BAB II

TINJAUAN PUSTAKA

1.1 Tinjauan studi

Penelitian yang sudah ada sebelumnya, yaitu :

1. Nur Afifah (2010), "Pembuatan Kamus Elektronik Kalimat Bahasa Indonesia dan Bahasa Jawa untuk Aplikasi Mobile Menggunakan Interpolation Search". Penelitian ini membangun aplikasi penerjemah kalimat tunggal bahasa Indonesia ke bahasa Jawa menggunakan metode rule-based.
2. Andri Hidayat (2011), "Aplikasi Penerjemah Dua Arah Bahasa Indonesia – Bahasa Melayu Sambas Berbasis Web dengan Menggunakan Decoder Moses". Dalam penelitian ini penulis menggunakan metode statistik namun perancangan aplikasi ini berbasis dari data web.
3. Rizky Aditya Nugroho (2015). "Penerjemahan Bahasa Indonesia dan Bahasa Jawa Menggunakan Metode Statistik Berbasis Frasa". Penerjemahan bahasa Indonesia dan bahasa Jawa dua arah pada makalah ini menggunakan metode statistik yang berbasis frasa, dengan sumber data yang diambil dari Alkitab.

1.2 Tinjauan pustaka

2.3.1 Natural language processing (NLP)

Natural Language Processing (NLP) merupakan salah satu cabang ilmu AI yang berfokus pada pengolahan bahasa natural. Bahasa natural adalah bahasa yang

secara umum digunakan oleh manusia dalam berkomunikasi satu sama lain. Bahasa yang diterima oleh komputer butuh untuk diproses dan dipahami terlebih dahulu supaya maksud dari user bisa dipahami dengan baik oleh komputer.

Ada berbagai terapan aplikasi dari NLP. Diantaranya adalah Chatbot (aplikasi yang membuat user bisa seolah-olah melakukan komunikasi dengan computer), Stemming atau Lemmatization (pemotongan kata dalam bahasa tertentu menjadi bentuk dasar pengenalan fungsi setiap kata dalam kalimat), Summarization (ringkasan dari bacaan), Translation Tools (menterjemahkan bahasa) dan aplikasi-aplikasi lain yang memungkinkan komputer mampu memahami instruksi bahasa yang diinputkan oleh user [*3*].

Perkembangan NLP menghasilkan kemungkinan dari interface bahasa natural menjadi knowledge base dan penterjemahan bahasa natural. Terdapat bahwa ada 3 (tiga) aspek utama pada teori pemahaman mengenai natural language [*3*]:

1. Syntax: menjelaskan bentuk dari bahasa. Syntax biasa dispesifikasikan oleh sebuah grammar. Natural language jauh lebih daripada formal language yang digunakan untuk logika kecerdasan buatan dan program komputer
2. Semantics: menjelaskan arti dari kalimat dalam satu bahasa. Meskipun teori semantics secara umum sudah ada, ketika membangun sistem natural language understanding untuk aplikasi tertentu, akan digunakan representasi yang paling sederhana.
3. Pragmatics: menjelaskan bagaimana pernyataan yang ada berhubungan dengan dunia. Untuk memahami bahasa, agen harus mempertimbangan lebih dari hanya sekedar kalimat. Agen harus melihat lebih ke dalam konteks kalimat, keadaan dunia, tujuan dari speaker dan listener, konvensi khusus, dan sejenisnya.
4. Morfologi. Adalah pengetahuan tentang kata dan bentuknya sehingga bisa dibedakan antara yang satu dengan yang lainnya. Bisa juga

didefinisikan asal usul sebuah kata itu bisa terjadi. Contoh : membangunkan –> bangun (kata dasar), mem- (prefix), -kan (suffix).

5. Fonetik. Adalah segala hal yang berhubungan dengan suara yang menghasilkan kata yang dapat dikenali. Fonetik digunakan dalam pengembangan NLP khususnya bidang speech based system .

2.3.2 Mesin penerjemah statistik

Mesin penerjemah statistik didasarkan pada pandangan bahwa setiap kalimat dalam bahasa memiliki kemungkinan terjemahan dalam bahasa lain. Sebuah kalimat dapat diterjemahkan dari satu bahasa ke dalam bahasa yang lain dalam banyak kemungkinan cara. Metode terjemahan statistik mengambil pandangan bahwa setiap kalimat dalam bahasa target adalah terjemahan yang dimungkinkan dari kalimat masukan.

Mesin penerjemah statistik mengasumsikan bahwa setiap kalimat T pada bahasa target merupakan sebuah kemungkinan hasil terjemahan dari kalimat S pada bahasa sumber. Melalui pendekatan bahawa teks yang diterjemahkan berdasarkan distribusi probabilitas P(S|T) dapat dilakukan dengan teorema Bayes yaitu:

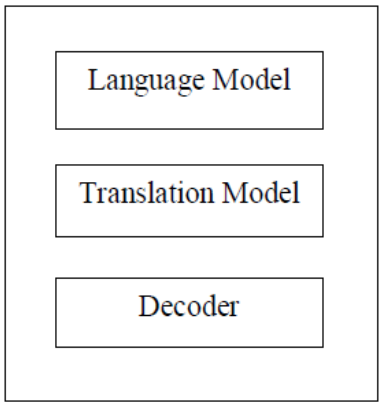
$$\mathrm{P}\ (\mathrm{T}|\mathrm{S}) = \frac{\mathrm{P}(\mathrm{T}).\mathrm{P}(\mathrm{S}|\mathrm{T})}{\mathrm{P}\ (\mathrm{S})} \qquad (1)$$

Disederhanakan menjadi,

$$\mathrm{P}\ (\mathrm{T}|\mathrm{S}) = \mathrm{P}(\mathrm{T}).\mathrm{P}(\mathrm{S}|\mathrm{T})$$

Karena nilai penyebut yang dinotasikan sebagai P(S) adalah pasti.

Dalam mesin penerjemah statistik, terdapat 3 komponen yang terlibat dalam proses penerjemahan dari satu bahasa ke bahasa lain yaitu : language model, translation model, dan decoder [*4*].



Gambar 1 : Komponen Mesin penerjemah statistik

2.3.2.1 Language model

Language model merupakan sumber pengetahuan yang penting dalam mesin penerjemah statistik. Language model digunakan pada aplikasi Natural. Language Processing. Dalam language model statistik, bagian-bagian yang merupakan elemen kunci adalah probabilitas dari rangkaian-rangkaian kata yang dituliskan sebagai P(w_1, w_2 ,…,w_n) atau P($w_{1,\ n}$). Untuk menghitung probabilitas kalimat, diperlukan untuk menghitung probabilitas kata, mengingat urutan kata yang mendahuluinya, dengan menggunakan aturan rantai, yaitu probabilitas kalimat P (T), dipecah sebagai probabilitas dari kata-kata individu P (K), seperti sebagai berikut [*5*] :

$$P(K) = P(w_1 w_2 \ldots w_n)$$
$$= P(w_1)P(w_2 w_1)P(w_3|w_1 w_2)\ldots P(w_n|w_1 w_2 \ldots w_{n-1}) \quad (2)$$

Dengan ‘K’ merupakan notasi ‘sentence’ dan w merupakan notasi ‘word’.

Rumusan tersebut dikenal dengan sebutan n-gram model. Model bahasa n-gram merupakan jenis probalilistik language model untuk memprediksi item berikutnya dalam urutan tersebut dalam bentuk (n-1). Probabilitas bersyarat dapat dihitung dari jumlah frekuensi n-gram [*4*] :

$$P(w_n|w_{n-1}) = \frac{\text{count}(w_{n-1}\ w_n)}{\text{count}(w_{n-1})} \quad (3)$$

Berikut merupakan contoh model bahasa n-gram, yaitu [*4*]:

1. Unigram (1-gram) : P(w_1),$P(w_2)$…P(w_n)
2. Bigram (2-gram) : P(w_1),$P(w_2 |w_1)$…P($w_n|w_{(n-1)}$)
3. Trigram (3-gram) : P($w_{1,n}$) = $P(w_1)$, $P(w_2|w_1)$, $P(w_3|w_{1,2})$…P($w_n|w_{n-2,n-1}$)

2.3.2.2 Translation model

Dalam mesin penerjemah komponen yang bertugas untuk mencari ketepatan adalah *translation model* (TM). P(S|T) digunakan sebagai notasi yang menunjukkan TM. Peluang untuk mendapatkan *translation model*, ditunjukkan dalam persamaan sebagai berikut :

$$P(S|T) = P(s_1|t_1)P(s_2t_2)\ldots P(s_nt_n) \quad (4)$$

Dengan,

s_n= kata-kata dari kalimat sumber pada posisi ke-n yang akan diketahui peluangnya.

t_n= kata-kata kalimat target pada posisi ke-n yang menerjemahkan kata s_n.

s_n yang terdapat pada persamaan tersebut merupakan kata apapun dari bahasa sumber pada posisi ke n, dan t_n adalah kata apapun dari bahasa target pada posisi ke n. Pada akhirnya, t_n merupakan terjemahan dari s_n.

Secara matematis untuk mendapatkan bobot relasi antara s_n dengan t_n adalah dalam persamaan 5. [*6*]

$$P(s_n|t_n) = \frac{count\ (t_n \text{ diterjemahkan sebagai } s_n)}{count\ (t_n \text{ diterjemahkan sebagai } s_y)} \quad (5)$$

Dengan,

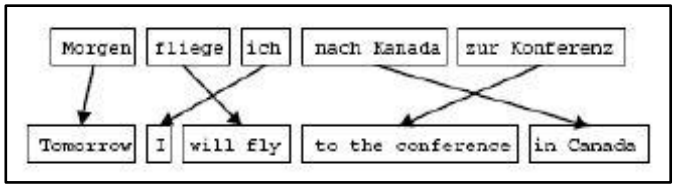
s_y = semua hasil penerjemahan yang mungkin. Sebagai contoh dalam tabel berikut.

Tabel 1. Relasi kata ‘makan’

Sumber = ‘makan’		
Target	***Count***	***Probability***
Mangan	4	0,2
Nedha	3	0,375
Dhahar	1	0,125
Total = 8		

Dari tabel xx dapat diketahui bahwa terjemahan dari kata ‘makan’ yang tertinggi adalah *‘mangan’,* sehingga penerjemahan kata ‘makan’ adalah *‘mangan’.*

Ada beberapa macam metode TM yang dapat diterapkan yakni pendekatan berbasis kata (*word based*), berbasis frasa (*phrase based*), dan kombinasi keduanya. Pendekatan berbasis kata sering kali tidak mampu menerjemahkan dengan baik konteks lokal suatu bahasa. Dalam TM berbasis frase, prosesnya dapat dibagi kedalam tiga bagian. Pertama, membuat kalimat sumber menjadi sebuah tabel frasa. Kedua, menerjemahkan setiap frase kedalam bahasa target. Kemudian, dilakukan tahap *reordering* [*5*].



Gambar 2 : Ilustrasi phrase based translation model

Proses dalam penerjemahan *phrase based translation model* dapat dipecah menjadi beberapa bagian, yaitu membagi kalimat bahasa sumber menjadi barisan frasa, menerjemahkan frasa ke dalam bahasa target, dan melakukan *reordering* [*5*].

Tabel penerjemah frase (*phrase translation table*) dibutuhkan untuk menerjemahkan frase dari bahasa sumber ke bahasa target. *Phrase alignment* atau tabel penerjemah frase diperoleh dengan melakukan dua langkah, yaitu membuat *word alignment* dari tiap pasangan kalimat dari

korpus paralel, selanjutnya mengekstrak pasangan frase yang konsisten dengan penjajaran kata tersebut (matrik penghubung kata) [*5*].

2.3.2.3 Decoder

Decoder bertugas menemukan teks dalam bahasa target yang memiliki probabilitas paling besar dengan pertimbangan faktor *translation model* dan language model [*4*]. Decoder disebut juga sebagai alogaritma pencarian. Bentuk matematikanya adalah sebagai berikut :

$$\hat{e} = arg_e max P(S|T)P(T) \quad (5)$$

Dengan $arg_e max$ adalah pencarian nilai probabilitas terbesar yang diperoleh dari *language model* dan *translation model.*

Philiph Koehn, dalam materi pengembanganya mengilustrasikan bagaimana proses dari decoding [*5*]. Berikut contoh prosesnya dengan kalimat ‘dia bersikap ramah tamah kepada banyak orang’.

Dia	bersikap	ramah	tamah	kepada	banyak	Orang

- Pilih kata bahasa target yang akan diterjemahkan

Dia	bersikap	ramah	tamah	kepada	banyak	Orang

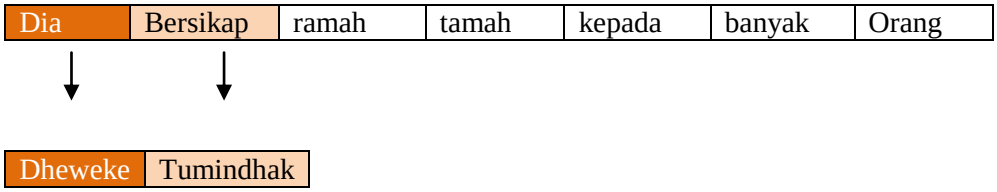
- Mencari kata yang cocok pada phrase translation tabel dan menambahkannya, kemudian menandainya sebagai kata yang telah diterjemahkan

Dia	bersikap	ramah	tamah	kepada	banyak	Orang

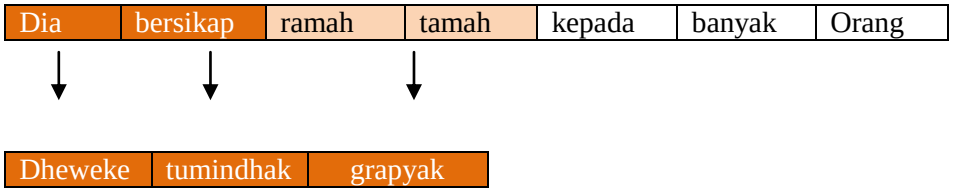
↓

Dheweke

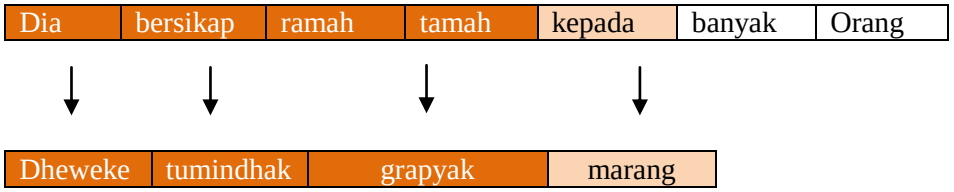
- Penerjemahan one to one



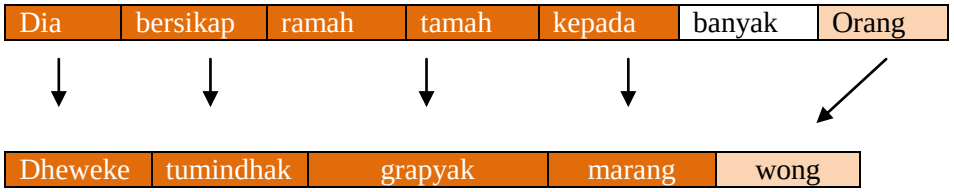
- Penerjemahan many to one



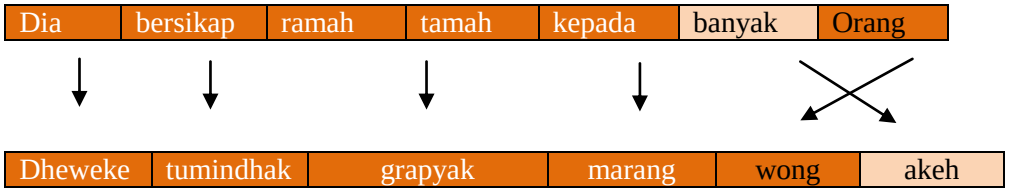
- Penerjemahan one to one



- Melakukan reordering



- Penerjemahan selesai



2.3.3 Penjajaran

Penjajaran kalimat adalah proses menata dua bahasa menjadi susunan baris-baris kalimat. Setelah itu dilanjutkan dengan penjajaran kata. Salah satu teknik penjajaran yang sering digunakan adalah teknik algoritme *expectation maximitation* (EM) dengan dua tahap, yaitu tahap *expectation* (ekspektasi) dan tahap *maximization* (maksimisasi). Untuk menghitung perkiraan kemiripan data berdasar pada suatu parameter terjadi pada tahap ekspektasi. Tahap maksimisasi digunakan untuk menghitung perkiraan kemiripan data berdasar hasil tahap ekspektasi. Parameter-parameter untuk menghitung perkiraan kemiripan data pada tahap ekspektasi didapatkan dari paralel korpus. Penjajaran ini sering digunakan karena tidak memerlukan pengetahuan leksikal. Semakin banyak parameter yang memiliki kemiripan, maka semakin besar pula peluang penjajaran.

2.3.4 Tuning

Model probabilitas merupakan model yang digunakan untuk mencari kemungkinan terbaik dari bahasa target di dalam mesin penerjemah statistik,. Nilai dari probabilitas tersebut ditentukan oleh empat parameter ,yaitu, *phrase translation table, language model, reordering model* atau *distortion model, word penalty*.

1. *Phrase translation table*

   Phrase translation table adalah parameter yang digunakan untuk memastikan bahwa frasa bahasa target dan frasa bahasa sumber mempunyai kualitas terjemahan yang sama baiknya.

2. *Language model*

   Language model adalah parameter untuk memastikan kefasihan (*fluency*) terjemahan dari bahasa target.

3. *Reordering model*

   Parameter reordering model digunakan untuk mengijinkan reordering dari kalimat masukan.

4. *Word penalty*

   Word penalty merupakan parameter yang digunakan untuk memastikan bahwa terjemahan mempunyai panjang kalimat yang tepat, tidak terlalu panjang dan tidak terlalu pendek.

Nilai probabilitas yang terbaik sehingga berpengaruh positif terhadap hasil terjemahan dipengaruhi oleh pemberian bobot dari masing-masing parameter [16]. Rentang bobot terbaik untuk parameter *phrase translation model, language model,* dan *reordering model* adalah 0.1 -1. Sedangkan rentang bobot terbaik *untuk word penalty* adalah -3 –3. Pencarian bobot terbaik pada tiap-tiap parameter tersebut disebut dengan *tuning* [*5*].

2.3.5 Evaluasi

Suatu permasalahan jika mencoba menerjemahkan satu kalimat dengan menggunakan beberapa mesin penerjemah, akan diperoleh berbagai jawaban yang berbeda. Mempertimbangkan hal-hal tersebut maka setiap pembangunan mesin penerjemah dibutuhkan tahap evaluasi terhadap mesin penerjemah tersebut. Dalam penelitian ini penulis menggunakan BLEU (*Billingual Evaluation Understudy*). BLEU bekerja dengan cara mengukur skor presisi dari n-gram termodifikasi (*modified n-gram precision score*) antara terjemahan otomatis dengan terjemahan rujukan dan menggunakan konstanta yang dinamakan *brevity penalty* (BP). Secara matematika dapat di peroleh persamaan seperti berikut

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \leq r \end{cases}$$

$$P_n = \frac{\sum_{C \in corpus\ n-gram\ \in C} \sum count_{clip}(n-gram)}{\sum_{C \in corpus\ n-gram \in C} \sum count\ (n-gram)}$$

$$BLEU = BP \cdot exp\ (\sum_{n-1}^{N} w_n \log p_n)$$

BP = *brevity penalty*, yakni penalti kandidat ketika penerjemahan kalimat tersebut (c) lebih panjang dibanding referensinya (r).

c = kandidat-kandidat MP/hiposis/panjang hasil

r = panjang efektif penerjamahan referensi

W*n* = I/N, bobot seragam dengan nilai umum (standar nilai N adalah 4)

N = panjang maksimum n-gram

*P*n = presisi n-gram termodifikasi

2.3.6 Korpus paralel

Korpus paralel adalah pasangan korpus yang berisi kalimat-kalimat dalam suatu bahasa dan terjemahannya. Korpus paralel merupakan bahan penting untuk melakukan eksperimen-eksperimen dalam bidang pemrosesan bahasa alami. Berikut contoh korpus paralel dalam sebuah tabel.

Tabel 2. Contoh corpus paralel

Bahasa Indonesia	Bahasa Jawa
Karena memiliki ibu yang cantik, Orlin memilki wajah yang cantik pula. Banyak orang yang menyukainya karena dia pandai.	Amarga nduweni ibu sing ayu, Orlin nduweni rupa sing uga ayu. Akeh wong sing seneng amarga deweke pinter.

2.3.7 Kerangka Penelitian