Pembangunan Sumber Daya Kamus Sinonim Bahasa Indonesia

Januar Saputra
Teknik Informatika
Universitas Komputer Indonesia
Bandung, Indonesia
januarsaputra1591@gmail.com

Ken Kinanti Purnamasari
Teknik Informatika
Universitas Komputer Indonesia
Bandung, Indonesia
guinev_gs@yahoo.co.nz

Sinonim adalah suatu kata yang memiliki bentuk yang berbeda namun memiliki arti atau pengertian yang sama atau mirip. Keberadaan sumber daya kamus sinonim saat ini masih dalam bentuk fisik atau buku dan belum berupa database yang dapat diakses secara digital atau opensource. Ditambah lagi belum adanya kamus sinonim digital yang hasil keluaran nya dapat diunduh secara gratis dan legal, dengan kata lain pengguna hanya memakai dan tidak mempunyai hak akses untuk mengubah atau menambahkan. Keberadaan sumber daya kamus bagi pengembang dibutuhkan, karena selain membantu dalam proses pengembangan kamus online itu sendiri sumber daya kamus digital juga dapat dimanfaatkan pada penelitian yang berhubungan dengan NLP.Sumber daya kamus sinonim sangat dibutuhkan sebagai data masukan dalam pengklasifikasian arti kata dalam proses disambiguation, Analis sentiment dan WordNet..

Dalam penelitian ini sumber masukan yang dikelola adalah kamus tesaurus bahasa indonesia format .txt. Tahap pendeteksian jeniskata dan sinonim terdiri dari lima langkah yaitu menggabungkan kata yang terpisah, menormalisasikan noise, menyamakan simbol majemuk, memisahkan baris lema dan sublema, dan ploting jeniskata dan sinonim sehingga menghasilkan format kata diikuti jenis kata dan sinonim selanjutnya disimpan kedalam database.

Hasil akhir dari pembangunan kamus sinonim adalah database kamus sinonim dalam tiga format dan kamus online sebagai implementasi. Pembangunansumber daya kamus sinonim menghasilkan 33.379 lema dengan persentase 49.984 %..

Kata kunci : Kamus sinonim, kamus online,Sumber daya NLP.

I. PENDAHULUAN

Sinonim adalah suatu kata yang memiliki bentuk yang berbeda namun memiliki arti atau pengertian yang sama atau mirip. Sinomin sendiri dapat juga disebut sebagai persamaan kata yang merupakan golongan kata dalam satuan bahasa berdasarkan kategori bentuk fungsi dan makna dalam susunan sistem gramatikal[1]. Pada penggunaannya sinonim memiliki kategori kata seperti kata yang lebih umum, intensitas makna, makna emotif, umum dan teknis, juga perbedaan tingkat kebakuan.

Keberadaan sumber daya kamus sinonim saat ini masih dalam bentuk fisik atau buku dan belum berupa *database* yang dapat diakses secara digital atau *opensource.* Ditambah lagi belum adanya kamus sinonim digital yang hasil keluaran nya dapat diunduh secara gratis dan legal, dengan kata lain pengguna hanya memakai dan tidak mempunyai hak akses untuk mengubah atau menambahkan. Keberadaan sumber daya kamus bagi pengembang dibutuhkan, karena selain membantu dalam proses pengembangan kamus online itu sendiri sumber daya kamus digital juga dapat dimanfaatkan pada penelitian yang berhubungan dengan NLP.Sumber daya kamus sinonim sangat dibutuhkan sebagai data masukan dalam pengklasifikasian arti kata dalam proses *disambiguation*, *Analis sentiment* dan *WordNet*.

Berdasarkan pada kebutuhan sumber daya kamus yang bersifat *open source* khususnya bahasa Indonesia, maka pembangunan *kamus sinonim* sebagai sumber daya NLP diharapkan dapat menjadi salah satu sumber daya yang dapat dipakai dalam pengembangan teknologi di bidang bahasa. Pembangunan *kamus sinonim* dilakukan dengan pendekatan bidang NLP. Maksud dari penelitian ini adalah implementasi pembangunan kamus sinonim digital sebagai sumber daya NLP Bahasa Indonesia. Adapun tujuannya adalah menyediakan sumber daya kamus sinonim digital bagi penelitian NLP Bahasa Indonesia bersifat *open source* dengan beberapa format seperti sql *.txt, .csv,* dan *.json*.

II. ISI PENELITIAN

Data yang digunakan untuk pemabangunan kamus sinonim Bahasa Indonesia menggunakan kamus tesaurus badan bahasa Indonesia berformat .txt.

- *Tesaurus*

Tesaurus adalah kamus Bahasa Indonesia yang menyajikan sebanyak mungkin kata dan kelompok kata serta sinonim atau padanan katanya dalam bahasaIndonesia. Dalam bahasa apa pun, tidak ada kesamaan makna yang mutlak. Karena itu, sinonim di dalam Tesaurus Indonesia disajikan dengan mempertimbangkan perbedaan-perbedaan yang ada berdasarkan dialek regional[2].

- *Sinonim*

Sinonim berasal dari bahasa Yunani yaitu syn ‘dengan’ dan ononom ‘nama’. Walaupun sinonim memiliki makna sama, akan tetapi persamaan makna itu hanya merupakan kemiripan atau kurang lebih sama. Jadi meskipun maknanya sama tetapi tetap memperlihatkan perbedaan-perbedaan, apalagi jika dihubungkan dengan pemakaian kata tersebut dalam kalimat. Kesamaan makna kata dalam bersinonim tidak mutlak seratus persen, tetapi masih terdapat perbedaan. Perbedaan makna dapat dilihat dengan memperhatikan dua hal yaitu makna dasar dengan makna tambahan dan nilai rasanya[1].

2.1 Identifikasi Masalah

Identifikasi masalah menurut latar belakang di atas adalah belum adanya sumber daya kamus sinonim digital bahasa Indonesia yang bersifat *open source*

2.2 Maksud dan Tujuan

Maksud dari penelitian ini adalah membangun kamus *digital* sinonim Bahasa Indonesia. Adapun tujuannya adalah menyediakan sumber daya kamus sinonim digital Bahasa Indonesia yang yang bersifat *open source.*

2.3 Batasan Masalah

Agar permasalahan yang dirumuskan dapat lebih fokus, maka dalam penelitian ini dibatasi dalam hal-hal sebagai berikut.

- Data yang diolah dalam membangun kamus digital sinonim adalah sekumpulan kata yang diambil dari Kamus Tesaurus Bahasa Indonesia Pusat Bahasa dengan format .txt.
- Proses pembangunan kamus digital sinonim yang dibangun hanya mengolah kata sinonim dan turunannya yang berimbuhan.
- Keluaran dari pembangunan sistem adalah kamus online sinonim dengan yang dapat diunduh dengan format .sql .*txt*, *.csv*, dan *.json.*

III. METODE DAN IMPLEMENTASI

Jenis penelitian bersifat deskriptif merupakan penelitian yang mempelajari masalah-masalah yang ada serta tata cara kerja yang berlaku. Tujuan dari penelitian deskriptif adalah untuk membuat deskipsi, gambaran atau lukisan secara sistematis, faktual dan akurat mengenai fakta-fakta, sifat-sifat serta hubungan antar fenomena yang diselidiki. Pada metode *rulesbased* ini peneliti menghimpun data, menyusunnya secara sistematis, faktual dan cermat. *Rulesbased Metode* penelitian ini memiliki dua tahapan yaitu pengumpulan data dan pembangunan perangkat lunak. Peneliti menggunakan metode ini karena sesuai dengan data dan hasil analisis pada penelitan sebelumnya. Pengembangan kamus sinonim yang akan dibangun berdasarkan pada fakta-fakta yang didapatkan pada saat analisis dan pengumpulan data.

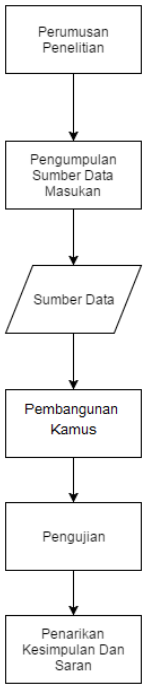


Fig. 1 Skema Metode Penelitian

3.1 Analisis Solusi

Dengan kata lain pada tahapan ini bertujuan untuk mengetahui mekanisme atau prosedur kerja dari proses yang akan dibangun dalam mengimplementasikan teknik pengumpulan data berupa sumber daya kamus sinonim.

Pada analisis solusi, sistem pembangunan kamus sinonim tahapan yang akan dijalankan memiliki tahapan pengambilan kata sinonim dan jenis kata. Penjelasan setiap tahapnya dapat dilihat berikut di bawah ini

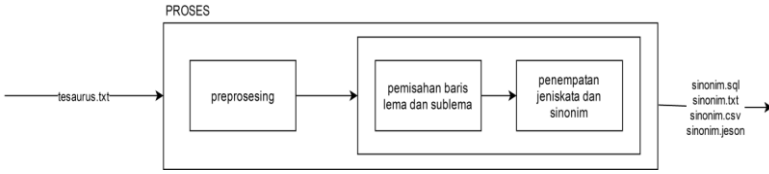


Fig. 2 Gambaran Tahapan Utama Sistem

Pada gambar 2 Gambaran Tahapan Utama Sistem akan dijelaskan sebagai berikut ini.

- Pada tahap awal user memasukan file teks yang merupakan Kamus Tesaurus Bahasa Indonesia yang berformat .txt ke dalam sistem

- Pada tahap preprocesing dilakukan proses yang perapihan kamus sebagai pendukung proses selanjutnya.
- Selanjutnya, proses pendeteksian jeniskata dan sinonim sebagai tahap akhir dimana diolah dengan mengambil kata,jenis kata dan kata sinonim sebagai contoh: *aba-aba n isyarat, aba-aba n petunjuk, aba-aba n kode*.
- Tahap terakhir yang dilakukan adalah konversi Format *database* menjadi beberapa format diantaranya sql .txt, .csv, .dan .json.

3.2 Analsis Data Masukan

Analisis data masukan pada sistem pembangunan sumber daya kamus sinonim yaitu menjelaskan proses data masukan berupa data file teks yang diterima dan akan diproses oleh sistem. Data masukan dalam pembangunan sumber daya kamus sinonim yang didapatkan dari Kamus Tesaurus Bahasa Indonesia dengan format *.txt*

3.3 Analisis Proses

Pendekatan sistematis dalam mengidentifikasi permasalahan, peluang, dan tujuan dalam merancang suatu sistem. Analisis proses yang dilakukan adalah pendeteksian kata sinonim sebagai proses utama dan konversi data.

3.4 Pengambilan Kata Kamus Pada Sinonim

Tahapan pengambilan kata sinonim dilakukan untuk mendapatkan struktur kalimat yang terdiri dari kata, jenis kata dan sinonim (misal: aba-aba *n isyarat*). Data masukan yang diolah berasal dari Kamus Tesaurus Bahasa Indonesia berformat *.txt*. Contoh data masukan yang diambil dari Kamus Tesaurus Bahasa Indonesia yang akan diolah adalah sebagai berikut.

aba-aba n arahan, instruksi, isyarat, kode, komando, pe- rintah, petunjuk, seruan, su- ruhan, tanda, titah

abad n era, kala, kurun, masa, periode, sepuluh dasawarsa, sepuluh de- kade, seratus tahun, zaman;

seabad n seera, sekurun, sezaman; berabad-abad adv beratus-ratus tahun, berkurun-kurun, bertahun-tahun, berza- man-zaman

abadi a awet, baka, daim, infinit, kekal, lang- geng, langsung, lestari, qadim, samad, te- tap, wujud; ant sementara

Data masukan pada tabel Data masukan adalah kamus tesaurus bahasa Indonesia dalam format .txt yang menjadi sumber data yang akan dimasukan dalam pengambilan kata sinonim. Dari data masukan, sistem yang dibangun akan mengambil kata,jenis dan sinonim misal: aba-aba *n* isyarat untuk dapat dijadikan proses selanjutnya. Tahap Pembentukan format kata yang dilakukan terdiri dari 10 tahapanya yaitu penggabungan kata terpisah, pembersihan noise, penyamaan simbol majemuk, pemisahan baris lema dan sublema dan penempatan jenis kata dan sinonim.

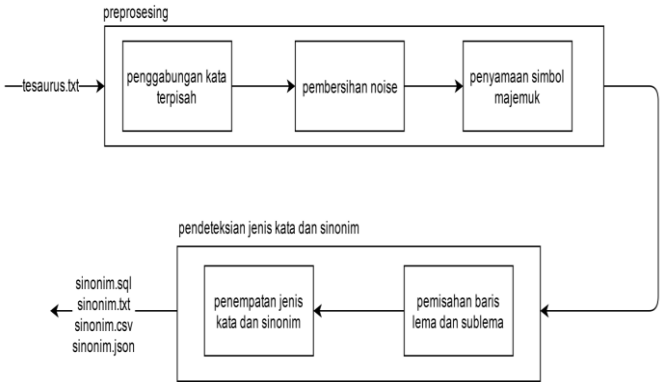


Fig. 3 Skema Pengambilan kata sinonim

Keluaran dari hasil pengambilan kata sinonim yang diharapkan adalah seperti pada gambar di bawah ini

abad n periode

abad n sepuluh dasawarsa

abad n sepuluh dekade

abad n seratus tahun

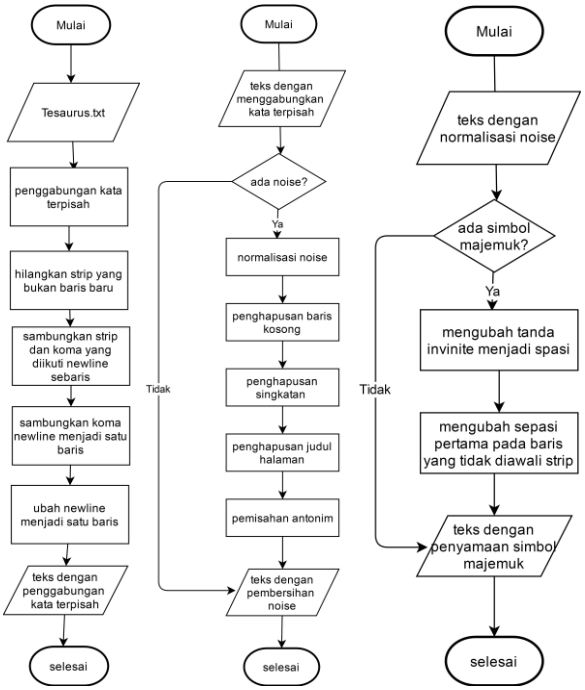
abad n zaman

abadi a awet

abadi a baka

abadi a daim

abadi a infinit



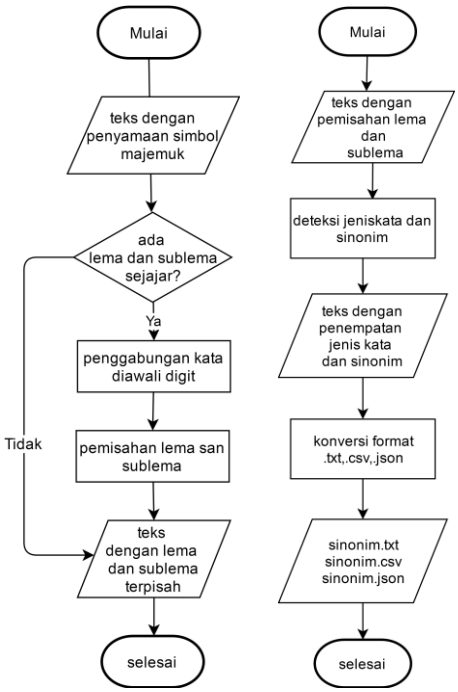


Fig. 4 Flowchart pendeteksian kata sinonim

3.5 Konversi Format

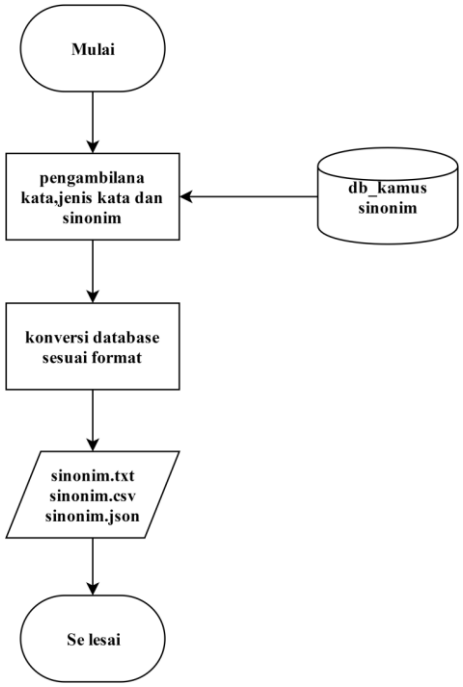


Fig. 5 *Flowchart* Konversi Format

IV. HASIL PENGUJIAN

Pada pengujian akurasi akan dihitung nilai akurasi dari jumlah lema pencarian yang dilakukan sistem berdasarkan data masukan. Dalam pengujiannya jumlah kata dan jenis kata yang diperoleh akan dibandingkan dengan jumlah lema dalam kamus tesaurus bahasa indonesia. Data kata yang diperoleh sistem adalah 33.379 lema, sedangkan jumlah lema dalam Tesaurus Bahas Indonesia adalah 66.779 lema.

Untuk memperoleh persentase hasil pengambilan kata dan jenis kata pada sistem yang dibangun, maka perhitungan yang dilakukan adalah jumlah lema yang dihasilkan sistem dibagi dengan jumlah lema pada Tesaurus Bahasa Indonesia. Nilai hasil bagi selanjutnya dikali dengan 100 % sebagai berikut.

- Jumlah lema Tesaurus Bahasa Indonesia : 66.779
- Jumlah lema yang dihasilkan sistem : 33.379
- Akurasi = $\frac{\text{jumlah lema yang dihasilkan sistem}}{\text{jumlah lema Tesaurus bahas indonesia}} \times 100\%$

  $= \frac{33.379}{66.779} \times 100\%$

  = 49.984 %

Penelitian yang dilakukan menghasilkan nilai akurasi sebesar 49.984 %.

Berikut adalah analisis hasil pengujian akurasi yang dilakukandalam pembangunan kamus sinonim.

Dalam proses pendeteksian, kata yang kondisinya kata, jeniskata lalu tidak diikuti dengan sinonim maka tidak dapat disimpan kedalam *database*. Karna itu data yang tidak dapat

disimpan mempengaruhi hasil dari akurasi. Di bawah ini merupakan contoh dalam kasus tersebut.

afiat a waras

afiksasi n derivasi

afiksasi n pengimbuhan

afiliasi n

afiliasi n aliansi

afiliasi n asosiasi

afiliasi n hubungan

afiliasi n koneksi

V. KESIMPULAN DAN SARAN

Sistem yang dibangun, dalam hal ini adalah pembangunan kamus sinonim Bahasa Indonesia mendapatkan Hasil pengujian pembangunan kamus sinonim menghasilkan 33.379 lema dengan persentase 49.984 %. Hal ini dapat dikatakan masih kurangnya dalam pendapatan jumlah lema. Disebabkan karena pada tahap pembentukan format kata tepatnya pada tahap ke lima yaitu plotting jeniskata dan sinonim tidak semua terdeteksi dengan baik.

Saran untuk penelitian berikutnya adalah peningkatan atau pembaruan pengkodean agar dapat mendeteksi format kata dengan lebih baik. Pengembangan dengan menyertakan kata antonim sebagai penambahan sumber daya. Semakin banyak kata yang termasuk dalam kategori, maka semakin baik dalam penggunaannya.

REFERENCES

[1] Departemen Pendidikan Nasional, Kamus Bahasa Indonesia, IV ed. Jakarta: Pusat Bahasa, 2008.

[2] Eko Endarmoko, Tesaurus Bahasa Indonesia. Jakarta: Gramedia Pustaka Utama, 2006.

[3] Nazir mohammad,Metode Penelitian. Bogor: Penerbit Ghalia Indonesia, 2014.

[4] Rosa A and M Salahudin, *Rekayasa Perangkat Lunak Terstruktur dan Berorientasi Objek*. Bandung: Informatika Bandung, 2014.