IDENTIFIKASI KESALAHAN TERHADAP HASIL TERJEMAHAN MESIN BAHASA INGGRIS KE BAHASA JAWA: KAJIAN SEMANTIK

Avi Meilawati dan Emi Nursanti
Universitas Negeri Yogyakarta
email: avimeilawati@gmail.com

Abstract

(Title: *Identification of Errors on English Engineering Translation Results to Java Language: Semantic Study*). This study uses linguistic principles to study the results of English to Javanese translations using an online translation engine. In particular, this study aims to identify translation errors and study them from a semantic perspective. Data is obtained by entering ten thousand samples of English sentences into the Google Translator translation engine which are then analyzed qualitatively. Analysis is focused on the level of words and phrases. The results of the study show four main problems in the semantic domain, namely the absence of lexicons, the inaccuracy of lexicons, word ambiguity, and idiom phrases. Suggestions were given to improve the quality of Javanese language SMT from a linguistic perspective, related to quantity, coverage, and quality of corporate data.

Keywords: semantics, machine translator, Javanese, English, linguistics

PENDAHULUAN

Machine Translation (MT) saat ini telah banyak memberikan manfaat praktis dan keilmuan bagi masyarakat luas. Salah satu jenis MT, yaitu Statistical Machine Translation (SMT), yang paling banyak digunakan secara luas saat ini adalah *Google Translator* (GT). GT mengembangkan database yang sangat besar dengan cara mengkodifikasi dan menganalisis jutaan kalimat. Sistem SMT yang dibangun menggunakan data monolingual corpora dan bilingual corpora. Berdasarkan *parralel corpora* ini sistem bekerja sedemikian rupa sehingga dapat memberikan hasil terjemahan sesuai dengan pasangan kalimat yang paling mendekati (Bird & Chiang,2012).

Saat ini 200 juta orang menggunakan GT setiap hari dan turut berperan menyempurnakan hasil terjemahan. Mereka dapat mamberi feedback dan meng-edit hasil terjemahan dalam bahasa target dan hasil edit-an ini akan tersimpan dalam database GT. GT telah mengakomodasi banyak bahasa di dunia, salah satunya bahasa Jawa (BJ) ke bahasa Inggris (BI) dan sebaliknya. Penelitian ini merupakan penelitian awal yang mengkaji secara khusus aspek semantik dari terjemahan teks BI ke dalam BJ. Dalam hal ini BI sebagai *Source Language* (SL) dan BJ sebagai *Target Language* (TL).

Sejauh ini belum ada penelitian yang secara khusus menyoroti penerjemahan online BJ. Penelitian ini diawali dengan kegelisahan terhadap banyaknya kekurangwajaran dalam hasil penerjemahan BI ke BJ dalam GT. Sebagai peutur BJ, kami melihat banyak kesalahan dalam berbagai level,baik semantik, sintaksis, maupun pragmatik atau *discourse*. Pada tulisan ini kami lebih fokus pada analisis semantik terhadap kesalahan-kesalahan tersebut.

Weaver (1955) mengatakan bahwa pendekatan semantik sangat diperlukan untuk menciptakan human like translation atau penerjemahan yang mendekati terjemahan manusia. Namun demikian, pembuatan Semantic Machine Translation akan lebih sulit sebab akan membutuhkan lebih banyak ahli linguistik dan programer dalam mewujudkannya. Dalam tulisan ini kami tidak akan membahas Semantic Machine Translation, tetapi kami akan mengkaji, seperti yang Weaver (1995) katakan, permasalahan semantik dalam hasil terjemahan GT yang sudah ada. Kami berharap dengan hasil penelitian ini, kami dapat memberikan pandangan-pandangan kritis dari sudut pandang linguistik.

Masalah semantik dalam penerjemahan sangat berkaitan dengan meaning transfer dari SL ke TL. Pada saat kita menganalisis dan

menerjemahkan kalimat, kita perlu melakukannya "from the standpoint of how it creates and expresses meaning." (Halliday, 2004, p.19). Ketidakstabilan makna sangat bisa terjadi karena penerjemah tidak dapat mempertemukan fitur kebahasaan dan konteks dari kata atau frasa dalam SL dan TL. Contoh dari kasus ini adalah kata ambigu, polisemy, dan oligosemy, dan saat terdapat unsur sosial budaya yang melekat dalam kata/frasa. Masalah semantik ini banyak terjadi dalam *machine translation*, karena sangat berkaitan dengan kualitas *parralel corpora* dan kemampuan komputer dalam membangun makna.

METODE

Data dikumpulkan dengan cara memasukkan dan menerjemahkan 5.000 kalimat BI ke BJ menggunakan GT. Kalimat-kalimat ini kami dapatkan dari website-website di Indonesia sehingga konteks kalimat dapat dipahami dengan mudah. Data dianalisis secara kualitatif dan dikelompokkan kedalam kategori yang muncul. Kategori-kategori ini muncul selama proses analisis (*emerging categories*) dan tidak ditentukan sebelumnya. Sebagai penelitian yang bersifat *data-driven*, berbagai kategori bisa muncul. Untuk itu, kategori yang dianggap paling signifikan yang akan diangkat dalam diskusi penelitian ini.

HASIL DAN PEMBAHASAN

Kajian semantik pada hasil terjemahan *online*

Makna merupakan masalah yang sangat kompleks tetapi sangat menentukan kualitas dan keakuratan penerjemahan. Penerjemahan oleh manusia saja sangat mungkin memiliki masalah semantik, apalagi penerjemahan mesin yang memiliki keterbatasan analisis makna.

Beberapa masalah semantik muncul dalam data yang dianalisis. Kami akan mendiskusikan empat masalah umum dan yang menganalisisnya dalam konteks terjemahan BI ke BJ. Perlu digarisbawahi bahwa BJ memiliki keunikan kebahasaan dan pemaknaan kata/frasa yang mungkin tidak ditemukan dalam bahasa lain.

No Parralel Data Found

Banyak ditemukan dalam data adanya hasil terjemahan masih mengandung leksikon BI. Hal ini disebabkan karena tidak adanya atau ditemukannya kata yang dicari dalam sistem. Tabel 1 menunjukkan contoh-contoh tidaklengkapan leksikon. Dalam data ini, hasil terjemahan BJ masih mengandung unsul kata/frasa BI.

Temuan pertama, yaitu ketidakmuncualan kata dalam BJ, terjadi dalam dua bentuk. Pertama, hilangnya kata/frasa dalam BJ. Da-

Tabel 1. Ketidaklengkapan kosa kata

Kalimat	BI	BJ
1	Padi has weeklyscheduled snorkeling tours available to the beautiful island of Nusa Lembongan and Penida.	Padi nduwe wisata *snorkel* [TOURS] sing dijadwalake saben minggu kanggo [BEAUTIFUL] Pulo Nusa Lembongan lan Penida sing ayu.
2	I always take jungle formula spray, just thought the sunscreen with repellant would help regards on the beach and general daytime movement	Aku tansah njupuk semprotan rumus *jungle*, mung ngira *sunscreen* karo *repellent* bakal bantuan *regards* ing pantai lan umum dinagerakan
3	One of the first online cupcake stores in the city has taken the rest of Indonesia by storm – now they even deliver to Bandung.	Salah siji saka toko *cupcake online* sing paling anyar ing kutha iki wis ngilangi negara Indonesia kanthi bada- saiki malah dikirim menyang Bandung

lam kalimat (1) kata *tours* dan *beautiful* tidak muncul sama sekali dalam BJ. Kedua, kata/frasa tersebut muncu tetap dalam bentuk linguistik yang sama. Kata *snorkel* (1), *sunscreen* dan *repellent* (2), dan *cupcake* (3) merupakan contoh-contoh dari fenomena ini.

Terdapat beberapa penyebab ketidakmunculan kata dalam TL. Yang paling utama adalah tidak adanya padanan kata tersebut dalam TL dan sulitnya menyesuaikan fitur makna. Dalam kalimat (1), kata *snorkelling* sulit dicari padanannya dalam BJ. Sama halnya dengan *sunscreen* dan *repellent* (3) dan *cupcake* (3). Kata *tours*, *beautiful*, *jungle* dan *regards* (2) sebenarnya masih bisa dan cukup mudah dicari padanan maknanya dalam BJ dengan cara memahami konteks bahasa aslinya, namun *input parralel corpora* tidak lengkap sehingga tidak muncul dalam terjemahan.

Masalah ketidakmunculan kata/frasa dalam TL bisa disebabkan oleh dua hal. Selain karena kurangnya jumlah data, kurangnya cakupan topik atau domain dari korpus data yang dikumpulkan merupakan penyebab utama ketidaktersediaan kata yang dicari pengguna.

Wrong Translation

Dalam analisis banyak ditemukan hasil terjemahan yang jauh dari makna aslinya, sebagaimana disajikan pada Tabel2.

Kata "*Credit*" dalam kalimat (4) berbeda dengan kredit dalam versi BJ. Demikian juga dengan *setting up* dan yang tidak sepadan dengan "nyetel". *Chillies* yang seharusnya diterjemahkan menjadi sambel juga diterjemahkan menjadi mangkok, sangat jauh dari makna aslinya. *Brings together* seharusnya diterjemahkan menjadi ngumpulke bukan ndadekke. Hal ini bisa disebabkan karena ketidak akuratan *parallel corpora* atau data terjemahan dari BI ke BJ. Kurang cermatnya proses terjemahan dan pengecekan ulang terjemahan (*proof reading*) bisa menjadi penyebab utama kesalahan makna dalam hasil terjemahan

Word Ambiguity

Ketidakakuratan data ini bisa juga disebabkan karena satu kata bisa saja bermakna ambigu atau berbeda dantara TL dan SL, seperti yang disajikan pada Tabel 3.

Tabel 2. Kesalahan penerjemahan kosa kata

Kalimat	BI	BJ
4	Widodo does not deserve all the blame for the perilous state of Southeast Asia's biggest economy. But he has not earned *credit* for *setting up* a *turnaround*, either	Widodo ora pantes kanggo nyalahke negara ekonomi Asia Tenggara. Nanging dheweke durung entuk kredit kanggo *nyetel turnaround*, uga.
5	This type of *grilled chicken* originates from the island of Lombok, and it's popular with spicy grilled chicken lovers through out Indonesia. When I saw the amount of *chilies* caked onto my ayam bakar Taliwang, I knew I was in for a *life-changing* grilled chicken experience, and it was true	Jenis *pitik* iki asal saka pulo Lombok, lan misuwur banget karo pecinta ayam bakar ing saindheng ing Indonesia. Nalika aku weruh jumlah *cangkir* sing disemprotake ing pitik Taliwang, aku sumurup, yen aku wis ngalami pengalaman pitik panggang sing *urip*, lan iki bener.
6	Sri Mulyani chaired the World Bank Group's Advisory Council on Gender and Development, which *brings together* global leaders and experts on gender issues, including from the private sector.	Sri Mulyani mimpin Dewan Penasehat Group Bank babagan Gender lan Pembangunan, sing *ndadekke* para pemimpin global lan para ahli babagan masalah gender, kalebu saka sektor swasta.

Kata-kata bercetak miring dalam kalimat (7), (8), dan (9) merupakas kata yang ambigu sebab dapat memiliki makna yang berbeda sesuai dengan konteks penggunaannya. interest dapat bermakna ketertarikan, bunga (*bank*), dan kepentingan. Dalam (7) seharusnya kata tersebut bermakna bunga bank namun diartikan menjadi kepentingan. *Can* dalam (8) harusnya bermakna kaleng dan bukan bisa. *Bank* seharusnya bermakna hulu, dan bukan institusi keuangan bank.

Dalam data base bisa saja, kata-kata ambigu tersebut diterjemahkan secara benar sesuai konteksnya, namun pertanyaannya adalah apakah komputer mampu mengenali perbedaan makna ini? Kalimat yang dimasukkan pengguna bisa sangat beragam dan kompleks.

Idiomatic Phrases

Machine Translation tidak memiliki pengetahuan umum dan kemampuan menerjemahkan sesuai dengan kenyataan di lapangan. Salah satunya adalah masalah-masalah kebahasaan yang berkaitan dengan *idiomatic expressions* yang sangat kental dalam penggunaan behasa dalam konteks sehari-hari. Oleh karena itu, SMT tidak dapat memahami dan menerjemahkan frasa-frasa idiom seperti "*raining cats* dan *dogs* dalam kalimat (10) yang diterjemahkan secara literal. Contoh rinci disajikan pada Tabel 4.

Idiom tidak dapat diterjemahkan secara literal, oleh karena itu kita sangat perlu memahami konteksnya dalam SL. Tantangan besar lainnya adalah sulitnya menemukan padanannya dalam TL. Yang terjadi biasanya adalah makna dipertahankan tetapi bentuk linguistiknya berbeda. Idiom lagi bukan lagi berbentuk idiom dalam TL. Tabel 3, menunjukkan contoh-contoh kesalahan penerjemahan idiom. Perlu penelitian lebih lanjut untuk penerjemahan idiom ini.

Masalah-masalah diatas bermuara pada prinsip-prinisip linguistik yang tidak bisa dilupakan dalam pemgembangan sebuah MT. Prinsip-prinsip linguistik ini sangat berkaitan dengan konteks penggunaan. Hal ini senada dengan penelitian-penelitian sebelumnya se-

Tabel 3. Kesalahan penerjemahan kata ambigu

Kalimat	BI	BJ
7	BNI Deposito is a term deposits that make your savings safe with an attractive *interest* rate. Benefit. Get a competitive interestrate.	BNI Deposito minangka simpenan istilah sing nggawe tabungan sampeyan aman kanthi tingkat *kapentingan* sing menarik. Mupangat. Njupuk tingkat kapentingan sing kompetitif.
8	Do you have a soda *can* to recycle?	Sampeyan duwe soda*bisa* kanggo daurulang?
9	River *banks* play an important role in causing flood in Jakarta.	*Bank-bank* ing pinggir kali nggawa peran penting ing nyebabake banjir ing Jakarta

Tabel 4. Kesalahan penerjemahan frasa idiom

Kalimat	BI	BJ
10	*Raining cats and dogs* in Jakarta tonight.	*Kucing lan asu* ing Jakarta bengi iki.
11	Some peoplesay that speaking Indonesian is just*a piece of cake*.	Sawetara wong ngomong yen basa Indonesia *mung sepotong kue*
12	I toldmy friend *not to judge a book by its cover*.	Aku ngomong kanca *ora ngadili buku kanthi tutupe*

perti yang diungkapkan oleh Van Eydne (2015):

> "… most of the hard problems of machine translation are of a linguistic nature. The choice of hardware and programming language(s), the definition of a formalism and a user language, the incorporation of world knowledge and statistical data, the maintenance of large dictionaries and terminology collections are all interesting and difficult problems in their own right, but whatever choices or solutions one proposes in these areas, they will have to be integrated in a system – which – by necessity – deals with linguistic data, both monolingual and bilingual" (Van Eydne, 2015, p. vi)

Selain masalah linguistik murni, salah satu yang sangat tipikal dalam BJ adalah kontek sosial budaya. Namun bahasan ini tidak akan kami kupas tuntaas dalam tulisan ini. Kualitas dan konten terjemahan yang memperhitungkan konteks budaya Jawa. Bahasa Jawa memiliki keunikan konteks budaya yang berbeda dari bahasa lain. Untuk memberikan hasil terjemahan yang tepat konteks dan budaya, maka produk SMT betul-betul memperhitungkan konteks budaya dari teks yang diterjemahkan.

Ada tiga prinsip yang kami ajukan untuk memperbaiki kualitas SMT BI ke BJ dari perspektif semantik, yaitu: *Quantity*, *Coverage*, and *Quality*. *Pertama*, quantity. Jumlah data yang besar semakin besar jumlah data semakin baik kualitas terjemahan yang dihasilkan dan semakin lengkap kosa kata. *Kedua*, *coverage*. Data sebaiknya mencakup area atau domain yang beragam, sehingga semakin lengkap datanya. Paling tidak ada 18 domain data yang dapat dikumpulkan. Kedelapan belas domain tersebut adalah: (a) kesehatan, (b) pendidikan, (c) ekonomi, (d) budaya, (e) seni, (f) hiburan/ *entertainment,* (g) bisnis, (h) politik, (i) fesyen dan gayahidup, (j) agama, (k) geografi, (l) agraria, (m) profesi/pekerjaan, (n) teknologi, (o) hukum, (p) kuliner (makanan dan minuman), (q) pariwisata, dan (r) sejarah. *Ketiga*, *quality*: c*ontinuous editing*. Untuk memperbaiki kualitas hasil terjemahan perlu dilakukan *proofreading data paralel* dan *editing* atau *feedback* secara berkelanjutan baik oleh pengguna dan pengelola. Pengguna memiliki peran yang sangat penting untuk penyempurnaan hasil terjemahan.

SIMPULAN

Pendekatan statistik yang digunakan oleh *machine translation system* seperti GT masih meninggalkan banyak masalah kebahasaan dalam hasil terjemahannya. Jika mempertahankan sistem statistik, untuk meningkatkan kualitas dan keakuratan hasil terjemahannya, *corpora* harus ditingkatkan baik secara kuantitas, *coverage*, maupun kualitasnya. Secara kuantitas, jumlah data harus terus ditambah dan domain data harus diperluas dan dilengkapi. Secara kualitas, keakuratan terjemahan *data source language* ke *target language* harus ditingkatkan. Peningkatan kualitas parallel corpora ini juga memberikan tantangan yang sangat besar mengingat bahasa Jawa sangat sarat dengan konteks sosial budaya.

DAFTAR PUSTAKA

Bird, S & Chiang, D. (2012). *Machine Translation for Language Preservation*. Mumbai.

Halliday, MAK. (2004*). An Introduction to Functional Grammar*. London: Arnold Publisher Hodder Headline Group.

Van Eydne, F. (2015). *Liuguistic Issues in Machine Translation*. Bloomsbury Publishing.

Weaver. (1955). Translation. In *Machine Transliation of Language*. Volume 14, Cambridge: MIT Press