JUMANJI
ISSN (p): 2598-8050
ISSN (e): 2598-8069
| Vol. 02 | No. 01 | Halaman 81 - 88
April 2018

# Analisis Pembangunan Korpus Berpasangan Untuk Pembangkitan Parafrasa Pada Makalah Ilmiah

**RIDWAN ILYAS[1], DWI HENDRATMO WIDYANTORO[2], MASAYU LEYLIA KHODRA[3]**

[1,2,3] Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung
rdwnilyas@gmail.com

**ABSTRAK**

*Pembangunan mesin yang dapat membangkitkan kalimat baru dengan tingkat semantik yang tinggi namun secara penulisan berbeda (parafrasa) membutuhkan sumberdaya bahasa berupa korpus parallel. Proses pembangunan korpus memerlukan analisis awal sesuai dengan domain dari mesin yang akan dibuat. Pada penelitian ini dilakukan analis dalam pembangunan korpus berpasangan pada makalah ilmiah. Kalimat-kalimat pada makalah ilmiah memiliki karakteristik yang berbeda dengan domain lain seperti berita atau media sosial. Dari hasil proses ekstraksi awal didapatkan 590.402 kalimat isi dan 23.584 kalimat abstrak. Hasil dari penelitian ini dapat menjadi kandidat korpus yang dilakukan dengan proses terkomputerisasi.*

***Kata kunci:*** *Analisis Korpus, Parafrasa, Makalah Ilmiah.*

## 1. PENDAHULUAN

Pembangkitan parafrasa adalah teknik menghasilkan teks baru yang informasinya sama dengan teks masukan, dengan penulisan yang berbeda. Kesamaan informasi antara teks masukan dan keluaran harus bersifat bolak balik. Pada makalah ilmiah parafrasa dapat ditemukan pada penjabaran abstrak ke sisi, penjabaran penelitian terkait pada bagian pembukaan, penjabaran metode dari bagian pembukaan dan kalimat sitasi antar makalah.

Pembangkitan parafrasa pada makalah harus mempertimbangkan beberapa kriteria. Kalimat pada makalah ilmiah umum bersifat argumentatif [1], sehingga satu kalimat terikat konteks dengan kalimat lain, baik dalam paragraf sebab akibat atau pun sebaliknya. Selain itu kalimat baru yang dihasilkan tidak boleh bersifat plagiat (Barrom-Cedeno dkk 2013), sehingga terdapat batas minum substitusi yang harus dilakukan. Makalah ilmiah banyak mengandung kalimat majemuk setara maupun bertingkat, sehingga bentuk kalimat lebih kompleks [3]. Pilihan leksikal pada pembangunan kalimat makalah ilmiah lebih terbatas dibanding bebas domain. Sebagai contoh, IEEE mengeluarkan beberapa aturan dalam penulisan makalah ilmiah yang membatasi penggunaan beberapa kata.

Parafrasa digunakan pada makalah ilmiah sebagai penghargaan bagi penulis dari makalah yang dirujuk [4]. Parafrasa digunakan dengan tujuan untuk menghindari plagiat. Tetapi Parafrasa dipakai dalam membangun kesimpulan atau analogi. Kalimat-kalimat sitasi suatu makalah pada beberapa makalah yang berbeda merupakan kalimat parafrasa, contoh (1) parafrasa pada kalimat sitasi [5]. Abstrak dari makalah ilmiah merupakan penulisan ulang dari poin-poin penting yang ada pada isi makalah, contoh (2) parafrasa abstrak dan isi makalah [3]. Kalimat-kalimat umum yang ada pada bagian pembukaan biasanya merupakan pernyataan yang juga ditemukan pada metode atau penelitian terkait.

(1) a. Miller et al. did not manage to verify whether saturation was reached.

b. Miller et al. did not verify whether saturation was reached.

(2) a. Tissue factor (TF) is a transmembrane glycoprotein and the major cellular trigger of blood coagulation

b. Tissue factor (TF) is a transmembrane glycoprotein and the main triggering element of blood coagulation

Pembangkitan parafrasa membutuhkan korpus berpasangan sebagai data latih. Korpus berpasangan dibangun dari kumpulan makalah ilmiah yang telah dievaluasi hasilnya oleh ahli/peneliti. Setiap makalah ilmiah diekstraksi menjadi kumpulan kalimat. Lalu digunakan 3 skenario pengumpulan korpus kalimat bermasangan yaitu:

1. Kumpulan kalimat-kalimat yang mensitasi makalah yang sama berpeluang sebagai kalimat parafrasa

2. Kalimat abstrak pada setiap makalah memiliki pasangan kalimat parafrasa pada isi makalah

3. Kalimat pada bagian pendahuluan memiliki pasangan kalimat parafrasa pada bagian metodologi

4. Kalimat pada bagian kesimpulan memiliki pasangan kalimat parafrasa pada bagian hasil eksperimen

5. Kalimat-kalimat yang menyatakan definisi dari satu istilah memiliki pasangan parafrasa dengan kalimat lain yang mendefinisikan masalah yang sama.

## 2. EKSTRASI PARAFRASA

Ekstraksi merupakan task untuk memproses korpus, baik itu korpus berpasangan (contoh: korpus terjemahan) atau korpus bebas (contoh: korpus berita), dengan menghasilkan pasangan parafrasa dalam bentuk leksikal, frasa, kalimat atau pola ekspresi bahasa . Teknik-teknik untuk mengumpulkan sinomim, hipernim atau hiponim dapat pula disebut parafrasa . Keluaran dari ekstrasi parafrasa berupa korpus yang berguna untuk *tasks* parafrasa yang lain.

### PPDB

PPDB adalah satu produk yang dihasilkan dari ektraksi parafrasa dengen memanfaatkan pararel korpus terjemahan mahasa Inggris dan Spanyol [6]. PPDB dikembangkan dengan *bilingual pivoting technique* dengan memanfaatkan kesamaan arti atas satu frasa yang diungkapkan secara berbeda. Tahapan pengembangan PPDB:

1. Mengekstraksi leksikal, frase dan sintak dari korpus multi bahasa dengan mempertimbangkan peluang kemunculan.

2. Menentukan nilai *similarity* setiap pasangan menggunakan Google n-*grams* dan *Annotated Gigaword Corpus.*

### PPDB2.0

PPDB dikembangkan menjadi versi PPDB2.0 dengan menambahkan hasil evaluasi manusia, relasi *entailment*, *similarity* berdasarkan *word embedding*, dan nilai perubahan gaya bahasa (*style score complaxcity* dan *formality)* [6]. PPDB berisi kurang lebih 100 juta pasangan parafrasa yang digunakan untuk beberapa bahasa populer dan dapat dipakai dengan berapa variasi ukuran. Pernambahan yang dilakukan pada PPDB2.0:

1. Merangking ulang pasangan parafrasa secara regresi menggunakan model yang dianotasi manual oleh ahli bahasa. Setiap unit dari parafrasa memiliki kemungkinan berbagai macam pasangan. Setiap pasangan memiliki kecenderungan dekat dan jauh, sehingga perlu adanya rangking.

2. Setiap pasangan parafrasa pada PPDB2.0 memiliki relasi *entailment* yaitu *hyponym* (□), *hypernym* (□), *non-entailing topical relatedness* (□), unrelatedness (#), dan *contradiction* (¬).

Setiap pasangan parafrasa memiliki nilai perubahan gaya bahasa (*styling score*) yang disesuaikan tergantung apikasi dari penempatan pasangan tersebut. Kegunakan dari *Syling score* dapat dipakai pada *Natural Language Generation* dalam memilih padanan yang sesuai. Nilai menunjukan dari yang paling komplek sampai yang paling sederhana. Standar penilaian ini telah dievaluasi pada penelitian yang lain. Setiap pasangan parafrasa memiliki nilai kedekatan *word embedding* menggunakan *Multi View Latent Semantik Analysis (MVLSA). MVLSA dipilih karena mampu mengukur bobot pada data yang muncul tidak tentu seperti pada dokumen multi bahasa atau dokumen embedding[7].*

Keunggulan dari korpus PPDB karena adanya banyak tambahan fitur yang dapat disesuaikan pada aplikasinya. Semakin besar data yang dipakai dapat meningkatkan *recall* namun menurunkan *precision*, begitu juga sebaliknya. Karena itu diperlukan teknik yang tepat dalam penggunakan korpus PPDB.

**MSRP**

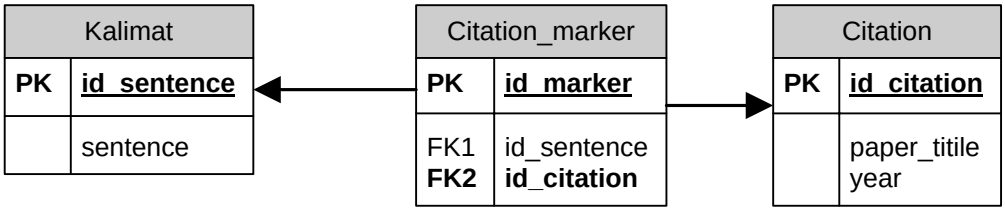
*Microsoft Research Paraphrase – Corpus* adalah sumber daya bahasa berisi pasangan kalimat yang secara semantik mimiliki informasi yang sama [8]. Korpus ini dibentuk dengan cara mengumpulkan 49.000 pasangan kalimat yang dinilai berdasarkan data latih menggunakan algoritma SVM. Data latih yang dipakai 10.000 (2968 positif, 7032 negatif) pasangan kalimat parafrasa hasil clustering dan dinilai oleh pakar. Dari hasi klasifikasi didapatkan 5801 pasangan kalimat parafrasa positif. Setelah dilakukan evaluasi ulang oleh pakar, dinyatakan hanya 3900 (67%) mengandung informasi yang sama. Dataset ini digunakan sebagai *gold* standar uji kualitas sistem identifikasi parafrasa ACL.

**PIT**

*Paraphrase In Twitter* adalah sumberdaya bahasa berisi pasangan kalimat yang diambil dari media sosial twitter [9]. Teknik yang digunakan untuk mengumpulkan parafrasa adalah dengan asumsi bahwa dua kalimat dinyatakan parafrasa jika ada dalam satu trending topik yang sama dan setidaknya memiliki 1 kata lain yang juga sama. Kesaan antar kalimat tidak dilihat dari term yang sama, namun juga dengan mempertimbangakan berbagai fitur antara lain POS Features, Topical Features, String Features. Dari hasil pengumpulan data didapatkan 18 ribu pasangan kalimat dimana 35% parafrasa dan 65% sisanya tidak. Dataset ini kemudian digunakan dalam kompetisi pembangunan mesin pendeteksi *semantic similarity* pada semeval2015.

Tabel 1. Perbandingan Korpus Parafrasa

| No | Peneliti (tahun) | Nama | Domain | Teknik | Fitur |
|---|---|---|---|---|---|
| 1 | Ganitkevitch dkk (2013) | PPDB | Bebas | Pivoting | Distribusi nilai |
| 2 | Pavlick dkk (2013) | PPDB 2.0 | Bebas | Pivoting | Rangking, Style Score, Relasi Entailment, WordEmbedding |
| 3 | Dolan dkk (2005) | MSRP | Berita | SVM | Yes or No |
| 4 | Xu dkk (2014) | PIT | Twitter | Multi Instance Lerrning | Yes or No, Bobot Similarity |




Gambar 1. Relasi antara Kalimat dengan Sitasi

## 3. METODE PENELITIAN

Pada penelitian ini sumber utama makalah ilmiah didapat dari ACL Anthology. Setelah melalui proses unduh didapat kurang lebih 40.000 makalah ilmiah yang terkait dengan bahasan *Computational Linguistic.* Makalah yang dikumpulkan bercampur antara prosiding dan jurnal.

Dalam prosesnya, tidak semua makalah dapat diekstraksi untuk dikumpulkan setiap kalimatnya.

Setiap makalah dalam format file pdf diekstraksi menjadi database relational yang berisikan identitas makalah, kumpulan kalimat isi, kumpulan sitasi, abstrak dan nama bagian makalah. Alat bantu yang digunakan dalam mengekstrasi file adalah dr. inventor framework. Dengan alat bantu ini maka kalimat-kalimat dapat dibagi menjadi tiga jenis yaitu kalimat biasa, kalimat abstrak dan kalimat sitasi. Khusus pada kalimat sitasi, maka setiap memiliki hubungan antara kalimat sitasi juga makalah sitasinya.

Kalimat-kalimat dari makalah ilmiah yang sudah dikumpulkan lalu kemudian dicari pasangan parafrasanya. Sampai dengan tulisan ini dibuat, sekenario yang dikerjakan adalah pada kalimat-kalimat sitasi untuk makalah yang sama. Kalimat yang mensitasi makalah yang sama diukur satu sama lain dengan menggunakan pendekatan *instance base similarity.*

Pada proses ini terdapat beberapa skenario penilaian dengan indikator teknik pemecahan kalimat, teknik pembobotan dan algoritma pengukuran similarity.

1. Representasi : n-gram, bigram, trigram

2. Pembobotan : tf-idf, binary, Tf

3. Algoritma: Cosine Similarity, Euclidean Distance, Jaccard Similarity

Setiap kalimat akan dicari 3 kalimat lain yang memiliki kedekatan tertinggi. Dari hasil tersebut tidak lantas dianggap sebagai pasangan parafrasa. Keluaran yang dihasilkan membantu bagi penliti dalam mengumpulkan pasangan parafrasa. Proses evaluasi skenario dilakukan dengan cara memiliki kombinasi mana yang paling membantu bagi penliti dalam mengumpulkan pasangan kalimat parafrasa.

## 4. HASIL DAN PEMBAHASAN

Satu buah makalah membutuhkan waktu ekstraksi kurang lebih 100 detik. Dalam satu hari komputer dijalankan untuk mengekstrasi 500 makalah. Sampai laporan ini dibuat, telah diekstrasi ≈12.500 makalah dari 40.000 makalah yang ada. Dari jumlah tersebut ≈8.000 makalah tidak dapat diekstraksi sehingga yang sudah diekstrasi sekitar ≈4.500 makalah. Dari angka tersebut terkumpul kalimat bagian isi sebanyak ≈590.402 dan abstrak ≈23.584.

Tidak semua kalimat adalah kalimat sitasi. Kalimat sitasi yang terkumpul ≈85.009 dengan jumlah yang beragam untuk setiap makalah yang disitasi. Sampai saat ini jumlah terbanyak satu makalah yang disitasi adalah 96 kalimat. Jumlah paling sedikit satu makalah disitasi oleh 1 kalimat.

Keluaran dari proses pencocokan adalah 1 cluster kalimat yang berisi 4 anggota. Dalam satu cluster tidak semua bernilai parafrasa. Bisa jadi seluruh kalimat pada satu cluster bersifat parafrasa (3). Namun bisa juga dalam satu cluster, keseluruhan kalimat tidak parafrasa.

(1) *a. We evaluate in truecase with BLEU and TER*

*b. Evaluation sets are translated using the cdec decoder and evaluated with the BLEU metric.*

*c. We use BLEU scores as the performance measure in our evaluation .*

*d. To evaluate the resulting translations, we use BLEU and NIST*

(2) *a. We pre-initialize the word embeddings by running the word2vec tool on the English Wikipedia dump and the jacana corpus as in*

*b. The word embedding was pretrained with the skip-gram model of word2vec using the dumped English Wikipedia data and the documents of the target insurance domain.*

*c. Word embeddings can be randomly initialized or pre-trained vectors, e.g. word2vec or dependency-based embeddings.*

*d. For word embeddings, we use an in-house Java re-implementation of word2vec to build 300- dimensional vector representations for all types that occur at least 10 times in our unannotated corpus.*

Dilakukan beberapa strategi untuk mengevaluasi korpus. Prose evaluasi dilakukan untuk mendapatkan konfigurasi terbaik. Dengan konfigurasi terbaik maka selanjutnya proses pengumpulan parafrasa dilakukan secara otomatis.

Tahapan strategi penilaian yang dilakukan terhadap skenario pengumpulan korpus:

1. Melakukan analisa dari makalah yang disitasi terbanyak
2. Melakukan analisa yang mewakili satu makalah yang memiliki tema khusus yang satu makalah yang memiliki tema umum
3. Melakukan skenario pengujian berdasarkan kombinasi representasi, pembobotan dan algoritma

Dari hasil analisa oleh tim peneliti, setelah melihat keluaran yang dihasikan dari proses *instance base similarity* adalah sebagai berikut:

1. Representasi :

   *n-gram* : menghasilkan nilai kedekatan yang cukup tinggi namun tidak memperhatikan urutan dari kata-kata pada kalimat

   *bigram*: lebih memperhatikan urutan kata pada selang 2 kata untuk setiap atribut. Menjadi representasi terbaik dari hasil analisis *trigram*: lebih kuat dalam memperhatikan urutan kata selang 3 kata untuk setiap atribut, namun sulit untuk menghasilkan similariti.

2. Pembobotan

   *Tf-idf*: mampu mereduksi kalimat-kalimat yang sering muncul sehingga dapat lebih menekankan pada kata-kata kunci. Menjadi teknik pembobotan paling tepat dari hasil analisis

   *Tf:* tidak memberikan tekatan kepenteingan sebaran kata pada setiap kalimat.

   *Binary*: merepresentasikan kata berdasarkan hadir atau tidaknya dalam kalimat. Mudah untuk dikomputasi namun tidak cukup bagus jiga digunakan untuk mengukur kedekatan kalimat.

3. Algoritma

   *Cosine Similarity*: tidak sensitif dengan panjang kalimat dan urutan dari representasi kalimat. Panjang kalimat pada setiap cluster beragam.

   *Euclidean Distance*: sensitif dengan panjang kalimat namun tidak terhadap representasi kalimat. Panjang kalimat pada setiap cluster beragam

   *Jaccard Similarity*: mampu mengukur kedekatan berdasarkan irisan kalimat sehingga setiap cluster relative memiliki panjang kalimat yang sama. Namun tidak sensitif terhadap urutan kalimat.

Dari proses analisis yang dilakukan oleh tim peneliti, maka skenario terbaik yang dapat membantu mengumpulkan korpus berpasangan adalah representasi *bigram*, pembobotan *tf-idf* dan algoritma *Jaccard Similarity*.

Sejauh ini telah terevaluasi 407 kalimat dari 6 target sitasi kalimat. Jumlah ini masih akan terus bertambah sampai waktu penelitian selesai dalam satu semester. Proses otomatisasi tidak benar-benar diterima sebagai metode untuk menghasilkan korpus kalimat karena tim peneliti ingin mendapatkan korpus yang paling tepat secara semantik.

Proses ekstraksi kalimat telah menghasilkan basis data yang cukup besar. Sekenario abstrak, silang antar bagian makalah dan pencarian kalimat defisnis belum dapat dilakukan karena proses evaluasi manual yang lama.

## 5. KESIMPULAN DAN RENCANA

Kesimpulan yang didapat dari proses pengumpulan korpus kalimat berpasangan parafrasa adalah:

1. Dari hasil proses ekstraksi makalah menjadi database didapatkan kalimat isi sebanyak ≈590.402 dan kalimat abstrak sebanyak ≈23.584.

2. Jumlah terbanyak satu makalah yang disitasi adalah 96 kalimat. Jumlah paling sedikit satu makalah disitasi oleh 1 kalimat. Jumlah target makalah yang disitasi sebanyak ≈23.584.

3. Skenario terbaik yang paling membantu dalam mengumpulkan korpus adalah representasi *bigram*, pembobotan *tf-idf* dan algoritma *Jaccard Similarity*.

4. Proses pengumpulan pasangan kalimat yang dapat dilakukan baru dari kalimat sitasi. Skenario antar bagian, abstrak ke isi dan pengumpulan kalimat definisi belum dapat dilaksanakan.

Rencana kegiatan penelitian selanjutnya:

1. Melanjutkan proses pengumpulan korpus bepasangan secara manual dengan bantuan skenario terbaik.

2. Mencoba algoritma yang dapat mengevaluasi kedekatan kalimat berdasarkan urutan representasi gram dalam kalimat.

3. Mecoba sekenario lain dan mengevaluasi secara kuantitatif setiap sekenario yang ada.

## DAFTAR PUSTAKA

[1] C. M. L. Lisa, "Merging Corpus Linguistics and Collaborative Knowledge," *English*, no. September, 2009.

[2] A. Barrom-Cedeno, M. Vila, dan A. Marti, "Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection," *Assoc. Comput. Linguist.*, no. November 2012, 2013.

[3] R. Kittredge, "Paraphrasing for condensation in journal abstracting," *J. Biomed. Inform.*, vol. 35, no. 4, hal. 265–277, 2002.

[4] L. Shi, "Rewriting and paraphrasing source texts in second language writing," *J. Second Lang. Writ.*, vol. 21, no. 2, hal. 134–148, 2012.

[5] S. Teufel, "Do ' Future Work ' sections have a purpose ? Citation links and entailment for global scientometric questions," in *Proceedings of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries*, 2017.

[6] E. Pavlick, P. Rastogi, J. Ganitkevitch, B. Van Durme, dan C. Callison-Burch, "PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification," *Proc. 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process. (Short Pap. Beijing, China, July 26-31, 2015*, hal. 425–430, 2015.

[7] K. Filippova, M. Mieskes, dan V. Nastase, "Cascaded Filtering for Topic-Driven Multi-Document Summarization," *Proc. Doc. Underst. Conf.*, hal. 30–35, 2007.

[8] W. B. Dolan dan C. Brockett, "Automatically Constructing a Corpus of Sentential Paraphrases," in *Proceedings of the Third International Workshop on Paraphrasing*, 2005, hal. 9–16.

[9] W. Xu, A. Ritter, C. Callison-burch, W. B. Dolan, dan Y. Ji, "Extracting Lexically Divergent Paraphrases from Twitter," *Trans. Assoc. Comput. Linguist. 2 (NAACL 2014)*, vol. 2, hal. 435–448, 2014.