**ANALISIS PERBANDINGAN ALGORITMA *SMOOTHING* DENGAN *LANGUAGE MODEL TOOLKIT* SRILM**

**UNTUK PENENTUAN AKURASI TERBAIK PADA**

**MESIN PENERJEMAH STATISTIK BAHASA**

**INDONESIA – BAHASA KHEK**

**SINGKAWANG**

**SKRIPSI**

Program Studi Sarjana Informatika

Jurusan Informatika

Oleh:

**WELLY**

NIM D1041151013



**FAKULTAS TEKNIK**

**UNIVERSITAS TANJUNGPURA**

**PONTIANAK**

**2021**

Halaman Pernyataan

Yang bertanda tangan di bawah ini:

Nama : Welly

NIM : D1041151013

menyatakan bahwa dalam skripsi yang berjudul ”Analisis Perbandingan Algoritma *Smoothing* Dengan *Language Model Toolkit* SRILM Untuk Penentuan Akurasi Terbaik Pada Mesin Penerjemah Statistik Bahasa Indonesia – Bahasa Khek Singkawang” tidak terdapat karya yang pernah diajukan untuk memperoleh gelar sarjana di suatu perguruan tinggi manapun. Sepanjang pengetahuan Saya, tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam Daftar Pustaka.

Demikian pernyataan ini dibuat dengan sebenar-benarnya. Saya sanggup menerima konsekuensi akademis dan hukum di kemudian hari apabila pernyataan yang dibuat ini tidak benar.

Pontianak, 21 Maret 2021

Welly

NIM. D1041151013

Halaman Pengesahan

**ANALISIS PERBANDINGAN ALGORITMA *SMOOTHING* DENGAN *LANGUAGE MODEL TOOLKIT* SRILM UNTUK PENENTUAN AKURASI TERBAIK PADA MESIN PENERJEMAH STATISTIK BAHASA INDONESIA – BAHASA KHEK SINGKAWANG**

Jurusan Informatika

Program Studi Sarjana Informatika

Oleh:

Welly

NIM D1041151013

Telah dipertahankan di depan Penguji Skripsi pada tanggal 04 November 2020

dan diterima sebagai salah satu persyaratan untuk memperoleh gelar sarjana.

Susunan Penguji Skripsi :

Dosen Pembimbing Utama : Dr. Herry Sujaini, S.T., M.T. (NIP 196806291997021001)

Dosen Pembimbing Kedua : Rudy Dwi Nyoto, S.T., M.Eng. (NIP 197803302005011002)

Dosen Penguji Utama : Dr Arif Bijaksana PN, S.T., M.T. (NIP. 197208081998021002)

Dosen Penguji Kedua : Hafiz Muhardi, S.T., M.Kom. (NIDK 8885370018)

|  |  |
| --- | --- |
| Pontianak, 23 November 2020  Dekan  Dr.rer.nat. Ir. R. M. Rustamaji, M.T.,IPU  NIP. 196801161994031003 | Wakil Dekan Bidang Akademik  Dr.-Ing. Ir. Slamet Widodo, M.T.,IPM  NIP. 196712231992031002 |

**HALAMAN PERSEMBAHAN**

Halaman Persembahan

**Kupersembahkan skripsi ini kepada kedua orang tua saya yang harus saya bahagiakan.**

Kupersembahkan skripsi ini kepada kalian atas kasih sayang dan bimbingan selama ini sehingga saya dapat menyelesaikan skripsi ini dengan baik. Semoga hasil dan perjuangan saya selama ini dapat berbuah hasil yang baik. Agar saya sanggup menghadapi dunia luar yang sebenarnya. Terima kasih sekali lagi sebesar-besarnya kepada Ayah dan Ibu.

**Tidak lupa, teman-teman seperjuangan saya (Fakultas Teknik Prodi Informatika, Angkatan 2015).**

Saya ingin mengucapkan terimakasih atas dukungan dari teman-teman saya yang membuat saya dapat menyelesaikan skripsi ini. Semngat untuk teman-teman yang masih berjuang dalam perkuliahan dan bagi yang lulus jangan lupa dengan kita semua.

**Untuk yang kuhormati para dosen ku dan dosen pembimbing ku.**

Pak Helfi Nasution, S.Kom., M.Cs sebagai dosen pembimbing akademik. Pak Dr. Herry Sujaini, S.T., M.T. dan Pak Rudy Dwi Nyoto, S.T., M.Eng sebagai dosen pembimbing skripsi. Terima kasih banyak atas bimbingannya, maaf jika selama ini sudah banyak merepotkan.

Dengan segala ketulusan hati,

Welly

Kata Pengantar

Puji dan syukur penulis panjatkan kehadirat Tuhan Yang Maha Esa yang senantiasa memberikan limpahan rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan penelitian ini yang berjudul “Analisis Perbandingan Algoritma *Smoothing* Dengan *Language Model Toolkit* SRILM Untuk Penentuan Akurasi Terbaik Pada Mesin Penerjemah Statistik Bahasa Indonesia – Bahasa Khek Singkawang”.

Dalam penyelesaian penelitian ini penulis secara langsung maupun tidak langsung telah mendapatkan bimbingan, pengarahan dan bantuan dari beberapa pihak. Untuk itu, pada kesempatan ini penulis ingin mengucapkan terima kasih kepada:

1. Bapak Dr. Herry Sujaini, S.T., M.T. selaku dosen pembimbing utama.
2. Bapak Rudy Dwi Nyoto, S.T., M.Eng. selaku dosen pembimbing pendamping.
3. Bapak Dr.Arif Bijaksana Putra Negara, S.T., M.T. selaku dosen penguji utama.
4. Bapak Hafiz Muhardi, S.T., M.Kom selaku dosen penguji pendamping.
5. Bapak Helfi Nasution, S.Kom., M.Cs selaku dosen pembimbing akademik.
6. Bapak Dr. Yus Sholva, S.T., M.T. selaku ketua jurusan Informatika.
7. Bapak Dr.rer.nat. Ir. R. M. Rustamaji, M.T. selaku Dekan Fakultas Teknik Universitas Tanjungpura.
8. Seluruh dosen pengajar dan staff administrasi, serta rekan-rekan di program studi Informatika Universitas Tanjungpura.
9. Seluruh teman-teman saya dan keluarga tercinta.

Penulis menyadari dalam penyusunan skripsi ini masih belum sempurna. Oleh karena itu, penulis mengharapkan kritik dan saran yang bersifat membangun demi sempurnanya skripsi ini sehingga dapat bermanfaat bagi masyarakat serta untuk pengembangan wawasan dan peningkatan ilmu pengetahuan bagi kita semua.

Pontianak, 21 Maret 2021

Penulis,

Welly

Abstrak

Bahasa merupakan alat komunikasi yang digunakan manusia dalam berkehidupan sosial. Setiap bahasa daerah memiliki ciri khas atau karakteristik tersendiri, tidak terkecuali dengan bahasa Khek Singkawang. Keberagaman bahasa terkadang membuat komunikasi antar manusia terhambat, karena ketidak pahaman antar orang yang berkomunikasi maka diperlukan penerjemah agar informasi tersampaikan. Mesin penerjemah statistik merupakan pendekatan mesin penerjemah berbasis pada pendekatan statistik terhadap analisis korpus *paralel*. Mesin penerjemah statistik pada penelitian ini menggunakan *language model toolkit* SRILM dengan membandingkan algoritma *smoothing* pada bahasa Indonesia dan bahasa Khek Singkawang dengan banyak data 2700 baris kalimat korpus *paralel*. Algoritma *smoothing* merupakan komponen yang dapat menentukan tingkat akurasi hasil terjemahan mesin penerjemah sehingga perlu dilakukan penelitian terhadap beberapa algoritma *smoothing* untuk mendapatkan hasil terjemahan yang terbaik. Pengujian dilakukan dengan membandingkan algoritma *smoothing* *default* (*Good-Turing*) dengan algoritma *smoothing* *Interpolation, Natural, Witten-Bell, Modified Kneser-Ney, Original Kneser-Ney*. Dalam perbandingannya dilakukan penambahan secara konsisten pada setiap mesin sebanyak 200 korpus dengan sepuluh pengujian. Hasil pengujian otomatis (*BLUE*) terhadap 3-gram pada algoritma *Interpolation* (25,34%)*, Good-turing* (25,30%)*, Natural* (25,24%)*, Witten-Bell* (24,70%)*, Modified Kneser-Ney* (25,14%)*, Original Kneser-Ney* (25,57%). Dan untuk hasil pengujian otomatis (*BLUE*) terhadap 5-gram pada algoritma *Interpolation* (25,19%)*, Good-turing* (25,26%)*, Natural* (25,44%)*, Witten-Bell* (24,82%)*, Modified Kneser-Ney* (25,02%)*, Original Kneser-Ney* (26,15%). Hasil pengujian otomatis (*BLUE*) terhadap 3-gram dan 5-gram menunjukkan algoritma *smoothing* *Original Kneser-Ney* pada 5-gram memiliki akurasi terjemahan lebih baik dibandingkan dengan algoritma *smoothing* lainnya. Dengan uji *BLUE Original Kneser-Ney* terhadap 5-gram memiliki akurasi terjemahan (25,16%). Sedangkan untuk uji manual oleh Yulia Magdalena, S.Kom memiliki nilai akurasi (67,81%), kemudian oleh Veronica, B.Ed memiliki nilai akuras (72,75%).

Kata kunci: bahasa, mesin penerjemah statistik, SRILM, algoritma *smoothing, interpolation, good-turing, natural, witten-bell, modified kneser-ney, original kneser-ney,* BLEU *score*

Abstract

A language is a communication tool that people use for social life. Each Local Language has a characteristic, not except for Khek Singkawang Language. The Diversity of languages sometimes hampers communication between peoples, because of misunderstanding between people who communicate, a translator is needed so that information is conveyed. Statistical Machine Translation is a machine translation approach based on a statistical approach to parallel corpus analysis. In this study, using the SRILM language model toolkit and comparing it with the smoothing algorithm in Indonesian and Khek Singkawang with lots of data from 2700 parallel corpus sentences. The smoothing algorithm is a component that can determine the level of accuracy of the translation results so it is necessary to conduct research on several smoothing algorithms to get the best translation results. Testing is done by comparing the default smoothing algorithm (Good-Turing) with the smoothing algorithms Interpolation, Natural, Witten-Bell, Modified Kneser-Ney, Original Kneser-Ney. In comparison, 200 corpuses were consistently added to each machine with ten tests. Automatic test results (BLUE) against 3-gram on the Interpolation algorithm (25.34), Good-turing (25.30), Natural (25.24), Witten-Bell (24.70), Modified Kneser-Ney (25.14), Original Kneser-Ney (25.57) ). And for automatic test results (BLUE) against 5-gram on the Interpolation algorithm (25.19), Good-turing (25.26), Natural (25.44), Witten-Bell (24.82), Modified Kneser-Ney (25.02), Original Kneser-Ney (26.15). Automatic test results (BLUE) for 3-gram and 5-gram show the Original Kneser-Ney smoothing algorithm at 5-gram has better translation accuracy than other smoothing algorithms. With the BLUE test, Original Kneser-Ney at 5-grams it has a translation accuracy (25.16%). For the manual test by Yulia Magdalena, S.Kom has an accuracy percentage (67.81%), then by Veronica, B.Ed has an accurate percentage (72.75%).

Keywords: language, statistical machine translator, SRILM, smoothing algorithm*,* interpolation, good-turing, natural, witten-bell, modified kneser-ney, original kneser-ney, BLEU score

Daftar Isi

[Halaman Pernyataan ii](#_Toc58536676)

[Halaman Pengesahan iii](#_Toc58536677)

[Halaman Persembahan iv](#_Toc58536678)

[Kata Pengantar v](#_Toc58536679)

[Abstrak vi](#_Toc58536680)

[Abstract vii](#_Toc58536681)

[Daftar Isi viii](#_Toc58536682)

[Daftar Tabel x](#_Toc58536683)

[Daftar Gambar xi](#_Toc58536684)

[Daftar Lampiran xiii](#_Toc58536685)

[Bab I Pendahuluan 1](#_Toc58536686)

[1.1 Latar Belakang 1](#_Toc58536687)

[1.2 Perumusan Masalah 3](#_Toc58536688)

[1.3 Tujuan Penelitian 3](#_Toc58536689)

[1.4 Pembatasan Masalah 3](#_Toc58536690)

[1.5 Sistematika Penulisan 3](#_Toc58536691)

[Bab II Tinjauan Pustaka 5](#_Toc58536692)

[2.1 Kajian Terkait 5](#_Toc58536693)

[2.2 Mesin Penerjemah Statistik 7](#_Toc58536694)

[2.2.1 Language Model 12](#_Toc58536695)

[2.2.2 Translation Model 20](#_Toc58536696)

[2.3 Moses 21](#_Toc58536697)

[2.3.1 Decoder 22](#_Toc58536698)

[2.4 Automatic Evaluation 24](#_Toc58536699)

[2.5 Case Folding 26](#_Toc58536700)

[2.6 Tokenizing 26](#_Toc58536701)

[2.7 Language Model 26](#_Toc58536702)

[2.8 Translasi Model 27](#_Toc58536703)

[2.9 Cara Pengujian 27](#_Toc58536704)

[2.10 K-Fold Cross Validation 28](#_Toc58536705)

[2.11 Bahasa Khek Singkawang 30](#_Toc58536706)

[Bab III Metodologi Penelitian 32](#_Toc58536707)

[3.1 Data dan Perangkat Penelitian 32](#_Toc58536708)

[3.1.1 Data Penelitian 32](#_Toc58536709)

[3.1.2 Perangkat Penelitian 32](#_Toc58536710)

[3.2 Langkah-langkah Penelitian 33](#_Toc58536711)

[3.2.1 Pengumpulan Data 33](#_Toc58536712)

[3.2.2 Pembuatan Korpus Teks Paralel 34](#_Toc58536713)

[3.2.3 Membangun Mesin Penerjemah Statistik 35](#_Toc58536714)

[3.2.4 Implementasi Mesin Penerjemah Statistik Bahasa Indonesia – Bahasa Khek Singkawang dengan *Language Model Toolkit* SRILM 36](#_Toc58536715)

[3.2.5 Pemodelan Bahasa Oleh SRILM 41](#_Toc58536716)

[3.2.6 Pengujian Terjemahan Hasil Mesin Translasi oleh BLEU dengan Membandingkan *Algoritma Smoothing* yang berbeda 44](#_Toc58536717)

[3.2.7 Analisis Hasil Pengujian 48](#_Toc58536718)

[3.2.8 Penarikan Kesimpulan 48](#_Toc58536719)

[Bab IV Hasil dan Analisis 49](#_Toc58536720)

[4.1 Hasil Penelitian 49](#_Toc58536721)

[4.2 Hasil Pembuatan Korpus Teks Paralel 49](#_Toc58536722)

[4.3 Membangun Mesin Penerjemah Statistik 50](#_Toc58536723)

[4.4 Implementasi Mesin Penerjemah Statistik Bahasa Indonesia ke Bahasa Khek Singkawang 50](#_Toc58536724)

[4.4.1 Implementasi *Language Model Toolkit* SRILM untuk Pemodelan Bahasa 51](#_Toc58536725)

[4.4.2 Implementasi Mgiza untuk Pemodelan Translasi 56](#_Toc58536726)

[4.5 Pengujian Hasil Terjemahan 61](#_Toc58536727)

[4.5.1 Pengujian Otomatis oleh *BLEU* 61](#_Toc58536728)

[4.5.2 Pengujian Terjemahan secara Manual 69](#_Toc58536729)

[4.6 Analisis Hasil Pengujian 70](#_Toc58536730)

[Bab V Kesimpulan dan Saran 72](#_Toc58536731)

[5.1 Kesimpulan 72](#_Toc58536732)

[5.2 Saran 72](#_Toc58536733)

[Daftar Pustaka 73](#_Toc58536734)

Daftar Tabel

[**Tabel 2. 1** Perbandingan Penelitian Kajian Terkait 6](#_Toc67378701)

[**Tabel 2. 2** Tabel probabilitas model bahasa sebelum dilakukan discounting methods & frequency smoothing [Zhang, 2015] 10](#_Toc67378702)

[**Tabel 2. 3** Tabel probabilitas model bahasa sesudah dilakukan discounting methods & frequency smoothing [Zhang, 2015] 11](#_Toc67378703)

[**Tabel 2. 4** Contoh Perhitungan modified unigram precission (Tanuwijaya, 2009) 25](#_Toc67378704)

[**Tabel 4. 1** Korpus paralel bahasa Indonesia – bahasa Khek Singkawang 50](#_Toc67378711)

[**Tabel 4. 2** Hasil Proses Cleaning dan Tokenisasi 51](#_Toc67378712)

[**Tabel 4. 3** Hasil Proses Case Folding 51](#_Toc67378713)

[**Tabel 4. 4** Hasil K-fold Cross Validation 63](#_Toc67378714)

[**Tabel 4. 5** Hasil Nilai Akurasi Pengujian Ahli Bahasa 69](#_Toc67378715)

Daftar Gambar

[**Gambar 2. 1** Komponen mesin penerjemah statistik 8](#_Toc67379784)

[**Gambar 2. 2** Proses pemodelan bahasa [Modifikasi : Bertoldi, 2008] 8](#_Toc67379785)

[**Gambar 2. 3** N-gram modelling dengan jumlah tiga n-gram[Bertoldi, 2008] 9](#_Toc67379786)

[**Gambar 2. 4** Contoh N-gram probabilities [Bertoldi, 2008] 10](#_Toc67379787)

[**Gambar 2. 5** Contoh format ARPA file [Bertoldi, 2008] 12](#_Toc67379788)

[**Gambar 2. 6** Contoh penerjemahan dengan word-based translation model 20](#_Toc67379789)

[**Gambar 2. 7** Contoh penerjemahan dengan phrase based translation model 21](#_Toc67379790)

[**Gambar 2. 8** Contoh case folding 26](file:///C:\Users\WELLY%20LIM\Desktop\Draft\TA_WELLY\Sidang%20Hasil\D1041151013-Welly-Laporan-Seminar-Hasil.docx#_Toc67379791)

[**Gambar 2. 9** Contoh tokenizing 26](file:///C:\Users\WELLY%20LIM\Desktop\Draft\TA_WELLY\Sidang%20Hasil\D1041151013-Welly-Laporan-Seminar-Hasil.docx#_Toc67379792)

[**Gambar 2. 10** Contoh hasil proses pembuatan ARPA file[Bertoldi, 2008] 27](#_Toc67379793)

[**Gambar 2. 11** Contoh file model translasi [Hadi, 2014] 27](#_Toc67379794)

[**Gambar 2. 12** Contoh grafik pengujian jumlah korpus 28](#_Toc67379795)

[**Gambar 2. 13** Contoh skema 10 fold cross validation 29](#_Toc67379796)

[**Gambar 3. 1** Diagram Alir Penelitian 33](#_Toc67378807)

[**Gambar 3. 2** Proses pembuatan korpus 34](#_Toc67378808)

[**Gambar 3. 3** Contoh korpus paralel bahasa Indonesia – bahasa Khek Singkawang 35](#_Toc67378809)

[**Gambar 3. 4** Arsitektur sistem Mesin Penerjemah Statistik bahasa Indonesia – bahasa Khek Singkawang [Modifikasi: Sujaini dan Negara, 2015] 36](#_Toc67378810)

[**Gambar 3. 5** Proses pemodelan bahasa 38](#_Toc67378811)

[**Gambar 3. 6** N-gram modelling dengan jumlah n-gram lima buah 39](#_Toc67378812)

[**Gambar 3. 7** N-gram probabilities dari korpus paralel bahasa khek singkawang 39](#_Toc67378813)

[**Gambar 3. 8** Contoh tabel model bahasa menggunakan Language Model Toolkit SRILM dengan bahasa Khek Singkawang sebagai bahasa target dengan format ARPA file 41](#_Toc67378814)

[**Gambar 3. 9** Proses Pemodelan bahasa menggunakan Language Model Toolkit SRILM dengan bahasa Khek Singkawang sebagai bahasa target [Modifikasi: Hadi,2014] 41](#_Toc67378815)

[**Gambar 3. 10** Proses pemodelan bahasa menggunakan language model toolkit srilm 42](#_Toc67378816)

[**Gambar 3. 11** Contoh grafik pengujian perbandingan nilai BLEU dengan 3 gram 45](#_Toc67378817)

[**Gambar 3. 12** Pengujian dengan metode K-Fold Cross Validation 46](#_Toc67378818)

[**Gambar 3. 13** Proses evaluasi secara manual oleh ahli bahasa 47](#_Toc67378819)

[**Gambar 4. 1** Perintah untuk melakukan proses cleaning, tokenizing dan case folding 50](#_Toc67378821)

[**Gambar 4. 2** Perintah membangun model bahasa dengan toolkit srilm 52](#_Toc67378822)

[**Gambar 4. 3** Tabel model bahasa dengan bahasa khek singkawang sebagai bahasa target dengan algoritma smoothing Interpolation 52](#_Toc67378823)

[**Gambar 4. 4** Tabel model bahasa dengan bahasa khek singkawang sebagai bahasa target dengan algoritma smoothing Good-turing 53](#_Toc67378824)

[**Gambar 4. 5** Tabel model bahasa dengan bahasa khek singkawang sebagai bahasa target dengan algoritma smoothing Natural 53](#_Toc67378825)

[**Gambar 4. 6** Tabel model bahasa dengan bahasa khek singkawang sebagai bahasa target dengan algoritma smoothing Witten-Bell 54](#_Toc67378826)

[**Gambar 4. 7** Tabel model bahasa dengan bahasa khek singkawang sebagai bahasa target dengan algoritma smoothing Modified Kneser-Ney 54](#_Toc67378827)

[**Gambar 4. 8** Tabel model bahasa dengan bahasa khek singkawang sebagai bahasa target dengan algoritma smoothing Original Kneser-Ney 55](#_Toc67378828)

[**Gambar 4. 9** Perintah membangun model translasi dengan algoritma smoothing pada language model toolkit srilm 56](#_Toc67378829)

[**Gambar 4. 10** File model translasi 56](#_Toc67378830)

[**Gambar 4. 11** Dokumen vocabulary corpus bahasa indonesia 57](#_Toc67378831)

[**Gambar 4. 12** Dokumen vocabulary corpus bahasa khek singkawang 57](#_Toc67378832)

[**Gambar 4. 13** Dokumen word alignment bahasa indonesia-khek singkawang dengan algoritma smoothing interpolate 58](#_Toc67378833)

[**Gambar 4. 14** Dokumen word alignment bahasa indonesia- khek singkawang dengan algoritma smoothing good-turing 58](#_Toc67378834)

[**Gambar 4. 15** Dokumen word alignment bahasa indonesia- khek singkawang dengan algoritma smoothing natural 58](#_Toc67378835)

[**Gambar 4. 16** Dokumen word alignment bahasa indonesia- khek singkawang dengan algoritma smoothing witten-bell 59](#_Toc67378836)

[**Gambar 4. 17** Dokumen word alignment bahasa indonesia- khek singkawang dengan algoritma smoothing modified kneser-ney 59](#_Toc67378837)

[**Gambar 4. 18** Dokumen word alignment bahasa indonesia- khek singkawang dengan algoritma smoothing original kneser-ney 60](#_Toc67378838)

[**Gambar 4. 19** Tabel model translasi 61](#_Toc67378839)

[**Gambar 4. 20** Grafik hasil pengujian otomatis oleh BLEU dengan 3 gram 62](#_Toc67378840)

[**Gambar 4. 21** Grafik hasil pengujian otomatis oleh BLEU dengan 5 gram 62](#_Toc67378841)

[**Gambar 4. 22** Grafik Perbandingan Algoritma Smoothing Good-Turing dan Interpolate oleh BLUE dengan 3 gram 64](#_Toc67378842)

[**Gambar 4. 23** Grafik Perbandingan Algoritma Smoothing Good-Turing dan Natural oleh BLUE dengan 3 gram 64](#_Toc67378843)

[**Gambar 4. 24** Grafik Perbandingan Algoritma Smoothing Good-Turing dan Witten-Bell oleh BLUE dengan 3 gram 65](#_Toc67378844)

[**Gambar 4. 25** Grafik Perbandingan Algoritma Smoothing Good-Turing dan M-Kneser-Ney oleh BLUE dengan 3 gram 65](#_Toc67378845)

[**Gambar 4. 26** Grafik Perbandingan Algoritma Smoothing Good-Turing dan O-Kneser-Ney oleh BLUE dengan 3 gram 66](#_Toc67378846)

[**Gambar 4. 27** Grafik Perbandingan Algoritma Smoothing Good-Turing dan Interpolate oleh BLUE dengan 5 gram 66](#_Toc67378847)

[**Gambar 4. 28** Grafik Perbandingan Algoritma Smoothing Good-Turing dan Natural oleh BLUE dengan 5 gram 67](#_Toc67378848)

[**Gambar 4. 29** Grafik Perbandingan Algoritma Smoothing Good-Turing dan Witten-Bell oleh BLUE dengan 5 gram 67](#_Toc67378849)

[**Gambar 4. 30** Grafik Perbandingan Algoritma Smoothing Good-Turing dan M-Kneser-Ney oleh BLUE dengan 5 gram 68](#_Toc67378850)

[**Gambar 4. 31** Grafik Perbandingan Algoritma Smoothing Good-Turing dan O-Kneser-Ney oleh BLUE dengan 5 gram 68](#_Toc67378851)

Daftar Lampiran

[Lampiran A Korpus Uji oleh Yulia Magdalena, S.Kom A-1](#_Toc57769181)

[Lampiran B Korpus Uji oleh Veronica, B.Ed B-1](#_Toc57769182)

# Pendahuluan

## Latar Belakang

Bahasa merupakan sistem lambang bunyi yang dipergunakan oleh para anggota suatu masyarakat untuk bekerja sama, berinteraksi, dan mengidentifikasi diri. Dengan bahasa seseorang dapat berkomunikasi dengan manusia lainnya menggunakan tanda, misalnya kata dan gerakan untuk mengetahui informasi yang dibutuhkan. Dengan bahasa juga seseorang dapat menyampaikan ide, mengenalkan diri, dan menceritakan pengalamannya kepada orang lain. Interaksi antar manusia yang baik terjadi jika diantara keduanya memahami apa yang disampaikan. Bahasa memainkan peran penting dalam setiap interaksi yang terjadi. Sebagaimana kita ketahui bahwa di Indonesia terdiri dari berbagai macam suku dan budaya, ini berbanding lurus dengan keberagaman bahasa yang dimiliki oleh bangsa Indonesia. Mengutip catatan dari kompas jumlah bahasa daerah yang ada di Indonesia sebanyak 746 bahasa daerah diseluruh wilayah Indonesia (Setyawan, 2011).

Dengan beragamnya bahasa daerah yang ada di Indonesia tidak mungkin untuk dikuasai secara keseluruhan. Setiap bahasa daerah memiliki ciri khas atau karakteristik tersendiri, tidak terkecuali dengan bahasa Khek Singkawang. Bahasa Khek Singkawang memiliki dialek yang sulit dipahami, susunan kata dalam kalimatnya pula sulit untuk dipahami. Bahasa Khek Singkawang memiliki keunikan, susunan katanya berbeda dengan bahasa Indonesia. Seperti kalimat “Saya akan membawakan kamu beberapa sekarang” dalam bahasa Khek Singkawangnya adalah “*Ngai boi tai pun nyi kia cak kin ha*”, perbedaan kata dalam bahasa Indonesia dan bahasa Khek Singkawang dapat diliat pada kata yang di garis bawahi. Inilah mengapa bahasa Khek Singkawang ini dipilih untuk dijadikan mesin penerjemah dari bahasa Indonesia ke bahasa Khek Singkawang, agar dapat membantu untuk memahami bahasa Khek Singkawang serta agar hasil penerjemahan yang dihasilkan memiliki tingkat akurasi yang baik.

Mesin penerjemah merupakan alat penerjemah otomatis pada sebuah teks dari satu bahasa ke bahasa yang lainnya (Manning dan Schutze, 2000). Mesin penerjemah memiliki keterbatasan dalam menerjemahkan suatu teks bahasa, sehingga dapat menyebabkan berkurang atau bahkan merubah arti dan maksud dari suatu kalimat. Ada beberapa cara yang dapat diterapkan agar dapat mengoptimalkan hasil dari mesin penerjemah salah satunya dengan pendekatan mesin penerjemah statistik atau *Statistical Machine Translation*. Mesin penerjemah statistik adalah sebuah pendekatan mesin penerjemah dengan hasil terjemahan dihasilkan atas dasar model statistik yang parameter-parameternya diambil dari hasil analisis korpus teks paralel (Tanuwijaya, 2009).

Dalam penyusunan mesin penerjemah statistik diperlukan salah satu komponen yang harus ada yaitu model bahasa (*Language Model).* Model bahasa (*Language Model)* merupakan model bahasa yang digunakan menentukan kesesuaian suatu kata, frasa, atau kalimat dalam bahasa tertentu. Dan termasuk juga persentase kemunculan kalimat yang diterjemahkan juga ditentukan oleh model bahasa (*Language Model*). Penelitian terkait model bahasa *(Language Model)* sudah banyak dilakukan salah satunya dilakukan oleh Yosep Jarob R pada tahun 2016, yaitu dengan judul “Uji Akurasi Mesin Penerjemahan Bahasa Indonesia-Dayak Taman dengan Penandaan Kata Dasar dan Imbuhan”, dalam penelitian ini digunakan model bahasa (*Language Model*) SRILM, dan didapatkan hasil nilai BLUE pada korpus uji 3110 sebesar 80,17%. Adapun dalam penelitian lainnya yaitu dalam penelitian Wenyang Zhang dengan judul “*Comparing the Effect of Smoothing and N-gram Order: Finding the Best Way to Combine the Smoothing and Order of N-gram*, dimana pada penelitian ini membandingkan *Algoritma Smoothing* dan *N-gram Order* untuk mencari kombinasi terbaik *algoritma Smoothing* dan *N-gram Order*: pengujian pertama, menghasilkan *witten-bell discounting* dan 3-gram dengan nilai *perplexity* sebesar 311.097. pengujian kedua menghasilkan *witten-bell* dan 5-gram dengan nilai *perplexity* sebesar 45.9748.

Berdasarkan penjelasan yang sudah dipaparkan, penelitian ini akan membuat mesin penerjemah statistik dengan model bahasa (*language model)* SRILM dengan membandingkan algoritma *smoothing* yang berbeda dan ditentukan mana dari jenis algoritma *smoothing* yang memiliki nilai akurasi paling tinggi dari mesin penerjemah statistik bahasa Indonesia ke bahasa Khek Singkawang dan mesin penerjemah statistik bahasa Khek Singkawang ke bahasa Indonesia berdasarkan evaluasi otomatis BLEU dan pengujian oleh ahli bahasa. Hasil yang akan didapatkan pada penelitian ini akan dijadikan acuan sehingga penelitian yang mungkin akan menggunakan *language model* yang sama tidak perlu lagi untuk membandingkan algoritma *smoothing* mana yang lebih baik untuk jenis *language model toolkit* SRILM.

## Perumusan Masalah

Penelitian terkait *SRILM* sudah banyak dilakukan pada penelitian-penelitian. Namun untuk penelitian yang membandingkan hasil nilai akurasi yang dihasilkan Algoritma *Smoothing* dari mesin penerjemah statistik untuk bahasa Indonesia ke bahasa daerah terhadap peningkatan akurasi mesin penerjemah belum pernah dilakukan. Dengan demikian pada penelitian ini dapat dirumuskan masalah, bagaimana perbandingan hasil nilai akurasi yang dihasilkan Algoritma *Smoothing* terhadap akurasi mesin penerjemah statistik bahasa Indonesia ke bahasa Khek Singkawang.

## Tujuan Penelitian

Adapun tujuan dari penelitian yang dilakukan adalah untuk mengetahui hasil perbandingan algoritma *smoothing* *default*(*Good-turing*) dengan algoritma *smoothing* *Interpolation, Natural, Witten-Bell, Modified Kneser-Ney, Original Kneser-Ney* dalam penentuan nilai akurasi yang lebih baik terhadap mesin penerjemah bahasa Indonesia ke bahasa Khek Singkawang.

## Pembatasan Masalah

Beberapa hal yang menjadi batasan dalam penelitian ini adalah sebagai berikut.

1. Korpus yang digunakan berasal dari novel Sherlock Holmes dan sumber lainnya yang kemudian akan di terjemahkan oleh ahli Bahasa dari bahasa Indonesia ke bahasa Khek Singkawang.
2. *Language Model* yang digunakan yaitu SRILM
3. Algoritma Smoothing yang digunakan yaitu: *Interpolation, Good-turing, Natural, Witten-Bell, Modified Kneser-Ney, Original Kneser-Ney*
4. Membandingkan Algoritma *Smoothing* berdasarkan *BLEU Score*

## Sistematika Penulisan

Sistematika penulisan penelitian ini disusun untuk memberikan gambaran umum tentang penelitian yang dijalankan. Sistematika laporan tugas akhir ini disusun dalam 5 (lima) bab yang terdiri dari Bab I Pendahuluan, Bab II Tinjauan Pustaka, Bab III Metodologi Penelitian, Bab IV Hasil dan Analisis, serta Bab V Penutup.

**Bab I Pendahuluan** adalah bab yang berisi latar belakang penelitian, perumusan masalah, tujuan penelitian, pembatasan masalah, dan sistematika penulisan.

**Bab II Tinjauan Pustaka** adalah bab yang membahas tentang gambaran umum dari penelitian yang didapat dari penelitian sebelumnya serta perangkat lunak yang melandasi pembangunan sistem dan landasan teori yang berhubungan dalam proses analisis permasalahan penelitian yang akan dilakukan.

**Bab III Metodologi Penelitian** adalah bab yang berisi tentang bahan penelitian, spesifikasi alat yang digunakan, metode yang akan digunakan pada penelitian, diagram alir penelitian serta perancangan pengujian yang akan dilakukan pada penelitian.

**Bab IV Hasil dan Analisis** adalah bab yang berisi hasil penelitian, penjelasan mengenai implementasi metode yang digunakan, hasil analisis serta analisis dari setiap pengujian. Bagian yang ditampilkan akan dilakukan analisis terlebih dahulu untuk mengarah kepada suatu kesimpulan.

**Bab V Penutup** adalah bab yang berisi kesimpulan dari penelitian yang telah dilakukan dan saran atau rekomendasi untuk perbaikan, pengembangan atau kesempurnaan dan kelengkapan penelitian yang telah dilakukan.

# Tinjauan Pustaka

## Kajian Terkait

Berikut adalah kajian-kajian terkait:

Jarob (2016) melakukan penelitian terkait untuk menguji akurasi mesin penerjemah bahasa Indonesia ke bahasa Dayak Taman dengan Penandaan Kata Dasar dan Imbuhan. Pada penelitian ini penulis menggunakan model bahasa (*language model*) yaitu SRILM yang merupakan salah satu model bahasa (*language model*) yang tercantum dalam *minimum requirements* dari moses yaitu toolkit dalam membuat mesin penerjemah statistik. Penelitian menggunakan korpus paralel sebanyak 3110 kalimat. Pengujian dilakukan dengan dua cara yaitu pengujian otomatis dengan BLEU dan pengujian oleh ahli bahasa Dayak Taman. Untuk hasil pengujian otomatis dengan BLEU dihasilkan nilai akurasi sebesar 80,17% dan untuk pengujian oleh ahli bahasa Dayak Taman sebesar 79,26% terkait mesin penerjemah bahasa Indonesia ke bahasa Dayak Taman.

Hadi (2014) melakukan penelitian terkait untuk menguji akurasi mesin penerjemah bahasa Indonesia ke bahasa Melayu Sambas dan mesin penerjemah bahasa Melayu Sambas ke bahasa Indonesia. Pada penelitian kali ini penulis menggunakan model bahasa (*language model*) SRILM, yang merupakan salah satu model bahasa (*language model*) yang tercantum dalam *minimum requirements* dari moses yaitu toolkit dalam membuat mesin penerjemah statistik. Penelitian ini menggunakan korpus paralel dari Bahasa Indonesia dan Bahasa Melayu Sambas. Pengujian dilakukan dengan dua cara yaitu pengujian otomatis dengan BLEU dan pengujian oleh ahli bahasa Melayu Sambas. Untuk hasil pengujian otomatis dengan BLEU, dari bahasa Indonesia ke bahasa Melayu Sambas pada pengujian pertama dihasilkan nilai akurasi sebesar 55%, untuk pengujian kedua dihasilkan nilai akurasi sebesar 49%. Untuk pengujian ahli bahasa mesin penerjemah bahasa Indonesia ke bahasa Melayu Sambas pada pengujian pertama dihasilkan nilai akurasi rata-rata sebesar 70,24%. Pada pengujian kedua mesin penerjemah bahasa Indonesia ke bahasa Melayu Sambas dihasilkan nilai akurasi rata-rata sebesar 63,76%.

Guo dkk. (2014) dalam penelitian yang berjudul, “*A Tunable Language Model for Statistical Machine Translation”* melakukan penelitian untuk menguji akursi mesin penerjemah bahasa Cina ke bahasa Inggris dengan model bahasa (*language model*) SRILM dan KENLM dengan model Kneser Ney yang telah dimodifikasi. Penelitian ini menggunakan korpus paralel bahasa Cina dan bahasa Inggris lebih dari 1.600.000 kalimat. Untuk hasil pengujian dilakukan dengan cara pengujian otomatis BLEU. Untuk hasil terjemahan dari bahasa Cina ke bahasa Inggris pada percobaaan pertama dihasilkan nilai akurasi sebesar 92,37% dan untuk percobaan kedua dihasilkan nilai akurasi sebesar 92,32%. Nilai akurasi dari model bahasa (*language model*) SRILM dan KENLM pada percobaan pertama dan kedua menghasilkan nilai akurasi yang sama.

Zhang (2015) dalam penelitian yang berjudul,” *Comparing the Effect of Smoothing and N-gram Order: Finding the Best Way to Combine the Smoothing and Order of N-gram*” melakukan penelitian untuk membandingkan *Algoritma Smoothing* dan *N-gram Order* yang berbeda untuk mencari kombinasi terbaik *algoritma Smoothing* dan *N-gram Order*. Pengujian pertama, menghasilkan *witten-bell discounting* dan 3-gram dengan nilai *perplexity* sebesar 311.097 dan pengujian kedua menghasilkan *witten-bell* dan 5-gram dengan nilai *perplexity* sebesar 45.9748.

Perbandingan penelitian-penelitian yang terdapat pada kajian terkait diperlihatkan pada Tabel 2.1.

**Tabel 2. 1** Perbandingan Penelitian Kajian Terkait

| **No.** | **Penulis/tahun** | **Judul** | **Keterangan** |
| --- | --- | --- | --- |
| 1. | Jarob (2016) | Uji Akurasi Penerjemahan Bahasa Indonesia – Dayak Taman dengan Penandaan Kata Dasar dan Imbuhan. | Penelitian menggunakan korpus paralel sebanyak 3110 kalimat. Pengujian dilakukan dengan dua cara yaitu pengujian otomatis dengan BLEU dan pengujian oleh ahli bahasa Dayak Taman. Untuk hasil pengujian otomatis dengan BLEU dihasilkan nilai akurasi sebesar 80,17% dan untuk pengujian oleh ahli bahasa Dayak Taman sebesar 79,26% terkait mesin penerjemah bahasa Indonesia ke bahasa Dayak Taman. |
| 2. | Hadi (2014) | Uji Akurasi Mesin Penerjemah Statistik (MPS) Bahasa Indonesia ke Bahasa Melayu Sambas dan Mesin Penerjemah Statistik (MPS) Bahasa Melayu Sambas ke Bahasa Indonesia | Penelitian ini menggunakan korpus paralel dari Bahasa Indonesia dan Bahasa Melayu Sambas. Pengujian dilakukan dengan dua cara yaitu pengujian otomatis dengan BLEU dan pengujian oleh ahli bahasa Melayu Sambas. Untuk hasil pengujian otomatis dengan BLEU, dari bahasa Indonesia ke bahasa Melayu Sambas pada pengujian pertama dihasilkan nilai akurasi sebesar 55%, untuk pengujian kedua dihasilkan nilai akurasi sebesar 49%. Untuk pengujian ahli bahasa mesin penerjemah bahasa Indonesia ke bahasa Melayu Sambas pada pengujian pertama dihasilkan nilai akurasi rata-rata sebesar 70,24%. Pada pengujian kedua mesin penerjemah bahasa Indonesia ke bahasa Melayu Sambas dihasilkan nilai akurasi rata-rata sebesar 63,76%. |
| 3. | Gou dkk (2014) | *A Tunable Language Model for Statistical Machine Translation* | Penelitian ini menggunakan korpus paralel bahasa Cina dan bahasa Inggris lebih dari 1.600.000 kalimat. Untuk hasil pengujian dilakukan dengan cara pengujian otomatis BLEU. Untuk hasil terjemahan dari bahasa Cina ke bahasa Inggris pada percobaaan pertama dihasilkan nilai akurasi sebesar 92,37% dan untuk percobaan kedua dihasilkan nilai akurasi sebesar 92,32%. Nilai akurasi dari model bahasa (language model) SRILM dan KENLM pada percobaan pertama dan kedua menghasilkan nilai akurasi yang sama. |
| 4. | Zhang (2015) | *Comparing the Effect of Smoothing and N-gram Order: Finding the Best Way to Combine the Smoothing and Order of N-gram* | Penelitian ini membandingkan Algoritma Smoothing dan N-gram Order yang berbeda untuk mencari kombinasi terbaik algoritma Smoothing dan N-gram Order. Pengujian pertama, menghasilkan witten-bell discounting dan 3-gram dengan nilai perplexity sebesar 311.097 dan pengujian kedua menghasilkan witten-bell dan 5-gram dengan nilai perplexity sebesar 45.9748. |

## Mesin Penerjemah Statistik

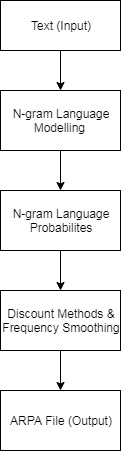
Mesin penerjemah statistik adalah suatu mesin penerjemah dimana penerjemahan dilakukan berbasis model statistik dengan parameter-parameter yang di ambil dari analisis korpus paralel. Pendekatan statistik yang digunakan adalah konsep probabilitas. Setiap pasangan kalimat (S,T) akan diberikan sebuah P(T|S) yang diinterpretasikan sebagai distribusi probabilitas yang sebuah penerjemah akan menghasilkan T dalam bahasa tujuan ketika diberikan S dalam bahasa sumber (Tanuwijaya & Manurung, 2009).

Mesin penerjemah statistik mengasumsikan bahwa setiap kalimat *T*  pada bahasa sasaran merupakan sebuah kemungkinan hasil terjemahan dari kalimat *S* pada bahasa sumber. Melalui pendekatan bahwa teks yang diterjemahkan berdasarkan distribusi probabilitas *P(T|S)* dapat dilakukan dengan teorema Bayes seperti dapat dilihat pada persamaan (2.1).

Pada mesin penerjemah statistik terdapat 3 komponen yang terlibat dalam proses penerjemahan dari bahasa asal ke bahasa target atau dari satu bahasa ke bahasa lainnya. Ketiga kompenen tersebuat adalah *language model*, *translation model*, dan decoder (Maning & Schutze, 2000), dapat di gambarkan seperti pada gambar 2.1



**Gambar 2. 1** Komponen mesin penerjemah statistik

****

**Gambar 2. 2** Proses pemodelan bahasa [Modifikasi : Bertoldi, 2008]

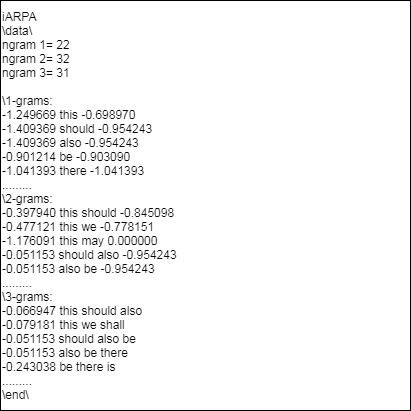
Gambar 2.2 menunjukkan tahapan-tahapan yang harus dilakukan dalam proses pemodelan bahasa antara lain yaitu: Text (Input), *N-gram Language Modelling*, *N-gram Language Probabilities, Discount Methods & Frequemcy Smoothing,* ARPA File (Output).

1. Text (Input)

Proses memasukkan korpus yang terlebih dahulu telah dilakukan proses *Preprocessing (cleaning, tokenization, case folding).*

1. *N-gram Language Modelling*

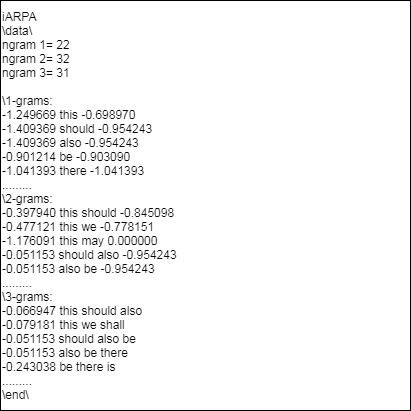
*N-gram Language Modelling* merupakan proses menentukan berapa banyak n-gram yang ingin digunakan dalam suatu pemodelan bahasa atau berapa banyak penggalan kata ataupun karakter dalam pembuatan pemodelan bahasa. Cara mengatur jumlah n-gram yang akan digunakan yaitu pada saat training dengan menuliskan pada *script order* diketik pada script seperti berikut “. . .-o 3 . . .”angka 3 menandakan bahwa kita menggunakan 3 N-gram dapat dilihat pada gambar 2.3.

****

**Gambar 2. 3** N-gram modelling dengan jumlah tiga n-gram[Bertoldi, 2008]

1. *N-gram Language Probabilities*

*N-gram Language Probabilities* merupakan proses untuk menghitung setiap kemungkinan atau probabilitas kemunculan kata atau karakter pada korpus dapat dilihat pada gambar 2.4.

****

**Gambar 2. 4** Contoh N-gram probabilities [Bertoldi, 2008]

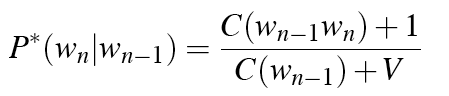
1. *Discount Methods & Frequency Smoothing*

*Discount Methods & Frequency Smoothing* digunakan untuk meminimalisir bahkan menghilangkan probabilitas yang bernilai “0” pada pemodelan bahasa dapat dilihat pada Tabel 2.2.

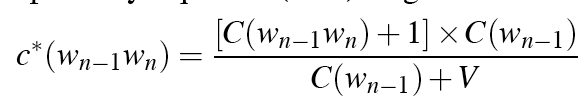
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | i | want | to | eat | chinese | food | lunch | spend |
| i | 5 | 827 | 0 | 9 | 0 | 0 | 0 | 2 |
| want | 2 | 0 | 608 | 1 | 6 | 6 | 5 | 1 |
| to | 2 | 0 | 4 | 686 | 2 | 0 | 6 | 211 |
| eat | 0 | 0 | 2 | 0 | 16 | 2 | 42 | 0 |
| chinese | 1 | 0 | 0 | 0 | 0 | 82 | 1 | 0 |
| food | 15 | 0 | 15 | 0 | 1 | 4 | 0 | 0 |
| lunch | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| spend | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

**Tabel 2. 2** Tabel probabilitas model bahasa sebelum dilakukan discounting methods & frequency smoothing [Zhang, 2015]

Sebagai contoh dengan menggunakan *Laplace-Smoothing* untuk memunculkan probabilitas yang bernilai “0” tadi dengan menggunakan persamaan 2.2 dan 2.3.



(2.2)



(2.3)

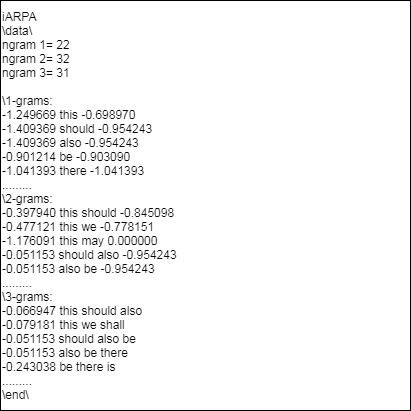
Barulah kita dapatkan nilai probabilitas dari rumus diatas yang dimana semuanya akan memiliki nilai dapat dilihat pada Tabel 2.3.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | i | want | to | eat | chinese | food | lunch | spend |
| i | 3.8 | 527 | 0.54 | 6.4 | 0.64 | 0.64 | 0.64 | 1.9 |
| want | 1.2 | 0.39 | 238 | 0.78 | 2.7 | 2.7 | 2.3 | 0.78 |
| to | 1.9 | 0.63 | 3.1 | 430 | 1.9 | 0.63 | 4.4 | 133 |
| eat | 0.34 | 0.34 | 1 | 0.34 | 5.8 | 1 | 16 | 0.34 |
| chinese | 0.2 | 0.098 | 0.098 | 0.098 | 0.098 | 8.2 | 0.2 | 0.098 |
| food | 6.9 | 0.43 | 6.9 | 0.43 | 0.86 | 2.2 | 0.43 | 0.43 |
| lunch | 0.57 | 0.19 | 0.19 | 0.19 | 0.19 | 0.38 | 0.19 | 0.19 |
| spend | 0.32 | 0.16 | 0.32 | 016 | 0.16 | 0.16 | 0.16 | 0.16 |

**Tabel 2. 3** Tabel probabilitas model bahasa sesudah dilakukan discounting methods & frequency smoothing [Zhang, 2015]

1. File *Language Model*

File yang sudah diproses dan melewati tahap *N-gram Modelling, N-gram Probabiities, Discount Methods & Frequency Smoothing* sehingga siap untuk diproses pada pemodelan translasi untuk membangun mesin penerjemah statistik dapat dilihat pada gambar 2.5.

****

**Gambar 2. 5** Contoh format ARPA file [Bertoldi, 2008]

### Language Model

*Language Model* digunakan pada aplikasi *Natural Language Processing* seperti *speech recognition, part-of-speech tagging,* dan *syntactic parsing*.

Dalam *Language Model* statistik, bagian-bagian yang merupakan elemen kunci adalah probabilitas dari rangkaian-rangkaian kata yang dituliskan sebagai P(W1, W2, … , Wn) atau P ( W1, n). *Language Model* menetapkan probabilitas P(W1, n) ke serangkaian *n* kata dengan *means* sebuah distribusi probabilitas. Rangkaian-rangkaian tersebut bisa berupa frase-frase atau kalimat-kalimat dan probabilitasnya dapat diperkirakan dari korpus dokumen-dokumen yang besar. Salah satu contoh pendekatan *language model* adalah ­*n-gram model.* Dalam sebuah *n-gram model*, probabilitas *w1,m* dari mengamati kalimat 𝑤1, 𝑤2, … , 𝑤𝑚 di aproksimasikan pada persamaan 2.4

Untuk mengestimasikan probabilitas dari mesin penerjemah disebut juga *Maximum Likehood*.

Probabilitas bersyarat (*conditional*) dapat dihitung dari perhitungan frekuensi n-gram dengan persamaan 2.5

Bentuk penulisan rumus lainnya untuk perhitungan probabilitas n-gram berdasarkan (Chen & Goodman, 1998) dapat dilihat pada persamaan 2.6

Berikut merupakan contoh model bahasa n-gram, yaitu:

* 1. Unigram (1-gram):

*P*(*W1, n*) = *P*(*W1*) *P*(*W2*) … *P(Wn*)

* 1. Bigram (2-gram):

*P*(*W1, n*) = *P*(*W1*) *P*(*W*2|*W1*) … *P*(*Wn*|*Wn-1*)

* 1. Trigram (3-gram):

*P*(*W1, n*) = *P*(*W1*) *P*(*W2*|*W1*) *P*(*W3*|*W1,2*) … *P*(*Wn*|*Wn-2,n-1*)

*Language model* P(e) berperan untuk menentukan apakah kalimat hasil terjemahan e adalah kalimat yang alami dan memiliki gramatikal yang baik. Semakin tinggi nilai probabilitas dari P(e) menunjukkan bahwa kalimat hasil terjemahan merupakan kalimat yang dibentuk dengan baik. Dan adapun pada penelitian kali ini akan memfokuskan untuk menggunakan model bahasa (*language model*) IRSTLM berdasarkan *minimum requirements* dari toolkit moses yaitu SRILM, KENLM, dan IRSTLM (Tanuwijaya & Manurung, 2009).

#### SRILM

Speech Research International Language Modelling Toolkit yang disingkat SRILM merupakan sebuah *open source toolkit* untuk mesin penerjemah statistik dan hal terkait. SRILM ditemukan oleh Andreas Stolcke yang merupakan seorang peneliti terkait pemrosesan bahasa alami khususnya bagian *speech technology and research laboratory,* di SRI Internasional Menlopark, California, U.S.A. SRILM pertama kali disusun dan diimplementasikan pada tahun 1995, serta diikuti dengan dirilisnya versi beta dari SRILM ke publik pada tahun 1999. Sejak saat itu SRILM telah digunakan secara luas dalam komunitas *speech* dan *natural language*. (Stolcke, dkk. 2011)

SRILM adalah sebuah *toolkit* untuk membangun dan mengevaluasi model bahasa (*language model*) secara statistik. Kebanyakan dari tipe model bahasa (*language model*) yang didukungnya berdasarkan pada statistik *N-gram,* termasuk standar model *back-off*, dengan berbagai algoritma perataan standar (jenis algoritma *smoothing* yang disupport oleh SRILM adalah *Good-Turing, Witten-Bell, Kneser-Ney, Ristad Natural Discounting, Ney Absolute discounting, Smooth by Adding Delta,* dan *Modified Kneser-Ney*.). Model berdasarkan kelas kata (diinduksi secara otomatis atau ditentukan secara eksternal) juga didukung. SRILM mengimplementasikan berbagai metode untuk menginterpolasi dan mengadaptasi model bahasa (*language model*), serta pemangkasan untuk menukar ukuran melawan kinerja. Sekali model bahasa (*language model*) di*training* (dilatih), itu dapat dievaluasi atau digunakan dalam berbagai tugas standar, seperti komputasi *perplexity*, *N-best­* dan *lattice scoring,* dan *text tagging and segmentation.*

Algoritma *smoothing* termasuk salah satu dari komponen penting dalam pemodelan bahasa untuk mengatasi permasalahan probabilitas kemunculan kata yang bernilai nol sehingga berdampak kepada nilai probabilitas nol terhadap n-gram sehingga nilainya tidak terdeteksi. Pada *toolkit* SRILM ada beberapa algoritma *smoothing* yang terintegrasi dengan *toolkit* ini diantaranya : *witten-bell, absolute discounting, ristad natural discounting, smooth by adding delta, modified kneser-ney*, *backoff* dan *interpolation*. Adapun pada penelitian kali ini digunakan enam algoritma *smoothing* yaitu : *interpolation, good-turing, natural, witten-bell, modified kneser-ney, original kneser-ney* berdasarkan penelitian Zhang (2015) yang berjudul “*Comparing the Effect of Smoothing and N-gram Order: Finding the Best Way to Combine the Smoothing and Order of N-gram”*.

1. Interpolation

Dalam praktiknya, interpolasi memiliki beberapa cara untuk memperkirakan distribusi probabilitas untuk variabel acak. Paling umum adalah memiliki sampel besar dengan informasi umum dan sampel kecil yang lebih spesifik, tetapi kurang dapat diandalkan. Jadi biasanya memiliki dua cara untuk memperkirakan distribusi probabilitas yang dihasilkan dalam dua fungsi p1 dan p2.

Dengan interpolasi kita dapat menggabungkan dua distribusi probabilitas p1 dan p2 untuk variabel X acak yang sama dengan memberikan bobot tetap dan menambahkannya. Jika kita beri bobot pertama 0 ≤ λ ≤ 1, maka kita ada 1 - λ tersisa untuk yang kedua:

Penerapan umum interpolasi muncul dari sampel data yang memperhitungkan kondisi yang berbeda. Misalnya, kita mungkin ingin untuk memprediksi cuaca besok M berdasarkan cuaca hari ini T, dan juga berdasarkan hari kalender D (untuk mempertimbangkan tipikal pola cuaca musiman).

Kami sekarang ingin memperkirakan distribusi probabilitas bersyarat p (m | t, d) dari statistik cuaca. Pengkondisian pada hari-hari tertentu mengarah untuk sampel data kecil dalam perkiraan (misalnya, mungkin hampir tidak ada setiap hari hujan pada tanggal 1 Agustus di Los Angeles). Karenanya, mungkin lebih praktis untuk menginterpolasi distribusi ini dengan p (m | t) yang lebih kuat:

Selain mempertimbangkan kondisi yang berbeda, kita mungkin juga memiliki perbedaan metodologi untuk memperkirakan distribusi probabilitas. Strategi masuk komunitas pembelajaran mesin untuk menggabungkan beberapa hasil dari metode ini disebut kombinasi pengklasifikasi atau pembelajaran ensemble dan dapat dilakukan dengan interpolasi sederhana.

1. Good-Turing Estimate

Good Turing dalam Chen dan Goodman (1998), The Good-Turing Estimate adalah sentral dari banyak Teknik *smoothing*. Untuk iterasi algoritma Good-Turing Estimate sebanyak *r* kali, estimasi perhitungan iterasi algoritma sebanyak *r\** kali dituliskan dalam rumus berikut ini.

Dan dimana adalah jumlah n-gram yang diiterasikan sebanyak *r* kali pada data yang di*training*. Untuk mengkonversikan perhitungan ini ke sebuah probabilitas, diperlukan proses normalisasi untuk sebuah n-gram dengan perhitungan *r*, notasi rumusnya sebagai berikut.

N = . Dengan penjelasan rumus sebagai berikut.

1. Witten-Bell

Bell, Cleary, Witten dalam Chen dan Goodman (1998), dikembangkan dengan tugas untuk memadatkan suatu teks, dan dapat dipertimbangkan sebagai Jelinek-Mercer *smoothing* praktis.

1. *Ristad Natural Discounting*

Algoritma *smoothing ristad natural discounting* ditemukan oleh Eric Sven Ristad pada tahun 1995 dalam jurnal yang berjudul “*A Natural Law of Succession”*. *Ristad Natural Discounting Law* dibuat untuk menyempurnakan *Laplace Law of Succession* dan *Lidstone Law of Succesion.* (Ristad, 1995)

1. Kneser-Ney

Kneser dan Ney dalam Chen dan Goodman (1998), telah memperkenalkan sebuah terobosan positif untuk algoritma *discounting*, dimana distribusi *low-order* digabungkan distribusi *high-order* yang dibangun dengan cara yang baru. Pada algoritma sebelumnya, distribusi *lower-order* umumnya dianggap versi yang lebih halus dari distribusi *lower-order* maximum likehood. Namun, distribusi *lower-order* adalah faktor signifikan dalam gabungan yang dimana didalamnya hanya sedikit atau tidak ada perhitungan yang disajikan distribusi *high-order*. Akibatnya harus dioptimalkan agar memiliki kinerja yang baik dalam situasi ini.

Untuk mewujudkan contoh yang nyata, dengan pertimbangan untuk membangun model bigram pada data dimana ada kata yang sangat umum, misalnya kata fransisco, dimana kata ini adalah kata setelah san. Karena probabilitas *c*(FRANCISCO), maka probabilitas unigram *p*(FRANCISCO) akan tinggi dan algoritma seperti *absolute discounting* akan menetapkan probabilitas yang relatif tinggi untuk kata FRANCISCO setelah riwayat baru bigram . Namun secara imtuitif seharusnya nilai probabilitasnya tidak tinggi karena dalam data *training* kata FRANCISCO hanya diikuti oleh satu riwayat kata. Yaitu, kemungkinan kata FRANCISCO seharusnya menerima probabilitas unigram yang rendah karena satu-satunya kata itu muncul adalah ketika sebelum kata FRANCISCO diawali dengan kata SAN, dalam hal ini probabilitas bigram memodelkan probabilitasnya dengan baik.

Memperluas garis penalaran ini, kemungkinan probabilitas unigram yang digunakan tidak sebanding dengan jumlah kemunculan kata, tetapi sebaliknya dengan jumlah kata yang berbeda yang mengikuti. Untuk memberikan argument intuitif, melewati data pelatihan secara berurutan dan membangun model bigram pada data sebelumnya untuk memprediksi kata saat ini. Kemudian, setiap kali bigram saat ini tidak terdapat dalam data sebelumnya, probabilitas unigram akan menjadi faktor besar dalam probabilitas bigram saat ini. Jika menetapkan hitungan ke unigram yang sesuai setiap kali peristiwa semacam itu terjadi, maka jumlah hitungan yang ditetapkan untuk setiap unigram berdasarkan jumlah kata yang berbeda yang mengikuti. Faktanya, pada Kneser-Ney *smoothing* probabilitas unigram dalam model bigram dihitung dengan cara ini. Perhitungan pada paper ini berbeda dengan paper aslinya. (Chen dan Goodman, 1998)

Alasan yang diberikan pada paper asli adalah bahwa pemilihan distribusi *low-order* sehingga distribusi *high-ordeir smoothed* harus sesuai dengan marginal dari data pelatihan. Untuk contohnya, untuk model bigram dipilihlah distribusi *smoothed* yang memenuhi Batasan unigram marginal untuk semua w1:

(2.14)

Sisi kiri persamaan ini adalah unigram marginal unigram dari distribusi bigram , dan sisi kanan persamaan ini adalah frekuensi unigram yang ditemukan dalam data pelatihan. (Chen dan Goodman, 1998)

Disini, disajikan distribusi yang berbeda dari distribusi yang dihasilkan daripada distribusi yang disajikan oleh Kneser-Ney (1995). Diasumsikan bahwa model memiliki bentuk yang diberikan dalam persamaan. (Chen dan Goodman, 1998)

Adapun persamaan diatas adalah persamaan yang berlawanan dengan yang ada di *paper* yang asli.

Dimana , dipilih untuk menghasilkan jumlah distribusi menjadi 1. Jadi, interpolasi dilakukan dengan distribusi *lower-order,* tidak hanya dilakukan dengan kata-kata yang memiliki nilai 0 dalam distribusi *higher-order*. (Chen dan Goodman, 1998)

1. Modified Kneser-Ney

Modified Kneser-Ney ditemukan oleh dua orang penemu bernama Stanley F. Chen dan Joshua Goodman pada tahun 1998, yang membedakan algoritma Kneser-Ney dengan Modified Kneser-Ney adalah pada penggunaan variable *discount* D. Pada algoritma Kneser-Ney untuk semua jumlah probabilitas bukan nol dinyatakan dalam *discount* D, sedangkan pada Modified Kneser-Ney terdapat tiga jenis parameter D yaitu *D1, D2,* dan *D3+*, yang diaplikasikan menjadi *n-gram* dengan, satu, dua, dan tiga atau bahkan lebih dengan masing-masing perhitungannya. Dengan kata lain daripada menggunakan persamaan (22) digunakan persamaan seperti ini

Dimana

Untuk membuat jumlah distribusi menjadi 1, maka

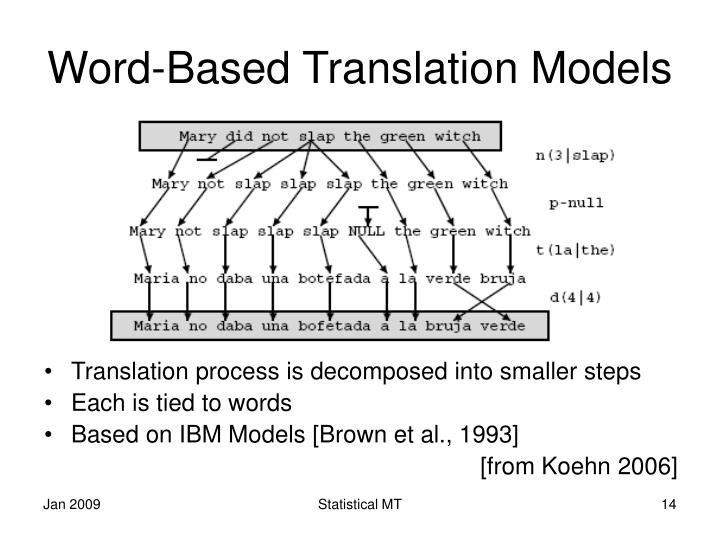
(2.18)

Dimana dan didefinisikan secara analog menjadi .

Seperti Ney, Essen, dan Kneser (1994) telah mengembangkan untuk estimasi *D* yang lebih optimal untuk *absolute discounting* dan Kneser-Ney *smoothing* sebagai fungsi dari perhitungan data *training* seperti pada persamaan (2.12), ini memungkinkan untuk membuat persamaan analog untuk mengestimasikan nilai yang optimal untuk *D1, D2,* dan *D3*. Hubungan analog algoritma Modified Kneser-Ney *smoothing* seperti berikut.

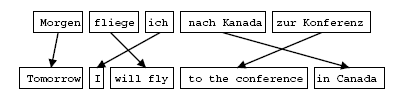
### Translation Model

*Translation Model* merupakan salah satu komponen penting pada *statistical machine translation* dalam proses penerjemahan. *Translation Model* digunakan untuk memasangkan teks *input* dalam bahasa sumberdengan teks *output* dalam bahasa sasaran. Dalam mesin penerjemah statistik terdapat dua model penerjemahan, yaitu *word-based translation model* (model translasi berbasis kata) dan *phrase-based translation model* (model translasi berbasis frase) (Hansel Tanuwijaya & Hisar M Manurung, 2009).



**Gambar 2. 6** Contoh penerjemahan dengan word-based translation model

Proses yang terjadi pada *word based translation model* adalah penerjemahan kata, *reordering*, dan *insertion*. (Koehn, Phillip, 2005) dapat dilihat pada gambar 2.6.



**Gambar 2. 7** Contoh penerjemahan dengan phrase based translation model

Menurut Koehn & Phillip (2005) pada H.Tanuwijaya menyatakan Proses penerjemahan dalam *phrase-based translation model* dapat dipecah menjadi beberapa bagian yaitu: membagi kalimat bahasa asal menjadi barisan frase, menerjemahkan setiap frase ke bahasa tujuan, dan *reordering*. Contoh penerjemahan menggunakan *phrase-based translation model* seperti yang diperlihatkan pada gambar 2.7.

*Translation model* berbasis frase saat ini lebih dipilih karena pada *translation model* tingkat kata menghilangkan banyak konteks local selama proses translasi yang berarti hasil terjemahannya lebih baik dan berkualitas dibanding hasil terjemahan dari *word based translation model.*

## Moses

Moses merupakan *software* gratis implementasi dari pendekatan dengan menggunakan metode statistik (*statistical machine translation*). Pendekatan ini adalah pendekatan yang sering dipakai pada saat ini, dan telah dipakai oleh sistem penerjemahan oleh Google dan Microsoft. Pada *statistical machine translation* (SMT), sistem penerjemahan melakukan *training* pada sekumpulan korpus paralel. Korpus paralel adalah kumpulan dari kalimat-kalimat dalam dua bahasa yang berbeda, yang merupakan kalimat-kalimat selaras, setiap kalimat dalam satu bahasa cocok dengan kalimat yang diterjemahkan dalam bahasa lain. Moses dirilis dibawah lisensi LGPL (*Lesser General Public License*) dan tersedia sebagai kode sumber dan binary untuk Windows dan Linux. Perkembangannya didukung oleh proyek Euro Matrix, dengan pendanaan oleh *European Commision* (Koehn, 2007).

Moses Decoder bekerja untuk menemukan skor tertinggi dari kalimat pada bahasa target berdasarkan model translasi yang sesuai dengan bahasa sumber. Model bahasa dibangun dengan korpus monolingual bahasa target menggunakan SRILM, KenLM ataupun IRSTLM sesuai dengan *minimum requirements* model bahasa (*language model*) dari moses. Selain itu, dengan menggunakan mgiza dibangun model translasi dan susunan kata untuk penerjemahan yang ditraining dari korpus parallel (<http://www.statmt.org/moses/?n=Moses.Overview>).

### Decoder

Fungsi *decoder* adalah untuk mencari teks dalam bahasa sasaran (*target language*) yang memiliki probabilitas paling besar dengan pertimbangan faktor *translation model* dan *language model*. Perhitungan (hasil terjemahan) dapat dituliskan seperti persamaan 2.18

=

Keterangan:

T = Hasil terjemahan

argr max = Nilai maksimal

P(T|S) = Probabilitas bahasa target terhadap bahasa sumber

P(S|T) = Probabilitas bahasa sumber terhadap bahasa target

P(S) = Probabilitas bahasa sumber

P(T) = Probabilitas bahasa target

Fungsi *max* mencari T (bahasa sasaran) yang dapat memberikan nilai probabilitas terbesar yang diperoleh. Proses *decoding* menggunakan algoritma *beam search. Beam search*  adalah algoritma pencarian heuristik yang merupakan optimasi dari pencarian *best-first search* yang mengurangi kebutuhan memorinya. Terdapat dua konsep penting dalam algoritma *beam search* yang digunakan, yaitu konsep pemangkasan (*pruning*) dan estimasi *future cost*. Berikut ini adalah contoh proses penerjemahan pada *phrase-based statistical machine translation* :

1. Memilih kata-kata asing dalam bahasa sumber

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Maria | No | dio | Una | bofetada | a | la | bruja | verde |

1. Mencari terjemahan frase bahasa sasaran
2. Menambahkan frase bahasa sumber pada akhir terjemahan parsial

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Maria | No | dio | Una | bofetada | a | la | bruja | Verde |

|  |
| --- |
| Mary |

1. Menandai bahasa sumber yang telah diterjemahkan

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Maria | no | dio | una | Bofetada | a | la | bruja | vendo |

|  |
| --- |
| Mary |

1. Penerjemahan dari satu- ke- banyak (one-to-many)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Maria | No | dio | una | Bofetada | A | la | bruja | verde |

|  |  |
| --- | --- |
| Mary | did not |

1. Penerjemahan dari banyak- ke- satu (many-to-one)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Maria | No | dio una bofetada | A | la | bruja | verde |

|  |  |  |
| --- | --- | --- |
| Mary | did not | Slap |

1. Reordering (penyususnan kembali)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Maria | No | dio una bofetada | A | la | bruja | verde |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mary | did not | Slap | the | Green |

Keterangan :

* Blok berwarna biru merupakan proses pencarian frasa
* Blok berwarna kuning merupakan frasa sudah diterjemahkan. (Manning dan Schutze, 2000)

## Automatic Evaluation

Suatu mesin penerjemah statistik membutuhkan suatu sistem evaluasi otomatis untuk menentukan kualitas terjemahan. Kualitas mesin penerjemah statistik secara umum dinilai dari hasil terjemahan yang dihasilkan. Penilaian dapat dilakukan secara manual dan otomatis. Penilaian secara manual merupakan cara penilaian yang terbaik karena memberikan nilai akurasi yang lebih tinggi. Namun penilaian secara manual memiliki kekurangan yaitu membutuhkan sumber daya manusia (ahli bahasa) yang banyak dan tentunya membutuhkan waktu yang lama.

Sistem evaluasi otomatis yang populer saat ini adalah BLEU (*Bilingual* *Evaluation Understudy*). BLEU adalah sebuah algoritma yang berfungsi untuk mengevaluasi kualitas dari sebuah hasil terjemahan yang telah diterjemahkan oleh mesin dari satu bahasa alami ke bahasa lain. Ide utama dibalik ini adalah “semakin dekat terjemahan sebuah mesin dengan terjemahan manusia, maka akan semakin baik” (Papineni, 2002). BLEU mengukur *modified n-gram precission score* antara hasil terjemahan otomatis dengan tejemahan rujukan dan menggunakan konstanta yang dinamakan *brevity penalty.*

Untuk menghitung nilai *precission score* dapat dilakukan dengan menghitung jumlah kata pada hasil terjemahan (*unigram*) yang sesuai dengan rujukan dan dibagi dengan jumlah kata (*unigram*) yang ada pada hasil terjemahan. Namun sayangnya mesin penerjemah statistik dapat menghasilkan kata-kata dengan berlebihan, sehingga menghasilkan terjemahan yang mustahil tetapi memiliki *precission* yang tinggi. Dalam penelitian yang dilakukan oleh Papineni, dihasilkan beberapa perubahan yang dikenal dengan metode *modified* n-*gram* *precision.* Untuk menghitungnya, pertama kali hitung berapa kali jumlah maksimal dari kata yang muncul dalam terjemahan rujukan tunggal. Selanjutnya gabungkan jumlah total dari setiap kalimat terjemahan dengan jumlah maksimal rujukan. Contoh perhitungan dengan *modified unigram precision* diperlihatkan pada Tabel 2.4.

**Tabel 2. 4** Contoh Perhitungan modified unigram precission (Tanuwijaya, 2009)

|  |  |
| --- | --- |
| **Hasil Terjemahan** | **Rujukan** |
| *the the the the the the the* | *The cat is on the mat* |

Nilai *precision unigram* adalah 7/7. Sedangkan nilai *modified unigram* 2/7. Jumlah maksimum kata “*the*“ pada rujukan adalah 2 dan jumlah *unigram* pada hasil terjemahan adalah 7 (Tanuwijaya, 2009). Nilai BLEU didapat dari hasil perkalian antara *brevity penalty* dengan rata-rata geometri dari *modified precision score*. Semakin tinggi nilai BLEU, maka semakin akurat dengan rujukan. Nilai dari BLEU berada pada rentang 0 sampai 1. Suatu terjemahan akan mencapai nilai 1 jika terjemahan tersebut identik dengan terjemahan rujukan. Oleh karena itu, meskipun dengan penerjemahan oleh manusia tidak mungkin akan menghasilkan nilai 1. Sangat penting untuk diketahui bahwa semakin banyak terjemahan rujukan per kalimatnya, maka akan semakin tinggi nilainya. Untuk menghasilkan nilai BLEU yang tinggi, panjang kalimat hasil terjemahan harus mendekati panjang dari kalimat referensi dan kalimat hasil terjemahan harus memiliki kata dan urutan yang sama dengan kalimat referensi. Rumus BLEU sebagai berikut dapat dilihat pada persamaan 2.19, 2.20 dan 2.21

(2.21)

(2.22)

(2.23)

Keterangan:

BP = *brevity penalty*

c = jumlah kata dari hasil terjemahan otomatis

r = jumlah kata rujukan

𝑃𝑛 = *modified precission score*

𝑤𝑛 = 1/N (standar nilai N untuk BLEU adalah 4)

𝑝𝑛 = jumlah *n-gram* hasil terjemahan yang sesuai dengan rujukan dibagi jumlah n-gram hasil terjemahan

## Case Folding

*Case folding* merupakan tahapan mengubah huruf dalam dokumen menjadi huruf kecil. Karakter selain huruf dihilangkan dan dianggap pembatas (*delimiter*) (Marlinda & Rianto, 2013). Contoh penggunaan *case folding* yaitu pada gambar 2.8.

**Gambar 2. 8** Contoh case folding

Teks output

Teks input

Siapa nama anak itu

siapa nama anak itu

## Tokenizing

Tahap *tokenizing* adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya (Marlinda & Rianto, 2013). Contoh dari tahap *tokenizing* seperti terlihat pada gambar 2.9.

siapa

nama

anak

itu

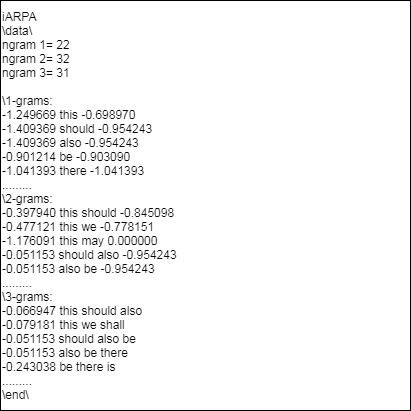
siapa nama anak itu

Teks Output

**Gambar 2. 9** Contoh tokenizing

## Language Model

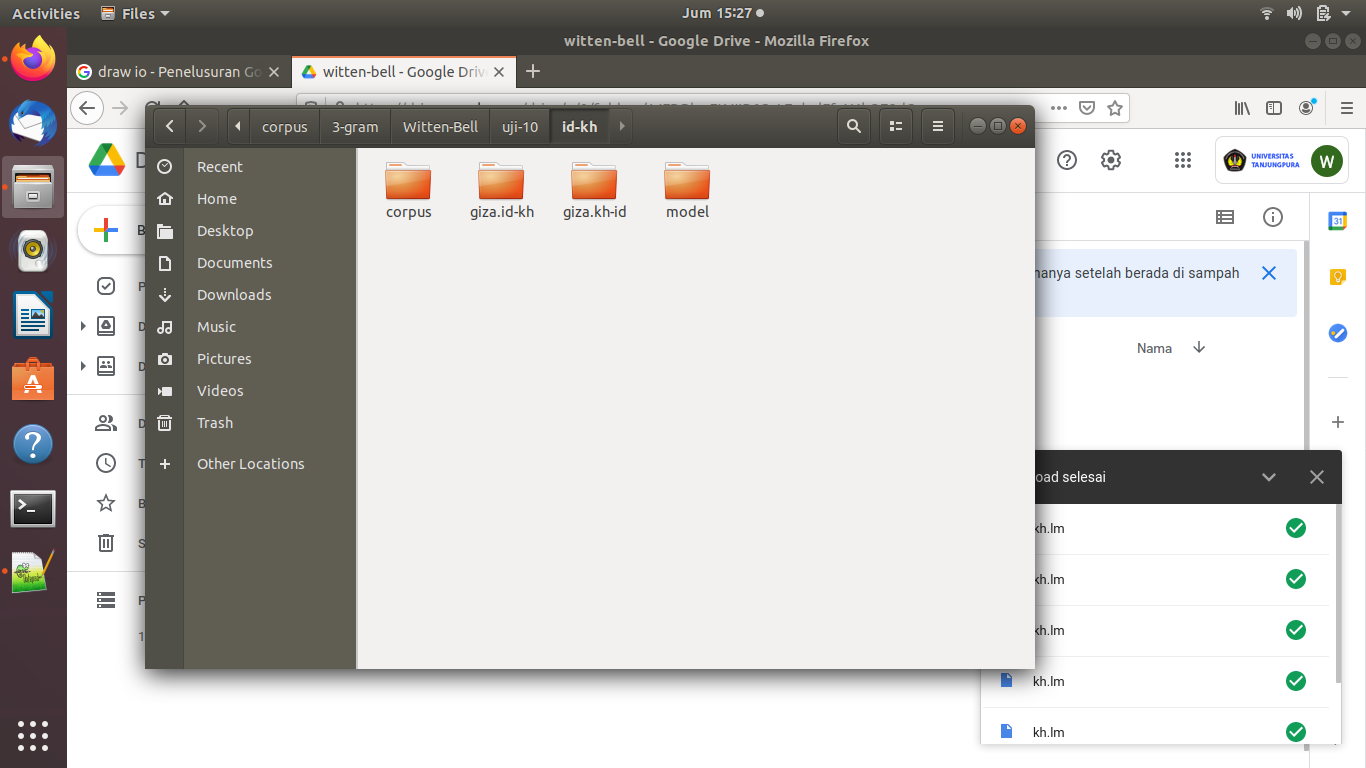
Adapun tahap ini merupakan tahap dimana mengubah korpus yang sudah dilakukan tahap *case folding* dan *Tokenizing* menjadi sebuah language model (ARPA *file*) dapat dilihat pada gambar 2.10



**Gambar 2. 10** Contoh hasil proses pembuatan ARPA file[Bertoldi, 2008]

## Translasi Model

Pada tahap ini kita melakukan *translasi model* dengan korpus yang sudah kita lakukan tahap preprocessing dan ARPA *file* adapun hasil dari pembuatan *translasi model* dapat dilihat pada gambar 2.11.



**Gambar 2. 11** Contoh file model translasi [Hadi, 2014]

## Cara Pengujian

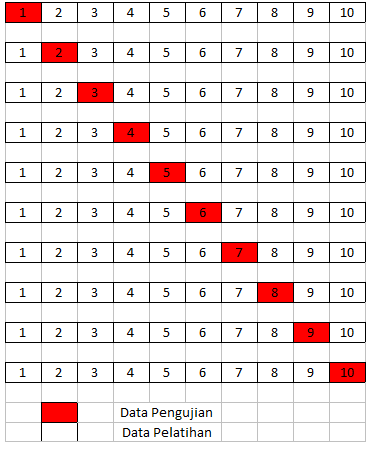
Adapun cara pengujian yang akan dilakukan dengan menguji korpus yang sudah ada dengan enam Algoritma smoothing yang digunakan untuk melihat hasil grafik yang dihasilkan. Dilakukannya pengujian beberapa kali dengan menambahkan korpus training secara konsisten hingga hasilnya yang sudah terlihat paralel maka hasilnya sudah dapat kita simpulkan untuk berhenti pada jumlah tersebut. Berikut cara pengujian dapat dilihat di gambar 2.12.

**Gambar 2. 12** Contoh grafik pengujian jumlah korpus

Adapun untuk penelitian ini akan dilakukan pengujian seperti gambar 2.12 dengan melakukan percobaan beberapa kali sampai nilai *BLEU score* yang dihasilkan terlihat paralel maka dari situ pengujiannya sudah bisa kita simpulkan dan untuk nilai akurasi yang akan kita jadikan perbandingan yaitu terletak pada nilai akhir akurasi yang muncul pada pengujian dan nilai *BLEU Score* Algoritma *smoothing* paling tinggi pada perbandingan diatas akan dijadikan bahan untuk metode K-Fold Cross Validation.

## K-Fold Cross Validation

Cross-validation (CV) adalah metode statistik yang dapat digunakan untuk mengevaluasi kinerja model atau algoritma dimana data dipisahkan menjadi dua subset yaitu data proses pembelajaran dan data validasi / evaluasi. Model atau algoritma dilatih oleh subset pembelajaran dan divalidasi oleh subset validasi. Selanjutnya pemilihan jenis CV dapat didasarkan pada ukuran dataset. Biasanya CV K-fold digunakan karena dapat mengurangi waktu komputasi dengan tetap menjaga keakuratan estimasi dapat dilihat pada gambar 2.13.



**Gambar 2. 13** Contoh skema 10 fold cross validation

Adapun untuk penelitian ini akan digunakan skema 27 Fold Cross Validation, sehingga akan dilakukan 27 kali percobaan pengujian akurasi setiap jenis *Language Model Toolkit*. Jadi dalam pemilihan data untuk korpus uji dari jumlah total korpus kalimat sebanyak 2.700 dibagi menjadi 27 kali percobaan didapatkan 100 kalimat untuk setiap percobaannya. Misalnya pada percobaan 1 korpus ujinya diambil dari kalimat 1 sampai 100 sedangkan untuk korpus yang di*-training* diambil dari kalimat 101 sampai 2.700, percobaan 2 korpus ujinya diambil dari kalimat 101 sampai 200 sedangkan korpus yang di*-training* diambil dari kalimat 1 sampai 100 dan 201 sampai 2.700 begitu seterusnya sampai percobaan ke 27. Adapun untuk perhitungan rata-rata nilai akurasi dari algoritma yang nilai score BLEU nya paling tinggi pada *language model toolkit SRILM* dapat dinyatakan dengan persamaan sebagai berikut:

Salah satu alasan utama untuk menggunakan validasi silang daripada menggunakan validasi konvensional (misalnya mempartisi kumpulan data menjadi dua set, yaitu 70% untuk pelatihan dan 30% untuk pengujian) adalah bahwa tidak ada cukup data yang tersedia untuk mempartisinya menjadi pelatihan terpisah dan data pengujian tanpa kehilangan pemodelan atau kemampuan pengujian yang signifikan. Dalam kasus ini, cara yang adil untuk memprediksi model prediksi dengan tepat adalah dengan menggunakan validasi silang sebagai teknik umum yang kuat. (Grossman, 2010).

Singkatnya, validasi silang menggabungkan (rata-rata) ukuran kecocokan (prediksi error) dengan melihat nilai rata-rata pada setiap putaran untuk mendapatkan perkiraan kinerja model prediksi yang lebih akurat.

## Bahasa Khek Singkawang

Bahasa Khek pada dasarnya memang bahasa dari daratan China, tetapi ia berbeda dengan bahasa Mandarin, yang menjadi salah satu bahasa internasional walau terdapat kosakata yang mirip-mirip. Menurut tokoh masyarakat Tionghoa Pontianak, Andreas Acui Simanjaya, bahasa Khek memang menjadi bahasa pergaulan di keluarga-keluarga Tionghoa di Kalimantan Barat. Bahasa Khek sendiri asal-muasalnya dari para leluhurnya, orang Hakka yang merupakan bagian dari suku Han yang tersebar di kawasan pegunungan Provinsi Guangdong, Fujian, dan Guangxi di China sebelah utara. Mereka beremigrasi ke Asia Tenggara, termasuk Indonesia. Bahasa Khek ini berkembang pada ranah oralitas saja. Artinya, ini menjadi sekadar bahasa tutur dan pergaulan di keluarga warga Tionghoa dan interaksi antar-Tionghoa. (Ikhsan, 2015)

Dalam proses pembelajaran, penulis menemukan dalam kedua bahasa yang memiliki rumpun yang sama ini, yaitu Bahasa Mandarin dan Bahasa Khek terdapat persamaan dan perbedaan. Persamaan yang sering didapati, misalnya :

1. Bahasan Mandarin : 慢慢走 (man man zou)

Bahasa Khek : man man cheu

Bahasa Indonesia : jalan pelan pelan

1. Bahasan Mandarin : 水 (shui)

Bahasa Khek : sui

Bahasa Indonesia : air

Selain persamaan, ada juga perbedaaannya. Perbedaannya dari segi tata bahasa misalnya :

1. Bahasan Mandarin : 你先走 (ni xian zou)

Bahasa Khek : *nyi cheu sian / nyi sian cheu*

Bahasa Indonesia : kamu duluan

1. Bahasan Mandarin : 慢慢儿骑 (man man er qi)

Bahasa Khek : *khi man man / man man khi*

Bahasa Indonesia : berkendara perlahan

Bahasa Khek Singkawang memiliki dialek yang sulit dipahami, susunan kata dalam kalimatnya pula sulit untuk dipahami. Bahasa Khek Singkawang memiliki keunikan, susunan katanya berbeda dengan bahasa Indonesia.

Contoh :

* kami tidak terlalu lapar

*ngai tew em boi an tu si ki .*

* saya tidak bisa menemukan tas saya .

*ngai em hiau chim nga su pau .*

Kata ngai memiliki arti saya, tetapi kata ngai dapat ditambahkan kata lain seperti ngai tew dan akan memiliki makna yang berbeda dari sebelumnya.

* kamu mirip dia

*nyi chiong ki*

* kue itu seperti kue ibu ku

*ka pan chiong nga mak pan*

Kata chiong dapat memiliki beberapa makna yang berbeda misalnya seperti/mirip/kayak.

# Metodologi Penelitian

## Data dan Perangkat Penelitian

### Data Penelitian

Data Penelitian yang digunakan berupa dokumen-dokumen bahasa Indonesia seperti novel, cerita rakyat, dan juga artikel-artikel berita yang kemudian dari dokumen-dokumen dari bahasa Indonesia tersebut diolah menjadi korpus setelah itu korpus bahasa Indonesia diterjemahkan ke dalam bahasa Khek Singkawang untuk diolah menjadi korpus bahasa Khek Singkawang.

### Perangkat Penelitian

Perangkat penelitian yang digunakan dalam penelitian ini berupa perangkat keras dan perangkat lunak. Adapun alat yang digunakan adalah sebagai berikut.

1. Perangkat Keras

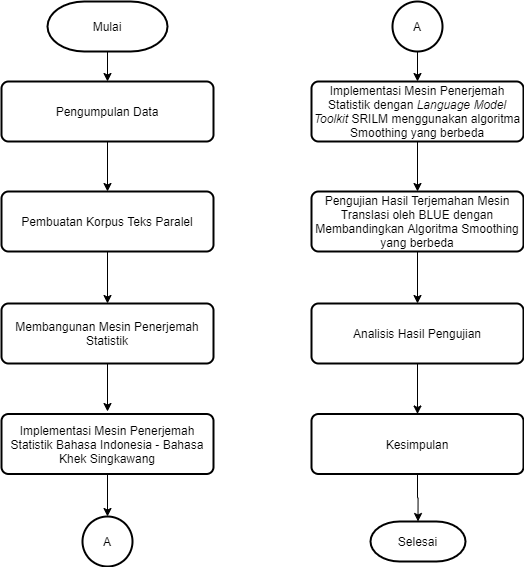
Pada dasarnya, perangkat keras adalah komponen komputer yang dapat dilihat secara langsung atau berbentuk nyata dan berfungsi untuk mendukung proses komputerisasi yang ada. Dalam penelitian ini, Laptop Lenovo Ideapad 100-141BD yang digunakan dengan spesifikasi sebagai berikut:

1. Prosesor Intel Core i3-5005U CPU 2.00 GHz Up to 3,5 GHz
2. Grafis Nvidia GeForce 920MX VRAM 2GB
3. Memori RAM 4 GB DDR3
4. Storage Hard Disk 500 GB HDD 5400 RPM
5. Perangkat Lunak

Berbeda dengan perangkat keras, perangkat lunak adalah program komputer yang berfungsi sebagai sarana interaksi antara pengguna dan perangkat keras. Adapun perangkat lunak yang digunakan adalah sebagai berikut.

1. Sistem Operasi Linux Ubuntu 18.04 64 Bit
2. SRILM untuk pemodelan Bahasa
3. Mgiza untuk pemodelan translasi
4. Moses untuk decoding
5. BLEU untuk pengujian otomatis akurasi
6. Microsoft Word dan Notepade++ untuk teks editor
7. Terminal untuk menjalankan command-line

## Langkah-langkah Penelitian



**Gambar 3. 1** Diagram Alir Penelitian

Gambar 3.1 merupakan langkah-langkah yang harus dilakukan dalam penelitian ini.

### Pengumpulan Data

Untuk mengumpulkan data digunakan metode wawancara dan observasi untuk memperoleh data yang diperlukan dalam penelitian. Langkah-langkah yang dilakukan adalah sebagai berikut.

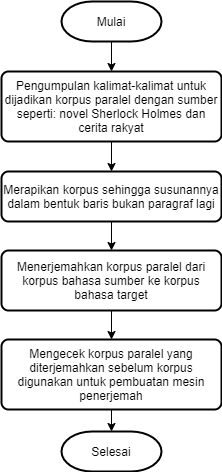
1. Wawancara

Diperlukan ahli bahasa sebagai narasumber dalam pengujian hasil translasi mesin penerjemah pada penelitian ini. Penulis melakukan wawancara kepada orang yang akan menjadi ahli bahasa dengan cara berdiskusi sejauh mana mereka mengerti bahasa Khek Singkawang dan kesediaan mereka untuk menjadi ahli bahasa dalam penelitian ini.

1. Observasi

Mencari dokumen-dokumen bahasa Indonesia yang nantinya akan diserahkan kepada ahli bahasa untuk diterjemahkan ke dalam bahasa Khek Singkawang yang akan dijadikan korpus dalam penelitian, yaitu korpus bahasa Indonesia dan korpus bahasa Khek Singkawang.

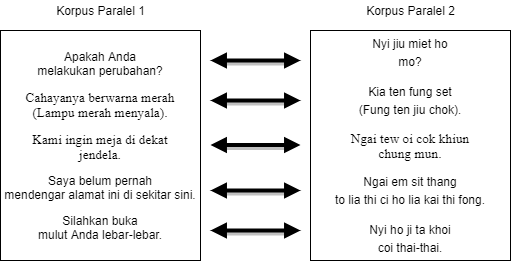
### Pembuatan Korpus Teks Paralel



**Gambar 3. 2** Proses pembuatan korpus

Korpus merupakan kumpulan dari beberapa teks sebagai sumber penelitian bahasa dan sastra. Penelitian yang akan dilakukan terdapat dua buah korpus paralel yang digunakan yaitu korpus bahasa Indonesia dan korpus bahasa Khek Singkawang. Adapun korpus paralel pada penelitian ini berupa dokumen-dokumen bahasa Indonesia yang kemudian diterjemahkan ke dalam bahasa Khek Singkawang proses pembuatan korpus terlihat pada gambar 3.2.

Korpus paralel yang digunakan disimpan dengan nama yang sama, tetapi berbeda format berkasnya. Korpus paralel yang dibuat dalam bentuk file teks disimpan dengan format .id untuk bahasa Indonesia dan .kh untuk bahasa Khek Singkawang. Contoh korpus paralel yang dibuat dalam bentuk file teks dapat dilihat pada Gambar 3.3



**Gambar 3. 3** Contoh korpus paralel bahasa Indonesia – bahasa Khek Singkawang

### Membangun Mesin Penerjemah Statistik

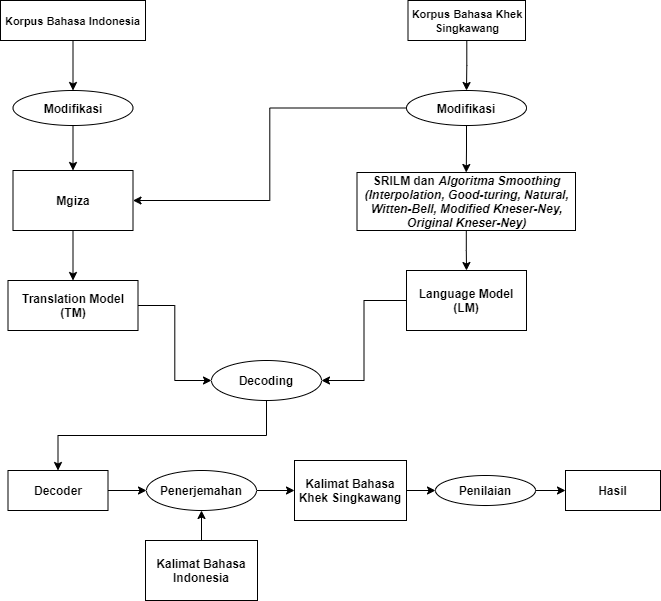
Membangun mesin penerjemah statistik dilakukan dengan cara melakukan instalasi perangkat lunak. Adapun perangkat lunak yang akan di-*install* untuk membangun mesin penerjemah statistik adalah sebagai berikut.

1. Moses decoder untuk membangun mesin penerjemah.
2. Mgiza untuk *compile* Moses.
3. SRILM untuk pemodelan bahasa.
4. Multi-bleu.perl untuk pengujian akurasi.

Seluruh perangkat lunak yang digunakan dalam membangun mesin penerjemah statistik merupakan perangkat lunak yang bersifat *open source* dan di-*install* pada system operasi Ubuntu.

### Implementasi Mesin Penerjemah Statistik Bahasa Indonesia – Bahasa Khek Singkawang dengan *Language Model Toolkit* SRILM

Arsitektur sistem pada penelitian ini terdiri dari beberapa proses pemodelan yaitu pemodelan bahasa, pemodelan translasi, *decoding*, dan evaluasi hasil terjemahan Arsitektur sistem mesin penerjemah statistik ditunjukkan pada Gambar 3.4



**Gambar 3. 4** Arsitektur sistem Mesin Penerjemah Statistik bahasa Indonesia – bahasa Khek Singkawang [Modifikasi: Sujaini dan Negara, 2015]

Korpus paralel yang digunakan terdiri dari dua buah korpus yaitu korpus bahasa Indonesia dan bahasa Khek Singkawang. Korpus paralel digunakan pada proses TM training dimana sebelumnya dilakukan modifikasi untuk menyesuaikan format korpus yang didukung untuk pembuatan model. Pada proses LM training yang digunakan adalah korpus Khek Singkawang saja. Proses modifikasi korpus yang dilakukan diantaranya adalah proses menghilangkan tanda baca titik di akhir kalimat, menghapus spasi berlebih, menjadikan huruf pada korpus menjadi lowercase atau huruf kecil semua, dan mentokenisasi kata dan tanda baca.

Selanjutnya korpus yang telah dimodifikasi digunakan dalam proses translation model training oleh Mgiza yang akan menghasilkan translation model. Proses pemodelan menggunakan GIZA++ menghasilkan vocabulary corpus, word alignment, dan lexical table. Langkah selanjutnya, korpus Bahasa Khek Singkawang digunakan dalam proses language model training oleh SRILM dan menghasilkan tabel model bahasa dengan n-gram data. Language model dan translation model yang telah didapatkan dari proses training akan digunakan dalam proses penerjemahan oleh moses dengan menggunakan input bahasa Indonesia dengan terjemahannya dalam bahasa Khek Singkawang.

Gambar 3.4 merupakan arsitektur sistem mesin penerjemah statistik bahasa Indonesia ke bahasa Khek Singkawang yang terdiri dari beberapa tahapan, yaitu pemodelan bahasa, pemodelan translasi, *decoding,* dan proses evaluasi hasil terjemahan yang mana proses evaluasi hasil terjemahan akan dijelaskan pada poin selanjutnya. Berikut penjelasan dari tahapan pemodelan bahasa, pemodelan translasi dan proses *decoding* tersebut.



**Gambar 3. 5** Proses pemodelan bahasa

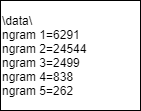
Gambar 3.5 merupakan proses pemodelan bahasa diantaranya ada empat proses yang harus dilalui yaitu: Text (Input), *N-gram Modelling*, *N-gram Probabilities*, *Discont Methods & Frequency Smoothing*, ARPA File (Output).

1. Text (Input)

Merupakan proses memasukkan korpus yang telah diproses untuk dilakukan pemodelan bahasa.

1. *N-gram Modelling*

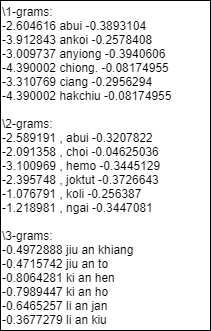
*N-gram Modelling* merupakan proses menentukan berapa banyak n-gram yang diinginkan dalam suatu pemodelan bahasa atau berapa banyak penggalan kata ataupun karakter yang diinginkan dalam pembuatan pemodelan bahasa. Cara mengatur jumlah n-gram dilakukan pada saat training dengan menambahkan *script order* diketik pada *script*, “. . . -o 5 . . .” seperti menandakan berarti jumlah pembagian n-gram hanya 5 buah seperti terlihat pada gambar 3.6.



**Gambar 3. 6** N-gram modelling dengan jumlah n-gram lima buah

1. *N-gram Probabilities*

*N-gram Probabilities* merupakan proses untuk menghitung setiap kemungkinan atau probabilitas kemunculan kata atau karakter pada korpus seperti terlihat pada gambar 3.7.

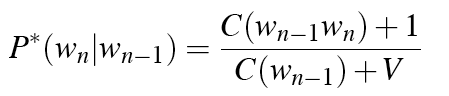


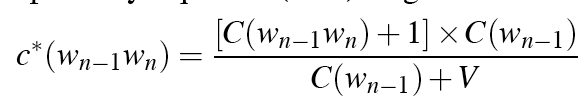
**Gambar 3. 7** N-gram probabilities dari korpus paralel bahasa khek singkawang

1. *Discount Methods* & *Frequency Smoothing*

*Discount Methods* & *Frequency Smoothing* digunakan untuk meminimalisir bahkan menghilangkan probabilitas bernilai “0” pada pemodelan bahasa.

Setelah itu sebagai contoh dengan digunakan *Laplace-smoothing* untuk memunculkan probabilitas yang bernilai “0” tadi dengan menggunakan beberapa rumus.

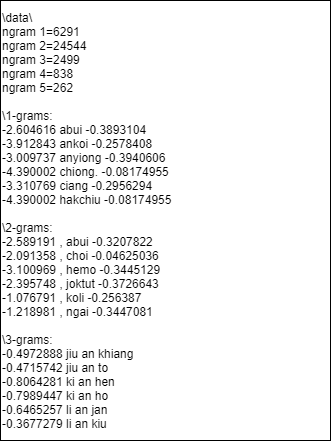




Pada Akhirnya barulah didapatkan hasil dari probabilitas yang semuanya memiliki nilai tanpa terkecuali.

1. ARPA File (Output)

File yang telah diproses sedemikian rupa setelah melewati tahap *N-gram Modelling, N-gram Probabilities, Discount Methods* & *Frequency Smoothing* sehingga siap untuk diproses pada pemodelan translasi untuk membangun mesin penerjemah statistik adapun contoh format ARPA file dapat dilihat pada gambar 3.8.



**Gambar 3. 8** Contoh tabel model bahasa menggunakan Language Model Toolkit SRILM dengan bahasa Khek Singkawang sebagai bahasa target dengan format ARPA file

### Pemodelan Bahasa Oleh SRILM

Pemodelan bahasa oleh SRILM (SRI Language Model) dilakukan pada bahasa target dan menghasilkan table model bahasa dengan n-gram data. Model bahasa n-gram memiliki nilai probabilitas dalam bahasa target. Proses pemodelan bahasa oleh SRILM dapat dilihat pada gambar 3.9 (Jarob, 2016).



**Gambar 3. 9** Proses Pemodelan bahasa menggunakan Language Model Toolkit SRILM dengan bahasa Khek Singkawang sebagai bahasa target [Modifikasi: Hadi,2014]

Proses pemodelan bahasa oleh LM *toolkit* SRILM dapat dilihat pada gambar 3.10. (Chen, 2008).



**Gambar 3. 10** Proses pemodelan bahasa menggunakan language model toolkit srilm

Algoritma pemodelan *language model* pada LM *toolkit* SRILM berdasarkan jurnal dari Berlin Chen pada tahun 2008, dengan jurnal yang berjudul “*Introduction to SRILM Toolkit*”.

1. Corpus

Sekumpulan perbendaharaan kata dari bahasa sumber dan bahasa target yang sudah di *preprocessing* (*cleaning, tokenizing,* dan *case folding)* atau bisa dibilang perbendaharaan kata yang sudah di*training* sehingga siap untuk diproses

1. N-gram count

Perhitungan jumlah setiap kata yang ada pada perbendaharaan kata untuk pemrosesan *Language Model,* lalu setiap kata diklasifikan berdasarkan N-gram. N-gram disini yang dimaksud pembagian kata atau karakter berdasarkan jumlahnya misal 1-gram untuk satu kata atau karakter dan seterusnya.

1. Lexicon

Masukkan (input) perbendaharaan kata *Out Of Vocabulary* (kata yang tidak dikenal) dengan cara menambahkan *command* “- *vocab”* adapun untuk tahapan ini bisa dilewati jika tidak memiliki *corpus* OOV. Sehingga kata yang tergolong OOV dimasukkan ke dalam kategori kata yang dikenali secara otomatis karena tidak ada *input* dari kumpulan kata OOV.

1. Count

Perhitungan probabilitas setiap kata yang ada pada perbendaharaan kata berdasarkan klasifikasi N-gram.

Unigram (1-gram):

*P*(*W1, n*) =

*P*(*W1*) *P*(*W2*) … *P(Wn*)

Bigram (2-gram):

*P*(*W1, n*) =

*P*(*W1*) *P*(*W*2|*W1*) … *P*(*Wn*|*Wn-1*)

Trigram (3-gram):

*P*(*W1, n*) =

*P*(*W1*) *P*(*W2*|*W1*) *P*(*W3*|*W1,2*) … *P*(*Wn*|*Wn-2,n-1*)

1. N-gram count file

Konversi nilai probabilitas setiap kata berdasarkan klasifikasi N-gram dengan cara menjadikan nilai probabilitas setiap kata dengan *log*.

1. Language Model (ARPA file)

Mengumpulkan hasil probabilitas setiap kata yang dikonversi dengan *log* berdasarkan klasifikasi N-gram dan juga menampilkan nilai probabilitas setiap kata dalam format *backoff* dengan tujuan mengefektifkan memori, dimana hasil akhirnya berdasarkan algoritma *smoothing* yang digunakan.

### Pengujian Terjemahan Hasil Mesin Translasi oleh BLEU dengan Membandingkan *Algoritma Smoothing* yang berbeda

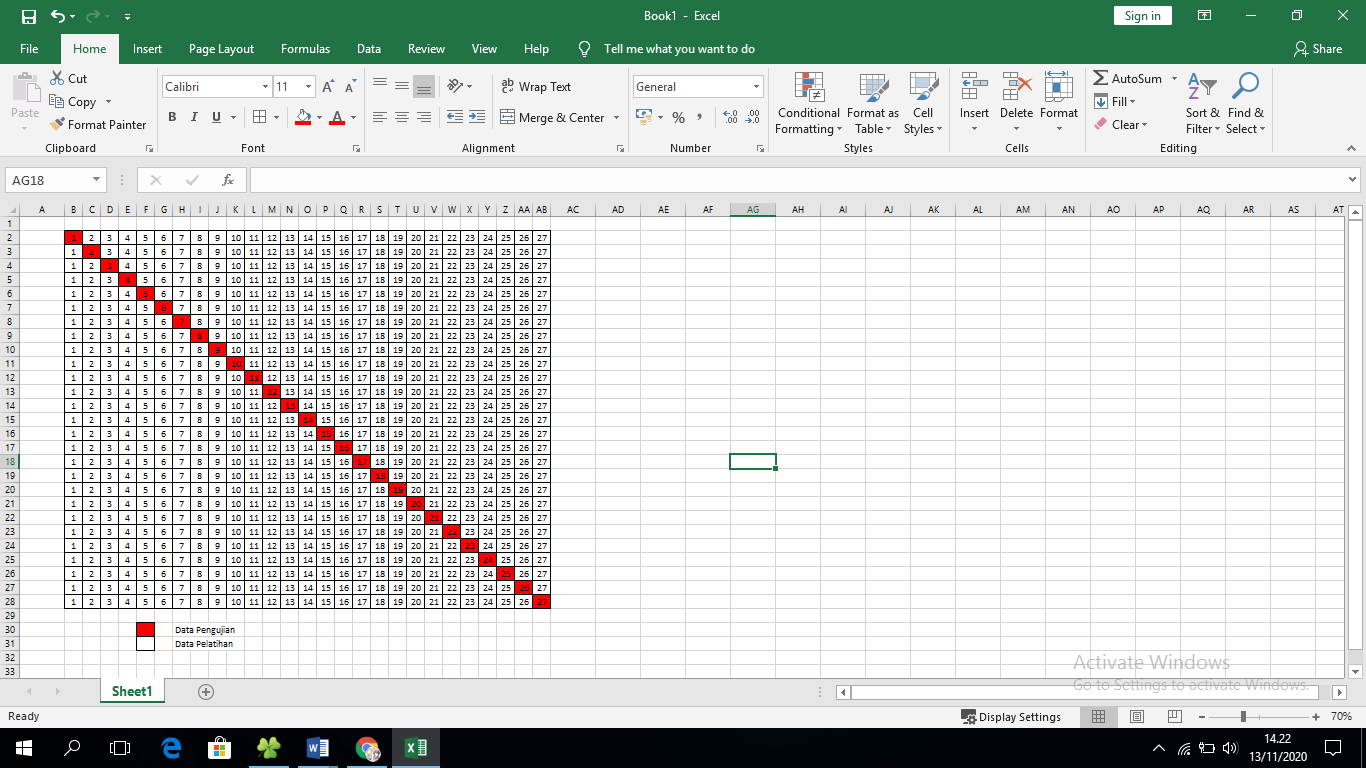
Pengujian hasil terjemahan dilakukan dengan dua cara yaitu otomatis dan manual. Pengujian otomatis dilakukan dengan menggunakan BLEU, serta digunakan metode *K-fold Cross Validation*. Penggunaan metode *K-fold Cross Validation* bertujuan agar nilai yang dihasilkan lebih akurat. Pengujian secara manual adalah metode pengujian yang memiliki tingkat akurasi paling baik, namun akan memakan waktu yang lama karena dilakukan secara manual oleh ahli bahasa. Penjelasan masing-masing metode evaluasi.

#### Evaluasi Otomatis

Pengujian dilakukan untuk mendapatkan nilai akurasi terjemahan mesin translasi. Pengujian dilakukan secara otomatis menggunakan BLEU. BLEU (*Billingual Evaluation Understudy)* mengukur *modified* n-gram *precision score* antara hasil terjemahan otomatis dengan terjemahan rujukan dan menggunakan konstanta yang disebut *brevity penalty.* Nilai BLEU didapat dari hasil perkalian antara *brevity penalty* dengan rata-rata geometri dari *modified precision score.* Dibawah ini merupakan grafik hasil pengujian menggunakan 3 gram dengan 10 kali percobaan dan penambahan jumlah korpus secara konstan yaitu 200 korpus.

**Gambar 3. 11** Contoh grafik pengujian perbandingan nilai BLEU dengan 3 gram

Dari pengujian otomatis menggunakan BLEU diatas diambil algoritma dengan nilai paling tinggi seperti terlihat pada gambar 3.11 nilai *Score* BLEU paling tinggi akan digunakan pada metode *K-Fold Cross Validation* untuk memastikan keakuratan dari hasil uji mesin penerjemah yang diuji berdasarkan jenis *language model toolkit SRILM* dan algoritma *smoothing* yang berbeda pada penelitian ini menggunakan total 2.700 kalimat korpus paralel bahasa Indonesia dan bahasa khek singkawang dengan format 100 kalimat uji dan 2.600 kalimat yang di-*training*.



**Gambar 3. 12** Pengujian dengan metode K-Fold Cross Validation

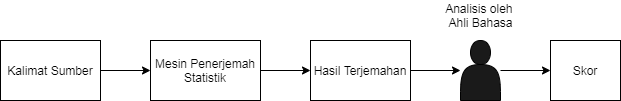
Gambar 3.12 merupakan pengujian dengan metode *K-Fold Cross Validation* menggunakan 2.700 kalimat dimana dilakukan 27 kali percobaan dan per percobaan 100 kalimat sebagai korpus uji

Keterangan:

1. Data Uji = Kalimat 1 – 100
2. Data Uji = Kalimat 101 – 200
3. Data Uji = Kalimat 201 – 300
4. Data Uji = Kalimat 301 – 400
5. Data Uji = Kalimat 401 – 500
6. Data Uji = Kalimat 501 – 600
7. Data Uji = Kalimat 601 – 700
8. Data Uji = Kalimat 701 – 800
9. Data Uji = Kalimat 801 – 900
10. Data Uji = Kalimat 901 – 1.000
11. Data Uji = Kalimat 1.001 – 1.100
12. Data Uji = Kalimat 1.101 – 1.200
13. Data Uji = Kalimat 1.201 – 1.300
14. Data Uji = Kalimat 1.301 – 1.400
15. Data Uji = Kalimat 1.401 – 1.500
16. Data Uji = Kalimat 1.501 – 1.600
17. Data Uji = Kalimat 1.601 – 1.700
18. Data Uji = Kalimat 1.701 – 1.800
19. Data Uji = Kalimat 1.801 – 1.900
20. Data Uji = Kalimat 1.901 – 2.000
21. Data Uji = Kalimat 2.001 – 2.100
22. Data Uji = Kalimat 2.101 – 2.200
23. Data Uji = Kalimat 2.201 – 2.300
24. Data Uji = Kalimat 2.301 – 2.400
25. Data Uji = Kalimat 2.401 – 2.500
26. Data Uji = Kalimat 2.501 – 2.600
27. Data Uji = Kalimat 2.601 – 2.700

#### Evaluasi Manual

Pengujian manual dilakukan oleh ahli bahasa Khek Singkawang. Pengujian ahli bahasa dilakukan apabila telah dilakukan pengujian hasil evaluasi mesin penerjemah statistik dengan BLEU menggunakan metode *K-Fold Cross Validation*. Pengujian ahli bahasa digunakan untuk mengetahui apakah hasil terjemahan dari mesin penerjemah statistik mendekati atau bahkan sudah sesuai dengan terjemahan dari ahli bahasa. Proses evaluasi ahli bahasa dapat dilihat pada gambar 3.13 (Modifikasi: Jarob, 2016).



**Gambar 3. 13** Proses evaluasi secara manual oleh ahli bahasa

### Analisis Hasil Pengujian

Analisis hasil pengujian dilakukan untuk menentukan algoritma *smoothing* mana yang paling tinggi nilai akurasi terjemahannya pada jenis *language model toolkit* SRILM serta faktor yang mempengaruhi mengapa akurasi algoritma tersebut dapat memiliki akurasi yang terbaik berdasarkan kesetaraan penambahan jumlah korpus pada setiap mesin secara konstan, nilai yang dihasilkan pada percobaan yang terakhir akan dilakukan perbandingan sehingga kita dapatkan nilai tertinggi dari algoritma *smoothing* yang digunakan. Setelah didapatkan nilai tertinggi dari algoritma yang digunakan maka dilakukan metode *K-fold Cross Validation* untuk melihat nilai rata-rata akurasi dari keseluruhan korpus yang digunakan, dengan algoritma dan jumlah penggunaan n-gram pada mesin penerjemah.

### Penarikan Kesimpulan

Penarikan kesimpulan dilakukan dengan melihat hasil analisis yang dilakukan dari perbandingan nilai akurasi mesin penerjemah statistik dengan algoritma *smoothing* yang digunakan, dilihat mana algoritma *smoothing* yang menghasilkan nilai akurasi terbaik.

# Hasil dan Analisis

## Hasil Penelitian

Berikut adalah hasil-hasil penelitian yang telah dibuat pada bab sebeumnya. Hasil dari penelitian ini terdiri dari pembuatan korpus teks paralel, membangun mesin penerjemah statistik, dan implementasi mesin penerjemah statistik bahasa Indonesia ke bahasa Khek Singkawang.

## Hasil Pembuatan Korpus Teks Paralel

Korpus teks paralel yang sudah dibuat yang terdiri dari beberapa cerita rakyat dan novel Sherlock Holmes dengan jumlah sebanyak 2.700 kalimat yang dimana dari 2.700 tersebut di ambil 100 kalimat untuk dijadikan sebagai korpus uji kemudian di simpan dengan format korpus.id untuk (bahasa Indonesia) dan korpus.kh untuk (bahasa Khek Singkawang). Hasil pembuatan korpus paralel dapat dilihat pada Tabel 4.1.

| **Bahasa Indonesia** | **Bahasa Khek Singkawang** |
| --- | --- |
| Apakah Anda melakukan perubahan?  Cahayanya berwarna merah (Lampu merah menyala).  Kami ingin meja di dekat jendela.  Saya belum pernah mendengar alamat ini di sekitar sini.  Apakah bus ini berhenti di Stoner Avenue?  Silahkan buka mulut Anda lebar-lebar.  Saya memesan kursi pada bagian bebas asap rokok, tapi ini adalah kursi untuk perokok.  Saya sudah membayar uang muka untuk makan dan hotel.  Saya akan memberitahunya untuk menelepon Anda segera setelah ia kembali.  Apakah tidak apa-apa kalau kita hanya makan makanan ringan? | Nyi jiu miet ho mo?  Kia ten fung set (Fung ten jiu chok).  Ngai tew oi cok khiun chung mun.  Ngai em sit thang to lia thi ci ho lia kai thi fong.  Lia cha tiam choi Stoner Avenue he mo?  Nyi ho ji ta khoi coi thai-thai.  Ngai kau tai mo jan kai cho bui, tapi ngai cho to nyin po jan kai bui.  Ngai ban ho thin thew loi sit ka li kon.  Ngai boi kong gi ti loi ta thien pun nyi he gi to con.  Boi co mai mo he chit ka chiang sit ci ki? |

**Tabel 4. 1** Korpus paralel bahasa Indonesia – bahasa Khek Singkawang

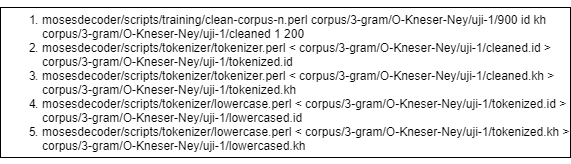
## Membangun Mesin Penerjemah Statistik

Sebagai langkah dalam membangun mesin penerjemah statistik harus dilakukan proses instalasi perangkat lunak yang akan digunakan. Adapun perangkat lunak yang harus di install adalah sebagai berikut:

1. Mosesdecoder untuk membangun mesin penerjemah.
2. Mgiza untuk pemodelan translasi.
3. *Language Model Toolkit* SRILM untuk pemodelan bahasa.
4. *Multi-bleu.perl* untuk oengujian nilai akurasi mesin penerjemah dengan *BLEU score*

## Implementasi Mesin Penerjemah Statistik Bahasa Indonesia ke Bahasa Khek Singkawang

Pada tahap ini terlebih dahulu kita lakukan proses *cleaning, tokenizing dan case folding* pada korpus teks paralel yang sudah dibuat. Berikut merupakan perintah untuk melakukan proses *cleaning, tokenizing* dan *case folding* yang dapat dilihat pada gambar 4.1.



**Gambar 4. 1** Perintah untuk melakukan proses cleaning, tokenizing dan case folding

Baris 1 merupakan perintah untuk melakukan proses *cleaning,* setelah dilakukan proses *cleaning* selanjutnya melakukan proses *tokenizing* untuk menghapus tanda baca titik di akhir kalimat dan menyisipkan spasi antar kata dan tanda baca dengan perintah pada baris 2 dan 3. Tahap berikutnya adalah dilakukan proses *case folding* yaitu mengubah huruf kapital yang terdapat pada korpus menjadi huruf kecil. Adapun perbandingan hasil dari proses sebelum dan sesudah dilakukan proses tersebut dapat dilihat pada tabel 4.2 dan 4.3.

**Tabel 4. 2** Hasil Proses Cleaning dan Tokenisasi

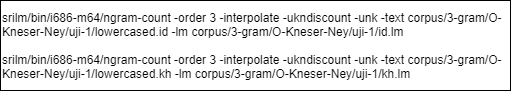
|  |  |
| --- | --- |
| Sebelum *Cleaning* | Sesudah Cleaning dan Tokenizing |
| Kepada majelis penonton tv yang terhormat Pak Harfan berulang kali menyampaikan bahwa semua itu adalah ide Mahar, dan bahwa Mahar itu adalah muridnya | Kepada majelis penonton tv yang terhormat Pak Harfan berulang kali menyampaikan bahwa semua itu adalah ide Mahar , dan bahwa Mahar itu adalah muridnya |

**Tabel 4. 3** Hasil Proses Case Folding

|  |  |
| --- | --- |
| Sebelum Case Folding | Sesudah Case Folding |
| Kepada majelis penonton tv yang terhormat Pak Harfan berulang kali menyampaikan bahwa semua itu adalah ide Mahar , dan bahwa Mahar itu adalah muridnya | kepada majelis penonton tv yang terhormat pak harfan berulang kali menyampaikan bahwa semua itu adalah ide mahar , dan bahwa mahar itu adalah muridnya |

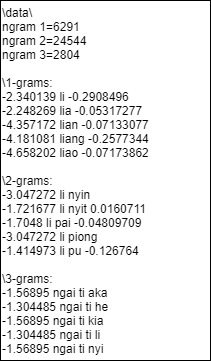
### Implementasi *Language Model Toolkit* SRILM untuk Pemodelan Bahasa

Model bahasa digunakan sebagai sumber pengetahuan berbasis teks dengan nilai-nilai probabilistik. Model bahasa dibangun dengan *language model toolkit* SRILM. Berikut perintah untuk membangun model bahasa dengan *language model toolkit* SRILM pada gambar 4.2.

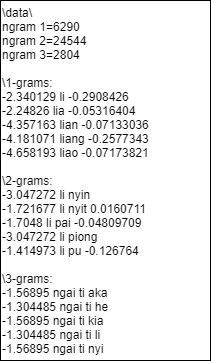


**Gambar 4. 2** Perintah membangun model bahasa dengan toolkit srilm

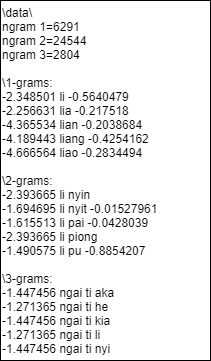
Pada gambar 4.2 merupakan perintah untuk membangun model bahasa dengan algoritma: Interpolation, Good-turing, Natural, Witten-Bell, Modified Kneser-Ney, Original Kneser-Ney.



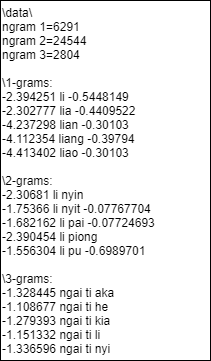
**Gambar 4. 3** Tabel model bahasa dengan bahasa khek singkawang sebagai bahasa target dengan algoritma smoothing Interpolation



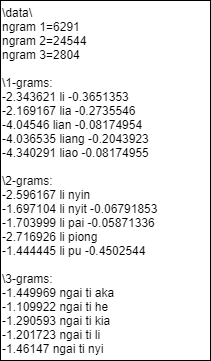
**Gambar 4. 4** Tabel model bahasa dengan bahasa khek singkawang sebagai bahasa target dengan algoritma smoothing Good-turing



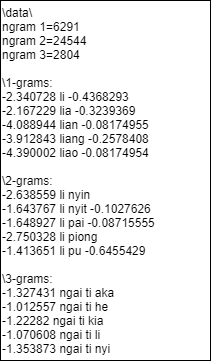
**Gambar 4. 5** Tabel model bahasa dengan bahasa khek singkawang sebagai bahasa target dengan algoritma smoothing Natural



**Gambar 4. 6** Tabel model bahasa dengan bahasa khek singkawang sebagai bahasa target dengan algoritma smoothing Witten-Bell



**Gambar 4. 7** Tabel model bahasa dengan bahasa khek singkawang sebagai bahasa target dengan algoritma smoothing Modified Kneser-Ney



**Gambar 4. 8** Tabel model bahasa dengan bahasa khek singkawang sebagai bahasa target dengan algoritma smoothing Original Kneser-Ney

Berdasarkan Gambar 4.3, 4.4, 4.5, 4.6, 4.7, dan 4.8 model Bahasa menghasilkan n-gram data yang terdiri dari n-gram 1, n-gram 2, n-gram 3. Unigram (n-gram 1) mempunyai data satu token, bigram (n-gram 2) mempunyai data dua token dan trigram (n-gram 3) mempunyai data tiga token dan masing-masing data dari n- gram semua algoritma *smoothing* memiliki jumlah yang sama.

Dari segi probabilitas kata pada pemodelan bahasa hampir secara keseluruhan memiliki perbedaan satu dengan yang lainnya, hanya saja pada algoritma *smoothing interpolate* dan *good-turing* pada bigram *file* memiliki nilai probabilitas yang sama. Adapun setelah dilihat pada bigram *file* algoritma *smoothing interpolate, good-turing*, *natural, witten-bell, modified kneser-ney,* dan *original kneser-ney* pada kalimat “li nyin” dan “li piong” tidak ditemukan berarti probabilitasnya adalah nol.

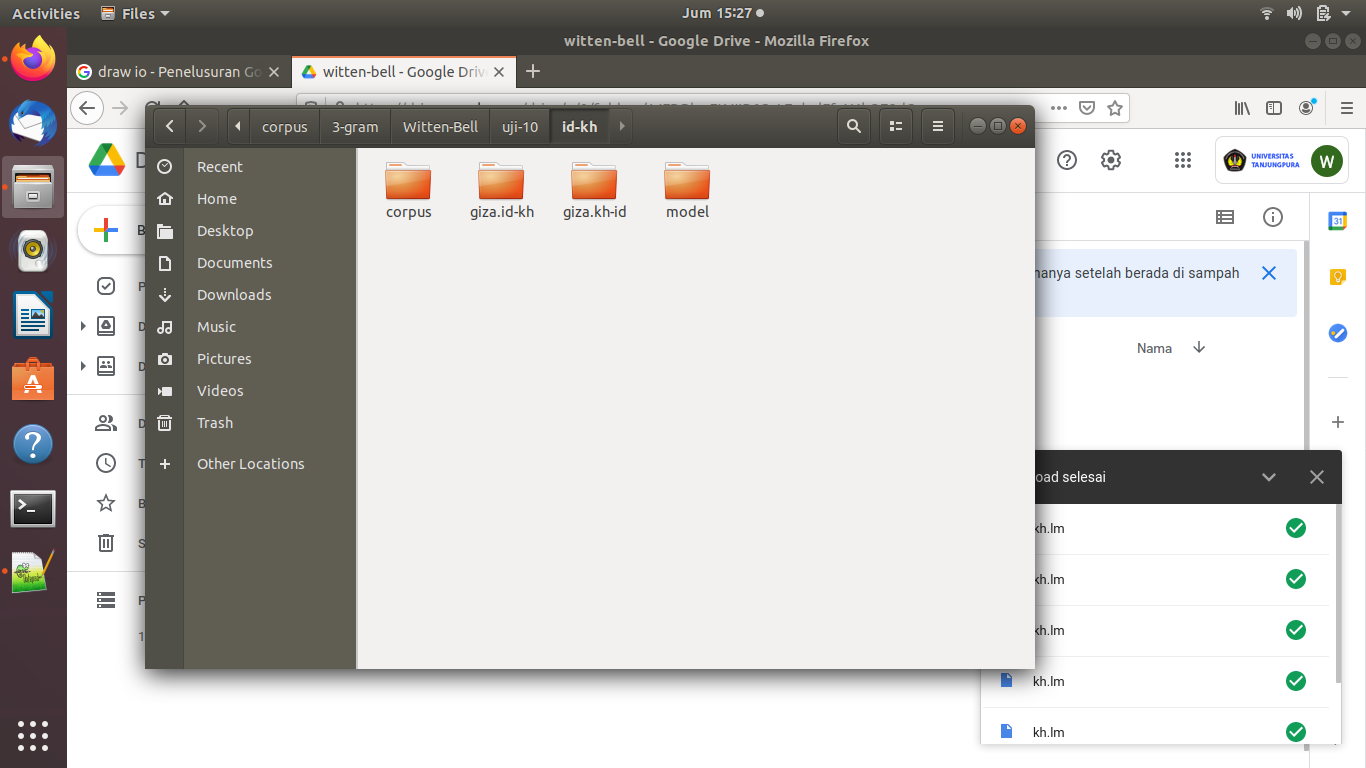
### Implementasi Mgiza untuk Pemodelan Translasi

Pemodelan translasi digunakan untuk memasangkan teks input dalam bahasa sumber dengan teks output dalam bahasa target. Model translasi dibangun dengan tools mgiza. Adapun perintah untuk membangun model translasi dapat dilihat pada gambar 4.9.



**Gambar 4. 9** Perintah membangun model translasi dengan algoritma smoothing pada language model toolkit srilm

Proses pemodelan translasi oleh mgiza menghasilkan dokumen *vocabulary corpus, word alignment* dan *lexical model table.* Dokumen-dokumen tersebut terdapat didalam folder pada gambar 4.10 dengan nama folder “*corpus, giza.id-kh, giza.kh-id, model*”.



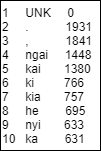
**Gambar 4. 10** File model translasi

#### Vocabulary Corpus

Dokumen *vocabulary corpus* berisi setiap kata pada masing-masing korpus dimana setiap kata tersebut memiliki *uniq id* yang diikuti oleh kata (token) dan frekuensi kemunculannya. *Vocabuary corpus* pada mesin penerjemah statistik dapat dilihat pada gambar 4.11 dan 4.12.



**Gambar 4. 11** Dokumen vocabulary corpus bahasa indonesia

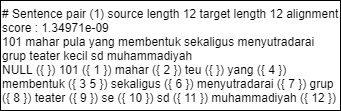


**Gambar 4. 12** Dokumen vocabulary corpus bahasa khek singkawang

Angka 1 sampai 10 pada dokumen *vocabulary corpus* merupakan *uniq id* untuk setiap data token, sedangkan angka disebelah kanan token menunjukkan frekuensi kemunculan. *Vocabulary corpus* yang dihasilkan mesin penerjemah bahasa Indonesia*-*bahasa Khek Singkawang terdiri dari 7926 token untuk korpus bahasa Indonesia dan 6290 token untuk bahasa Khek Singkawang.

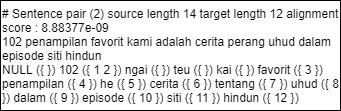
#### Word Alignment *SRILM*

Proses *word alignment* akan menghasilkan dokumen korpus *alignment* yaitu kalimat bahasa target dipetakan pada kalimat bahasa sumber. Dokumen *word alignment* yang terdapat pada mesin penerjemah dapat dilihat pada gambar 4.13, 4.14, 4.15, 4.16, 4.17 dan 4.18 di bawah ini.



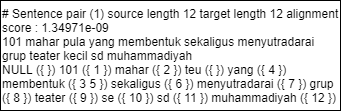
**Gambar 4. 13** Dokumen word alignment bahasa indonesia-khek singkawang dengan algoritma smoothing interpolate

Dokumen *alignment* pada gambar 4.13 Baris pertama berisi letak kalimat target (1) dalam korpus, panjang kalimat sumber (12), panjang kalimat target (12) dan nilai *alignment* untuk algoritma *smoothing interpolate* (1.34971e-09).



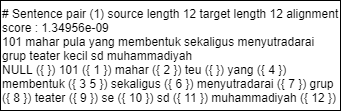
**Gambar 4. 14** Dokumen word alignment bahasa indonesia- khek singkawang dengan algoritma smoothing good-turing

Dokumen *alignment* pada gambar 4.14 Baris pertama berisi letak kalimat target (2) dalam korpus, panjang kalimat sumber (14), panjang kalimat target (12) dan nilai *alignment* untuk algoritma *smoothing good-turing* (8.88377e-09).



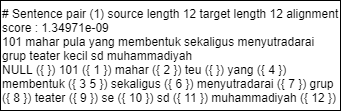
**Gambar 4. 15** Dokumen word alignment bahasa indonesia- khek singkawang dengan algoritma smoothing natural

Dokumen *alignment* pada gambar 4.15 Baris pertama berisi letak kalimat target (1) dalam korpus, panjang kalimat sumber (12), panjang kalimat target (12) dan nilai *alignment* untuk algoritma *smoothing natural* (1.34971e-09).



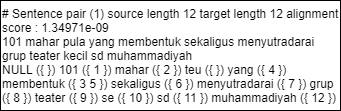
**Gambar 4. 16** Dokumen word alignment bahasa indonesia- khek singkawang dengan algoritma smoothing witten-bell

Dokumen *alignment* pada gambar 4.16 Baris pertama berisi letak kalimat target (1) dalam korpus, panjang kalimat sumber (12), panjang kalimat target (12) dan nilai *alignment* untuk algoritma *smoothing witten-bell* (1.34956e-09).



**Gambar 4. 17** Dokumen word alignment bahasa indonesia- khek singkawang dengan algoritma smoothing modified kneser-ney

Dokumen *alignment* pada gambar 4.17 Baris pertama berisi letak kalimat target (1) dalam korpus, panjang kalimat sumber (12), panjang kalimat target (12) dan nilai *alignment* untuk algoritma *smoothing modified kneser-ney* (1.34971e-09).



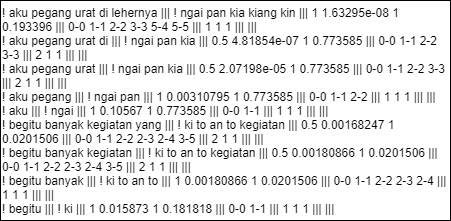
**Gambar 4. 18** Dokumen word alignment bahasa indonesia- khek singkawang dengan algoritma smoothing original kneser-ney

Dokumen *alignment* pada gambar 4.18 Baris pertama berisi letak kalimat target (1) dalam korpus, panjang kalimat sumber (12), panjang kalimat target (12) dan nilai *alignment* untuk algoritma *smoothing original kneser-ney* (1.34971e-09).

Berdasarkan keenam gambar diatas nilai *aligment* tidak berbeda jauh bahkan semuanya sama yang memiliki nilai alignment yaitu (1.34971e-09). Baris ketiga merupakan bahasa sumber dan baris kelima merupakan alignment kalimat bahasa target terhadap kalimat bahasa sumber. Secara sederhana penerapan *alignment* yaitu dengan kalimat bahasa target di-*align* ke kalimat bahasa sumber. Makna kata ”se” ({ 10 }) pada kalimat bahasa target, di-*align* ke kata kesepuluh bahasa sumber yaitu “kecil”.

#### Tabel Model Translasi SRILM

Proses pemodelan translasi oleh mgiza akan menghasilkan table model translasi yang terdiri dari table kata yang berisi sekumpulan kata-kata yang telah dipasangkan antara bahasa sumber dengan bahasa target. Table model translasi frasa yang dihasilkan dapat dilihat pada gambar 4.19 dibawah ini.



**Gambar 4. 19** Tabel model translasi

Setiap baris dalam tabel aturan menjelaskan satu aturan terjemahan. Ini terdiri dari lima komponen yang dipisahkan oleh tiga bar:

* 1. kalimat sumber
  2. kalimat target
  3. skor: di sini hanya satu, tetapi biasanya beberapa skor digunakan
  4. penyelarasan antara non-terminal (menggunakan posisi kata yang dimulai dengan 0, sebagai target sumber), dan
  5. jumlah frekuensi frase sumber & target (untuk tujuan *debugging*; tidak digunakan selama *decoding*).

## Pengujian Hasil Terjemahan

Pengujian hasil terjemahan dilakukan dengan dua cara. Cara pertama dilakukan pengujian otomatis dengan BLEU dan kedua pengujian secara manual oleh ahli bahasa Khek Singkawang.

### Pengujian Otomatis oleh *BLEU*

Pada pengujian otomatis oleh BLEU disini menggunakan metode penambahan secara konsisten dengan jumlah 200 korpus training yang dimana 100 korpus sebagai korpus uji dan 900 korpus training pada uji pertama, uji kedua sampai uji kesepuluh dilakukan penambahan korpus seperti pada grafik hasil pengujian pada gambar 4.20 dan 4.21 yang dimana untuk nilai tertinggi pada uji terakhir adalah *Original Kneser-Ney* dengan nilai skor BLEU 25,57 pada pengujian 3 gram dan untuk pengujian 5 gram dengan nilai skor BLEU 26,15 yang dapat dilihat pada gambar 4.20 dan 4.21 dibawah.

**Gambar 4. 20** Grafik hasil pengujian otomatis oleh BLEU dengan 3 gram

**Gambar 4. 21** Grafik hasil pengujian otomatis oleh BLEU dengan 5 gram

Setelah ditemukannya hasil dengan nilai tertinggi pada Algoritma *Smoothing* *Original Kneser-Ney* dengan 5 gram, hasil tersebut yang akan digunakan pada *k-fold cross validation*, untuk mengetahui nilai rata-rata dari algoritma *smoothing* tersebut, adapun nilai yang diperoleh dari pengujian menggunakan metode *k-fold cross validation* adalah 28,04 untuk bahasa indonesia – khek singkawang seperti pada tabel 4.4.

**Tabel 4. 4** Hasil K-fold Cross Validation

|  |  |  |
| --- | --- | --- |
| **Uji** | **Data Uji** | **Indonesia – Khek Singkawang** |
| 1 | 1-100 | 24,98 |
| 2 | 101 – 200 | 28,63 |
| 3 | 201 – 300 | 27,41 |
| 4 | 301 – 400 | 24,29 |
| 5 | 401 – 500 | 25,59 |
| 6 | 501 – 600 | 24,95 |
| 7 | 601 – 700 | 23,77 |
| 8 | 701 – 800 | 26,98 |
| 9 | 801 – 900 | 35,82 |
| 10 | 901 – 1.000 | 37,64 |
| 11 | 1001 – 1.100 | 36,47 |
| 12 | 1101 – 1.200 | 35,60 |
| 13 | 1201 – 1.300 | 30,70 |
| 14 | 1301 – 1.400 | 36,27 |
| 15 | 1401 – 1.500 | **38,68** |
| 16 | 1501 – 1.600 | 19,58 |
| 17 | 1601 – 1.700 | 20,77 |
| 18 | 1701 – 1.800 | 26,18 |
| 19 | 1801 – 1.900 | 22,75 |
| 20 | 1901 – 2.000 | 23,53 |
| 21 | 2001 – 2.100 | 26,11 |
| 22 | 2101 – 2.200 | 23,67 |
| 23 | 2201 – 2.300 | 28,71 |
| 24 | 2301 – 2.400 | 30,01 |
| 25 | 2401 – 2.500 | 28,24 |
| 26 | 2501 – 2.600 | 23,93 |
| 27 | 2601 – 2.700 | 25,74 |
| **Total** | 2700 | 757 |
| **Rata-rata** | **100** | **28,04** |

Berikut hasil perbandingan algoritma *smoothing* *default*(*Good-turing*) dengan algoritma *smoothing* *Interpolation, Natural, Witten-Bell, Modified Kneser-Ney, Original Kneser-Ney* pada 3-gram dan 5-gram yang dapat dilihat pada Gambar 4.22, 4.23, 4.24, 4.25, 4.26, 4.27, 4.28, 4.29, 4.30, dan 4.31.

**Gambar 4. 22** Grafik Perbandingan Algoritma Smoothing Good-Turing dan Interpolate oleh BLUE dengan 3 gram

**Gambar 4. 23** Grafik Perbandingan Algoritma Smoothing Good-Turing dan Natural oleh BLUE dengan 3 gram

**Gambar 4. 24** Grafik Perbandingan Algoritma Smoothing Good-Turing dan Witten-Bell oleh BLUE dengan 3 gram

**Gambar 4. 25** Grafik Perbandingan Algoritma Smoothing Good-Turing dan M-Kneser-Ney oleh BLUE dengan 3 gram

**Gambar 4. 26** Grafik Perbandingan Algoritma Smoothing Good-Turing dan O-Kneser-Ney oleh BLUE dengan 3 gram

**Gambar 4. 27** Grafik Perbandingan Algoritma Smoothing Good-Turing dan Interpolate oleh BLUE dengan 5 gram

**Gambar 4. 28** Grafik Perbandingan Algoritma Smoothing Good-Turing dan Natural oleh BLUE dengan 5 gram

**Gambar 4. 29** Grafik Perbandingan Algoritma Smoothing Good-Turing dan Witten-Bell oleh BLUE dengan 5 gram

**Gambar 4. 30** Grafik Perbandingan Algoritma Smoothing Good-Turing dan M-Kneser-Ney oleh BLUE dengan 5 gram

**Gambar 4. 31** Grafik Perbandingan Algoritma Smoothing Good-Turing dan O-Kneser-Ney oleh BLUE dengan 5 gram

### Pengujian Terjemahan secara Manual

Pengujian manual dilakukan oleh ahli bahasa Khek Singkawang, pengujian manual diambil berdasarkan hasil pengujian otomatis menggunakan metode *k-fold cross validation* dimana dari hasil pengujian sebanyak 27 kali diambil nilai dengan hasil tertinggi. Adapun nilai tertinggi didapat pada uji kelima belas dengan korpus uji kalimat 1401 sampai 1500 dengan nilai 38,68%. Setelah didapatkan nilai tertiggi maka pada data uji dengan nilai tertinggi itu akan digunakan sebagai data untuk dilakukan pengujian manual oleh ahli bahasa yang dimana terlihat pada pengujian kelima belas dengan nilai 38,68% dari data ini nanti akan dihitung persentase akurasi yang dihasilkan pada pengujian manual dengan persamaan :

P =

P = *Persentase* akurasi

C = Jumlah kata yang diterjemahkan dengan tepat menurut penilaian dari ahli bahasa

R = Jumlah kata hasil terjemahan

**Tabel 4. 5** Hasil Nilai Akurasi Pengujian Ahli Bahasa

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | Nama Ahli Bahasa Khek Singkawang | C | R | P=C/R\*100% |
| 1. | Yulia Magdalena, S.Kom | 672 | 991 | 67,81% |
| 2. | Veronica, B.Ed | 721 | 991 | 72,75% |

Hasil terjemahan yang dilakukan pengujian manual oleh Yulia Magdalena, S.Kom diperlihatkan pada lampiran A dan pengujian manual oleh Veronica, B.Ed diperlihatkan pada lampiran B.

## Analisis Hasil Pengujian

Berikut ini merupakan analisis hasil pengujian yang telah dilakukan

1. Pengujian otomatis dengan metode penambahan korpus secara konsisten 200 kalimat *training* dengan 100 korpus uji dan 900 korpus *training* pada pengujian pertama sampai pengujian ke sepuluh dengan jumlah total korpus yang digunakan dari sepuluh kali pengujian adalah 2700 korpus, dimana algoritma *smoothing* yang digunakan yaitu: *Interpolation, Good-turing, Natural, Witten-Bell, Modified Kneser-Ney, Original Kneser-Ney.* Nilai BLEU *score* terbaik didapat oleh algoritma *smoothing Original Kneser-Ney* dengan nilai akurasi 26,15% dengan menggunakan *N-gram* *order* 5-gram.
2. Penilaian Ahli Bahasa berbanding lurus dengan penilaian otomatis terhadap akurasi dari mesin penerjemah statistik dimana hasil pengujian dengan metode *K-Fold Cross Validation* dengan algoritma *smoothing original kneser-ney* pada rentang kalimat 1401 sampai 1500 didapatkan akurasi sebesar 38,68% berbanding lurus dengan akurasi oleh ahli bahasa dengan akurasi masing-masing sebesar 67,81%% oleh Yulia Magdalena, S.Kom dan 72,75% oleh Veronica, B.Ed.
3. Pemodelan bahasa menggunakan *language model toolkit* yang sama yaitu SRILM dengan menggunakan algoritma *smoothing* yang berbeda untuk menghasilkan nilai BLEU *score* yang menjadi acuan analisis pada penelitian kali ini.
4. Penilaian ahli bahasa sangat dipengaruhi oleh pengetahuan dan pemahaman ahli bahasa
5. Dari Gambar 4.21 dapat kita lihat hasil pengujian algoritma *modified Kneser-ney* dengan 5gram dari pegujian pertama sampai kesembilan dengan jumlah kalimat 2500 tidak menghasilkan nilai *BLEU* dan pada pengujian kesepuluh dengan jumlah kalimat 2700 di dapat nilai *BLEU* sebesar 25.02%. Dapat disimpulkan pengujian algoritma *modified Kneser-ney* dengan 5gram dapat menghasilkan nilai *BLEU* dengan jumlah kalimat diatas 2500.
6. Berdasarkan referensi jurnal-jurnal terkait penelitian pemodelan bahasa pengujian lebih sering menggunakan *perplexity* (tingkat *error* dari korpus) daripada dengan *BLEU*, pada penelitian lanjutan terkait pemodelan bahasa khususnya *algoritma smoothing* dapat digunakan pengujian dengan *perplexity* untuk mengetahui apa perbedannya dengan pengujian otomatis *BLEU* serta dampaknya terhadap nilai akurasi mesin penerjemah statistik.

# Kesimpulan dan Saran

## Kesimpulan

Berdasarkan hasil analisis dan pengujian yang sudah dilakukan dapat diambil kesimpulan sebagai berikut:

1. Berdasarkan pengujian dengan metode penambahan korpus secara konsisten algoritma *smoothing* terbaik adalah algoritma *smoothing original kenser-ney* menggunakan 3 gram dengan nilai akurasi 25,57% dan algoritma *smoothing original kneser-ney* menggunakan 5 gram dengan nilai akurasi 26,15%.
2. Penggunaan algoritma *smoothing* dalam membangun mesin penerjemah statistik dapat meningkatkan nilai akurasi mesin penerjemah, namun dapat pula menyebabkan penurunan kualitas terjemahan diantaranya dapat disebabkan oleh kuantitas korpus *training* ataupun kualitas dari korpus.

## Saran

Adapun saran yang dapat diberikan sebagai pengembangan dari penelitian ini adalah:

1. Perlunya ahli bahasa yang benar-benar mengerti dan menguasai bahasa yang dijadikan sebagai bahasa target agar keakuratan bahasa tersebut mendekati hasil terjemahan dari penutur ahli bahasa tersebut, kenapa point ini penting untuk diperhatikan dikarenakan bahasa daerah pada dasarnya memiliki ciri khas masing-masing tergantung dimana bahasa tersebut digunakan, bahasa khek singkawang sendiri memiliki beragam perbedaan mulai dari intonasi, penyebutan, bahkan penulisan untuk menyeragamkan hal tersebut ahli bahasa sangat berperan penting.
2. Perlunya dilakukan perbaikan lebih lanjut dari segi korpus terkait kualitas korpus dan kuantitas korpus agar nilai akurasi yang dihasilkan lebih baik.
3. Perlu mencoba parameter lain sebagai patokan dalam hasil pengujian mesin penerjemah statistik selain *BLEU Score* seperti dengan *perplexity* untuk mengetahui perbedaanya dengan pengujian otomatis seperti *BLEU*.

Daftar Pustaka

Bertoldi, Nicola. 2008. *A Tutorial On The IRSTLM Library*. FBK-irst,Trento, Italy

Chen, S. F., and Goodman, J. 1998. *An Empirical Study of Smoothing Techniques for Language Modelling.* Cambridge, Computer Science G roup Harvard University.

Federico, M.; Bertoldi, N., and Cettolo, M. 2008. *IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models.* Povo, Interspeech 2008 incorporating SST08.

Guo, Junfei; Liu, Juan; Han, Qi; Maletti, Andreas. 2014. *A Tunable Language Model for Statistical Machine Translation.* Vancouver, Al-Onaizan & Simard (Eds.) Proceedings of AMTA 2014, vol 1: MT Researchers

Grossman, Robert; Seni, Giovanni; Elder, John; Agarwal, Nitin; Liu, Huan 2010. *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*. Morgan & Claypool Publisher.

Hadi, I. 2014. *Uji Akurasi Mesin Penerjemah Statistik (MPS) Bahasa Indonesia ke Bahasa Melayu Sambas dan Mesin Penerjemah Statistik (MPS) Bahasa Melayu Sambas ke Bahasa Indonesia*. Pontianak, Jurnal Sistem dan Teknologi Informasi (JUSTIN) Vol. 2 No. 3.

Hadary, F., dkk. 2015. *Pedoman Penulisan Proposal Skripsi*. Pontianak: Fakultas Teknik Universitas Tanjungpura.

Heafield, K. 2011. *KenLM: Faster and Smaller Language Model Queries.* Pittsburgh, Association for Computational Linguistics Stroudsburg.

Tanggok, M. Ikhsan. 2015. *AGAMA DAN KEBUDAYAAN ORANG HAKKA DI SINGKAWANG Memuja Leluhur dan Menanti Datangnya Rezeki.* Jakarta, PT Kompas Media Nusantara.

Jarob, Yosep R; Sujaini, Herry; Safriadi, Novi. 2016*.* *Uji Akurasi Penerjemahan Bahasa Indonesia – Dayak Taman dengan Penandaan Kata Dasar dan Imbuhan*. Pontianak, Jurnal Edukasi dan Penelitian Informatika (JEPIN) Vol. 2 No. 2

Kanchan, K. B.; Das, P.; Hannan, A., and Shikhar S. 2014. *Assamese-English Billingual Machine Translation*. Guwahati, International Journal on Natural Language Computing (IJNLC) Vol. 3, No. 3.

Koehn, P., dkk. 2007. *Moses: Open Source Toolkit for Statistical Machine Translation.* Prague, Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics.

Koehn, P. 2009. *Moses: User Manual and Code Guide*. Cambridgeshire, Cambridge University Press.

Manning, C. D., and Schutze, H. 2000. *Foundations Of Statistical Natural Language Processing*. London : The MIT Press Cambridge Massachusetts.

Marlinda, Linda dan Rianto, Harsih. 2013. *Pembelajaran Bahasa Indonesia Berbasis Web Menggunakan Metode Maximum Marginal Relevance*.Manajemen Informatika. Jakarta: Sesindo.

Papineni, K.; Ruokos, S.; Ward, T., and Zhu, W. 2002. *BLEU: A Method For Automatic Evaluation of Machine Translation.* USA: IBM TJ Watson Research Center.

Ristad, Eric. 1995. *A Natural Law of Succesion.* Research Report CS-TR-495-95

Tanuwijaya, H., dan Manurung, Hisar M. 2009. *Penerjemahan Inggris-Indonesia Menggunakan Mesin Penerjemah Statistik Dengan Word Reordering dan Phrase Reordering*. Jakarta, Jurnal ilmu Komputer dan Informasi Vol 2 No 1.

Setyawan, Aan. 2011. *Bahasa Daerah dalam Perspektif Kebudayaan dan Sosilinguistik: Peran dan Pengaruhnya dalam Pergeseran dan Pemertahanan Bahasa.* Semarang, International Seminar Language Maintenance and Shift

Stolcke, Andreas; Zheng, Jing; Wang, Wen; Abrash, Victor. 2011 *SRILM at Sixteen: Update and Outlook.* Big Island, Proc. IEEE Automatic Speech Recognition and Understanding Workshop

Sujaini, Herry., Negara, Arif Bijaksana Putra. 2015. *Analysis of Extended Word Similarity Clustering based Algorithm on Cognate Language.* Gujarat: ESRSA Publications Pvt. Ltd.

Zhang, W. 2015. *Comparing the Effect of Smoothing and N-gram Order :Finding the Best Way to Combine the Smoothing and Order of N-gram.* Melbourne, Florida : Master of Science in Computer Engineering.

1. Korpus Uji oleh Yulia Magdalena, S.Kom

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | | | | |  |
| **No** | **Bahasa Indonesia** | **Hasil Terjemahan** | **Terjemahan Menurut Ahli Bahasa** | **Jumlah Kata Benar** | **Jumlah Kata Hasil Terjemahan** |  |
| 1 | tolong bawakan saya sebuah pengering . | then siw bawakan ngai jit cak pengering . | then siw kiam pun ngai jit cak pengering . | 6 | 8 |  |
| 2 | tolong , bisakah anda memberi saya air ? saya sangat haus . | then siw , nyi hiau pun ngai sui ? ngai an haus . | then siw , nyi hiau pun ngai sui mo ? ngai an tu si hot . | 11 | 13 |  |
| 3 | dimana counter jal ? | abui counter jal mo ? | abui counter jal ? | 4 | 5 |  |
| 4 | kami tidak terlalu lapar . | ngai teu mo an tushiki . | ngai tew em boi an tu si ki . | 5 | 6 |  |
| 5 | saya tidak bisa menemukan tas saya . | ngai em hiau chim to nga su pau . | ngai em hiau chim nga su pau . | 8 | 9 |  |
| 6 | perlihatkan saya sepatu warna cokelat . | ngai khon pun ngai sepatu cokelat set . | khon pun ngai sepatu cokelat set . | 7 | 8 |  |
| 7 | dimana dan kapan itu pergi ? | abui kak kiang si ka hi mo ? | abui ka kiang si ka hi ? | 7 | 8 |  |
| 8 | adat istiadat berbeda antara satu negara dengan negara lainnya , dan adat istiadat jepang tampaknya sangat berbeda dari adat istiadat barat . | adat istiadat pi jit jong kiet kak kiet kai , koli adat istiadat nyip pun khon to an beda choi adat istiadat barat . | fung siuk jit cak kiet eng jit jong ka phet cak kiet , ka nyip pun fung siuk khon to an eng jit jong ka mui kiet fung siuk . | 7 | 24 |  |
| 9 | saya akan membayar lima puluh dolar . | ngai oi ban eng sip dollar . | ngai oi ban eng sip dollar . | 7 | 7 |  |
| 10 | tolong simpan kopor ini sampai jam lima sore . bisa saya minta karcisnya ? | then siw piong lia kap pit teu ng tiam sore . ngai hiau tho karcisnya mo ? | then siw piong lia kap pit teu eng tiam ha ciw . ngai hiau tho ka phiau mo ? | 16 | 17 |  |
| 11 | saya ingin menyewanya untuk lima hari . | ngai oi ci ki loi eng nyit . | ngai oi ci gi loi eng nyit . | 7 | 8 |  |
| 12 | untuk makanan saja , harganya empat puluh dolar . sebotol anggur harganya sekitar sepuluh dolar . | loi sit ti , harganya si sip dollar . sebotol fu tho harganya khiun sip dollar . | loi sit aja , ka ka si sip dollar . jit tut fu tho ka ka cham em to sip dollar . | 12 | 17 |  |
| 13 | ini adalah penerbangan ke miami . | lia he pui ki hi miami . | lia jiu he hi miami pui ki . | 2 | 7 |  |
| 14 | terima kasih atas bantuan anda . saya merasa seperti mau muntah . apakah anda punya obat untuk mabuk udara ? | kam chia tangko thensiu ngai chi to chiong . nyi oi muntah . he nyi jiu jok loi cui udara mo ? | kam chia nyi then siw ngai . ngai chi to chiong oi euw . nyi jiu jok loi hin cha mo ? | 14 | 22 |  |
| 15 | saya punya sebotol wiski . | ngai jiu sebotol wiski . | ngai jiu jit tut wiski . | 4 | 5 |  |
| 16 | saya akan datang untuk menemui anda pukul sepuluh pada hari jumat . | ngai boi loi nyi nyi sip pai ng tiam . | ngai oi loi loi nyi nyi sip tiam pai eng . | 6 | 10 |  |
| 17 | tanda tangan di bagian belakang cek . | tanda kia poi heu kai tan . | fa ho choi poi mui tan . | 3 | 7 |  |
| 18 | maaf , saya menekan nomor yang salah . | tui ngchu , ngai tun ho kai salah . | tui em chu , ngai cit sala ho ma . | 6 | 9 |  |
| 19 | tolong , panggilan jarak jauh ke jepang . | then siw , ta thien an jan hi nyip pun . | then siw , ta thien hi nyip pun . | 9 | 11 |  |
| 20 | makanan saya sudah siap apa belum ? | sit ngai siap ho ma’ai mang mo ? | nga cu sit cu ho mang ? | 2 | 8 |  |
| 21 | saya ingin melakukan reservasi untuk bermain golf . | ngai oi co reservasi loi kau golf . | ngai oi co reservasi loi kau golf . | 8 | 8 |  |
| 22 | koin apa yang bisa digunakan untuk mesin penjual otomatis ini ? | koin hemai hiau digunakan loi mesin penjual otomatis lia mo ? | ma'ai thiet lui hiau jung loi lia otomatis mai ki ? | 3 | 11 |  |
| 23 | saya ingin wedang jahe . | ngai oi wedang jahe . | ngai oi wedang kiong ma . | 4 | 5 |  |
| 24 | tolong , saya ingin beberapa kemeja . | then siw , ngai oi kia kemeja . | then siw , ngai oi kia neu sam . | 7 | 8 |  |
| 25 | saya ingin beberapa makanan bayi . | ngai oi kia sit bayi . | ngai oi kia onga sit tung si . | 4 | 6 |  |
| 26 | pegang acar ini . | pan acar lia . | pan lia acar . | 2 | 4 |  |
| 27 | saya tidak merasa lebih baik . | ngai mo chi to ko ho . | ngai mo chi to hau ho . | 6 | 7 |  |
| 28 | maukah anda menunjukkan kepada saya pilihan yang lain ? | nyi oi pun khon pun ngai pilihan phet cak mo ? | nyi oi khon pun ngai phet cak mo ? | 9 | 11 |  |
| 29 | gaun ini terbuat dari nilon ? | khiun lia terbuat choi nilon mo ? | lia gaun nilon co loi he mo ? | 2 | 7 |  |
| 30 | oke , tentu . bayar ini di depan ketika anda siap . gunakan waktu anda , oke ? | oke , he . bayar lia mien chien kai cisi nyi siap . nyi jung si kien , oke ? | oke , ho wa . ban lia choi mien chien he nyi cang liau . jung nya si kien , oke ? | 11 | 20 |  |
| 31 | ya , dua blok ke arah itu . | he , nyi blok hi chu kai . | he , liong blok hi kai sak . | 5 | 8 |  |
| 32 | sekarang jam dua belas . | liha liong cak tiam cung . | kin ha sip nyi tiam . | 0 | 6 |  |
| 33 | sekarang saya tahu apa yang salah . | liha ngai ti hemai salah . | kin ha ngai ti he ma'ai sala . | 5 | 6 |  |
| 34 | gerbang mana yang harus saya lalui untuk transit saya ? | lew abui ngai oi lalui transit pun ngai mo ? | abui mun ngai oi ko loi nga transit ? | 4 | 10 |  |
| 35 | white beach ini indah . | white beach lia an ciang kai . | lia white beach an ciang . | 3 | 7 |  |
| 36 | tolong lakukan reservasi . | then siw co reservasi . | then siw co reservasi . | 5 | 5 |  |
| 37 | saya akan membawa kembalian anda nanti . | ngai oi tai nyi kaha tai ban con lui . | ngai oi tai nya ban con ka ha . | 6 | 10 |  |
| 38 | saya belum melakukan reservasi , tapi apakah ada kursi yang tersedia ? | ngai mang co reservasi , tan he jiu ten kai tersedia mo ? | ngai mang co reservasi , tapi han jiu ten mo ? | 9 | 13 |  |
| 39 | tolong , saya ingin pita mesin tulis . | then siw , ngai oi pita mesin sia si . | then siw , ngai oi sia ki pita . | 6 | 10 |  |
| 40 | tolong buka pintu dengan kunci cadangan . | then siw khoi mun kak cadangan so si . | then siw khoi mun jung cadangan so si . | 8 | 9 |  |
| 41 | roma tidak dibangun dalam sehari . | roma mo dibangun lok jit nyit . | roma mo tin hong jit nyit . | 5 | 7 |  |
| 42 | saya mau secangkir kopi lagi . | ngai oi jit pui kopi liau . | ngai oi jit pui kopi kak lin . | 6 | 7 |  |
| 43 | saya ingin hamburger untuk pergi . | ngai oi hamburger loi hi . | ngai oi hamburger loi hi . | 6 | 6 |  |
| 44 | dapatkah anda memeriksa ini ? | nyi hiau cha lia mo ? | nyi hiau cha lia mo ? | 6 | 6 |  |
| 45 | dompet saya telah dicuri . | nga dompet tin theu het liau . | nga dompet theu het liau . | 6 | 7 |  |
| 46 | dapatkah anda memintanya untuk menelepon kembali ? | nyi hiau pun tho ki tathien pun to con mo ? | nyi hiau tho gi loi ta thien to con mo ? | 9 | 11 |  |
| 47 | kami menyediakan makan di mobil dan kedai kopi di kereta . | menyediakan choi cha kak ngai teu kai sit kopi tiam choi cha . | ngai tew cu cang sit choi mo ka ka kopi tiam choi cha . | 6 | 13 |  |
| 48 | kamar saya belum dibersihkan . | nga kien tu mang dibersihkan . | nga kien tu mang met chiang . | 5 | 6 |  |
| 49 | ini ikeda yang sedang berbicara . saya telah memesan satu meja untuk pukul enam tiga puluh malam ini . tapi saya minta maaf saya akan datang terlambat setengah jam . | lia ikeda yang sedang kong boi tin cok ngai kau tai jit cak loi liuk tiam sam sip li pu ja . tan he ngai tho tui ngchu ngai boi loi thet man jit pan tiam . | lia ikeda kong boi . ngai jiu kau tai jit cak cok loi liuk tiam pan li pu ja . tapi ngai tui em chu ngai boi loi thet man pan tiam cung . | 24 | 37 |  |
| 50 | anda dapat naik bus di halte bus lima . | nyi hiau song cha choi halte cha ng . | nyi hiau song thai cha choi thi eng thai cha cham . | 7 | 9 |  |
| 51 | jika anda menemukannya , tolong , dapatkah anda menyimpannya untuk saya ? | he nyi chim gi , then siw , nyi hiau piong gi pun ngai mo ? | he nyi chim to gi , then siw , nyi hiau piong gi pun ngai mo ? | 16 | 16 |  |
| 52 | mana yang lebih baik , penerbangan pagi atau penerbangan sore ? | abui ko ho , pui ki co atau pui ki ciw mo ? | abui ko ho , co pui ki he ha ciw pui ki ? | 9 | 13 |  |
| 53 | apakah ada arena luncur es di dekat sini ? | he jiu arena luncur siat khiun ali mo ? | jiu arena luncur es khiun ali mo ? | 7 | 9 |  |
| 54 | terima kasih banyak . simpan kembaliannya . | kam chia to-to . piong ka cau con lui . | kam chia to-to . kiam ka to con lui . | 7 | 10 |  |
| 55 | tiga puluh delapan dolar lima puluh sen sehari , dengan jarak tempuh tidak terbatas . bensin tidak termasuk . | pat sam sip dollar eng sip sen jit nyit , kak li hang mo thon . bensin mo lok . | sam sip pat dollar eng sip sen jit nyit , ka hang mo thon . thien jiu mo son . | 14 | 20 |  |
| 56 | apakah anda menaruh perangko pada kartu pos ? | nyi jiu perangko menaruh kak pos phai mo ? | nyi jiu piong perangko choi pos phai mo ? | 7 | 9 |  |
| 57 | tiga malam . | sam pu ja . | sam ja . | 3 | 4 |  |
| 58 | jane , temanmu susan di sini untuk menemuimu . | jane , temanmu susan ali pun nyi nyi . | jane , nya phen jiu susan choi ali loi nyi nyi . | 7 | 9 |  |
| 59 | apakah ini ke edinburgh ? | he lia hi edinburgh ? | lia hi edinburgh he mo ? | 4 | 5 |  |
| 60 | kapan anda akan pergi ? | kiang si nyi oi hi mo ? | kiang si nyi oi hi ? | 6 | 7 |  |
| 61 | apakah saya harus mentransfer ? | he ngai oi mentransfer mo ? | ngai oi transfer mo ? | 4 | 6 |  |
| 62 | jika demikian , anda harus mengambil ukuran sepuluh sesuai standar amerika . | he anyiong , nyi oi kiam pien sip sesuai standar mui kiet . | he nyong hin , nyi oi kiam thi sip ho ma cau mui kiet standar . | 9 | 13 |  |
| 63 | sebenarnya , saudara-saudara saya sudah menikah , dan saya satu-satunya yang masih tinggal dengan orang tua saya . | chi sit , saudara-saudara ngai kau li , koli ngai han jitcak-jitcak het ka jit sa lo ngai . | cau li , nga ci oi kau liau , ka ngai tan sa yang choi het ka nga ja oi . | 7 | 19 |  |
| 64 | tidak ada handuk . | mo jiu handuk . | mo mang . | 2 | 4 |  |
| 65 | apakah anda keberatan orang lain mendapatkan itu untuk saya ? | he nyi boi chung to phet nyin tet to kai pun ngai mo ? | nyi chung to mo he phet sa nyin tet ka pun ngai ? | 9 | 14 |  |
| 66 | apakah ini yang anda maksud ? | lia he nyi kia mo ? | lia he nyia ji si mo ? | 5 | 6 |  |
| 67 | ini adalah kunci saya . | lia he ngai so si . | lia jiu he nga so si . | 5 | 6 |  |
| 68 | dapatkah anda memanggil nomor tersebut lagi ? | nyi hiau se tersebut ho liau ? | nyi hiau ta thien kai cak ho ma kak lin mo ? | 3 | 7 |  |
| 69 | mmm , bagus . permisi , dapatkah anda memeriksa salad saya ? | mmm , khiang . cia mun , nyi hiau cha nga salad mo ? | mmm , khiang . cia mun , nyi hiau cha nga salad mo ? | 14 | 14 |  |
| 70 | oh , anda telah melewatinya . silahkan kembali lagi . | oh , nyi ho ko het gi . ho ji fan to con . | oh , nyi ko het gi liau . ho ji fan loi . | 11 | 14 |  |
| 71 | saya pikir ini adalah adat jepang yang unik . | ngai siong to lia he adat nyip pun an unik . | ngai siong lia jiu he nyip pun fung siuk an unik . | 8 | 11 |  |
| 72 | apakah anda punya keju ? | he nyi jiu keju mo ? | nyi jiu keju mo ? | 5 | 6 |  |
| 73 | tolong , bungkus ini satu per satu , dan kenakan pita pada masing-masing bungkusan . | then siw , bungkus lia jit sa kok sa , koli kenakan pita choi kok sa bungkusan . | then siw , pau lia cak kok cak , ka jung pita pun kok cak pau . | 8 | 18 |  |
| 74 | tolong , bolehkah saya meminta daftar anggur ? | then siw , ngai ho ji tho daftar fu tho ciu mo ? | then siw , ngai ho ji tho daftar fu tho mo ? | 12 | 13 |  |
| 75 | saya lihat dulu . ya benar , makan siang hari ini akan dilayani lebih cepat . | ngai khon ha . he chok , sit ciu ci nyit lia boi dilayani ko mang . | ngai khon sen . he chok , sit tong ciu li nyit boi hau mang tin ta li . | 10 | 17 |  |
| 76 | anda berasal dari negara mana ? | nyi he choi kiet abui ? | nyi abui kiet loi ? | 3 | 6 |  |
| 77 | dapatkah anda mengubah ini menjadi seperempatan ? | nyi hiau bon lia pien seperempatan mo ? | nyi hiau bon lia pien si cha mo ? | 7 | 8 |  |
| 78 | itu tujuh . | chit ka . | ka chit . | 1 | 3 |  |
| 79 | yang putih ya . | phak kai ha . | phak kai dia . | 3 | 4 |  |
| 80 | bagaimana cara mematikan lampu baca ? | apan hin kia cara mematikan baca ten mo ? | nyong pan hin met kim thuk ten ? | 3 | 9 |  |
| 81 | saya mau dua porsi mie morisoba . | ngai oi liong porsi mie morisoba . | ngai oi liong phan morisoba mien . | 5 | 7 |  |
| 82 | saya bekerja di suatu perusahaan pialang saham . | ngai co she kai choi jit cak perusahaann pialang saham . | ngai co choi jit cak pialang saham nyin hong . | 5 | 11 |  |
| 83 | ya , ada satu kamar kosong . ada berapa orang ? | he , jiu jit kien khung . jiu jit sa nyin ? | he , jiu jit kien khung . jiu to sa nyin ? | 11 | 12 |  |
| 84 | anda tidak diizinkan untuk parkir di sini . | nyi em mo ji loi parkir choi ali . | nyi em mo ji loi parkir choi ali . | 9 | 9 |  |
| 85 | apakah ada toko penukaran uang dekat sini ? | he jiu tiam penukaran lui khiun ali mo ? | jiu kau lui tiam khiun ali mo ? | 6 | 9 |  |
| 86 | nama saya takeshi. nama anda siapa ? | nga miang takeshi. nyia miang he asa ? | nga miang takeshi . nyia miang ma sa ? | 7 | 8 |  |
| 87 | gunakan ini seminggu sekali . lumurilah dengan tipis setelah mencuci wajah anda , hindari jangan sampai terkena mata dan rambut . | jung lia jit cak lipai jit lin . lumurilah kak apunf kohet mencuci mien nyi , hindari ngmo teu tin to mukcu koli mo . | jung lia jit li pai jit lin . ko phok-phok he nyi se li mien , em mo teu tin to muk cu ka mo . | 18 | 25 |  |
| 88 | saya terluka oleh kaca di sini . | ngai terluka kiang kai choi ali . | ngai kot to tin to kiang choi ali . | 5 | 7 |  |
| 89 | saya seharusnya bertanya kepada pelatih , tapi bisakah anda katakan kepada saya mana mesin yang baik untuk melatih punggung saya ? | ngai cau li mun pun pelatih , tan he nyi hiau kong pun ngai abui mesin ho kai pun melatih punggung ngai mo ? | ngai cau li mun pelatih , tapi nyi hiau kong pun ngai thang abui ki hau ho loi lien nga jau kut mo ? | 15 | 24 |  |
| 90 | tolong berhenti di sisi persimpangan berikutnya . | then siw tiam choi sisi lin con ben . | then siw thin choi sun ka ha con ben . | 6 | 9 |  |
| 91 | apakah ada pramugari yang bisa berbahasa jepang ? | pramugari he jiu nyin hiau kong nyip pun mo ? | jiu pramugari hiau kong nyip pun boi mo ? | 6 | 10 |  |
| 92 | tolong , saya ingin limun . | then siw , ngai oi limun . | then siw , ngai oi son kam . | 5 | 7 |  |
| 93 | saya sakit . | ngai thung . | ngai pot phiang . | 2 | 3 |  |
| 94 | tolong , saya ingin susu . | then siw , ngai oi susu . | then siw , ngai oi nen . | 6 | 7 |  |
| 95 | berapa ons parfum saya bisa bebas bea ? | to ons parfum ngai hiau bebas bea mo ? | jit to ons nga hiong fa sui hiau bebas pajak ? | 5 | 9 |  |
| 96 | lengan ini terlalu panjang bagi saya . | kia lia an kiu pun ngai . | lia siu thet chong pun ngai . | 3 | 7 |  |
| 97 | tolong , dapatkah rambut saya dipotong , dikeriting dan dirapikan ? | then siw , hiau mo ngai dipotong dikeriting , koli dirapikan mo ? | then siw , hiau mo nga mo na cien , na kiw ka na rapikan ? | 6 | 13 |  |
| 98 | anda tidak harus menggantinya , tapi mobil ini untuk manusia gua . | nyi oi bon gi mo , tan he lia cha pun nyin gua . | nyi em si bon gi , tapi lia cha loi gua nyin . | 7 | 14 |  |
| 99 | hanya pariwisata . | chiang pariwisata . | chiang li hang . | 2 | 3 |  |
| 100 | kalau begitu , lift ada di sana . | he anyiong , jiu lift choi kai . | he nyong hin , lift jiu choi aun . | 4 | 8 |  |
|  |  |  |  | 672 | 991 |

1. Korpus Uji oleh Veronica, B.Ed

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | | | | |  |
| **No** | **Bahasa Indonesia** | **Hasil Terjemahan** | **Terjemahan Menurut Ahli Bahasa** | **Jumlah Kata Benar** | **Jumlah Kata Hasil Terjemahan** |  |
| 1 | tolong bawakan saya sebuah pengering . | then siw bawakan ngai jit cak pengering . | then siw kiam pun ngai jit cak pengering . | 6 | 8 |  |
| 2 | tolong , bisakah anda memberi saya air ? saya sangat haus . | then siw , nyi hiau pun ngai sui ? ngai an haus . | then siw , nyi hiau pun ngai sui mo ? ngai an tu si hot . | 12 | 13 |  |
| 3 | dimana counter jal ? | abui counter jal mo ? | abui counter jal ? | 4 | 5 |  |
| 4 | kami tidak terlalu lapar . | ngai teu mo an tushiki . | ngai tew em boi an tu si ki . | 5 | 6 |  |
| 5 | saya tidak bisa menemukan tas saya . | ngai em hiau chim to nga su pau . | ngai em hiau chim nga su pau . | 8 | 9 |  |
| 6 | perlihatkan saya sepatu warna cokelat . | ngai khon pun ngai sepatu cokelat set . | khon pun ngai sepatu cokelat set . | 6 | 8 |  |
| 7 | dimana dan kapan itu pergi ? | abui kak kiang si ka hi mo ? | abui ka kiang si ka hi ? | 5 | 8 |  |
| 8 | adat istiadat berbeda antara satu negara dengan negara lainnya , dan adat istiadat jepang tampaknya sangat berbeda dari adat istiadat barat . | adat istiadat pi jit jong kiet kak kiet kai , koli adat istiadat nyip pun khon to an beda choi adat istiadat barat . | fung siuk jit cak kiet eng jit jong ka phet cak kiet , ka nyip pun fung siuk khon to an eng jit jong ka mui kiet fung siuk . | 8 | 24 |  |
| 9 | saya akan membayar lima puluh dolar . | ngai oi ban eng sip dollar . | ngai oi ban eng sip dollar . | 7 | 7 |  |
| 10 | tolong simpan kopor ini sampai jam lima sore . bisa saya minta karcisnya ? | then siw piong lia kap pit teu ng tiam sore . ngai hiau tho karcisnya mo ? | then siw piong lia kap pit teu eng tiam ha ciw . ngai hiau tho ka phiau mo ? | 15 | 17 |  |
| 11 | saya ingin menyewanya untuk lima hari . | ngai oi ci ki loi eng nyit . | ngai oi ci gi loi eng nyit . | 7 | 8 |  |
| 12 | untuk makanan saja , harganya empat puluh dolar . sebotol anggur harganya sekitar sepuluh dolar . | loi sit ti , harganya si sip dollar . sebotol fu tho harganya khiun sip dollar . | loi sit aja , ka ka si sip dollar . jit tut fu tho ka ka cham em to sip dollar . | 13 | 17 |  |
| 13 | ini adalah penerbangan ke miami . | lia he pui ki hi miami . | lia jiu he hi miami pui ki . | 6 | 7 |  |
| 14 | terima kasih atas bantuan anda . saya merasa seperti mau muntah . apakah anda punya obat untuk mabuk udara ? | kam chia tangko thensiu ngai chi to chiong . nyi oi muntah . he nyi jiu jok loi cui udara mo ? | kam chia nyi then siw ngai . ngai chi to chiong oi euw . nyi jiu jok loi hin cha mo ? | 13 | 22 |  |
| 15 | saya punya sebotol wiski . | ngai jiu sebotol wiski . | ngai jiu jit tut wiski . | 4 | 5 |  |
| 16 | saya akan datang untuk menemui anda pukul sepuluh pada hari jumat . | ngai boi loi nyi nyi sip pai ng tiam . | ngai oi loi loi nyi nyi sip tiam pai eng . | 9 | 10 |  |
| 17 | tanda tangan di bagian belakang cek . | tanda kia poi heu kai tan . | fa ho choi poi mui tan . | 5 | 7 |  |
| 18 | maaf , saya menekan nomor yang salah . | tui ngchu , ngai tun ho kai salah . | tui em chu , ngai cit sala ho ma . | 5 | 9 |  |
| 19 | tolong , panggilan jarak jauh ke jepang . | then siw , ta thien an jan hi nyip pun . | then siw , ta thien hi nyip pun . | 9 | 11 |  |
| 20 | makanan saya sudah siap apa belum ? | sit ngai siap ho ma’ai mang mo ? | nga cu sit cu ho mang ? | 4 | 8 |  |
| 21 | saya ingin melakukan reservasi untuk bermain golf . | ngai oi co reservasi loi kau golf . | ngai oi co reservasi loi kau golf . | 8 | 8 |  |
| 22 | koin apa yang bisa digunakan untuk mesin penjual otomatis ini ? | koin hemai hiau digunakan loi mesin penjual otomatis lia mo ? | ma'ai thiet lui hiau jung loi lia otomatis mai ki ? | 3 | 11 |  |
| 23 | saya ingin wedang jahe . | ngai oi wedang jahe . | ngai oi wedang kiong ma . | 3 | 5 |  |
| 24 | tolong , saya ingin beberapa kemeja . | then siw , ngai oi kia kemeja . | then siw , ngai oi kia neu sam . | 7 | 8 |  |
| 25 | saya ingin beberapa makanan bayi . | ngai oi kia sit bayi . | ngai oi kia onga sit tung si . | 5 | 6 |  |
| 26 | pegang acar ini . | pan acar lia . | pan lia acar . | 2 | 4 |  |
| 27 | saya tidak merasa lebih baik . | ngai mo chi to ko ho . | ngai mo chi to hau ho . | 6 | 7 |  |
| 28 | maukah anda menunjukkan kepada saya pilihan yang lain ? | nyi oi pun khon pun ngai pilihan phet cak mo ? | nyi oi khon pun ngai phet cak mo ? | 9 | 11 |  |
| 29 | gaun ini terbuat dari nilon ? | khiun lia terbuat choi nilon mo ? | lia gaun nilon co loi he mo ? | 4 | 7 |  |
| 30 | oke , tentu . bayar ini di depan ketika anda siap . gunakan waktu anda , oke ? | oke , he . bayar lia mien chien kai cisi nyi siap . nyi jung si kien , oke ? | oke , ho wa . ban lia choi mien chien he nyi cang liau . jung nya si kien , oke ? | 13 | 20 |  |
| 31 | ya , dua blok ke arah itu . | he , nyi blok hi chu kai . | he , liong blok hi kai sak . | 6 | 8 |  |
| 32 | sekarang jam dua belas . | liha liong cak tiam cung . | kin ha sip nyi tiam . | 2 | 6 |  |
| 33 | sekarang saya tahu apa yang salah . | liha ngai ti hemai salah . | kin ha ngai ti he ma'ai sala . | 4 | 6 |  |
| 34 | gerbang mana yang harus saya lalui untuk transit saya ? | lew abui ngai oi lalui transit pun ngai mo ? | abui mun ngai oi ko loi nga transit ? | 4 | 10 |  |
| 35 | white beach ini indah . | white beach lia an ciang kai . | lia white beach an ciang . | 5 | 7 |  |
| 36 | tolong lakukan reservasi . | then siw co reservasi . | then siw co reservasi . | 5 | 5 |  |
| 37 | saya akan membawa kembalian anda nanti . | ngai oi tai nyi kaha tai ban con lui . | ngai oi tai nya ban con ka ha . | 7 | 10 |  |
| 38 | saya belum melakukan reservasi , tapi apakah ada kursi yang tersedia ? | ngai mang co reservasi , tan he jiu ten kai tersedia mo ? | ngai mang co reservasi , tapi han jiu ten mo ? | 9 | 13 |  |
| 39 | tolong , saya ingin pita mesin tulis . | then siw , ngai oi pita mesin sia si . | then siw , ngai oi sia ki pita . | 7 | 10 |  |
| 40 | tolong buka pintu dengan kunci cadangan . | then siw khoi mun kak cadangan so si . | then siw khoi mun jung cadangan so si . | 8 | 9 |  |
| 41 | roma tidak dibangun dalam sehari . | roma mo dibangun lok jit nyit . | roma mo tin hong jit nyit . | 5 | 7 |  |
| 42 | saya mau secangkir kopi lagi . | ngai oi jit pui kopi liau . | ngai oi jit pui kopi kak lin . | 6 | 7 |  |
| 43 | saya ingin hamburger untuk pergi . | ngai oi hamburger loi hi . | ngai oi hamburger loi hi . | 6 | 6 |  |
| 44 | dapatkah anda memeriksa ini ? | nyi hiau cha lia mo ? | nyi hiau cha lia mo ? | 6 | 6 |  |
| 45 | dompet saya telah dicuri . | nga dompet tin theu het liau . | nga dompet theu het liau . | 6 | 7 |  |
| 46 | dapatkah anda memintanya untuk menelepon kembali ? | nyi hiau pun tho ki tathien pun to con mo ? | nyi hiau tho gi loi ta thien to con mo ? | 9 | 11 |  |
| 47 | kami menyediakan makan di mobil dan kedai kopi di kereta . | menyediakan choi cha kak ngai teu kai sit kopi tiam choi cha . | ngai tew cu cang sit choi mo ka ka kopi tiam choi cha . | 9 | 13 |  |
| 48 | kamar saya belum dibersihkan . | nga kien tu mang dibersihkan . | nga kien tu mang met chiang . | 5 | 6 |  |
| 49 | ini ikeda yang sedang berbicara . saya telah memesan satu meja untuk pukul enam tiga puluh malam ini . tapi saya minta maaf saya akan datang terlambat setengah jam . | lia ikeda yang sedang kong boi tin cok ngai kau tai jit cak loi liuk tiam sam sip li pu ja . tan he ngai tho tui ngchu ngai boi loi thet man jit pan tiam . | lia ikeda kong boi . ngai jiu kau tai jit cak cok loi liuk tiam pan li pu ja . tapi ngai tui em chu ngai boi loi thet man pan tiam cung . | 29 | 37 |  |
| 50 | anda dapat naik bus di halte bus lima . | nyi hiau song cha choi halte cha ng . | nyi hiau song thai cha choi thi eng thai cha cham . | 7 | 9 |  |
| 51 | jika anda menemukannya , tolong , dapatkah anda menyimpannya untuk saya ? | he nyi chim gi , then siw , nyi hiau piong gi pun ngai mo ? | he nyi chim to gi , then siw , nyi hiau piong gi pun ngai mo ? | 16 | 16 |  |
| 52 | mana yang lebih baik , penerbangan pagi atau penerbangan sore ? | abui ko ho , pui ki co atau pui ki ciw mo ? | abui ko ho , co pui ki he ha ciw pui ki ? | 9 | 13 |  |
| 53 | apakah ada arena luncur es di dekat sini ? | he jiu arena luncur siat khiun ali mo ? | jiu arena luncur es khiun ali mo ? | 7 | 9 |  |
| 54 | terima kasih banyak . simpan kembaliannya . | kam chia to-to . piong ka cau con lui . | kam chia to-to . kiam ka to con lui . | 9 | 10 |  |
| 55 | tiga puluh delapan dolar lima puluh sen sehari , dengan jarak tempuh tidak terbatas . bensin tidak termasuk . | pat sam sip dollar eng sip sen jit nyit , kak li hang mo thon . bensin mo lok . | sam sip pat dollar eng sip sen jit nyit , ka hang mo thon . thien jiu mo son . | 16 | 20 |  |
| 56 | apakah anda menaruh perangko pada kartu pos ? | nyi jiu perangko menaruh kak pos phai mo ? | nyi jiu piong perangko choi pos phai mo ? | 5 | 9 |  |
| 57 | tiga malam . | sam pu ja . | sam ja . | 3 | 4 |  |
| 58 | jane , temanmu susan di sini untuk menemuimu . | jane , temanmu susan ali pun nyi nyi . | jane , nya phen jiu susan choi ali loi nyi nyi . | 6 | 9 |  |
| 59 | apakah ini ke edinburgh ? | he lia hi edinburgh ? | lia hi edinburgh he mo ? | 4 | 5 |  |
| 60 | kapan anda akan pergi ? | kiang si nyi oi hi mo ? | kiang si nyi oi hi ? | 6 | 7 |  |
| 61 | apakah saya harus mentransfer ? | he ngai oi mentransfer mo ? | ngai oi transfer mo ? | 4 | 6 |  |
| 62 | jika demikian , anda harus mengambil ukuran sepuluh sesuai standar amerika . | he anyiong , nyi oi kiam pien sip sesuai standar mui kiet . | he nyong hin , nyi oi kiam thi sip ho ma cau mui kiet standar . | 9 | 13 |  |
| 63 | sebenarnya , saudara-saudara saya sudah menikah , dan saya satu-satunya yang masih tinggal dengan orang tua saya . | chi sit , saudara-saudara ngai kau li , koli ngai han jitcak-jitcak het ka jit sa lo ngai . | cau li , nga ci oi kau liau , ka ngai tan sa yang choi het ka nga ja oi . | 10 | 19 |  |
| 64 | tidak ada handuk . | mo jiu handuk . | mo mang . | 2 | 4 |  |
| 65 | apakah anda keberatan orang lain mendapatkan itu untuk saya ? | he nyi boi chung to phet nyin tet to kai pun ngai mo ? | nyi chung to mo he phet sa nyin tet ka pun ngai ? | 11 | 14 |  |
| 66 | apakah ini yang anda maksud ? | lia he nyi kia mo ? | lia he nyia ji si mo ? | 3 | 6 |  |
| 67 | ini adalah kunci saya . | lia he ngai so si . | lia jiu he nga so si . | 5 | 6 |  |
| 68 | dapatkah anda memanggil nomor tersebut lagi ? | nyi hiau se tersebut ho liau ? | nyi hiau ta thien kai cak ho ma kak lin mo ? | 3 | 7 |  |
| 69 | mmm , bagus . permisi , dapatkah anda memeriksa salad saya ? | mmm , khiang . cia mun , nyi hiau cha nga salad mo ? | mmm , khiang . cia mun , nyi hiau cha nga salad mo ? | 14 | 14 |  |
| 70 | oh , anda telah melewatinya . silahkan kembali lagi . | oh , nyi ho ko het gi . ho ji fan to con . | oh , nyi ko het gi liau . ho ji fan loi . | 11 | 14 |  |
| 71 | saya pikir ini adalah adat jepang yang unik . | ngai siong to lia he adat nyip pun an unik . | ngai siong lia jiu he nyip pun fung siuk an unik . | 8 | 11 |  |
| 72 | apakah anda punya keju ? | he nyi jiu keju mo ? | nyi jiu keju mo ? | 5 | 6 |  |
| 73 | tolong , bungkus ini satu per satu , dan kenakan pita pada masing-masing bungkusan . | then siw , bungkus lia jit sa kok sa , koli kenakan pita choi kok sa bungkusan . | then siw , pau lia cak kok cak , ka jung pita pun kok cak pau . | 8 | 18 |  |
| 74 | tolong , bolehkah saya meminta daftar anggur ? | then siw , ngai ho ji tho daftar fu tho ciu mo ? | then siw , ngai ho ji tho daftar fu tho mo ? | 12 | 13 |  |
| 75 | saya lihat dulu . ya benar , makan siang hari ini akan dilayani lebih cepat . | ngai khon ha . he chok , sit ciu ci nyit lia boi dilayani ko mang . | ngai khon sen . he chok , sit tong ciu li nyit boi hau mang tin ta li . | 12 | 17 |  |
| 76 | anda berasal dari negara mana ? | nyi he choi kiet abui ? | nyi abui kiet loi ? | 4 | 6 |  |
| 77 | dapatkah anda mengubah ini menjadi seperempatan ? | nyi hiau bon lia pien seperempatan mo ? | nyi hiau bon lia pien si cha mo ? | 7 | 8 |  |
| 78 | itu tujuh . | chit ka . | ka chit . | 3 | 3 |  |
| 79 | yang putih ya . | phak kai ha . | phak kai dia . | 3 | 4 |  |
| 80 | bagaimana cara mematikan lampu baca ? | apan hin kia cara mematikan baca ten mo ? | nyong pan hin met kim thuk ten ? | 3 | 9 |  |
| 81 | saya mau dua porsi mie morisoba . | ngai oi liong porsi mie morisoba . | ngai oi liong phan morisoba mien . | 5 | 7 |  |
| 82 | saya bekerja di suatu perusahaan pialang saham . | ngai co she kai choi jit cak perusahaann pialang saham . | ngai co choi jit cak pialang saham nyin hong . | 8 | 11 |  |
| 83 | ya , ada satu kamar kosong . ada berapa orang ? | he , jiu jit kien khung . jiu jit sa nyin ? | he , jiu jit kien khung . jiu to sa nyin ? | 11 | 12 |  |
| 84 | anda tidak diizinkan untuk parkir di sini . | nyi em mo ji loi parkir choi ali . | nyi em mo ji loi parkir choi ali . | 9 | 9 |  |
| 85 | apakah ada toko penukaran uang dekat sini ? | he jiu tiam penukaran lui khiun ali mo ? | jiu kau lui tiam khiun ali mo ? | 6 | 9 |  |
| 86 | nama saya takeshi. nama anda siapa ? | nga miang takeshi. nyia miang he asa ? | nga miang takeshi . nyia miang ma sa ? | 8 | 8 |  |
| 87 | gunakan ini seminggu sekali . lumurilah dengan tipis setelah mencuci wajah anda , hindari jangan sampai terkena mata dan rambut . | jung lia jit cak lipai jit lin . lumurilah kak apunf kohet mencuci mien nyi , hindari ngmo teu tin to mukcu koli mo . | jung lia jit li pai jit lin . ko phok-phok he nyi se li mien , em mo teu tin to muk cu ka mo . | 17 | 25 |  |
| 88 | saya terluka oleh kaca di sini . | ngai terluka kiang kai choi ali . | ngai kot to tin to kiang choi ali . | 5 | 7 |  |
| 89 | saya seharusnya bertanya kepada pelatih , tapi bisakah anda katakan kepada saya mana mesin yang baik untuk melatih punggung saya ? | ngai cau li mun pun pelatih , tan he nyi hiau kong pun ngai abui mesin ho kai pun melatih punggung ngai mo ? | ngai cau li mun pelatih , tapi nyi hiau kong pun ngai thang abui ki hau ho loi lien nga jau kut mo ? | 17 | 24 |  |
| 90 | tolong berhenti di sisi persimpangan berikutnya . | then siw tiam choi sisi lin con ben . | then siw thin choi sun ka ha con ben . | 6 | 9 |  |
| 91 | apakah ada pramugari yang bisa berbahasa jepang ? | pramugari he jiu nyin hiau kong nyip pun mo ? | jiu pramugari hiau kong nyip pun boi mo ? | 7 | 10 |  |
| 92 | tolong , saya ingin limun . | then siw , ngai oi limun . | then siw , ngai oi son kam . | 6 | 7 |  |
| 93 | saya sakit . | ngai thung . | ngai pot phiang . | 2 | 3 |  |
| 94 | tolong , saya ingin susu . | then siw , ngai oi susu . | then siw , ngai oi nen . | 6 | 7 |  |
| 95 | berapa ons parfum saya bisa bebas bea ? | to ons parfum ngai hiau bebas bea mo ? | jit to ons nga hiong fa sui hiau bebas pajak ? | 5 | 9 |  |
| 96 | lengan ini terlalu panjang bagi saya . | kia lia an kiu pun ngai . | lia siu thet chong pun ngai . | 4 | 7 |  |
| 97 | tolong , dapatkah rambut saya dipotong , dikeriting dan dirapikan ? | then siw , hiau mo ngai dipotong dikeriting , koli dirapikan mo ? | then siw , hiau mo nga mo na cien , na kiw ka na rapikan ? | 6 | 13 |  |
| 98 | anda tidak harus menggantinya , tapi mobil ini untuk manusia gua . | nyi oi bon gi mo , tan he lia cha pun nyin gua . | nyi em si bon gi , tapi lia cha loi gua nyin . | 12 | 14 |  |
| 99 | hanya pariwisata . | chiang pariwisata . | chiang li hang . | 2 | 3 |  |
| 100 | kalau begitu , lift ada di sana . | he anyiong , jiu lift choi kai . | he nyong hin , lift jiu choi aun . | 6 | 8 |  |
|  |  |  |  | 721 | 991 |