

# APLIKASI WORDNET INDONESIA BERDASARKAN KAMUS THESAURUS BAHASA INDONESIA MENGGUNAKAN ALGORITMA RULE BASED TEXT PARSING

Dzulfie Zamzami <sup>1</sup>, Dr.Eng.Faisal Rahutomo,ST.,M.Kom <sup>2</sup>, Dwi Puspitasari, S.Kom., M.Kom. <sup>3</sup>

<sup>1,2</sup>Jurusan Teknologi Informasi, Program Studi Teknik Informatika, Politeknik Negeri Malang

<sup>1</sup>[dzulfiez@gmail.com](mailto:dzulfiez@gmail.com), <sup>2</sup> [faisal.polinema@gmail.com](mailto:faisal.polinema@gmail.com), <sup>3</sup> [dwi\\_sti@gmail.com](mailto:dwi_sti@gmail.com)

---

## Abstrak

WordNet adalah database bahasa yang digunakan untuk mencari *synonym set* (*synset*) pada sebuah kata (lema) yang nantinya akan berelasi dari satu lema dengan lema lainnya. Untuk pembuatan WordNet Bahasa Indonesia dari Kamus Tesaurus Bahasa Indonesia yang memang sebelumnya belum pernah dibuat. Algoritma metode yang dipakai untuk pembuatan WordNet Bahasa Indonesia ini adalah *Rule Based Text Parsing*, yang berfungsi untuk mengurai semua tanda baca yang ada dalam Kamus Tesaurus Bahasa Indonesia dan pada akhirnya akan masuk kedalam *database* dan menunjukkan relasi antar *synset*. WordNet yang berupa kesinoniman kata di Indonesia masih belum ada, untuk menekan kelestarian kata Bahasa Indonesia, maka dibentuklah Aplikasi Wordnet Indonesia Berdasarkan Kamus Tesaurus Bahasa Indonesia Menggunakan Algoritma Rule Based Text Parsing. Penelitian ini bertujuan untuk membentuk WordNet Indonesia dengan cara *parsing* Kamus Tesaurus Indonesia menggunakan algoritma metode *Rule Based Text Parsing* yang didasarkan dengan teori dan referensi yang telah ada.

**Kata kunci : WordNet, Indonesia, Synonym Set.**

---

## 1. Pendahuluan

WordNet adalah hasil proyek penelitian di Princeton University yang bertujuan untuk memodelkan pengetahuan leksikal pembicara asli bahasa inggris. Informasi di dalam WordNet diorganisasikan ke dalam kelompok logikal yang disebut *synset*. Tiap-tiap *synset* berisikan bentuk *synonym* kata dan pointer semantik yang menjelaskan hubungan antara satu *synset* dengan *synset* lainnya (Miller,1993). WordNet adalah basis data leksikal *online* yang menyediakan tempat penyimpanan leksikal bahasa inggris. WordNet didesain untuk menyediakan hubungan antara empat label kelas bagian perkataan (*Parts of Speech*, POS) – *noun*, *verb*, *adjective* dan *adverb*. Unit terkecil WordNet adalah *synset* yang merepresentasikan makna spesifik sebuah kata. Ia mengandung kata, penjelasannya, dan *synonym*. Makna spesifik satu kata di satu label kelas POS disebut *sense*. Tiap-tiap *sense* sebuah kata berada di dalam *synset* yang berbeda. *Synset* sebanding dengan *sense*, yaitu struktur yang mengandung sekumpulan *term* dengan makna *synonym*. Manfaat dari WordNet itu sendiri merupakan tempat dari basis pengetahuan untuk memberikan penjelasan yang lebih spesifik kepada penggunaannya agar tidak mengalami kerancuan atau menjauh dari arti/makna kata sebenarnya.

Di Indonesia basis data seperti WordNet ini masih belum ada. Untuk pencarian makna yang sama dalam sebuah kata yang ada di Indonesia masih menggunakan kamus konvensional.

Penggunaan WordNet di Indonesia masih belum dirasakan oleh warga di Indonesia yang ingin menggunakan WordNet sebagai basis data pengetahuan. Di Indonesia terdapat banyak sekali bahasa dan memiliki arti yang berbeda ataupun serupa. Untuk memanfaatkan perkembangan teknologi yang semakin maju, maka WordNet Indonesia diciptakan dengan penelitian ini. Karena untuk pencarian sinonim sebuah kata dalam kamus konvensional yang telah ada akan memakan waktu lama saat membuka lembar demi lembar dalam pencariannya. Sedangkan dengan adanya WordNet Indonesia, maka pengguna hanya harus mengetik kata yang ingin dicari sinonimnya. Keuntungan dalam penelitian ini berdampak sangat positif bagi pengguna khususnya warga di Indonesia.

Penelitian ini dibentuk dengan memasukkan data dari isi Kamus Tesaurus Bahasa Indonesia dengan tujuan membentuk WordNet Indonesia yang merupakan basis data yang berisikan stuktur terkecil atau biasa disebut *synset* (*synonym set*) yang berfungsi sebagai pencarian atau pencocokan simantik (makna) yang belum ada sebelumnya. Keuntungan dari penelitian ini berupa penghematan waktu bagi pengguna untuk mencari atau mencocokkan makna kata yang belum atau tidak diketahui sebelumnya. Penelitian ini juga dapat berfungsi sebagai media pembelajaran bagi siswa/i, mahasiswa/i, pengajar ataupun khalayak umum yang pastinya lebih efektif dan efisien dibandingkan kamus konvensional yang berupa buku tebal. Dilihat dari segi ekonomis/biaya, maka penelitian sangat

berguna karena tidak memakan biaya dibandingkan Kamus Thesaurus Bahasa Indonesia lainnya yang masih konvensional dan tidak membutuhkan ruang penyimpanan yang banyak.

## 2. Metodologi Penelitian

Metode penelitian yang digunakan adalah metode *Rule Based Text Parsing*. *Rule Based Text Parsing* merupakan salah satu metode untuk memarsing data menggunakan aturan dasar yang telah dibuat sebelumnya.

### 2.1 Konsep

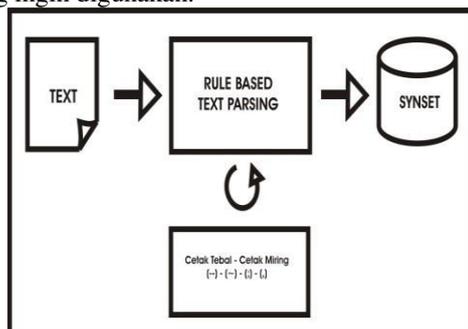
Penelitian ini memiliki konsep yang sederhana, yaitu dengan mengurai isi dari Kamus Tesaurus Bahasa Indonesia. Namun sebelum adanya proses penguraian isi Kamus Tesaurus Bahasa Indonesia. Penulis memberikan beberapa simbol-simbol untuk beberapa bagian pada isi Kamus Tesaurus Bahasa Indonesia. Seperti penambahan simbol (\*) untuk lema yang bercetak tebal, simbol (#) untuk tipe kata yang menjelaskan apakah dalam satu *synset* merupakan noun, adverb, verb dan lain sebagainya.

Penelitian ini bertujuan untuk membentuk Wornet Indonesia dengan cara mengubah Kamus Tesaurus Indonesia menggunakan metode *Rule Based Text Parsing* yang didasarkan dengan teori dan referensi yang telah ada.

### 2.2 Gambaran Aplikasi

Penggunaan penelitian ini berfungsi untuk mencari atau mencocokkan kesamaan arti kata tersebut digunakan untuk membuat kalimat-kalimat yang pada nantinya akan difungsikan sebagai kalimat yang mempunyai arti kata yang lebih detail atau lebih spesifik. Dalam menciptakan bahasa yang baku dan benar maka dibutuhkan kata-kata yang benar untuk memberikan makna yang sebenarnya. Sebagai contoh; “Andi memberi *aba-aba* tugas-tugas untuk anggotanya”. Dari kalimat berikut, *aba-aba* mempunyai arti kata (sinonim) seperti berikut ; “Andi *mengarahkan* tugas-tugas untuk anggotanya”.

Selain berguna untuk mencari kata sinonim sebagai mana mestinya, penelitian ini juga berfungsi untuk melestarikan berbagai kosa kata yang hampir punah. Penelitian berguna mencari setiap sinonim kata dengan hanya menginputkan setiap kata yang ingin dicari dan akan muncul setiap sinonim kata yang ingin digunakan.

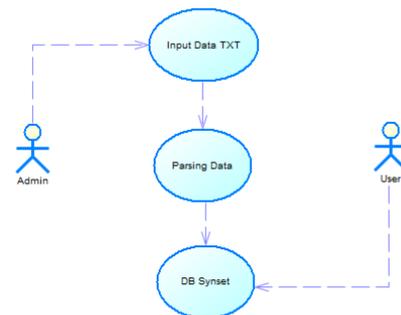


Gambar 1. Gambaran Aplikasi Wordnet Indonesia

Dari gambaran umum diatas dapat dijelaskan bahwa penelitian ini memiliki langkah-langkah sebagai berikut:

1. Kamus thesaurus yang berbentuk file pdf berisi 768 halaman.
2. Ekstraksi lema dan sublema dari kamus thesaurus satu persatu.
3. Memasukan algoritma untuk *Rule Based Text Parsing* dalam pembacaan aturan dasar baca kamus thesaurus.
4. Hasil dari semua langkah diatas adalah *synset* yang menghubungkan arti kata dari satu kata dengan kata yang lainnya.

Dalam penelitian ini terdapat 2 aktor yang masing-masing mempunyai fungsi berbeda. Actor pertama adalah admin, tugas admin disini untuk mengedit data dan membuat program agar isi dari Kamus Thesaurus bisa diekstraksi dan dinikmati oleh umum. Actor kedua adalah user, user hanya bertugas untuk mencari sinonim dari sebuah lema yang pada akhirnya user mendapat *synset* dari aplikasi tersebut. User dalam penelitian ini tidak mempunyai batasan umur, jabatan, kelamin dan lain sebagainya. Agar mempermudah dalam menjelaskan, peneliti membuat *Use Case* untuk ekstraksi Kamus Thesaurus Bahasa Indonesia untuk membentuk WordNet Indonesia dengan metode *ruled base text parsing*.



Gambar 2. Use Case Diagram WordNet

*Capture* diatas menjelaskan tahapan yang dilakukan admin dalam *input* data untuk membentuk *database* yang pada akhirnya dapat diakses oleh user. Namun, tidak sesederhana seperti *input* data yang lain, sebelum melakukan proses *input*, admin juga melakukan proses editing dari isi Kamus Thesaurus.

### 2.3 Perancangan

Perancangan pada penelitian ini dilakukan dengan cara dilakukan dengan beberapa tahapan, antara lain :

1. Mengubah kamus yang sebelumnya berformat .PDF menjadi data TXT per-abad.
2. Penambahan simbol pada bagian tertentu di isi kamus disetiap abad.

3. Pembuatan kode untuk mengurai setiap tanda pada isi setiap Kamus Tesaurus Bahasa Indonesia yang nantinya akan menjadi *synset* setiap lema dan akan berelasi antara satu lema dengan lema lainnya.
4. Desain *database* yang nantinya akan menjadi inti penelitian ini yaitu menciptakan bank bahasa Indonesia dengan harapan berguna dengan baik untuk khalayak umum dan menjadi salah satu aplikasi yang menuturkan Bahasa Indonesia sebagai pusat bahasa.



Gambar 3. Interface WordNet Indonesia.

## 2.4 Pembuatan

Proses pembuatan pertama pada penelitian ini adalah pembuatan kode untuk mengurai setiap data yang telah ditentukan sebelumnya antara cetak tebal, cetak miring dan isi dari sinonim itu sendiri. Dalam proses *parsing* data yang telah diubah menjadi format TXT akan terurai sesuai ketentuannya. Dengan begitu setelah proses *parsing* akan dilanjutkan dengan *insert* kedalam *database* yang telah disiapkan.

Algoritma yang dibuat untuk WordNet Bahasa Indonesia, memiliki tahapan sebagai berikut :

- 1 Mengambil dan membaca data pada folder TXT yang telah dibuat sebelumnya (C:\xampp\htdocs\test\_wordnet\TXT),
- 2 Mengganti karakter spasi dengan tanda baca (;), hal ini dimaksudkan untuk mendapatkan token-token pada setiap synset,
- 3 Membaca setiap tanda baca dimulai dari kiri ke kanan dengan cara membandingkan antara token ke-1 atau  $A_0$  sampai dengan token ke-n atau  $A_n$  sehingga bertemu tanda titik koma (;),
- 4 Jika saat proses perbandingan bertemu dengan tanda baca bintang (\*) maka kata dengan tanda bintang (\*) pada awalan kata adalah sebuah lema yang akan masuk ke dalam tabel lema,
- 5 Jika tidak ada tanda bintang (\*), maka tanda baca pagar (#) pada awalan label kelas kata akan masuk ke dalam tabel synset mengikuti sebuah lema yang berada disisi kirinya,
- 6 Menghapus spasi yang masih ada, karena spasi antara tanda baca koma (,) dan kata yang akan menjadi sinonim dapat menjadi sinonim baru yang hanya berisi spasi,

Jika tidak ada tanda baca bintang (\*) dan pagar (#), program akan membaca tanda baca koma (,) sebagai sinonim yang akan masuk ke dalam tabel sinonim sampai bertemu dengan tanda titik koma (;).

Selain membuat kode program proses *parsing*, dibutuhkan pula *interface* bagi pengguna yang ingin menjalankan aplikasi ini. Gambar dibawah ini akan memvisualisasikan *interfaces* untuk WordNet Indonesia.

### 3. Uji Coba

Dalam pengujian *error* dapat ditentukan dengan mengambil 52 sampel dari semua isi WordNet Indonesia yang masing-masing sampel berisi 2 synset yang terdiri dari abjad A sampai Z. Dari hasil yang keluar dari semua sampel,, dicocokkan kembali secara manual dengan isi synset yang berada pada data asli berformat .TXT. Menghitung hasil dari pencarian dengan dibagi dari semua sampel yang ditentukan.

Dari analisa secara manual yang mengambil 52 sampel dari 2 lema yang mewakili semua lema setiap abjad, maka didapatkan hasil untuk menjelaskan pengujian penelitian ini. Keberhasilan penelitian ini dapat dihitung menggunakan rumus :

$$\frac{71}{52} \cdot 100\%$$

Gambar 4. Hasil *Error* Pada Tanda Koma (,)

Dengan keterangan

N : data yang *error* perbandingan manual

52 : semua sampel

Maka hasil akhir dari perhitungan manual untuk mendapatkan sinonim dari sebuah lema adalah :

$$\frac{5}{52} \cdot 100\% = 10\% \text{ Kesalahan}$$

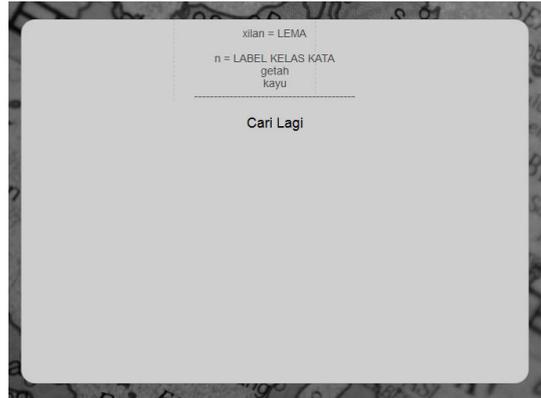
$$\frac{47}{52} \cdot 100\% = 90\% \text{ Keberhasilan}$$

Perhitungan diatas menjelaskan bahwa *error* yang didapat sekitar 5%. Ada sebab mengapa terjadinya *Error*, antara lain:

Terdapat kata gabungan yang sebenarnya adalah satu makna, namun program hanya dapat membaca bahwa kata gabungan tersebut adalah 2 lema. Hal ini dikarenakan program menggunakan variabel `preg_replace('/\s/', ';')` yang dimana setiap ada spasi dalam kata gabungan akan digantikan titik koma (;). Titik koma itu sendiri adalah batas antar synset.

```
*xilan #n getah kayu;
*xilem #n pembuluh kayu;
```

Gambar 5. *Error* WordNet Indonesia



Gambar 6. Hasil *Error* WordNet Indonesia

Selain melakukan uji coba manual untuk memeriksa relasi *synset*. Dilakukan pula uji coba secara langsung kepada 10 responden untuk memakai WordNet Bahasa Indonesia ini dan mendapatkan hasil sebagai berikut :

$$\frac{47}{50} \cdot 100\% = 94\% \text{ Mendukung}$$

$$\frac{3}{50} \cdot 100\% = 6\% \text{ Kurang Mendukung}$$

### 4. Kesimpulan dan Saran

Solusi yang ditawarkan dari penelitian ini dapat menemukan sinonim dari sebuah kata Bahasa Indonesia dan aplikasi yang dibangun pada artikel ini dapat membantu pengguna mencari dan menemukan sinonim set dari sebuah kata dengan menggunakan kata kunci.

Namun aplikasi ini belum sempurna, maka dari itu Materi yang digunakan hanya berupa sinonim. Jadi untuk pengembangannya diharapkan memberikan materi yang lebih luas serta sinonim set yang telah dibuat dapat digabungkan dengan sinonim set yang lain sehingga memberikan informasi yang lebih banyak.

### Daftar Pustaka:

- Dang Tuan Nguyen, Khoa Dang Nguyen, Ha Thanh Le. 2013. *Semantic Parsing Of Simple Sentences In Unification-Based Vietnamese Grammar*. International Journal on Natural Language Computing (IJNLC).
- Wayan Simri Wicaksana, Linta\g Yuniar, Lily Wulandari. 2005. *Pentingnya Peranan Bahasa dalam Interoperabilitas Informasi berbasisan Komputer karena Keragaman Semantik*. Latest Version Available : [http://ftp.gunadarma.ac.id/research/WorkGroupInformationSystem/DissertationS3IT\\_Gundarma\\_Wayan/MyPublication/2005-02\\_PESAT05\\_Pentingnya\\_IWS.pdf](http://ftp.gunadarma.ac.id/research/WorkGroupInformationSystem/DissertationS3IT_Gundarma_Wayan/MyPublication/2005-02_PESAT05_Pentingnya_IWS.pdf). [14 April 2016]
- Wicaksana, Wayan Simri. 2006. *Membandingkan Pendekatan Latent Semantic terhadap WordNet untuk Semantic Similarity*. Latest Version Available : [http://iwayan.staff.gunadarma.ac.id/Publications/files/708/2006\\_Kommit\\_LatentSemantic\\_IWS.pdf](http://iwayan.staff.gunadarma.ac.id/Publications/files/708/2006_Kommit_LatentSemantic_IWS.pdf). [14 April 2016]