PENGEMBANGAN PIRANTI PENELITIAN SISTEM TEMU KEMBALI INFORMASI BAHASA INDONESIA

Faisal Rahutomo[1], Erfan Rohadi[2]
Program Studi Teknik Informatika, Jurusan Teknologi Informasi, Politeknik Negeri Malang
Jl. Soekarno Hatta 9, Malang, 65141
Telp : (0341) 404424-404425, Fax : (0341) 404420
E-mail : [1]faisal.polinema@gmail.com, [2]fan_da@yahoo.com

Abstrak

Piranti penelitian sistem temu kembali informasi (STKI) Bahasa Indonesia belum tersedia secara terbuka. Piranti ini terdiri atas data leksikal bahasa Indonesia, data uji performa, metrik ukur, dan kumpulan metoda pembanding. Meskipun telah banyak dilakukan penelitian STKI bahasa Indonesia, hasilnya tidak bisa diperbandingkan satu sama lain. Peneliti lain tidak bisa membandingkan hasil usulan mereka karena ketiadaan sumberdaya ini. Makalah ini memaparkan sebuah proyek yang sedang berjalan untuk mengembangan piranti penelitian sistem temu kembali informasi bahasa Indonesia. Hasil dari penelitian ini diharapkan mengembangkan kualitas dan kuantitas penelitian STKI bahasa Indonesia. Hasil penelitian ini juga diharapkan dapat memicu tumbuh kembangnya penelitian STKI bahasa daerah, mengingat Bahasa Indonesia adalah bahasa nasional dari berbagai suku dengan bahasa daerahnya masing-masing. Sumberdaya penelitian STKI Bahasa Indonesia dapat digunakan sebagai jembatan antara sumberdaya bahasa daerah satu dengan lainnya. Ketersediaan sumberdaya komputasi Bahasa Indonesia yang memadai secara tidak langsung mendukung kelestarian budaya dan bahasa bangsa Indonesia.

Kata kunci: *sistem temu kembali informasi, bahasa Indonesia, piranti penelitian.*

Abstract

Research platform for Indonesian-based information retrieval system is not openly yet. The platform contains Indonesian lexical data, test collections, performance metrics, and methods implementations. Researchers can not compare their results with the other proposals because of this gap, eventhough many proposals are available in this area of research. This paper exposes an ongoing project of building Indonesian-based information retrieval research platform. The aim of this project is to strengthen researches of Indonesian-based information retrieval. Further target of this project is to promote information retrieval reserches of Indonesian local language; due to the fact, Indonesian is a national language of a country consists of many tribes with their own local languages. Based on the logic, Indonesian can serve as a language bridge. The availability of a mature Indonesian-based research platform will support Indonesian language and culture preservation.

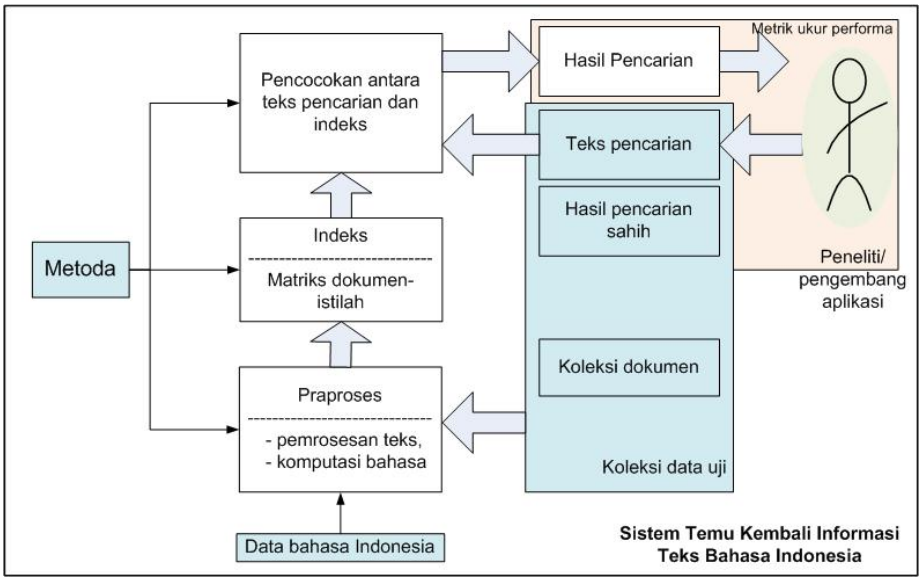
Kata kunci: *information retrieval system, Indonesian, research platform.*

1. PENDAHULUAN

Piranti pendukung penelitian sistem temu kembali informasi (STKI) Bahasa Indonesia belum tersedia secara terbuka. Piranti pendukung penelitian ini terdiri atas koleksi data bahasa, data uji performa, metoda pembanding, dan metrik ukur performa [1] sebagaimana terpapar di dalam Gambar 1. Piranti ini penting untuk digunakan sebagai standar pembanding antara satu usulan metoda dengan usulan metoda yang lain. Peneliti yang mengajukan sebuah usulan metoda yang baru hanya dapat membandingkan usulannya dengan usulan yang lain bila sistemnya diuji di dalam lingkungan data uji dan metrik ukur yang sama. Suatu usulan metoda yang baru tidak bisa disebut lebih baik bila tidak diperbandingkan dengan lingkungan kerja yang sama.

Piranti penelitian serupa telah disediakan oleh komunitas sistem temu kembali internasional dalam berbagai bahasa, terutama bahasa Inggris [2][3][4][5][6][7]. Sayangnya piranti serupa belum tersedia secara terbuka untuk sistem berbahasa Indonesia. Fakta tersebut terjadi meskipun penelitian STKI bahasa Indonesia telah banyak dilakukan semacam [8][9][10][11][12]. Merujuk ke Gambar 1, bagan yang berwarna biru muda belum tersedia secara terbuka. Uji ulang penelitian-penelitian yang telah dilakukan di bidang ini oleh peneliti lain tidak dimungkinkan kecuali oleh kelompok peneliti tersebut sendiri. Pembandingan performa usulan metoda-metoda

baru dengan metoda yang telah diusulkan sebelumnya sulit dilakukan karena tidak adanya pembanding yang jelas.



Gambar 1. Diagram Penelitian STKI

Pengembangan piranti pendukung penelitian STKI berbahasa Indonesia sangatlah penting dengan beberapa sebab. Pertama, karakteristik bahasa Indonesia tidak sama persis dengan karakteristik bahasa lain yang telah tersedia piranti penelitiannya. Ketersediaan piranti penelitian STKI Bahasa Indonesia diharapkan akan menumbuh kembangkan penelitian di bidang ini. Penelitian-penelitian yang baik akan mengembangkan aplikasi-aplikasi terkait baik aplikasi komersial maupun nonkomersial. Kedua, Bahasa Indonesia adalah bahasa nasional dan bahasa persatuan Negara Kesatuan Republik Indonesia (NKRI). Sedangkan NKRI sejatinya terdiri atas berbagai suku bangsa dengan bahasa suku atau daerahnya masing-masing. Warga dengan suku yang berbeda dapat berkomunikasi satu dengan lainnya dengan menggunakan bahasa Indonesia. Logika komputasi yang setara juga dapat digunakan. Komputasi antara satu bahasa daerah dengan bahasa daerah yang lain dimungkinkan apabila sumberdaya Bahasa Indonesia tersedia secara memadai. Dengan tersedianya piranti STKI berbahasa Indonesia diharapkan dapat memicu berkembangnya piranti STKI bahasa-bahasa daerah. Secara tidak langsung diharapkan memicu pelestarian keanekaragaman budaya dan bahasa di NKRI.

2. PLATFORM PENELITIAN STKI

Piranti penelitian STKI terdiri atas data leksikal bahasa Indonesia, data uji performa, metrik ukur, dan kumpulan metoda pembanding. Sebelum masuk ke dalam materi yang diusulkan, makalah ini akan membahas piranti penelitian STKI yang telah tersedia di dalam bahasa Inggris dan bahasa Indonesia. Kemudian lubang yang ingin diisi di dalam makalah ini ditunjukkan di bagian akhir bab ini.

2.1 Sumberdaya yang Tersedia

STKI adalah area penelitian yang membawa manfaat besar di era internet saat ini. Katakanlah Google, Yahoo, Bing, adalah mesin pencari papan atas yang memudahkan pengguna internet mencari konten yang terkait dengan ketertarikan mereka. Tumpukan informasi teks dan non teks yang terus bertambah secara cepat tiap harinya pasti membutuhkan sebuah mesin pencari untuk pelayanan yang terbaik bagi pengguna informasi.

Piranti yang diperlukan untuk mendukung penelitian STKI adalah data koleksi, metrik ukur performa, metoda-metoda pembanding, dan sumberdaya komputasi bahasa. Komunitas peneliti STKI internasional telah saling menyediakan sumberdaya ini secara terbuka baik yang gratis maupun berbayar.

Terdapat banyak Komunitas STKI internasional. Tiap-tiap komunitas menyediakan koleksi data uji masing-masing untuk kasus yang berbeda-beda. Koleksi data ini juga seringkali disebut sebagai *corpus*. Komunitas tersebut antara lain:
- CLEF (*cross-language evaluation forum*),
- DUC (*document understanding conferences*),
- FIRE (*forum for information retrieval evaluation*),
- INEX (*initiative for the evaluation of XML retrieval*),

- MIREX (*music information retrieval evaluation exchange*),
- ACM SIGIR (*ACM Special Interest Group on Information Retrieval*),
- NTCIR (*NII-NACSIS test collection for IR systems*), dan
- TREC (*text retrieval conference*)

Komunitas peneliti juga telah menyediakan metrik kinerja STKI. Metrik tersebut adalah *precision* dan *recall*. Untuk mengkombinasikan keduanya dipergunakan parameter *F-score* yang merupakan rata-rata harmonic *precision* dan *recall*. Tujuan utama mesin pencari adalah memaksimalkan *precision* dan *recall* ini dengan nilai maksimalnya adalah 1 [1].

Untuk metoda, komunitas peneliti internasional juga menyediakan piranti penelitian yang lengkap. Metoda praproses *tokenizing*, *stemming*, dan pembobotan tersedia dengan berbagai macam implementasinya. Contohnya algoritma Porter stemmer tersedia implementasinya secara gratis [13]. Proses *tokenizing* juga tidaklah sulit diimplementasikan. Pembobotan TF dan IDF terdapat pula implementasinya secara terbuka [14]. Pencocokan dokumen dan indeks dengan berbagai metrik semacam cosine dan jaccard juga tersedia [15]. Piranti penelitian latent semantic analysis (LSA) tersedia dengan berbagai implementasi [16]. Metoda explicit semantic analysis (ESA) yang menggunakan matriks dokumen-istilah Wikipedia juga tersedia [17].

Komunitas peneliti internasional juga menyediakan piranti bahasa yang lengkap secara terbuka. Salah satu piranti yang sangat berguna adalah WordNet [3]. WordNet adalah hasil proyek penelitian di Princeton University yang bertujuan untuk memodelkan pengetahuan leksikal pembicara asli bahasa inggris. Informasi di dalam WordNet diorganisasikan ke dalam kelompok logikal yang disebut *synset*. Tiap-tiap *synset* berisikan bentuk sinonim kata dan pointer semantik yang menjelaskan hubungan antara satu *synset* dengan *synset* lainnya [3].

2.2 Penelitian STKI di Indonesia

Penelitian STKI di Indonesia dipelopori oleh Universitas Indonesia (UI). Penyebutan institusi ini tidak mengecilkan hasil penelitian dari institusi lainnya. Sebagai kampus terkemuka di Indonesia, UI telah melakukan banyak penelitian di bidang ini yang menghasilkan berbagai publikasi ilmiah bereputasi [18]. Penelitian tersebut dilakukan oleh kelompok peneliti laboratorium perolehan informasi (http://ir.cs.ui.ac.id/). Penelitian yang dilakukan terkait dengan: *Computational Linguistics*, *Cross Language IR*, *Geographic IR*, *Image Retrieval*, *Music Retrieval*, *Question Answering*, dan *Summarization*. UI juga telah menyediakan beberapa tool dan layanan web (*webservice*) dalam bentuk *Application Program Interface* (API) untuk digunakan secara publik [19] dan [20].

Sayangnya, data-data dasar komputasi Bahasa Indonesia berupa daftar kata dasar, kata majemuk, kata jadian, kata tak penting untuk tahap penyaringan kata tidak tersedia secara terbuka. Padahal data-data dasar tersebut menentukan praproses sistem temu kembali informasi. Praproses yang berbeda akan memberikan hasil yang berbeda.

Di sisi yang lain potongan program dalam bentuk API yang disediakan juga tidak efisien digunakan dalam penelitian STKI. Pada umumnya penelitian di bidang ini melakukan komputasi ribuan hingga jutaan kali terkait fungsi API tertentu. Proses ini tidak efisien bila dilakukan dengan mengakses API *webservice* tersebut.

Peneliti berusaha mengikuti perkembangan penelitian STKI berbahasa Indonesia yang dilakukan dan meyakini, sumberdaya dan piranti penelitian yang dimaksud di dalam usulan ini tidak tersedia secara terbuka. Tidak ada lagi yang mengembangkan sumberdaya yang dimaksud hingga bulan April 2015. Bilamana peneliti lain ingin mengevaluasi ulang penelitian terbaru semacam [11] atau [12], mereka akan menghadapi jalan buntu. Rangkuman sumberdaya STKI Bahasa Indonesia yang tersedia tercantum di dalam Tabel 1. Berdasarkan akses internet bulan April 2015.

Tabel 1. Sumberdaya STKI Bahasa Indonesia yang Tersedia

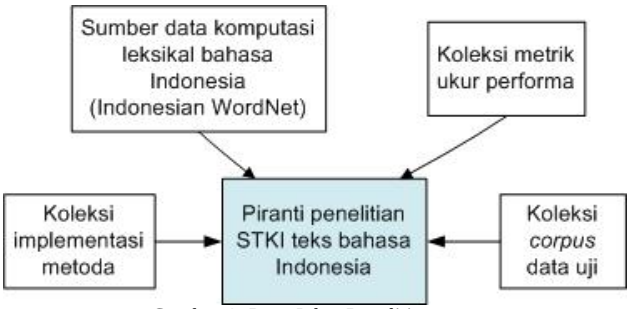
No.	Nama Proyek	Status
1	KBBI online	Database tidak tersedia
2	Wordnet Indonesia	Database tertutup
3	Kateglo (kamus thesaurus dan glosary bahasa Indonesia)	Database tertutup, API terbatas
4	Stop word list	Validitas tidak jelas
5	Stopword list	Tersedia tetapi perlu divalidasi ulang
6	*Corpus* Indonesia	Web tidak aktif
7	SEALANG	Database tertutup

2.3 Usulan

Penelitian ini ingin menjembatani komunitas peneliti di Indonesia untuk mengembangkan piranti penelitian STKI secara terbuka baik berbayar atau gratis. Penelitian ini ingin menyediakan piranti-piranti untuk penelitian STKI Bahasa Indonesia sebagaimana yang telah dilakukan komunitas peneliti internasional. Piranti-piranti yang ingin dibangun di dalam penelitian ini adalah data leksikal komputasi bahasa Indonesia, koleksi data uji standar, koleksi implementasi metoda, dan koleksi metrik ukur performa STKI.

3. PENGEMBANGAN PLATFORM STKI BAHASA INDONESIA

Untuk membangun piranti yang dimaksud, diperlukan beberapa bagian: Koleksi data leksikal komputasi bahasa Indonesia, koleksi data uji standar STKI bahasa Indonesia, koleksi implementasi metoda, dan koleksi implementasi metrik ukur performa. Peta jalan yang dituju oleh usulan penelitian ini diilustrasikan oleh Gambar 2. Piranti penelitian STKI Bahasa Indonesia ini berupa perangkat lunak yang terkandung di dalamnya empat item koleksi yang dimaksud. Setelan rinci perangkat lunak ini untuk memilah data uji yang diperlukan, metoda yang digunakan, dan metrik yang ingin dipakai. Sumber data leksikal Bahasa Indonesia berperan penting di dalam tahap praproses dan pengindeksan data uji. Peneliti yang mengajukan usulan baru dapat mengubah-ubah setiap detail yang diperlukan, sesuai kebutuhan masing-masing. Piranti ini juga memungkinkan penambahan item dan uji coba dari item dimaksud. Sumber-sumber data dan koleksi yang ingin dihasilkan oleh penelitian ini adalah artifak yang berharga bagi dunia STKI bahasa Indonesia. Pengembang aplikasi STKI juga dapat menggunakan bagian-bagian dari piranti ini untuk aplikasi berbayar atau gratis. Penjelasan mengenai pengembangan platform ini ada di sub bab selanjutnya. Karena keterbatasan tempat, penjelasan detail yang lain akan dipaparkan pada publikasi penulis selanjutnya.



Gambar 2. Peta Jalan Penelitian

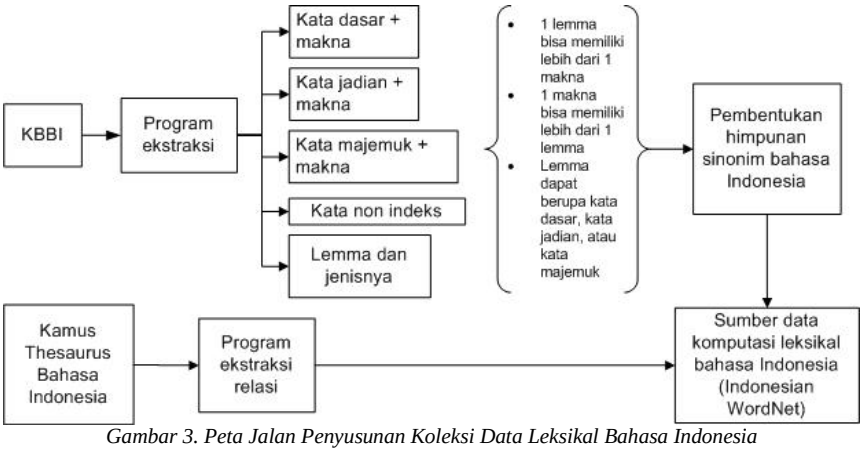
3.1 Koleksi Data Leksikal Bahasa Indonesia

Koleksi data leksikal ini bermanfaat untuk tahap praproses STKI dan pencocokan teks pencarian dengan indeks. Tahapan ini memproses teks-teks ke dalam sebuah indeks STKI. Data yang diperlukan pada tahapan ini adalah data bahasa Indonesia berupa daftar kata dasar, kata majemuk, kata jadian, dan kata yang tidak perlu diindeks. Tahapan-tahapan dalam praproses STKI yang memerlukan data leksikal komputasi Bahasa Indonesia adalah: *tokenizing*, *stemming*, dan *filtering*. Pembobotan istilah juga dapat menggunakan relasi-relasi semantik antara himpunan sinonim yang ada.

Beberapa metoda pencocokan semantik [21] juga memerlukan sumberdaya leksikal bahasa semacam ini. Hubungan antara teks dapat dicari hingga relasi maknanya, bukan hanya relasi bentuk teksnya. Penelitian kemiripan semantik antara teks Bahasa Indonesia mungkin untuk ditelaah lebih lanjut dan diperbandingkan satu dengan lainnya bila sumberdayanya tersedia.

Berbeda dengan penelitian yang telah dilakukan sebelumnya [22][9] usulan penelitian ini mengajukan proses ekstraksi dari versi elektronik kamus besar bahasa Indonesia (KBBI) dan kamus thesaurus bahasa Indonesia. Diagram peta jalan penelitian ini diilustrasikan oleh Gambar 3. Banyak hal yang dapat diperoleh dari proses ekstraksi ini: daftar kata dasar, kata jadian, kata majemuk, dan kata non indeks. Pada tahapan selanjutnya, dapat disusun daftar lemma bahasa Indonesia. Daftar ini bermanfaat untuk menguak fakta bahwa sebuah lemma dapat memiliki banyak makna. Begitupun sebaliknya, sebuah makna bisa diwakili oleh lebih dari satu lemma.

Ekstraksi tersebut memungkinkan peneliti untuk menyusun himpunan sinonim bahasa Indonesia, sebuah komponen dasar terbentuknya jejaring makna layaknya WordNet Indonesia. Relasi antara makna tersebut dapat diekstraksi dari kamus thesaurus bahasa Indonesia. Mengingat luasnya buku KBBI dan kamus thesaurus bahasa Indonesia, seluruh proses ekstraksi dilakukan dengan perangkat lunak khusus yang dibangun untuk menyelesaikan proyek ini.



Gambar 3. Peta Jalan Penyusunan Koleksi Data Leksikal Bahasa Indonesia

3.2 Koleksi Data Uji Performa

Koleksi data uji digunakan untuk menguji performa sebuah metoda. Koleksi data uji STKI memiliki karakteristik yang unik. Koleksi ini terdiri atas koleksi dokumen teks, koleksi teks pencarian, dan kumpulan penilaian manusia yang merelasikan teks pencarian dengan koleksi dokumen yang benar. Penelitian ini mengusulkan beberapa *corpus*: *corpus* berita, *corpus* rangkuman teks, *corpus* deskripsi video, dan *corpus* tanya jawab.

3.2.1 *Corpus* STKI Berita

Penyusunan data uji STKI berita dilakukan dengan melakukan akses berkelanjutan selama 6 bulan 6 portal berita online: Vivanews, Okezone, Kompas, Tempo, JPNN, dan detik. Ide penyusunan ini serupa dengan ide [23] yang sayangnya tidak tersedia secara terbuka. Hasil akses disimpan dalam bentuk halaman HTML yang telah dikelompokkan berdasarkan topik berita portal tersebut. Selama proses akses ini, ide-ide teks pencarian dapat digali berikut halaman relevan yang terkait dengan teks pencarian. Dengan demikian dapat terbentuk koleksi data uji berita yang terdiri atas: koleksi dokumen, koleksi teks pencarian, dan koleksi relasi relevan antara dokumen dan teks pencarian. Validasi koleksi data uji ini dilakukan oleh beberapa orang penilai.

3.2.2 *Corpus* STKI Rangkuman Teks

Data uji STKI rangkuman teks dapat diperoleh dari makalah-makalah ilmiah. Usulan penelitian ini merencanakan mengikuti dan membeli prosiding 10 seminar nasional dengan topik yang berbeda-beda sepanjang tahun. Dari kegiatan ini, dapat diperoleh ratusan makalah dari berbagai topik yang ada. Secara intuitif abstrak makalah atau kesimpulan dan saran makalah merupakan rangkuman dari makalah itu sendiri. Dengan demikian dapat diperoleh koleksi data uji STKI rangkuman.

3.2.3 *Corpus* STKI Semantik

Penyusunan data uji STKI semantik mengikuti langkah yang dilakukan oleh peneliti sebelumnya [2]. Sekelompok orang dipekerjakan untuk mengekspreksikan kata-kata yang tepat atas tayangan sebuah video singkat. Dengan proses ini, peneliti dapat memperoleh sekelompok teks yang mengacu pada makna yang sama meskipun diekspresikan dengan bentuk teks yang berbeda-beda.

3.2.4 *Corpus* STKI Tanya Jawab

Data uji STKI tanya jawab disusun dengan langkah yang serupa dengan data uji STKI semantik. Sebagaimana penelitian serupa dilakukan [24] dan [8]. Sayangnya peneliti [8] tidak menyediakan sumberdaya penelitiannya yang berbahasa Indonesia sebagaimana [24] menyediakan dalam bahasa Inggris. Penyusunan *corpus* ini dilakukan oleh sekelompok orang yang cukup terdidik yang dipekerjakan untuk menjawab pertanyaan-pertanyaan paparan dengan topik pengetahuan umum. Dengan langkah ini dapat diperoleh kumpulan teks yang relevan antara satu pertanyaan dengan beberapa teks jawaban. Dikarenakan proses yang rumit dari penyusunan koleksi data uji yang diusulkan di dalam penelitian ini, peneliti mengembangkan perangkat lunak yang sesuai untuk tiap-tiap koleksi yang dibangun.

3.3 Koleksi Implementasi Metoda

Koleksi implementasi metoda yang dimaksud dapat berupa metoda tahapan praproses, metoda pencocokan dokumen dengan teks pencarian, ataupun metoda STKI dalam mengolah indeks. Koleksi yang telah tersedia secara terbuka dapat dikumpulkan. Metoda-metoda yang belum ada implementasinya programnya dan hanya tersedia berupa ide tertulis dalam makalah-makalah ilmiah perlu dikembangkan implementasinya.

3.4 Koleksi Implementasi Metrik

Koleksi implementasi metrik ukur performa. Koleksi metrik ukur performa ini telah disediakan komunitas peneliti STKI internasional. Implementasinya di dalam piranti pemrograman perlu dibangun di dalam penelitian ini.

4. KESIMPULAN DAN SARAN

Makalah ini telah memaparkan sebuah proyek yang sedang berjalan. Proyek tersebut bertujuan membangun sebuah piranti penelitian STKI berbahasa Indonesia. Hasil penelitian ini direncanakan dapat diakses secara terbuka. Hasil yang ingin dicapai adalah bergairahnya penelitian STKI berbahasa Indonesia dengan validitas yang tinggi dan dapat dipertanggungjawabkan. Piranti yang ingin dibangun terdiri atas koleksi data leksikal bahasa Indonesia, koleksi data uji, koleksi metrik ukur, dan koleksi metoda pembanding. Makalah ini telah memaparkan peta jalan pengembangan piranti tersebut. Publikasi selanjutnya berupa laporan tahapan pengembangan yang dicapai di dalam proyek ini.

5. DAFTAR RUJUKAN

[1] Yates, R.B. dan Neto, B.R., 1999, *Modern Information Retrieval*, Addison Wesley Longman Limited, New York.

[2] Chen, D.L. dan Dolan, W.B., 2011, Collecting highly parallel data for paraphrase evaluation, In: *the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

[3] Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K., 1993, Introduction to WordNet: An On-line Lexical Database.

[4] Porter, M.F., 1980, An algorithm for suffix stripping, *Program*, 14 no. 3, pp 130-137.

[5] Quirk, C., Brockett, C., dan Dolan, W., 2004, Monolingual machine translation for paraphrase generation, In: *the 2004 Conference on Empirical Methods in Natural Language Processing*.

[6] Salton, G., Buckley, C., 1988, Term-weighting Approaches in Automatic Text Retrieval, *Information Processing & Management*, Vol.24, No.5, pp.513-523.

[7] Salton, G. , Wong, A., Yang C.S., 1975, A Vector Space Model for Automatic Indexing, *Communication of the ACM*, Vol.18, No.11, pp.613-620.

[8] Larasati, S.D. dan Manurung, R., 2007, Towards a semantic analysis of bahasa indonesia for question answering, In: *the 10th Conference of the Pacific Association for Computational Linguistics (PACLING 2007)*.

[9] Margaretha, E., Franky, dan Manurung, R., 2008, English-to-Indonesian Lexical Mapping using Latent Semantic Analysis, In: *the 2nd International MALINDO Workshop, Cyberjaya, Malaysia*.

[10] Sari, S., Manurung, R., dan Adriani, M. (2010), *Indonesian WordNet Sense Disambiguation using Cosine Similarity and Singular Value Decomposition*.

[11] Martadinata, P., Distiawan, B., Manurung, R., Adriani, M., 2015, Building Indonesian Local Language Detection Tools using Wikipedia Data, In: *2nd International Workshop on Worldwide Language Service Infrastructure, Kyoto, Japan*.

[12] Wicaksono, Farizki, A., Vania, C., Distiawan, B., Adriani, M., 2014, Automatically Building a Corpus for Sentiment Analysis on Indonesian Tweets, In: *28th Pacific Asia Conference on Language Information and Computing*. Phuket, Thailand.

[13] http://nlp.stanford.edu/index.shtml

[14] http://php-nlp-tools.com/

[15] http://php-nlp-tools.com/

[16] http://lsa.colorado.edu/

[17] http://www.cs.technion.ac.il/~gabr/resources/ code/esa/esa.html

[18] Adriani, M., dan Manurung, R, 2008, A survey of bahasa Indonesia NLP research conducted at the University of Indonesia, In: *the 2nd International MALINDO Workshop*.

[19] http://bahasa.cs.ui.ac.id/resources_id.php

[20] http://bahasa.cs.ui.ac.id/webapps_id. php

[21] Yang, L., Bhavsar, V.C., Boley, H., 2008, On Semantic Concept Similarity Methods, In: *International Conference On Information & Communication Technology And System*. Surabaya, Indonesia.

[22] Putra, D.D., Arfan, A., dan Manurung, R., 2008, Building an Indonesian WordNet, In: *the 2nd International MALINDO Workshop*.
[23] Manurung, R., Distiawan, B., Putra, D.D., 2010, Developing an Online Indonesian Corpora Repository, In: *the 24th Pacific Asia Conference on Language, Information and Computation*.
[24] Mohler, M., Bunescu, R., dan Mihalcea, R., 2011, Learning to grade short answer questions using semantic similarity measures and dependency graph alignments, In: *the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.