis data atau disebut sebagai *WordNet*.

Dalam tugas akhir ini, homograf digunakan sebagai kata yang bermakna ambigu dan pendekatan yang digunakan adalah *unsupervised* karena dapat menghilangkan ambiguitas makna kata tanpa menggunakan penandaan kata yang bermakna ambigu, dilengkapi dengan alat bantu yaitu pengurai. Pengurai yang digunakan adalah hasil penelitian [10]. Pengurai tersebut menghasilkan pohon urai, kata dan kelas kata. Kelas kata dan kata hasil penguraian akan digunakan sebagai perbandingan terhadap kelas kata dari kata yang tersimpan dalam basis data. Kata yang bermakna ambigu biasanya memiliki kelas kata lebih dari satu. Jika kata yang bermakna ambigu memiliki kelas kata hanya satu buah, maka makna yang dipilih dari basis data adalah makna yang berdasarkan kelas kata yang dimilikinya. Jika kata yang bermakna ambigu memiliki kelas kata lebih dari satu, maka akan dilihat kesamaan dari kelas kata-nya. Misalnya kata yang

bermakna ambigu tersebut memiliki dua buah kelas kata, dan kedua-duanya adalah kata benda. Maka akan dilakukan proses penentuan makna dengan menggunakan algoritma Lesk yang diterapkan pada basis data semantik peristilahan atau *WordNet*. Dari penerapan algoritma Lesk akan diperoleh skor dari setiap makna yang menjadi kandidat. Kemudian dilakukan perbandingan terhadap skor – skor tersebut, dan skor terbesar yang akan dipilih sebagai acuan pemilihan makna dari makna kandidat. Jika misalnya kata yang bermakna ambigu tersebut memiliki dua buah kelas kata, yaitu kata benda dan kata sifat, maka makna yang dipilih dari basis data adalah makna yang berdasarkan kelas kata hasil penguraian.

1.2 Perumusan masalah

Titik berat tugas akhir adalah pembahasan tentang bagaimana cara menentukan makna yang harus dipilih untuk menghilangkan ambiguitas makna dari kata dalam kalimat berbahasa Indonesia.Adapun hal – hal yang mempengaruhi titik berat pembahasan, yaitu struktur kalimat bahasa Indonesia, jenis kelas kata yang akan digunakan, penguraian kalimat bahasa Indonesia, cara membaca hasil penguraian dan kamus yang akan digunakan.

Batasan masalah untuk tugas akhir ini adalah sebagai berikut :

1. Kalimat yang digunakan adalah kalimat yang sesuai dengan tata bahasa baku bahasa Indonesia
2. Kalimat yang digunakan berupa kalimat tertulis dan deklaratif.
3. Kalimat aktif dan pasif.
4. Tidak menangani frasa ambigu dan kata ambigu berimbuhan.
5. Aspek semantik peristilahan yang digunakan adalah homograf.
6. Asumsi basis data sudah ada dan lengkap.
7. Referensi kamus yang digunakan adalah Kamus Besar Bahasa Indonesia dan Kamus Lengkap Bahasa Indonesia. Penggunaan kedua kamus bertujuan untuk saling melengkapi data yang dimiliki oleh masing-masing kamus.

1.3 Tujuan

Penelitian ini bertujuan untuk membuat suatu aplikasi penghilangan ambiguitas makna kata dengan cara menggunakan pengurai, *WordNet* dan algoritma Lesk.

1.4 Metodologi penyelesaian masalah

Metode yang digunakan dalam penyelesaian tugas akhir ini adalah :

1. Mempelajari struktur kalimat dalam tata bahasa baku bahasa Indonesia,mempelajari teori pengurai, mempelajari jurnal-jurnal yang berkaitan dengan penghilangan ambiguitas makna kata dalam kalimat. Mempelajari cara kerja PC-PATR dan teori yang diimplementasikan oleh PC-PATR
2. Mengumpulkan kalimat – kalimat yang akan digunakan sebagai kasus uji.
3 Mengumpulkan kata - kata yang termasuk homograf beserta makna dan kelas kata dari homograf tersebut.
4 Mengumpulkan kata – kata yang biasanya berhubungan dengan homograf.

5 Mengumpulkan kata – kata yang biasanya berelasi dengan kata – kata yang berhubungan dengan homograf guna membangun *WordNet*.
6. Menganalisa algoritma Lesk dan menerapkannya ke dalam *WordNet*.
7. Merancang basis data, merancang antar muka.
8. Melakukan implementasi dari perancangan menggunakan PHP dan MySQL.
9. Melakukan pengujian hasil implementasi dan pengujian terhadap kalimat.
10. Evaluasi terhadap hasil pengujian

5. Penutup

5.1 Kesimpulan

1. Perangkat Lunak Penghilangan Ambiguitas Makna Kata dalam Kalimat Berbahasa Indonesia dengan menggunakan Pengurai, WordNet dan Algoritma Lesk telah berhasil dibuat sesuai dengan spesifikasi yang ditetapkan.
2. Perangkat lunak tidak dapat menentukan makna kata, jika skor yang dihasilkan sama besar, hal tersebut terjadi karena algoritma Lesk merujuk skor terbesar sebagai acuan pemilihan makna
3. Perangkat lunak tidak dapat menentukan makna kata jika tidak menemukan kata yang berelasi dengan kata dalam kalimat, hal tersebut terjadi karena basis data tidak lengkap dan hasil penguraian yang tidak sesuai dengan isi basis data
4. Perangkat lunak berhasil menentukan makna kata: jika menghasilkan sebuah skor terbesar, jika kata yang berelasi dengan kata dalam kalimat sesuai dengan isi basis data, dan jika hasil penguraian sama dengan isi basis data.
5. Kelengkapan basis data berpengaruh terhadap besar skor yang akan dihasilkan, sehingga mempengaruhi pemilihan makna.

5.2 Saran

1 Diharapkan untuk pengembangan lebih lanjut, dapat melakukan penghilangan ambiguitas makna bila skor tertinggi yang dihasilkan lebih dari satu, yaitu dengan cara memodifikasi algoritma Lesk.
2. Untuk pengembangan lebih lanjut, dapat dilengkapi dengan kata ambigu berupa frasa ambigu, kata berimbuhan yang ambigu dan semantik peristilahan selain homograf.
3. Hasil penguraian dapat juga menyebabkan tidak terdapatnya kata dari kalimat yang berelasi dengan kata yang berelasi dengan homograf., karena perbedaan pemisahan kata seperti yang terjadi pada pengujian butir 4.3.2.6. Untuk pengembangan selanjutnya, perlu diperbaiki aturan sintak yang diterapkan pada PC-PATR
4. Untuk pengembangan terhadap kalimat dalam satu paragraf, diperlukan pengurai lain sebagai pengganti.

6. Daftar Pustaka

[1] Aho,Alfred V. Ravi Sethi, Jeffery D. Ullman, 1988 , *Compilers Principles, Techniques and Tools* , America, Addison Wesley.

[2] Allen, J. 1994, *Natural Language Understanding,* America, The Benjamin/Cumming Publishing Company Inc.

[3] Alwi, H ; Soenjono Darhjowidjojo, Hans Lapoliwa, Anton M. Moeliono, 2003, *Tata Bahasa Baku Bahasa Indonesia* , Jakarta, Departemen Pendidikan dan Kebudayaan Republik Indonesia

[4] Banerjee,S. , 2002, *Adapting the Lesk Algorithm for Word Sense Disambiguation to WordNet* , http://www.d.umn.edu/~tpederse/Pubs/ banerjee.pdf didownload pada tanggal 29 November 2005.

[5] Ekedahl, Jonas , Koraljka Golub, 2004, *Word Sense Disambiguation Using WordNet and the Lesk Algorithm* ,www.cs.lth.se/EDA171/ Reports/2004/jonas/_koraljka.pdf didownload pada tanggal 29 November 2006

[6] Fajri, E.Z. , Ratu Aprilia Senja, 2002, *Kamus Lengkap Bahasa Indonesia* ,Yogyakarta Difa e

[7] Gelbukh,Alexander .,Grigori Sidorov, San-Yong Han, 2003 , *Evolutionary Approach to Natural Language Word Sense Disambiguation through Global Coherence Optimization* , http://www.data.cicling.org/lab/Publications/2003/WSI-GA.pdf didownload pada tanggal 29 November 2006

[8] Haryanto, Steven 2003 , *Regex* , Jakarta ,Dian Rakyat

[9] Ide, Nancy , Jean Veronis, 1998, *Word Sense Disambiguation: The State of the Art* , New York , www.up.univ-mrs.fr/ ~veronis/pdf/ 1998wsd.pdf, didownload tanggal 29 November 2006.

[10] Joice, 2002, *Pengembangan Lanjut Pengurai Struktur Kalimat Bahasa Indonesia Yang Menggunakan Constraint-Based-Formalism*, Jakarta, Fakultas Ilmu Komputer Universitas Indonesia

[11] Pateda, Mansoer 2001, *Semantik Leksikal*, Jakarta, Rineka Cipta

[12] PHP Documentation Group, 2005 , *PHP Manual* , www.php.net, didownload tanggal 29 November 2006

[13] Suyanto, 2002 , *Intelijensia Buatan*,Bandung, Jurusan Teknik Informatika STT Telkom

[14] Tim Penyusun Kamus Pusat Pembinaan dan Pengembangan Bahasa, 1997, *Kamus Besar Bahasa Indonesia Edisi Kedua*, Jakarta, Balai Pustaka

[15] http://www.e-psikologi.com/dewasa/210803.htm didownload pada tanggal 13 Januari 2007

[16] http://www.gizi.net/cgi-bin/berita/fullnews.cgi?newsid1052189868,77383, didownload pada tanggal 13 Januari 2007

[17] http://www.inmypad.com/2006/09/karakter-berdasarkan-inisial-nama/ didownload pada tanggal 13 Januari 2007

[18] http://io.ppi-jepang.org/article.php?id=14 didownload pada tanggal 13

Januari 2007
[19] http://ms.wikipedia.org/wiki/ didownload pada tanggal 13 Januari 2007
[20] http://www.sedap-sekejap.com/artikel/2001/edisi11/files/reka.htm
didownload pada tanggal 13 Januari 2007