LAPORAN KEMAJUAN II PENELITIAN DISERTASI

***RECOGNISING TEXTUAL ENTAILMENT***

Oleh

**YUDI WIBISONO**

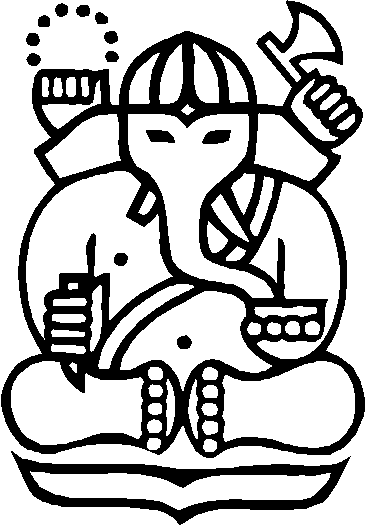
**NIM 33211005**

Promotor:

**Dr. Ir. Dwi H. Widyantoro**

Co Promotor:

**Dr. Nur Ulfa Maulidevi, ST., M.Sc**



**PROGRAM PASCASARJANA**

**INSTITUT TEKNOLOGI BANDUNG**

**2014**

Daftar Isi

[2 Pendahuluan 3](#_Toc406470114)

[3 Hasil semester sebelumnya 6](#_Toc406470115)

[4 Hasil semester ini 9](#_Toc406470116)

[4.1 Recognising Textual Entailment (RTE) 10](#_Toc406470117)

[4.1.1 Penelitian Terkait RTE 10](#_Toc406470118)

[4.1.2 Eksperimen 13](#_Toc406470119)

[A. Preprocessing 14](#_Toc406470120)

[B. Ekstraksi Kalimat 15](#_Toc406470121)

[C. Ekstraksi komponen kalimat: subjek, predikat, and objek 17](#_Toc406470122)

[D. Ekstraksi Fitur dan Classifier 17](#_Toc406470123)

[4.1.3 Hasil Eksperimen 18](#_Toc406470124)

[5 Publikasi 20](#_Toc406470125)

[6 Daftar Pustaka 20](#_Toc406470126)

# Pendahuluan

*Recognising textual entailment* (RTE) adalah suatu task dalam pemrosesan teks untuk mengenali apakah suatu fragmen teks dapat diinferensi dari fragment teks yang lain berdasarkan pengetahuan umum. Contoh berikut adalah dua teks yang memiliki hubungan T1 entails T2, sehingga jika T1 benar maka T2 juga benar:

T1: Most of Grameen Bank's borrowers use the small loans to start businesses by, for instance, buying a cow to produce milk, or a rickshaw, or materials to make cloth or pottery.

T2: Grameen Bank's loans are used to buy a cow.

Pada contoh tersebut walaupun T1 entails T2, tetapi belum tentu sebaliknya. Ini disebabkan entailment dapat bersifat satu arah. Ini perbedaan utama antara entailment dengan task pengenalan parapharase.

RTE dapat digunakan dalam aplikasi pemrosesan bahasa alami seperti perangkuman teks, information retrieval (IR), question and answering (QA), dan machine translation Dalam perangkuman teks, RTE dapat digunakan untuk membuang kalimat yang redundan dan dapat meningkatkan kinerja perangkuman sampai dengan 6.78% (Lloret, 2008). Sedangkan dalam aplikasi QA RTE dapat digunakan untuk memilih kandidat jawaban. Harabagiu (2006), mengurutkan kandidat jawaban berdasarkan probabilitas entailment-nya dengan pertanyaan. Metode ini meningkatkan kinerja sebesar 20%. Magnini (2014) menggunakan entailment graph untuki menganalisis interaksi customer seperti pada email dan media sosial.

# Penelitian Terkait RTE

Penelitian mengenai RTE mulai berkembang setelah dimulainya PASCAL RTE Challenge pada tahun 2005. Tujuan dari RTE Challenge adalah menyediakan dataset yang dapat dijadikan benchmark untuk task RTE. Dalam tantangan ini, disediakan pasangan teks yang disebut Hypothesis (H) dan Text (H), dan peserta diminta untuk memprediksi apakah H dapat diturunkan dari T (T entails H). Kinerja sistem dilihat berdasarkan akurasi prediksi kelas entailment.

Pendekatan untuk RTE dapat dibagi menjadi empat kategori:

1. Pemrosesan bahasa alami seperti *part of speech tagger*, *dependency parser*, *coreference*, dan *syntatic alignment*.
2. Pendekatan berdasarkan statistik menggunakan machine learning. Fitur yang digunakan untuk klasifikasi dapat berupa *shallow features* seperti frekuensi kemunculan kata sampai dengan fitur-fitur yang diperoleh dari pemrosean bahasa alami seperti dependency tree.
3. Menggunakan representasi pengetahuan dan penentuan entailment dilakukan menggunakan inferensi.
4. Kombinasi dari ketiga metode diatas.

Untuk menangani masalah variasi bahasa, dapat digunakan sumber eksternal seperti Wordnet, DIRT, Wikipedia, VerbNet dan Framenet.Although external resources could increase accuracy, it needs more processing power and time.

Untuk memudahkan membandingkan kinerja setiap pendekatan, Tabel 1 memperlihatkan analisis hasil peserta RTE Challenge ke-3. Secara umum, akurasi masih rendah (rata-rata 0.6) dengan akurasi tertinggi mencapai 0.8000 tetapi membutukan tahapan yang panjang. Hal ini menunjukkan task RTE masih dapat dikaji lebih dalam lagi dan dapat memberikan kontribusi baru.

Tabel : Analisis hasil RTE-3 Challenge

| **Penulis** | **Teknik** | ***Machine Learning Model*** | **External Resources** | **Akurasi** |
| --- | --- | --- | --- | --- |
| Burchardt | Analisis linguistik (LFG, Framenet). LFG diproyeksikan ke framenet dalam bentuk graph. Kecocokan graph (node dan edge), disebut semantic overlap. | Fitur: semantic overlap (total 47 fitur) | Wordnet, SUMO | 0.6250 |
| Burchardt | Word overlap antara T dan H. Kata yang digunakan hanya kata benda, kata kerja dan kata sifat, kata keterangan. | Satu fitur: word overlap. Algoritma: J48 | - | 0.6262 |
| Bobrow | Pemetaan dependency structure ke konsep yang berisi Agent, Theme, and Goal. Penentuan entailment dilakukan dengan alignment konsep tersebut. Dilakukan normalisasi sintaks (pasif ke aktif), penanganan negasi, aspek temporal. | - | Wordnet, Verbnet | 0.5112 |
| Tatu | Named entity, coreference, semantic parsing, normalisasi ekspresi waktu. Bangkitkan variasi lexical chain. Konversi T-H menjadi logical representation form, penanganan negasi. Penentuan dilakukan dengan pembuktian. | - | Extended Wordnet | 0.7225 |
| Iftene | Dependency tree, named entitiy. Tree edit distance antar dependency tree T dan H. Transformasi tree hypothesis menggunakan ekstrenal resources. Hitung nilai fitness (lokal dan global), jika melewati threshold maka dianggap entail. | - | Wordnet, Dirt, Wikipedia, Acronim database | 0.6913 |
| Bar-Haim | Dependency tree. Generate berbagai variasi tree dari Text menggunakan entailment rule, hitung cost function antara tree tersebut dengan H. Entailment rule mencakup: syntatic, polarity, negation, lexical, syntatic rule. | - | - | 0.6112 |
| Harmeling | Mirip dengan Bar-Haim, tetapi menggunakan wordnet. Parameter diperoleh dengan learning | - | Wordnet | 0.5775 |
| Li | Dependency tree, word sense disambiguation, lexical semantic similarity (LSS), named entitiy, dependent content word pair, negasi. Fitur paling penting adalah LSS dan named entitity. | J48 | Wordnet | 0.6488 |
| Hickl | Dependency tree, named entitiy, coreference, normalisasi temporal dan spasial. Heuristik digunakan untuk mengekstrak *discourse commitment* (DC) dari T dan H. DC adalah proposisi yang bernilai true untuk T dan H. DC terbaik dipilih menggunakan weighted *bipartite matching algorithm* (allignment). Selanjutnya entailment ditentukan dengan decision tree classifier antara pasangan-pasangan DC dari T dan DC dari H. Jika dinyatakan entail, proses dilanjutkan dengan pemeriksaan kontradiksi. Heuristik yang digunakan untuk mengestrak DC: sentence segmentation, dekomposisi, supplemental expression, relation extraction, coreference resolution. | Decision tree | Wordnet | 0.8000 |

### Eksperimen

Untuk eksperimen, data diambil dari RTE Challenge ke-3. Ekstraksi kalimat dan ekstraksi subyek predikat objek (SPO) digunakan untuk menentukan entailment antara T dan H. Untuk ekstraksi kalimat sendiri, digunakan ekstraksi anak kalimat, kalimat dalam frase preposisi, dan transformasi dari kalimat pasif ke aktif.

Dataset pada RTE3 diambil dari web based system. Untuk menghindari masalah copyright, sebagian data diambil dari dataset kompetisi di bidang information retrieval dan summarization. Sebagian lagi diambil dari Wikinews atau Wikipedia (Giampiccolo 2007).

Jumlah record pada RTE3 adalah 1600 pasangan Text (T) dan Hypothesis (H), yang terdiri atas 800 pasang untuk pengembangan atau pelatihan dan 800 pasang untuk test. Dari setiap 800 pasangan tersebut, terbagi atas 400 pasang contoh entails positif (T entails H) dan 400 pasang entails negatif (T not entails H). Ukuran evaluasi adalah akurasi dari dataset test.

Tabel berikut adalah beberapa contoh dari pasangan T-H dari RTE 3

Tabel Contoh pasangan pada dataset RTE3

|  |  |  |
| --- | --- | --- |
| **Text (T)** | **Hypothesis (H)** | **Entailment** |
| A U.S. soldier accused of participating in the rape of an Iraqi girl and then killing her and her family was sentenced to 90 years in jail on Thursday, media reports said. | A U.S. soldier was sentenced to 90 years in jail. | TRUE |
| While the House has been extremely active in passing bills addressing the key priorities of entrepreneurs, most of the legislation has hit a brick wall over in the Senate. | 80% approve of Mr. Bush. | FALSE |
| "For Japanese yakuza, the most important thing is staying alive, and making money is second," the yakuza said. | Yakuza are the Japanese mafia. | FALSE |
| Typhoon Xangsane lashed the Philippine capital on Thursday, grounding flights, halting vessels and closing schools and markets after triggering fatal flash floods in the centre of the country. | A typhoon batters the Philippines. | TRUE |
| The royal tombs at Vergina were identified as the ancient capital of Aigai in the 1930s. A royal tomb identified as possibly Philip II of Macedonia, Alexander the Great's father, was found in 1977 by Manolis Andronikos. | Alexander the Great's Tomb remains undiscovered. | FALSE |
| Zhao, who died at 85, was a reformer who was removed from office, and imprisoned, in 1989 after denouncing the use of force against student protesters in Tiananmen Square. | Zhao was arrested in 1989. | TRUE |

Gambar 1 memperlihatkan arsistektur sistem yang digunakan untuk RTE. Pada langkah pertama dilakukan praproses dengan Stanford Parser (Klein, 2003) untuk menghasilkan struktur sintaks berbentuk pohon sintaks. Kemudian ekstraksi kalimat diaplikasikan pada T dan H, sehingga T dan H akan menghasilkan beberapa sub kalimat. Kemudian dilakukan ekstraksi subyek-predikat-objek (SPO) untuk setiap sub kalimat tersebut. Fitur lalu diekstrak dari SPO sub kalimat. Berikut akan dibahas secara lebih rinci setiap bagian.

Preprocessing

Ekstraksi Kalimat

Subject, Predicate, Object Extraction

Features Extraction

Classifier

Text

Hypothesis

**Gambar 1 : Arsitektur Sistem**

## Preprocessing

Pada tahap praproses, dihasilkan *part of speech* tag dari T dan H. Simbol seperti koma juga dihapus. Tabel berikut memperlihatkan daftar sebagian part of speech tag (Buchholz,2002)

|  |  |  |
| --- | --- | --- |
| **Tag** | **Deskripsi** | **Contoh** |
| NP | Noun Phrase | the strange bird |
| DT | Determiner | the, a, these |
| NN | Noun, singular or mass | tiger, chair, laughter |
| VP | Verb phrase | was looking |
| VBZ | |  |  | | --- | --- | |  |  |   Verb, 3rd person singular present | she **thinks** |
| VBP | verb, non-3rd person singular present | I **think** |
| VBG | verb, gerund or present participle | **thinking** is fun |
| VBD | verb, past tense | they **thought** |
| VBN | verb, past participle | a **sunken** ship |
| NNPS | Noun, proper plural | we met two **Christmases** ago |
| SBAR | Subordinating conjunction | **whether** or not |
| PP | prepositional phrase | in between, at home |
| S | Subordinating sentence |  |
| MD | verb, modal auxillary | may, should |

Contoh hasil dari part of speech parser adalah sebagai berikut:

“A typhoon batters the Philippines.”

(ROOT

(S

(NP

(DT **A**)

(NN **typhoon**)

)

(VP (VBZ **batters**)

(NP

(DT **the**)

(NNPS **Philippines**)

)

)

)

)

## Ekstraksi Kalimat

Ekstraksi kalimat diperlukan karena hypothesis (H) dapat cocok hanya pada sebagian text (T). Sebagai contoh:

T: Ebola hemorrhagic fever is a fatal disease caused by a new virus which has no known cure. When Ebola Epidemic breaks out in Zaire in the spring of 1995, it was widely perceived as a threat to the West. Public attention was intense.

H: Ebola Epidemic breaks out in Zaire.

Pada contoh tersebut hypothesis “Ebola Epidemic breaks out in Zaire” diturunkan dari sebagian kalimat kedua pada text yaitu “When **Ebola Epidemic breaks out in Zaire**, it was widely …”

Dengan memanfaatkan POS tag dan pohon sintaks, ekstraksi kalimat dilakukan sebagai berikut:

1. Deteksi kalimat, dengan mengambil kalimat-kalimat dari teks yang dipisahkan oleh titik.
2. Ekstraksi klausa, dengan mendeteksi tag “SBAR” di part of speech tag. SBAR adalah tag yang menyatakan kata sambung yang menghubungkan antara dua klausa yang tidak sederajat. Contoh kata SBAR adalah: after, before, because, when.
3. Ekstraksi kalimat dalam frase preposisi, dengan mendeteksi tag “S” yang menyatakan subkalimat atau “VP” (verb phrase) yang berada di dalam di dalam “PP” (prepositional phrase).
4. Buat kalimat aktif dari kalimat pasif.

Dengan menggunakan langkah-langkah tersebut, berikut contoh ekstraksi kalimat dari sebuah teks.

“Ebola hemorrhagic fever is a fatal disease caused by a new virus which has no known cure. When a new epidemic was detected in Zaire in the spring of 1995, it was widely perceived as a threat to the West. Public attention was intense.”

Akan menghasilkan sub kalimat seperti pada Tabel 3

Tabel Hasil Ekstraksi Kalimat

| **Kalimat** | **Langkah** |
| --- | --- |
| Ebola hemorrhagic fever is a fatal disease caused by a new virus which has no known cure. | Sentence detection |
| When a new epidemic was detected in Zaire in the spring of 1995, it was widely perceived as a threat to the West | Sentence detection |
| Public attention was intense | Sentence detection |
| a new epidemic was detected in Zaire in the spring of 1995 | Subordinate clause |
| a new virus which has no known cure | Sentence in prepositional phrase |
| [\*] detected in Zaire in the spring of 1995 | Passive to active sentence |

## Ekstraksi komponen kalimat: subjek, predikat, and objek

Setiap kalimat yang merupakan hasil tahapan ekstraksi kalimat diproses kembali untuk mendapatkan subjek, predikat, dan objek dengan mengaplikasikan aturan heuristik dari pohon sintaks. Tabel 4 memperlihatkan contoh hasil ekstraksi SPO.

Contoh hasil ekstraksi subjek, predikat, dan objek

Tabel Contoh ekstraksi Subyek-Predikat-Objek

|  |  |  |  |
| --- | --- | --- | --- |
| **Kalimat** | **Subjek** | **Predikat** | **Objek** |
| A unique feature of previous Ebola outbreaks has been the relative sparing of children. | A unique feature of previous Ebola outbreaks | has been | sparing of children |
| The number of the confirmed Ebola cases has risen slightly to 26 in Gabon and to 16 in Congo Brazzaville | The number of the confirmed Ebola cases | has risen | 26 in Gabon and to 16 in Congo Brazzaville |

Subyek diekstraksi dengan memilih frase noun dengan level tertinggi pada pohon sintaks. Predikat diekstraksi dari frase verb pertama setelah subjek, sedangkan objek diekstraksi dari frase noun level tertinggi setelah predikat.

## Ekstraksi Fitur dan Classifier

Fitur leksikal digunakan dengan pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF). Kata dengan TF-IDF tinggi memiliki hubungan yang kuat dengan kalimatnya.

Setelah ekstraksi kalimat, setiap T dan H dapat ditransformasi menjadi beberapa kalimat. *Cosine similarity* digunakan untuk menghitung jarak antara bagian kalimat (subyek, predikat, objek) dari semua kalimat H ke semua kalimat T. Nilai *cosine similarity* antara T dan H diasumsikan terkait dengan hubungan entailment-nya. Selain itu, *cosine similarity* digunakan juga untuk menghitung jarak antara T dan H tanpa menggunakan bagian SPO. Tabel 5 menunjukkan fitur lengkapnya.

Feature-set

Tabel Fitur

| **Feature** | **Deskripsi** |
| --- | --- |
| best\_t\_h\_spo | Nilai cosine similarity terbaik antara semua subkalimat T dan H dengan menggunakan ektraksi subject – predikat – objek (SPO). Atau  best\_t\_h\_spo = max (cosine\_sim(subt,subh), cosine\_sim(predt,predh), cosine\_sim(objt,objh))  ;    ;  Dimana  t adalah subkalimat yang dihasilkan dari Teks (T)  h adalah subkalimat yang dihasilkan dari Hypothesis (H)  subt : adalah subject yang diektrak dari sub kalimat t  Predt : adalah predikat yang diektrak dari sub kalimat t  Objt : adalah objek yang diektrak dari sub kalimat t |
| best\_t\_h | Nilai *cosine similarity* terbaik antara subkalimat tanpa SPO |
| avg\_t\_h\_spo | Nilai rata-rata cosine similarity menggunakan SPO |
| t\_h | *Cosine similarity* antara H dan T langsung (tanpa menggunakan ekstraksi kalimat) |

Setelah semua fitur diekstrak, dibangun model atau classifier berbasis pembelajaran mesin untuk mengestimasi fungsi yang dapat menentukan apakah T entails H.

### Hasil Eksperimen

Menggunakan semua fitur pada Tabel 5 akurasi yang diperoleh adalah 0.62, tetapi jika menggunakan hanya fitur best\_t\_h and t\_h yang tidak menggunakan ekstraksi subyek predikat, objek (SPO), akurasi naik menjadi 0.63 Ini menunjukkan ekstraksi kalimat dapat digunakan tetapi ekstraksi SPO tidak memiliki peran signifikan.

Sebagai contoh, pasangan T-H:

T: They cite scholarly research showing that young women at co-ed schools tend to defer to men in intellectual debate.

H: Young women at co-ed colleges tend to defer to men.

T diekstrak menjadi dua subkalimat:

T1: They cite scholarly research showing that young women at co-ed schools tend to defer to men in intellectual debate.

T2: young women at co-ed schools tend to defer to men in intellectual debate

sedangkan H diekstrak menjadi satu kalimat:

H1: Young women at co-ed colleges tend to defer to men.

Dapat dilihat pasangan T2 dan H1 memiliki subyek dan predikat yang sama dan objek yang mirip, tetapi setelah rata-rata kesamaan SPO hanya 0.32, ini mungkin disebabkan jumlah kata yang jauh lebih sedikit pada subyek, predikat dan objek dibandingkan kalimat lengkap. Permasalahan yang lain adalah jumlah data yang sedikit (800 pasang), mempersulit machine learning sulit untuk mendapatkan model yang tepat.

Dibandingkan dengan metode lain yang sama-sama tidak menggunakan resources eksternal, akurasi sistem ini lebih baik akurasinya daripada (Malakasiotis, 2007) dan (Marsi, 2007) walaupun masih lebih rendah daripada (Li 2007).

# Publikasi

Yudi Wibisono, Dwi Widyantoro, Nur Ulfa Maulidevi, “Sentence Extraction in Recognition Textual Entailment Task”, IcoDSE, 2014.

Yudi Wibisono, Dwi Widyantoro, Nur Ulfa Maulidevi, “Rancangan Sistem Pembangkit Anotasi untuk Kredibilitas dan Reliabilitas Informasi dalam Jejaring Sosial Online”, KNSI, 2014.

# Daftar Pustaka

Burchardt, Aljoscha, et al. "A semantic approach to textual entailment: System evaluation and task analysis." Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Association for Computational Linguistics, 2007.

Bobrow, Daniel G., et al. "Precision-focused textual inference." Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Association for Computational Linguistics, 2007.

Tatu, Marta, and Dan Moldovan. "Cogex at RTE3." Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Association for Computational Linguistics, 2007.

Iftene, Adrian, and Alexandra Balahur-Dobrescu. "Hypothesis transformation and semantic variability rules used in recognizing textual entailment." Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Association for Computational Linguistics, 2007.

Bar-Haim, Roy, et al. "Semantic inference at the lexical-syntactic level for textual entailment recognition." Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Association for Computational Linguistics, 2007.

Harmeling, Stefan. "An extensible probabilistic transformation-based approach to the third recognizing textual entailment challenge." Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Association for Computational Linguistics, 2007.

Li, Baoli, et al. "Machine learning based semantic inference: Experiments and Observations at RTE-3." Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Association for Computational Linguistics, 2007.

Lloret, Elena, et al. "A Text Summarization Approach under the Influence of Textual Entailment." NLPCS. 2008.

Harabagiu, Sanda, and Andrew Hickl. "Methods for using textual entailment in open-domain question answering." Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006.

Magnini, Bernardo, et al. "Entailment Graphs for Text Analytics in the Excitement Project." Text, Speech and Dialogue. Springer International Publishing, 2014.

Giampiccolo, Danilo, et al. "The third pascal recognizing textual entailment challenge." *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*. Association for Computational Linguistics, 2007.

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430

Malakasiotis, Prodromos, and Ion Androutsopoulos. "Learning textual entailment using SVMs and string similarity measures." Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Association for Computational Linguistics, 2007.

Marsi, Erwin, Emiel Krahmer, and Wauter Bosma. "Dependency-based paraphrasing for recognizing textual entailment." Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Association for Computational Linguistics, 2007

Buchholz, S. (2002). *Memory-Based Grammatical Relation Finding*. ILK, Tilburg Universit