Pengumpulan Korpus Paralel Bahasa Indonesia-Sunda dari Wikipedia Menggunakan Metode Pointwise Mutual Information

Indonesian-Sundanese Parallel Corpus Retrieval from Wikipedia Using Pointwise Mutual Information Method

Arizal Firdaus[1], Arie Ardiyanti Suryani[2], Kurniawan Nur Ramadhani[3]

[123]Program Studi S1 Teknik Informatika, Fakultas Informatika, Universitas Telkom
[1]firdausarizal23@studemts.telkomuniversity.ac.id, [2]ardiyanti@telkomuniversity.ac.id,
[3]kurniawannr@telkomuniversity.ac.id

Abstak

Pengumpulan korpus paralel sedang gencar dilakukan untuk keperluan studi dan pengembangan NLP. Namun, untuk pasangan kalimat beberapa bahasa, khususnya Bahasa Indonesia-Sunda, jumlah korpus paralel yang tersedia masih sangat sedikit. Sedangkan untuk mengumpulkan korpus paralel secara manual memerlukan waktu yang lama dan biaya yang mahal. Dengan alasan tersebut, pengumpulan korpus paralel akan lebih efektif dan efisien jika dikumpulkan secara otomatis. Dalam tugas akhir ini, akan dilakukan penelitian pengumpulan korpus paralel pada Wikipedia meggunakan metode *Pointwise Mutual Information* (PMI) untuk menentukan *sentence similarity*. Pengambilan data dari artikel Wikipedia bahasa Indonesia dan Sunda dengan memanfaatkan fasilitas *interlanguage link* dan MediaWIki API. Dengan metode ini, diharapkan didapat korpus paralel yang cukup baik dengan efisien.

Kata kunci: korpus paralel, Wikipedia, pointwise mutual information, interlanguage link, MediaWiki API

Abstract

Parallel corpus retrieval is currently done for NLP research and development. However, for some language pairs, especially Indonesian-Sundanese, parallel corpus number is still very few. Moreover, manual parallel corpus retrieval requires time and expensive fee. For that reason, parallel corpus retrieval will be more effective and efficient if retrieved automatically. In this thesis, parallel corpus retrieval research at Wikipedia using Pointwise Mutual Information (PMI) for determining sentence similarity will be done. Interlanguage link and MediaWiki API will be used for data retrieval from Wikipedia, which are Indonesian and Sundanese articles. By this method, hopefully good parallel corpus will be obtained efficiently

Keywords: parallel corpus, wikipedia, pointwise mutual information, interlanguage link, MediaWiki API

1. Pendahuluan

Pada saat ini, *Natural Language Processing* (NLP) menjadi salah satu topik penelitian teknologi informasi yang sedang populer. NLP dapat didefinisikan sebagai pemrosesan bahasa alami secara otomatis. Salah satu komponen yang penting pada NLP adalah korpus paralel. Korpus paralel dapat digunakan untuk keperluan NLP seperti membuat mesin translasi berdasarkan statistik atau yang biasa disebut *Statistical Machine Translation* (SMT), tata bahasa, dan sosiolinguistik. SMT menggunakan korpus paralel untuk melakukan proses *learning* untuk menemukan keterkaitan antar teks sumber dengan teks yang dituju. Pengumpulan korpus paralel sedang gencar dilakukan untuk keperluan studi dan pengembangan NLP. Korpus paralel adalah pasangan dokumen teks yang berisi pasangan kalimat yang saling menerjemahkan satu sama lain secara harfiah [1]. Korpus paralel seperti itu sudah sangat banyak dan memiliki kualitas yang baik untuk beberapa bahasa yang sudah sering dipakai sebagai objek penelitian NLP seperti korpus paralel bahasa Inggris-Perancis yang berjumlah dua juta pasang kalimat di situs Europarl. Selain Europarl, terdapat situs-situs lain penyedia korpus paralel seperti OPUS, Multi-UN, dan Mircotopia. Namun, untuk sebagian yang lain seperti korpus paralel Bahasa Indonesia-Sunda masih sangat sedikit. Pengumpulan korpus paralel secara manual memakan waktu yang sangat lama dan biaya yang tidak sedikit sehingga akan lebih efektif jika pengumpulan korpus paralel dilakukan secara otomatis.

Penelitian ini bertujuan untuk mengumpulkan korpus paralel pasangan bahasa Sunda dan bahasa Indonesia dari Wikipedia. Hal pertama yang dilakukan adalah mengambil pasangan artikel bahasa Indonesia-Sunda yang setopik dengan memanfaatkan fitur *interlanguage link*. Setelah itu, melakukan *preprocessing* untuk merapikan pasangan artikel yang diambil. Artikel yang dihasilkan dari tahap preprocessing akan dipakai untuk membuat kamus sama kata, kamus PMI, dan korpus paralel. Kamus manual [2] atau kamus yang ditulis secara manual berdasarkan buku kamus dan kamus sama kata dibutuhkan untuk membuat kamus PMI. Korpus paralel

dikumpulkan berdasarkan pasangan kata yang tersimpan pada semua kamus yang tersedia. Hasil korpus paralel yang terkumpul kemudian dinilai ketepatannya secara manual.

2. Penelitian Terkait

Penelitian ini menggunakan metode serupa yang sudah pernah dilakukan sebelumnya pada penelitian Joel Martin, Howard Johnson, Benoit Farley, & Anna Maclachlan [3] dan penelitian Rada Mihalcea, Courtney Corley, & Carlo Strapparava [4]. Penelitian pertama menggunakan sumber korpus berupa teks paralel dari Nunavut Hansards untuk mendapatkan kamus kata Bahasa Inggris-Inuktitut. Mereka menggunakan proses *sentence alignment* sebelum melakukan ekstraksi kamus kata. Penelitian tersebut berhasil mendapatkan kamus kata dengan nilai presisi sebesar 87%. Penelitian kedua menggunakan korpus parafrase Microsoft untuk melakukan pengujian *semantic* similarity dan menentukan bahwa kedua kata tersebut parafrase atau bukan. Penelitian tersebut berhasil mendapatkan kamus kata dengan nilai presisi sebesar 70,2%.

Perbedaan antara kedua penelitian tersebut dengan penelitian ini adalah penggunaan metode PMI pada kedua penelitian tersebut menjadi produk akhir dari penelitian, yaitu sebagai kamus kata. Sedangkan pada penelitian ini, kamus yang dihasilkan dari metode PMI akan dipakai untuk menentukan kandidat korpus paralel.

Pemakaian PMI sebagai metode penelitian yang dipakai pada penelitian ini didasari pada penelitian Rada Mihalcea, Courtney Corley, & Carlo Strapparava [4]. Penelitian tersebut menggunakan *Latent Semantic Analysis* (LSA) dan PMI sebagai pengukur *semantic similarity* berbasis korpus yang menghasilkan nilai presisi kamus kata PMI lebih besar dari presisi LSA dengan selisih 0,05%. Dengan mempertimbangkan nilai presisi PMI yang didapat lebih besar daripada LSA, maka metode tersebut dipilih untuk diuji pada penelitian ini. Selain itu, PMI relatif lebih sederhana untuk mencari arti kata yang berdekatan dengan kata lain yang sudah ada di kamus dibandingkan dengan LSA karena LSA harus menggunakan *matrix* sedangkan PMI cukup dengan menghitung frekuensi kemunculan kata. Namun PMI mempunyai kelemahan, yaitu jika kedua kata yang sedang dihitung mempunyai frekuensi kemunculan kata yang kecil, maka nilai PMI kedua kata tersebut dapat menjadi besar.

Pemakaian *interlanguage links* pada Wikipedia untuk mencari kesamaan informasi antar artikel yang ditulis dalam bahasa lain sudah digunakan dalam beberapa penelitian. Contohnya adalah penelitian Mohammadi dan GasemAghaee [5] yang memanfaatkan *interlanguage links* dan artikel wikpedia sebagai sumber data korpus *comparable* untuk menghasilkan korpus paralel dua bahasa, yaitu bahasa Inggris-Persia.

3. Wikipedia Sebagai Sumber Korpus Paralel

Korpus paralel adalah pasangan dokumen teks yang berisi pasangan kalimat yang saling menerjemahkan satu sama lain secara harfiah [1]. Korpus paralel mempunyai peran penting dalam mesin translasi dan pemrosesan berbagai bahasa alami, contohnya sebagai data training untuk keperluan SMT.

Penelitian ini memanfaatkan fasilitas yang terdapat pada Wikipedia, yaitu *interlanguage link*. *Interlanguage link* adalah fasilitas Wikipedia berupa tautan yang digunakan untuk mengarahkan pengguna ke artikel Wikipedia berbahasa lain yang mengandung judul atau topik yang sama. Misalnya, pada artikel Wikipedia bahasa Indonesia terdapat suatu artikel berjudul rumah, pada halaman tersebut terdapat *interlanguage links* yang akan mengarah kepada artikel bahasa Sunda yang berjudul *imah* dan begitu pula sebaliknya.

Artikel Wikipedia dapat diakses menggunakan API yang disediakan oleh MediaWiki (perangkat lunak berbasis PHP yang digunakan oleh Wikipedia) untuk mendapatkan daftar *interlanguage links* yang ada pada suatu artikel Wikipedia. API diakses dengan cara mengirimkan HTTP *request* melalui api.php ke *web service* MediaWiki.

4. Pointwise Mutual Information

Pointwise Mutual Information atau PMI adalah rasio probabilitas antara dua kejadian (w_1 dan w_2) terjadi secara bersamaan dengan gabungan probablitias dari dua kejadian yang terjadi secara independen [6], dalam hal ini yang dimaksud kejadian tersebut adalah kata. Semakin besar nilai PMI maka semakin besar kemungkinan dua kata tersebut dependen. Sebaliknya, semakin kecil nilai PMI maka semakin besar kemungkinan dua kata tersebut independen.

Jika diberi dua kata, maka nilai PMI kedua kata tersebut dapat dihitung dengan rumus:

$$PMI(w_1, w_2) = log2\ \frac{p(w_1 \& w_2)}{p(w_1) * p(w_2)} \qquad (1)$$

$p(w_1 \& w_2)$: Probabilitas kemunculan kedua kata secara bersamaan
$p(w_1)$: Probabilitas kemunculan kata pertama
$p(w_2)$: Probabilitas kemunculan kata kedua

5. Evaluasi terhadap Korpus Paralel

Untuk mengevaluasi dan menentukan bahwa korpus paralel yang tersedia saling mengartikan, dilakukan penghitungan presisi secara manual. Presisi adalah perbandingan dari jumlah data yang relevan dengan jumlah semua data yang relevan ditambah data yang tidak relevan [7]. Presisi biasanya dinyatakan dengan persentase [7]. Rumus presisi yang dipakai adalah sebagai berikut:

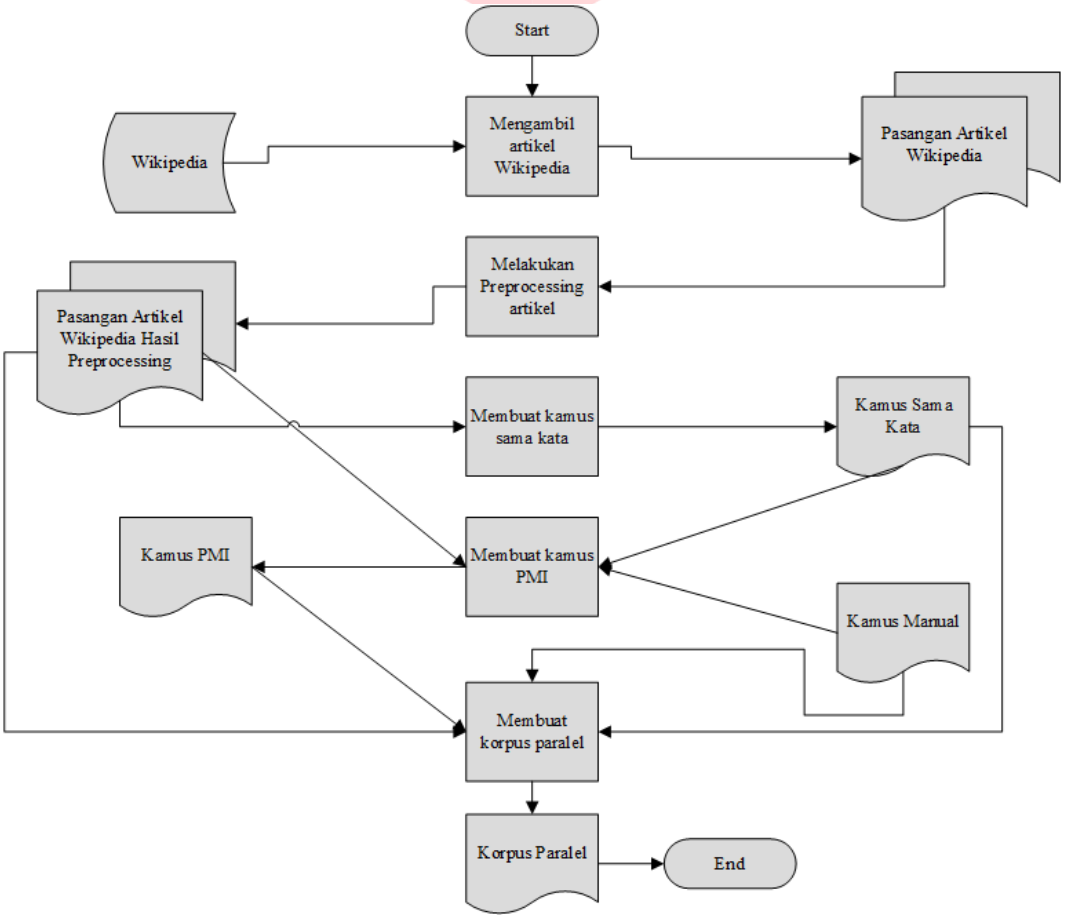
$$Presisi = \frac{CCP}{CP} * 100 \qquad (2)$$

CCP: Jumlah korpus paralel yang benar
CP: Jumlah seluruh kandidat korpus paralel

Sebagai tambahan untuk meningkatkan presisi korpus paralel, disisipkan penghitungan panjang kalimat. Pemakaian penghitungan panjang kalinat mengacu pada Gale dan Church [8] yang menyebutkan bahwa kalimat panjang cenderung diartikan ke kalimat panjang juga dalam bahasa lain dan kalimat pendek cenderung diartikan ke kalimat pendek juga dalam bahasa lain.

6. Pengumpulan Korpus Paralel

Penelitian ini bertujuan untuk membentuk korpus paralel bahasa Indonesia-Sunda dalam tingkat kalimat yang diambil dari artikel Wikipedia bahasa Indonesia dan Sunda. Sistem yang dibuat pada penelitian ini terdiri beberapa langkah-langkah pengerjaan yang digambarkan dengan gambar berikut.



Gambar 1 Flow chart langkah-langkah penelitian

6.1 Pengambilan artikel Wikipedia

Pada tahap ini, dilakukan proses pencarian dan pengunduhan pasangan artikel Wikipedia berjudul atau bertopik sama yang ada pada versi bahasa Indonesia dan Sunda. Pengambilan artikel dilakukan dengan memanfaatkan fasilitas *interlanguage links* yang terdapat pada artikel Wikipedia. Artikel diperoleh secara *random*.

Program akan memasuki artikel bahasa Sunda terlebih dahulu, kemudian mengecek *interlanguange links* bahasa Indonesia. Jika ada, maka pasangan artikel tersebut akan disimpan.

6.2 Preprocessing Artikel

Pada tahap ini, dilakukan "pembersihan" artikel yang sudah diunduh untuk merapikan artikel dan mengurangi kendala yang muncul saat melakukan tahap penerapan metode penelitian. Program akan mengecek artikel yang sudah diunduh kemudian membersihkan dan merapikan kalimat satu persatu sesuai dengan fungsi *preprocessing* yang ditetapkan. Proses ini menghasilkan dokumen *file* baru yang berisi artikel hasil *preprocessing* yang dinamai sama dengan *file* artikel asli. Artikel ini akan dipakai menjadi sumber untuk pembuatan kamus sama kata, kamus PMI, dan pengumpulan korpus paralel. Proses ini mempunyai fungsi sebagai berikut:

- Memisahkan kalimat menjadi satu kalimat per baris.
- Memperbaiki tanda titik yang bukan sebagai penanda akhir kalimat. Sebagai contoh pada awalnya tanda titik pada angka ribuan dan gelar dianggap sebagai akhir kalimat, setelah penerapan proses ini tanda titik pada angka ribuan dan gelar akan menjadi satu kalimat.
- Menghapus *whitespace*.
- Menghapus judul subbab.
- Menjadikan kalimat setelah tanda titik dua menjadi kalimat baru.
- Mengubah format dokumen menjadi .txt.
- Menghapus tanda titik yang lebih dari satu
- Membuang *link* rujukan.

6.3 Pembuatan Kamus Sama Kata

Pada tahap ini, dilakukan pengumpulan kata-kata yang sama pada setiap pasangan artikel menjadi satu *file* dokumen. Kamus ini akan dipakai sebagai sumber pembuatan kamus PMI dan pengumpulan korpus paralel. Cara pembuatan kamus sama kata memakai logika sederhana, yaitu dengan cara mengecek satu persatu kata pada artikel bahasa Sunda dicocokkan dengan kata pada artikel bahasa Indonesia yang bertopik sama. Jika ada kata yang sama, maka kata tersebut akan dimasukkan ke dalam kamus sama kata. Jika tidak, maka kata tersebut tidak akan dimasukkan ke dalam kamus sama kata dan program akan melanjutkan ke kata selanjutnya. Kamus ini berguna untuk menangani nama, istilah, dan kata bahasa Sunda yang memiliki penulisan yang sama dalam versi bahasa Indonesia. Kamus ini bersifat unik, sehingga tidak akan ada kata yang sama.

6.4 Pembuatan Kamus PMI

Pada tahap ini, dilakukan pembuatan kamus PMI yang akan digunakan untuk menerapkan metode PMI untuk pengambilan korpus paralel. Kamus PMI dibuat dengan bantuan kamus manual dan kamus sama kata. Kamus manual adalah kamus yang diketik secara manual dari buku Kamus Basa Sunda karangan Satjadibrata [2]. Pasangan kata yang dihasilkan oleh PMI sangat bergantung dengan skor sehingga diperlukan *threshold* untuk menyaring pasangan kata yang masuk ke dalam kamus PMI. *Threshold* tidak mempunyai patokan pasti sehingga *threshold* yang dipakai pada penelitian ini ditetapkan dengan mempertimbangkan *range* skor PMI pada kamus PMI. *Threshold* yang digunakan pada penelitian ini ada tiga, yaitu dua, enam, dan sepuluh. Artikel yang diperiksa skor PMI-nya hanya artikel Sunda saja. Artikel bahasa Sunda yang dipakai berjumlah 5.173 artikel yang terdiri dari 732.965 kata.

6.5 Pengumpulan Korpus Paralel

Pada tahap ini, diliakukan pengumpulan korpus paralel dari artikel yang sudah ditetapkan berdasarkan kata-kata yang ada pada kamus manual, kamus PMI, dan kamus sama kata. Penghitungan panjang kalimat digunakan untuk meningkatkan presisi korpus paralel. Garis besar proses ini adalah mencocokkan kata pada kalimat suatu pasangan artikel dengan kamus yang sudah disediakan. Bila pasangan kalimat mempunyai kata yang saling mengartikan sesuai dengan kamus, maka pasangan kalimat tersebut dapat dianggap paralel. Semakin banyak kata yang saling mengartikan pada pasangan kalimat tersebut maka semakin besar juga kemungkinan pasangan kalimat tersebut paralel.

Untuk pengujian pengaruh pemakaian kamus PMI terhadap kualitas korpus paralel, digunakan persamaan berikut:

$$\text{Skor} = (2 \times \text{nKM}) + (2 \times \text{nKSM}) + (1 \times \text{nKPMI}) \quad (3)$$

nKM : Jumlah pasangan kata di kalimat suatu artikel yang terdapat pada kamus *manual*

nKSM : Jumlah pasangan kata di kalimat suatu artikel yang terdapat pada kamus sama kata
nKPMI : Jumlah pasangan kata di kalimat suatu artikel yang terdapat pada kamus PMI

Bedasarkan persamaan di atas, bobot kata pada kamus PMI dibawah bobot kata pada kamus lainnya. Hal ini dikarenakan pasangan kata pada kamus PMI mempunyai tingkat presisi lebih rendah dibandingkan pasangan kata pada kamus manual dan kamus sama kata yang sudah pasti benar.

Untuk pengujian pengaruh pemakaian penggunaan panjang kalimat terhadap kualitas korpus paralel, digunakan persamaan berikut:

$$Skor = ((2 \times nKM) + (2 \times nKSM) + (1 \times nKPMI)) \times Rasio \quad (4)$$

nKM : Jumlah pasangan kata di kalimat suatu artikel yang terdapat pada kamus *manual.*
nKSM : Jumlah pasangan kata di kalimat suatu artikel yang terdapat pada kamus sama kata.
nKPMI : Jumlah pasangan kata di kalimat suatu artikel yang terdapat pada kamus PMI.
Rasio : Panjang kalimat pada pasangan kalimat yang memiliki kata lebih sedikit dibagi dengan panjang kalimat pada pasangan kalimat yang memiliki kata lebih banyak.

7. Evaluasi

Persiapan pasangan artikel yang diambil secara *manual* dari artikel yang sudah diunduh sebelumnya. Penelitian ini menggunakan 20 pasang korpus paralel yang terdiri dari 230 kalimat dan 20 pasang korpus *comparable* yang terdiri dari 167 kalimat. Korpus paralel adalah pasangan dokumen teks yang berisi pasangan kalimat yang saling menerjemahkan satu sama lain secara harfiah [1]. Adapun korpus *comparable* adalah pasangan dokumen teks setopik yang kalimatnya tidak saling menerjemahkan secara langsung satu sama lain [1]. Terdapat enam kali pengujian dengan kondisi yang berbeda pada masing-masing pasangan korpus paralel dan *comparable*. Kondisi pengujian tersebut adalah sebagai berikut:

1. Pengujian tanpa menggunakan kamus PMI dan penghitungan panjang kalimat
2. Pengujian tanpa menggunakan kamus PMI dan menggunakan penghitungan panjang kalimat
3. Pengujian menggunakan kamus PMI *threshold* 2, 6, dan 10 tanpa menggunakan penghitungan panjang kalimat
4. Pengujian menggunakan kamus PMI *threshold* 10 dan penghitungan panjang kalimat

Adapun alasan threshold PMI terbesar adalah 10 karena kamus PMI yang terbentuk dengan *threshold* lebih dari 10 menghasilkan korpus paralel yang tidak lebih baik dari kamus PMI *threshold* 10, sehingga kamus PMI *threshold* 10 dipilih karena kata yang dihasilkan di kamus lebih banyak.

Penelitian ini menggunakan enam kondisi yang berbeda untuk membandingkan besarnya presisi yang dicapai oleh kondisi tersebut. Dari enam kondisi tersebut, terdapat empat kondisi yang memakai kamus PMI dan tanpa menggunakan penghitungan panjang kalimat. Untuk meringkas hasil pengujian, hasil pengujian akan dibagi menjadi dua bagian, yaitu pengujian pada pasangan korpus paralel dan korpus *comparable*. Masing-masing pengujian menggunakan persamaan 3 dan 4.

7.1 Pengujian pada Korpus Paralel

Tabel 1: Presisi hasil pengujian pengaruh kamus PMI terhadap pasangan korpus paralel

	Tanpa PMI	**PMI Threshold 2**	**PMI Threshold 6**	**PMI Threshold 10**
5 Besar	99,13%	98,70%	98,70%	99,57%
Tertinggi	92,61%	94,78%	94,35%	93,91%

Tabel 2 : Presisi hasil pengujian pengaruh panjang kalimat terhadap pasangan korpus paralel

	Tanpa PMI	**Tanpa PMI Menggunakan Panjang Kalimat**	**PMI Threshold 10**	**PMI Threshold 10 Menggunakan Panjang kalimat**
5 Besar	99,13%	99,13%	99,57%	99,57%

Tertinggi	92,61%	97,83%	93,91%	97,83%

Dari tabel di atas, dapat dilihat bahwa pemakaian kamus PMI dapat meningkatkan kualitas korpus paralel yang dibentuk. Maksud dari 5 besar adalah kalimat bahasa Indonesia yang merupakan terjemahan dari suatu kalimat bahasa Sunda meripakan kandidat kalimat paralel *top*-5 dengan skor tertinggi yang ditampilkan pada data hasil uji. Sedangkan maksud dari tertinggi adalah kalimat bahasa Indonesia yang merupakan terjemahan dari suatu kalimat bahasa Sunda terletak pada urutan pertama, dengan kata lain sistem berhasil menentukan korpus paralel dengan tepat.

Kasus hasil uji yang belum dapat ditangani

Terdapat kasus yang belum dapat ditangani pada pengujian korpus paralel tanpa menggunakan panjang kalimat, yaitu nilai presisi korpus paralel yang dihasilkan menggunakan kamus PMI *threshold* 10 lebih kecil daripada nilai presisi korpus paralel yang dihasilkan menggunakan kamus PMI *threshold* 2. Menurut hasil pengujian pasangan artikel pada tabel 4.3, PMI *threshold* 10 mendapat skor terbesar pada kategori "5 besar" dan PMI *threshold* 2 mendapat skor terbesar pada kategori "tertinggi". Presisi korpus paralel menggunakan PMI meningkat pada kategori kategori "tertinggi" sebesar 1-3%. Hal ini tidak sesuai dengan perkiraan awal karena PMI *threshold* 10 diprediksi akan memiliki tingkat presisi tertinggi pada kategori "tertinggi". Hal ini dipengaruhi oleh *entry* kamus PMI yang berbeda. Pada beberapa kalimat di artikel yang diuji, terdapat kata yang ada di kamus PMI *threshold* 2, sedangkan di kamus PMI *threshold* 10 tidak ada.

Adapun hasil pengujian pengaruh panjang kalimat terhadap pasangan korpus paralel telah sesuai dengan ekspektasi seperti yang tertera pada tabel 4.4. Presisi korpus paralel setelah menggunakan penghitungan panjang kalimat pada kategori "tertinggi" meningkat sebesar 3-6%. Dengan menggunakan penghitungan panjang kalimat, kalimat yang bukan merupakan terjemahan kalimat yang sedang diperiksa skornya akan berkurang cukup besar karena panjang kalimat kalimat tersebut berbeda cukup jauh. Sebaliknya, kalimat yang merupakan terjemahan kalimat yang sedang diperiksa skornya akan tetap atau berkurang sedikit karena panjang kalimat kalimat tersebut sama atau berbeda sedikit.

Presisi korpus paralel yang dihasilkan oleh metode yang dipakai tinggi sehingga dapat disimpulkan metode yang dipakai andal dan diperkirakan akan mendapatkan presisi yang cukup tinggi ketika diuji terhadap korpus *comparable*.

7.2 Pengujian pada Pasangan Korpus Comparable

Tabel 3: Presisi hasil pengujian pengaruh kamus PMI terhadap pasangan korpus *comparable*

	Tanpa PMI	**PMI Threshold 2**	**PMI Threshold 6**	**PMI Threshold 10**
5 Besar	83,83%	85,03%	83,83%	83,83%
Tertinggi	66,48%	67,67%	68,26%	68,86%

Tabel 4: Presisi hasil pengujian pengaruh panjang kalimat terhadap pasangan korpus *comparable*

	Tanpa PMI	**Tanpa PMI Menggunakan Panjang Kalimat**	**PMI Threshold 10**	**PMI Threshold 10 Menggunakan Panjang kalimat**
5 Besar	83,83%	83,83%	83,83%	83,83%
Tertinggi	66,48%	69,46%	68,86%	70,66%

Menurut hasil pengujian pasangan artikel pada tabel 3, PMI *threshold* 2 mendapat skor terbesar pada kriteria "5 besar" dan PMI *threshold* 10 mendapat skor terbesar pada kategori "tertinggi". Presisi korpus paralel menggunakan PMI meningkat pada kategori "tertinggi" sebesar 1-3%. Hal ini sesuai dengan perkiraan awal karena PMI *threshold* 10 diprediksi akan memiliki tingkat presisi tertinggi pada kategori "tertinggi".

Kasus hasil uji yang belum dapat ditangani

Terdapat kasus yang belum dapat ditangani pada pengujian korpus *comparable* menggunakan panjang kalimat. Sebenarnya, hasil pengujian pengaruh panjang kalimat terhadap pasangan korpus *comparable* telah sesuai dengan ekspektasi seperti yang tertera pada tabel 4. Presisi korpus paralel juga setelah menggunakan penghitungan panjang kalimat meningkat sebesar 2-4% pada kategori "tertinggi". Namun, terdapat kendala yang cukup signifikan memengaruhi presisi korpus paralel pada pasangan korpus *comparable*, yaitu pasangan kalimat yang menerjemahkan secara tidak langsung. Pasangan kalimat yang menerjemahkan secara tidak langsung memiliki selisih panjang kalimat yang cukup jauh sehingga hal ini memengaruhi nilai presisi hasil uji.

8. Kesimpulan

Sistem yang digunakan mampu mengumpulan korpus paralel yang lebih baik dengan menggunakan kamus PMI dengan bantuan kamus sama kata dan kamus *manual*. Hasil analisis menunjukkan pada pengujian korpus paralel maupun *comparable* kategori tertinggi menggunakan metode PMI meningkatkan presisi sebesar 1-3% dibandingkan dengan tidak menggunakan metode PMI. Namun, terdapat kendala pada pengujian korpus paralel tanpa menggunakan panjang kalimat, yaitu presisi korpus paralel yang dihasilkan oleh kamus PMI threshold 10 lebih kecil dibandingkan dengan presisi korpus paralel yang dihasilkan oleh kamus PMI threshold 2. Hal ini dipengaruhi oleh *entry* kamus PMI yang berbeda. Pada beberapa kalimat di artikel yang diuji, terdapat kata yang ada di kamus PMI *threshold* 2, sedangkan di kamus PMI *threshold* 10 tidak ada.

Adapun penghitungan panjang kalimat meningkatkan presisi sebesar 3-6% pada pengujian korpus paralel dibandingkan dengan tidak menggunakan panjang kalimat dan 2-4% pengujian pada korpus *comparable* dibandingkan dengan tidak menggunakan panjang kalimat. Pemakaian penghitungan panjang kalimat mampu mengumpulkan korpus paralel yang lebih baik dengan cara mengubah pasangan kalimat yang hanya masuk kategori "5 besar" menjadi masuk ke dalam kategori "tertinggi" juga. Namun, kalimat yang mengartikan secara tidak langsung menjadi kendala pada pengujian ini karena penghitungan panjang kalimat sangat bergantung pada kesamaan panjang kalimat sedangkan kalimat yang mengartikan secara tidak langsung cenderung memiliki panjang kalimat yang berbeda cukup jauh.

Daftar Pustaka:

[1] R. C. Kulkarni, "Extraction of Parallel Corpora from Comparable Corpora," Department of Computer Science & Engineering, Indian Institute of Technology, India, 2013

[2] R. Satjadibrata, "Kamus Basa Sunda," Yogyakarta: Kementrian P.P dan K, 1987

[3] J. Martin, H. Johnson, B. Farley, and A. Maclachlan, "Aligning and Using an English-Inuktitut Parallel Corpus," in Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond - Volume 3, Association for Computational Linguistics Stroudsburg, PA, USA, 2003.

[4] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and Knowledge-based Measures of Text Semantic Similarity*,"* in Proceedings of the 21st national conference on Artificial intelligence - Volume 1, AAAI'06, 2006.

[5] M. Mohammadi and N. GhasemAghaee, "Building Bilingual Parallel Corpora Based on Wikipedia," in Computer Engineering and Applications (ICCEA), 2010 Second International Conference, Bali, Indonesia, 2010.

[6] N. Varma, "IdentifyingWord Translations in Parallel Corpora Using Measures of Association," Department of Computer Science, University of Minnesota, USA, 2002.

[7] M. Akhila and E. Sruthi, "Social Spam Detection in Microblog," in Recent Trends in Computational Intelligence & Image processing - RICP 2017, 2017

[8] W. A. Gale and K. W. Church, "A Program for Aligning Sentences in Bilingual Corpora," Computational Linguistics, vol. 19, no. 1, pp. 75-102, 1993.