

PENGARUH PART-OF-SPEECH PADA MESIN PENERJEMAH BAHASA INGGRIS-INDONESIA BERBASIS FACTORED TRANSLATION MODEL

Herry Sujaini¹, Kuspriyanto², Arry Akhmad Arman³, Ayu Purwarianti⁴

^{1,2,3,4}Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung

Jl. Ganesha no. 10 Bandung 40132

E-mail: herry_sujaini@yahoo.com, kuspriyanto@yahoo.com, arry.arman@yahoo.com, ayu@stei.itb.ac.id

ABSTRAKS

Tata Bahasa Indonesia memiliki banyak perbedaan dibandingkan dengan Bahasa Inggris. Fakta ini berakibat buruk pada kinerja Mesin Penerjemah Statistik (MPS) Inggris-Indonesia yang berbasis frase. Model MPS memiliki keterbatasan pada pemetaan frase atau blok dari sumber ke bahasa target tanpa penggunaan informasi linguistik. Menambahkan fitur linguistik, seperti part-of-speech (PoS) dapat berdampak pada kualitas terjemahan. Dalam tulisan ini, penggunaan penandaan PoS dimasukkan sebagai fitur linguistik dalam model penerjemah faktor (MPF) menggunakan sistem MPS Moses dan menggunakan BLEU sebagai alat evaluasi. Dari hasil penelitian, penggunaan PoS memiliki dampak terhadap meningkatnya kualitas terjemahan untuk bahasa Inggris-Indonesia, hal tersebut terlihat dari hasil eksperimen bahwa dengan menambahkan fitur PoS, akurasi mesin penerjemah meningkat sebesar 2%.

Kata Kunci: mps,mpf, pos, inggris-indonesia

1. PENDAHULUAN

1.1 Latar Belakang

Salah satu pendekatan Mesin Penerjemah (MP) adalah dengan menggunakan pendekatan statistik yang menggunakan konsep probabilitas yang biasa disebut Mesin Penerjemah Statistik (MPS). Setiap pasangan kalimat (Sc, Tg) akan diberikan sebuah $P(Tg|Sc)$ yang diinterpretasikan sebagai distribusi probabilitas dimana TM akan menghasilkan Tg dalam bahasa tujuan ketika diberikan Sc dalam bahasa sumber.

Translasi berbasis statistik yang berbasis frase (*phrase-based models*), terbatas pada pemetaan potongan teks tanpa adanya tambahan penggunaan informasi linguistik seperti morfologi, sintaksis, atau semantik. Informasi tambahan tersebut telah terbukti berharga dengan mengintegrasikannya dalam langkah-langkah pra-pengolahan atau pasca-pengolahan.

Beberapa penelitian terhadap beberapa bahasa telah memperlihatkan bahwa keakuratan TM semakin baik dengan tambahan fitur-fitur seperti *lemma*, *part-of-speech (PoS)*, *gender* dan lain-lain. Penelitian ini akan lebih difokuskan pada fitur PoS khususnya yang digunakan dalam Bahasa Indonesia sebagai bahasa target. Sedangkan untuk bahasa asalnya, penulis memilih bahasa Inggris sebagai bahasa Internasional dan memiliki sumber bahan penelitian (kopos paralel) yang lebih banyak tersedia dibandingkan dengan bahasa-bahasa lainnya. Untuk melihat pengaruh PoS terhadap hasil terjemahan, penelitian ini menggunakan model penerjemah faktor (MPF).

1.2 Permasalahan

PoS sebagai salah satu fitur pada mesin penerjemah dapat meningkatkan keakuratan hasil

terjemahan berbagai bahasa seperti yang telah dipaparkan oleh Koehn dkk. (2007), Razavian dkk. (2010), dan Youssef dkk. (2009). Permasalahannya adalah seberapa besar pengaruh faktor PoS terhadap peningkatan keakuratan terjemahan bahasa Inggris ke bahasa Indonesia.

1.3 Tujuan

Dari permasalahan di atas, dapat diungkapkan bahwa tujuan penelitian ini adalah untuk mengetahui seberapa besar pengaruh faktor PoS terhadap peningkatan keakuratan terjemahan bahasa Inggris ke bahasa Indonesia, khususnya pada MPS.

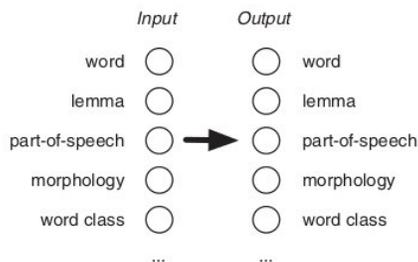
2. TINJAUAN PUSTAKA

2.1 Model Penerjemah Faktor (*Factored Translation Model*)

Penelitian awal untuk MPF telah dimulai oleh Lioma dan Ounis (2005), mereka memisahkan PoS dari korpus yang telah ditandai dengan PoS untuk meningkatkan sumber daya terjemahan pada MPS. Pendekatan ini telah terbukti berhasil untuk mengintegrasikan penandaan PoS sebagai faktor kelas kata (Shen dkk, 2006), CCG super tag (Birch dkk, 2007), dan penandaan morfologi untuk pemodelan bahasa (Badr dkk., 2008). Melengkapi ini, sisi sumber dapat diperkaya dengan markup tambahan, misalnya untuk lebih baik memprediksi infleksi yang tepat dalam bahasa keluaran yang kaya dengan morfologis (Avramidis dan Koehn, 2008). Model faktor yang lebih kompleks untuk menerjemahkan morfologi telah dieksplorasi untuk terjemahan Inggris-Ceko (Bojar, 2007)

Model-model penerjemahan faktor mengintegrasikan penandaan linguistik tambahan pada tingkat kata. Setiap jenis informasi tambahan pada tingkat kata disebut faktor. Gambar 1

mengilustrasikan tentang jenis informasi yang dapat berguna. Penerjemahan lemma dan faktor morfologi secara terpisah akan membantu dengan masalah data yang jarang dalam bahasa-bahasa yang kaya dengan morfologis. Informasi tambahan seperti sebagai *part-of-speech* mungkin dapat membantu dalam membuat keputusan pada *reordering* atau *grammatical coherence*. Kehadiran fitur morfologi pada sisi target memungkinkan untuk memeriksa aturan dalam frase kata benda atau antara subjek dan kata kerja (Koehn, 2010).



Gambar 1. Model Penerjemah Faktor

Seperti pada TM berbasis frase (*phrase based*), model penerjemahan faktor (*factored translation models*) dapat dilihat sebagai kombinasi dari beberapa komponen (*language model, reordering model, translation steps, generation steps*). Komponen-komponen ini mendefinisikan satu atau lebih fungsi fitur yang digabungkan dalam model *log-linier* : (Koehn, 2007)

$$p(e|f) = \frac{1}{Z} \exp \sum_{i=1}^n \lambda_i h_i(e, f) \quad (1)$$

Z adalah konstanta normal yang diabaikan pada tataran implementasi. Untuk menghitung probabilitas dari hasil terjemahan e dari kalimat input f, setiap fitur dari fungsi h_i dapat dievaluasi.

Untuk menghitung probabilitas translasi e diberikan kalimat input f, sebuah evaluasi dari setiap fitur fungsi h_i harus dilakukan dan kemudian dikalikan dengan bobot fitur λ_i . Misalnya, fungsi fitur untuk sebuah komponen *language model* bigram adalah :

$$h_m(e, f) = p_m e = p(e_1) p(e_2 | e_1) \dots p(e_m | e_{m-1}) \quad (2)$$

dimana m adalah jumlah kata e_i dalam kalimat e.

Mengingat fungsi fitur tersebut diperkenalkan oleh langkah-langkah terjemahan dan generasi model terjemahan faktor, terjemahan dari kalimat f ke kalimat e diuraikan menjadi satu set terjemahan frase (f_j, e_j).

Untuk komponen langkah terjemahan, masing-masing fungsi fitur h_t didefinisikan atas pasangan frase (f_j, e_j) diberi fungsi penilaian τ :

$$h_t(e|f) = \sum_j \tau(f_j, e_j) \quad (3)$$

Untuk komponen langkah generasi, masing-masing fungsi fitur h_g diberi fungsi penilaian γ yang didefinisikan sebagai output kata-kata e_k :

$$h_g(e|f) = \sum_j \gamma(e_k) \quad (4)$$

Fungsi fitur mengikuti dari fungsi penilaian (τ, γ) yang diperoleh selama pelatihan terjemahan dan tabel generasi. Fitur bobot λ_i dalam model log-linier ditentukan dengan metode *minimum error rate training*.

2.2 Part-of-Speech (PoS)

Kelas kata adalah penggolongan kata menurut bentuk, fungsi, dan maknanya. Kelas kata dalam Bahasa Indonesia terbagi atas Kata Kerja (Verba), Kata Sifat (Adjektiva), Kata Benda (Nomina), Kata Bilangan (Numeralia), Kata Ganti (Pronomina), Kata Keterangan (Adverbia), Kata Tunjuk (Demonstrativa), Kata Tanya (Interogativa), Kata Sandang (Artikula), Kata Depan (Preposisi), Kata Seru (Interjeksi), Kata Penghubung (Konjungsi), dan Kata Ulang (Reduplikasi) (Wahidah, 2008).

Setiap bahasa memiliki kelas kata yang berbeda-beda, kelas kata untuk bahasa Yunani telah didefinisikan oleh *Dionysius Thrax* pada tahun 100 SM yang terdiri atas delapan kelas kata, yaitu : *noun, verb, pronoun, preposition, adverb, conjunction, particle, dan article*.

Saat ini kelas kata untuk berbagai bahasa telah dikembangkan untuk keperluan komputerisasi, salah satunya adalah *Penn Treebank* oleh LINC Laboratory, Computer and Information Science, University of Pennsylvania (Marcus dkk., 1993)

Secara umum, satu set tag mengkodekan kedua fitur target klasifikasi, memberitahu pengguna informasi yang berguna tentang kelas gramatikal dari sebuah kata, dan fitur prediksi, encoding fitur yang akan berguna dalam memprediksi perilaku kata lain dalam konteks. Kedua tugas harus tumpang tindih, tetapi mereka tidak selalu identik (Manning, 1999).

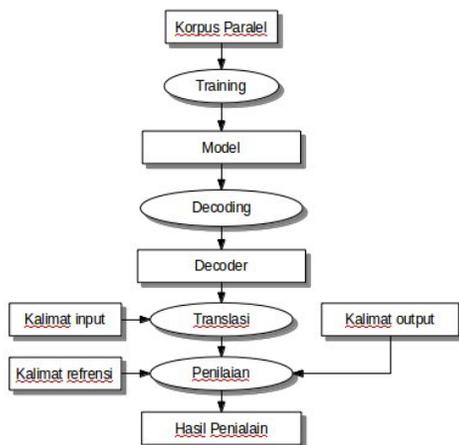
Berbagai set PoS telah digunakan dalam penelitian-penelitian bidang pemrosesan bahasa alami (PBA), diantaranya lewat PAN Localization Project, khusus untuk bahasa Indonesia telah dikembangkan PoS khusus untuk terjemahan ke bahasa Inggris pada tahun 2009 (Adriani, 2009). Kelas kata tersebut dibangun berdasarkan *Penn Treebank PoS tag sets* (Marcus dkk., 1993) terdiri dari 29 PoS tags. PoS utama Bahasa Indonesia adalah kata kerja (*verb*), kata sifat (*adjective*), kata keterangan (*adverb*), kata benda (*noun*), and kata tugas (*function words*), berdasarkan 5 (lima) PoS utama tersebut dan observasi data, Pisceldo dkk. (2009) mendefinisikan 37 tag untuk Bahasa Indonesia. Wicaksono dkk. (2010) dalam penelitiannya menggunakan 35 tag hasil dari modifikasi tagset yang dihasilkan oleh Adriani, (2009) dan Pisceldo dkk. (2009). Terakhir, Larasati

dkk. (2011) menggunakan hanya 19 tag dalam penelitiannya.

3. METODE PENELITIAN

3.1 Proses Penelitian

Proses penelitian dari persiapan sistem penerjemah sampai evaluasi terlihat pada gambar 2.

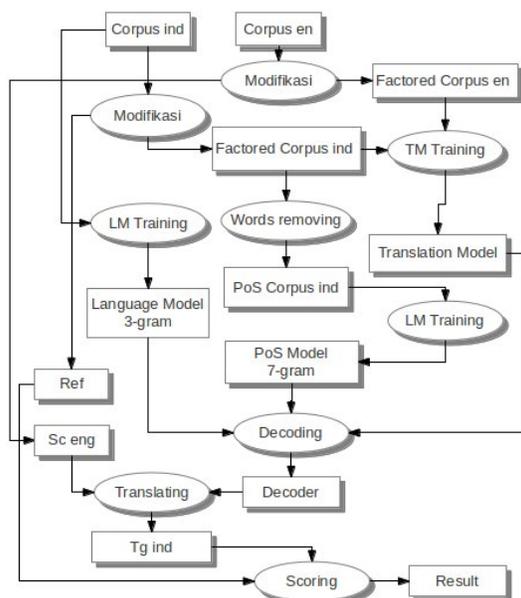


Gambar 2. Proses Penelitian

Tahap awal dari penelitian ini adalah pengolahan korpus, dimana korpus Identic dimodifikasi agar dapat digunakan pada penelitian ini. Selanjutnya dilakukan proses *training* untuk memperoleh model bahasa (*language model*), model PoS (*PoS model*), dan model translasi (*translation model*).

Untuk mengevaluasi sistem, dilakukan penilaian (*scoring*) dengan membandingkan hasil terjemahan terhadap referensi dengan menggunakan metode BLEU (Papineni, 2002).

Secara utuh, skema penelitian yang dilakukan dapat digambarkan seperti gambar 3.



Gambar 3. Detail Proses Penelitian

Korpus yang digunakan adalah korpus paralel, yaitu sepasang korpus yang berisi kalimat-kalimat dalam suatu bahasa dan terjemahannya. Pada penelitian korpus yang digunakan adalah korpus paralel bahasa Indonesia dan bahasa Inggris.

Proses modifikasi korpus dilakukan ditahap awal, untuk disesuaikan dengan format yang diperlukan pada proses *TM Training*. Kedua korpus yang telah dimodifikasi digunakan pada proses *TM training* untuk menghasilkan *translation model*. Selain itu, pada proses modifikasi dilakukan juga penghapusan faktor untuk mengambil kalimat asal bahasa Inggris yang digunakan sebagai input dalam proses penerjemahan dan kalimat asal bahasa Indonesia yang digunakan sebagai referensi dalam proses penilaian.

Proses penghapusan kata dilakukan terhadap korpus bahasa Indonesia untuk mengambil PoS nya saja. Hasilnya akan digunakan pada proses *LM training* dan menghasilkan *PoS model*. Korpus bahasa Indonesia yang belum ditandai akan dilakukan proses *LM training* dan akan menghasilkan *language model*. *Language model* pada penelitian ini menggunakan trigram, sedangkan *PoS model* menggunakan 7-gram

Translation model, *language model*, dan *PoS model* digunakan bersama-sama pada proses *decoding* untuk menghasilkan sebuah *decoder*. Selanjutnya *decoder* digunakan dalam proses *translating* dengan input korpus bahasa Inggris dan akan menghasilkan terjemahan dalam bahasa Indonesia.

Proses terakhir adalah proses *scoring* yang akan memberikan penilaian hasil terjemahan dengan membandingkannya terhadap korpus bahasa Indonesia.

3.2 Alat dan Bahan

Alat-alat (*tools*) yang digunakan pada penelitian ini adalah :

- Statistical Machine Translation : MOSES (Koehn, 2007)
- SRILM (Stolcke, 2011)
- GIZA++ (Och, 2003)
- BLEU (Papineni, 2002)

Bahan-bahan yang digunakan pada penelitian ini berupa korpus paralel yang diambil dari sumber korpus Identic (Larasati, 2012)

4. EKSPERIMEN

Pendekatan yang digunakan pada makalah ini adalah menggabungkan PoS sebagai faktor linguistik dalam model penerjemahan faktor seperti yang telah dijelaskan sebelumnya. Selain itu juga dilakukan eksperimen tanpa menggunakan PoS sebagai bahan perbandingan.

Eksperimen yang dilakukan menggunakan korpus paralel bahasa Indonesia-Inggris yang masing-masing terdiri atas 15.000 kalimat. Korpus identic yang digunakan pada eksperimen ini hanya

diambil informasi surface dan PoS untuk bahasa Indonesia serta surface untuk bahasa Inggris.

Sebagai contoh, pada korpus identic :

1 Ku aku aku<p>_PS1 PS1 p|P|S|1|aku|1|1
2 mencintai cinta meN+cinta<n>+i_VSA VSA
n|V|S|A|meN+cinta+i|1|1
3 mu kamu kamu<p>_PS2 PS3 p|P|S|3|kamu|1|1
4 . . .<z>_Z-- Z-- z|Z|-.|1|1
dimodifikasi menjadi :
Ku|P mencintai|V mu|P

5. HASIL EKSPERIMEN

Eksperimen yang dilakukan menggunakan 500 kalimat yang diambil secara random dari korpus, evaluasi dengan menggunakan metode BLEU terlihat pada tabel 1 berikut ini :

Tabel 1. Scoring dengan BLEU

Jumlah Kalimat	Hasil Scoring
1.000	BLEU = 31.63, 56.8/35.4/25.6/19.4 (BP=1.000, ratio=1.006, hyp_len=12794, ref_len=12712) BLEU = 31.61, 54.3/35.2/25.9/20.2 (BP=1.000, ratio=1.019, hyp_len=12949, ref_len=12712)
2.000	BLEU = 31.30, 56.2/35.3/25.3/19.1 (BP=1.000, ratio=1.010, hyp_len=12835, ref_len=12712) BLEU = 31.70, 54.6/35.3/26.0/20.1 (BP=1.000, ratio=1.009, hyp_len=12823, ref_len=12712)
5.000	BLEU = 31.86, 57.3/36.3/25.9/19.5 (BP=0.995, ratio=0.995, hyp_len=12652, ref_len=12712) BLEU = 31.56, 56.0/36.3/25.9/19.3 (BP=0.994, ratio=0.994, hyp_len=12631, ref_len=12712)
10.000	BLEU = 31.51, 57.8/36.2/25.6/18.7 (BP=0.996, ratio=0.996, hyp_len=12659, ref_len=12712) BLEU = 31.84, 57.3/36.9/26.6/20.0 (BP=0.977, ratio=0.977, hyp_len=12419, ref_len=12712)
12.000	BLEU = 32.21, 59.0/37.8/27.2/20.2 (BP=0.968, ratio=0.968, hyp_len=12310, ref_len=12712) BLEU = 32.72, 59.5/39.6/28.5/21.2 (BP=0.948, ratio=0.949, hyp_len=12064, ref_len=12712)
15.000	BLEU = 31.26, 58.7/37.1/26.2/19.3 (BP=0.966, ratio=0.966, hyp_len=12286, ref_len=12712) BLEU = 33.26, 59.9/40.3/29.2/21.9 (BP=0.943, ratio=0.945, hyp_len=12010, ref_len=12712)

Baris pertama adalah hasil penilaian hasil penerjemahan tanpa menggunakan PoS, sedangkan hasil kedua adalah penilaian hasil penerjemahan dengan menggunakan PoS.

Format hasil penilaian di atas adalah :

BLEU = FS, 1-gr/2-gr/3-gr/4-gr (BP, ratio, hyp_len, ref_len)

dimana :

FS = score akhir

1-gr = presisi 1-gram

2-gr = presisi 2-gram

3-gr = presisi 3-gram

4-gr = presisi 4-gram

BP (brevity penalti) = min (1, kata output / kata ref)

ratio = jumlah kata output/jumlah kata ref

hyp_len = jumlah kata hasil terjemahan

ref_len = jumlah kata referensi.

Beberapa contoh hasil terjemahan baik menggunakan PoS maupun tanpa PoS dapat dilihat pada tabel 2 berikut ini :

Tabel 2. Contoh Kalimat Hasil Terjemahan

No	Kalimat	Ket
1	as known widely, finance institution generally hardly touch this marginal sector	Input
	seperti diketahui bersama, lembaga keuangan pada umum nya jarang menyentuh sektor yang marginal ini	Ref
	seperti diketahui, lembaga keuangan secara umum hardly touch ini marginal sektor	Surf
	seperti diketahui bersama, lembaga keuangan pada umum nya jarang menyentuh sektor yang marginal ini	Surf +Pos
2	local residents said members of the hamas-led force had surrounded the home of the activist and gunfire ensued	Input
	para penduduk setempat mengatakan, para anggota pasukan pimpinan hamas telah mengepung rumah aktivis tersebut dan kemudian menyerangnya	Ref
	penduduk setempat mengatakan, anggota pasukan pimpinan hamas telah mengepung rumah pendahulu aktivis tersebut dan senapan ensued	Surf
	penduduk setempat mengatakan anggota pasukan pimpinan hamas telah mengepung rumah aktivis tersebut dan kemudian menyerangnya	Surf +Pos
3	china repudiated the report, spokesman qin gang said in a statement on the ministry 's website .	Input
	china mempermasalahkan laporan itu, kata jurubicara qin gang dalam suatu pernyataan di situs internet kementerian tersebut	Ref
	china mempermasalahkan laporan itu, kata jurubicara qin gang dalam sebuah pernyataan nya di situs internet kementerian	Surf
	china mempermasalahkan laporan itu, kata jurubicara qin gang dalam sebuah pernyataan nya di situs internet kementerian tersebut	Surf +Pos

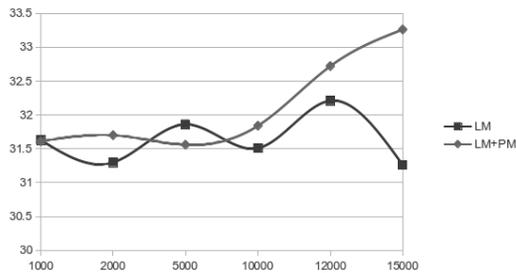
Input adalah kalimat dalam bahasa Inggris yang diterjemahkan, ref adalah kalimat referensi dalam bahasa Indonesia, Surf adalah hasil terjemahan tanpa menggunakan PoS, sedangkan Surf+PoS merupakan hasil terjemahan dengan menggunakan PoS.

Hasil eksperimen terhadap 15.000 kalimat, dengan menggunakan surface (tanpa PoS) dan Surface+PoS dapat dilihat seperti pada tabel 3 di bawah ini :

Tabel 3. Hasil Akhir Scoring dengan BLEU

Jumlah Kalimat	Surface (%)	Surface+PoS (%)
1000	31,63	31,61
2000	31,30	31,70
5000	31,86	31,56
10000	31,51	31,84
12000	32,21	32,72
15000	31,26	33,26

Grafik hasil eksperimen terlihat seperti gambar berikut :



Gambar 1. Perbandingan Hasil Scoring Surface VS Surface+PoS

Dari grafik di atas terlihat bahwa dengan menambahkan faktor PoS, sistem mesin penerjemah akan meningkat kualitasnya sejalan dengan jumlah kalimat yang digunakan pada korpus. Hasil eksperimen juga memperlihatkan bahwa dengan menggunakan korpus yang berisi 15.000 kalimat hasil scoring meningkat dari 31,26% menjadi 33,26%. Peningkatan ini cenderung semakin meningkat sejalan dengan besarnya korpus yang digunakan.

6. KESIMPULAN

Dari hasil penelitian yang dilakukan, dapat disimpulkan bahwa faktor PoS memberikan kontribusi terhadap peningkatan keakuratan terjemahan bahasa Inggris ke bahasa Indonesia sebesar 2% untuk korpus sebesar 15.000 kalimat.

Perlu dilakukan penelitian lebih jauh dengan korpus yang lebih besar dan menggunakan set PoS yang bervariasi untuk lebih mendalami pengaruh PoS terhadap peningkatan keakuratan terjemahan bahasa Inggris ke bahasa Indonesia.

PUSTAKA

- Adriani, Mirna (2009), "Developing Postag for Bahasa Indonesia", Diakses pada 1 April 2012 dari http://www.pan110n.net/Presentations/Laos/RegionalConference/CorpusCollection/Tagset_Tagging_for_Bahasa_Indonesia.pdf
- Avramidis, E. and Koehn, P. (2008). *Enriching Morphologically Poor Languages for Statistical Machine Translation*. In Proceedings of ACL-08: HLT, Columbus, Ohio
- Badr, I., Zbib, R., and Glass, J. (2008). *Segmentation for English-to-Arabic statistical machine translation*. Proceedings of ACL-08: HLT, Short Papers, Columbus, Ohio.
- Bojar, O. (2007). *English-To-Czech Factored Machine Translation*. In Proceedings of the Second Workshop on Statistical Machine Translation, Prague.

- Birch, A., Osborne, M., and Koehn, P. (2007). *CCG Supertags in Factored Statistical Machine Translation*. Proceedings of the Second Workshop on Statistical Machine Translation, Prague.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., dan Herbst, E. (2007). *Moses: Open Source Toolkit For Statistical Machine Translation*. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, Prague.
- Koehn, Phillip. (2010). *Statistical Machine Translation*. Cambridge University Press, New York.
- Koehn, Phillip dan Hieu Hoang. (2007). *Statistical Machine Translation*. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague.
- Larasati, Septina Dian. (2012). *IDENTIC Corpus: Morphologically Enriched Indonesian - English Parallel Corpus*, Diakses pada 1 April 2012 dari <http://ufal.mff.cuni.cz/~larasati/identic/documentation.html>
- Larasati, Septina Dian, Vladislav Kuboň, dan Daniel Zeman. (2011). *Indonesian Morphology Tool (MorphInd): Towards an Indonesian Corpus*. SFCM 2011. Springer CCIS proceedings of the Workshop on Systems and Frameworks for Computational Morphology, Zurich.
- Lioma, C. and Ounis, I. (2005). *Deploying Part-of-Speech Patterns to Enhance Statistical Phrase-Based Machine Translation Resources*. Proceedings of the ACL Workshop on Building and Using Parallel Texts, Ann Arbor, Michigan.
- Manning, Christopher D. and Hinrich Schütze. (1999). *Foundations of Statistical Natural Language Processing*, The MIT Press .
- Marcus, Mitchell P., Mary Ann Marcinkiewicz and Beatrice Santoroni. (1993). *Building a Large Annotated Corpus of English: The Penn Treebank*, Association for Computational Linguistics, 1993.
- Och, F. J. and Ney, H. (2003). *A Systematic Comparison Of Various Statistical Alignment Models*. Computational Linguistics, Volume 29, Number 1.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). *BLEU: A Method For Automatic Evaluation of Machine Translation*, In Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL).
- Pisceldo, Femphy, Mirna Adriani, dan Ruli Manurung. (2009). *Probabilistic Part-of-Speech Tagging for Bahasa Indonesia*, Third

- International Wokshop on Malay and Indonesian Language Engineering , Singapore.
- Razavian, Narges Sharif dan Stephan Vogel. (2010). *Fixed Length Word Suffix for Factored Statistical Machine Translation* , Proceedings of the ACL 2010 Conference Short Papers, Uppsala.
- Shen, W., Zens, R., Bertoldi, N., and Federico, M. (2006). *The JHU workshop 2006 IWSLT system*. Proceedings of the International Workshop on Spoken Language Translation, Kyoto.
- Stolcke, Andreas, Jing Zheng, Wen Wang, dan Victor Abrash, (2011), *SRILM at Sixteen: Update and Outlook*, Diakses pada 1 Mei 2012 dari <http://www.speech.sri.com/papers/asru2011-srilm.pdf>
- Waridah, Ernawati. (2008). *EYD dan Seputar Kebahasa-Indonesiaan*, Kawan Pustaka, Jakarta.
- Wicaksono, Alfian Farizki dan Ayu Purwarianti. (2010). *HMM Based Part-of-Speech Tagger for Bahasa Indonesia*, The 4th International Malindo Wokshop, Jakarta.
- Youssef, Islam, Mohamed Sakr dan Mohamed Kouta . (2009), *Linguistic Factors in Statistical Machine Translation Involving Arabic Language*, IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.11.