韓日機械翻訳における助詞の誤訳の問題

姜龍熙　kang@phiz.c.u-tokyo.ac.jp
東京大学大学院　言語情報科学専攻

1．序論

朝鮮語と日本語は、文法全体の類似性、特に語順が似ており、ともに助詞が存在するいう点で、機械翻訳を行う上で、有利な条件が比較的揃っている。しかし、機械翻訳のシステムには依然多くの問題が残されている。本研究では、誤訳の原因となるものとして助詞に注目し、そのシステムを考える。先行研究では、韓日の機械翻訳に関する論文は数少なく、主に日韓の機械翻訳に関する論文が多かった。本研究は、corpus から助詞を含む様々な文を予め集めて、日立情報ネットワークの構文ダイレクト式の HICOM/MT に機械翻訳を行った結果から、各助詞の誤訳のパターンを割り出すという方法を採る。なお、統語的に韓国語の助詞と日本語の助詞が必ずしも対応しない例もある。

例1）나는　전철를　탄다　　私は電車に乗る。
例2）나는　전철를　갈아탄다　　私は電車を乗り換える。
例3）나는　경부선으로　갈아탄다　　私は新幹線に乗り換える。

2．本論

2.1 先行研究

日本語と韓国語を対象とする機械翻訳システム（殆どのシステムがダイレクト方式を採用している）や論文は日韓の研究が多い。

両国語の類似点及び相違点に関する先行研究としては金政仁（１９９６）、金泰錫（１９９２、１９９３、１９９５）等の研究がある。また、日韓の様々な機械翻訳システムを corpus をもって客観的に評価した崔杞鮮（１９９６）の研究も注目すべきである。言語学の立場で本研究に参考となる論文としては、韓国語の資料の出現頻度と共起確率等の量と継起の順序に着目した野間（１９９３）と結合価理論(Valenzhteorie)から日本語文型を分析した新田（１９９５）がある。 本研究ではそれぞれ野間と新田に従い、corpus 分析を中心とする韓日機械翻訳のシステムを提案する。

2.2 corpus の分析

東亜日報と朝鮮日報の CD-ROM から予め名詞､代名詞＋助詞を含んでいる文や同音異義の用言の文等を３種類用意した。Ⅰ各助詞を含んでいる文（８００）Ⅱ朝鮮日報の news 記事文（１５０）
Ⅲ同音異義形の中で corpus 頻度を調べるために分節形で集めた文（１０００）

表1　機械翻訳にかけた韓国語の corpus の中の助詞形の分節の頻度

	은	는	이	가	을	를	으로	로	에게	에서	의	과	와	도	에
Ⅰ	431	953	550	350	758	462	271	251	71	125	910	172	148	273	497
Ⅱ	118	140	126	73	176	105	51	67	5	11	137	39	24	33	123
合計	549	1093	676	423	934	567	322	318	76	136	1047	211	172	306	620
%	7%	15%	9%	6%	13%	8%	4%	4%	1%	2%	14%	3%	2%	4%	8%

Ⅰ:corpus＋Ⅱ:news　合計;７４５０

上の結果から８％を越える助詞は同音異義の環境（用言の連体形、或いは名詞の一部）でその頻度が高くなっている。そこでⅢの corpus を品詞別に調べた結果は次の通りである。

	은	는	을	를	남은	남을	먹을
名詞＋助詞	８０％	３３％	９２％	９７％	０％（他人は）	２２％（他人を）	１％（墨を）
用言の連体形	２０％	６７％	7.5％	2.5％	７４％（残る）	７０％（残る）	９９％（食べる）

上の結果から名詞の中では、助詞によって使用頻度が大きく変わることが分かった。また、誤訳では頻度の低い用言の連体形を選んだ例が多かった。例えば、「우리는」は代名詞＋助詞の variant と動詞の우리다の連体形がある。

2.2.2助詞としての誤訳のタイプ

助詞の中には言語の環境（同音異義語、多義語で）、助詞として判断しやすいものと同音異義に判断されやすいものがある。
αタイプ：同音異義の中に漢語などがあり、助詞ではなく、同音異義として判断される確率が高い。
βタイプ：同音異義の中に用言の活用形があり、用言の活用形を含む語幹として判断される確率が高い。
γタイプ：助詞として判断されやすいが、多義語のため別の助詞の誤訳になる確率が高い。

2.2.3 助詞の同音異義環境（用言の活用形）
次は韓国語の用言の paradigm（姜１９９６）の一部を表す

	走る	食べる	する	くる	行く	分かる	作る	聞く	手伝う	違う
語幹類						ㄹ変則	ㅅ変則	ㄷ変則	ㅂ変則	ㄹ変則
S-1	달리	먹	하	오	가	알	짓	듣	돕	다르
語形										
現在連体	달리는	먹는	하는	오는	가는	아는	짓는	듣는	돕는	다르는
過去連体	달린	먹은	한	온	간	안	지은	들은	도운	다른
未来連体	달릴	먹을	할	올	갈	알	지을	들을	도울	다를
現在	달린	＊먹는	한	온	간	안	＊짓는	＊듣는	＊돕는	다른

下線は名詞＋助詞または代名詞＋助詞の同音異義環境（ blank を含んでいる。）を表す。
＊は単独形助詞との同音異義形に見えるのだが、現実には blank を含まないので同音異義にはならない。
ㄴ지라도，ㄴ지도，ㄹ지도，ㄹ지라도，ㄹ수록の形は現在連体と未来連体につく形態素であるが、blank を含まないので同音異義の環境としては認められない。

2.3 誤訳の原因及び類型

2.3.1 辞書形誤訳；辞書に情報を記載する際、明らかな誤りか、或いは情報が未登録の状態であるために起こる誤訳の類型。
2.3.2 分析の誤りによる誤訳の分類
構成要素の分析の誤り；多音節の形態素成分の、文字境界単位は正確に判断したが、複数の選択肢を決める過程で誤りが生じる類型。
構成単位の境界の分析の誤り；多音節の形態素成分の、文字境界単位を正確に認識しない誤りで生じる誤訳の類型。（頻度が高く、連鎖誤訳につながる恐れが多い。）
生成型誤訳；言語の統語的な差異で、日本語に生成する過程で変形、添削の作業をしなければならないことから生じる誤訳の類型。
2.3.3 影響別による分類
単独誤訳；翻訳される際の順序が後であるかまたは前の形態素であるが、誤訳の影響がその成分単独に終わる類型。
連鎖誤訳；翻訳をしていく過程で、先の形態素の誤りによって後の形態素までその影響を及ぼすと見られる類型。

2.4 誤訳の例（漢字→漢、連続→連、構成単位→単、構成要素→要）

a 인사가	b 코닥사가	c 우리는	d 두 나라는	e 정상회담을
人事が	コダック社が	我々は	両国は	首脳会談を
名詞＋助詞	固有名詞＋助詞	代名詞＋助詞	数詞＋名詞＋助詞	名詞＋助詞
あいさつが	鼻＃私家	抜き取る	２出ろという	情商会盛る
同音異義＋助詞	名詞＋未＋漢＋漢	用言の連体形	数詞＋用言の活用形	漢＋漢＋用言の連体形
要	単＋連＋未登録	要、単	単、要、連	単＋要＋連

2.5 誤訳の環境

上にあげた例を含め、助詞誤訳の環境を整理すると次の表２になる。

表２　　助詞の言語環境

助詞	誤訳のタイプ（α β γ）	助詞の形態分析の誤りの可能性	同音異義の誤りの環境	多議語の誤りの環境	誤訳の類型
는	β	○	用言の現在連体形、動詞「늘다」（増える）	×	単、連（用言の連体）
은	α β	○	用言の現在連体形、漢語「은」（銀、恩）	同上	単、要（銀）、形容詞が多い
이	α	○	名詞の一部	が、では、の、に、を	単、生（否定のでは）
가	α	○	名詞の一部	同上	単、生（否定）
을	α β	○	用言の未来連体形、漢語「을」（乙）	を、が、に	単、要（乙）、連
를	β	○	用言の未来連体形	同上	単、連（用言）
으로	γ	×	×	で、に、へ、から	要、で（到発）→に
로	α γ	○	副詞形、名詞の一部、漢語	上同	単（副詞→助詞）
에게	γ	×	×	に、へ	要、へ→に
에서	γ	○	×	で、から、に	要、から→で
의	α	○	名詞の一部、漢語	×	要、名詞→助詞
과	α	○	名詞の一部、漢語	×	単
와	α	○	名詞の一部、漢語	×	単
도	α γ	○	名詞の一部、漢語	と、も	単、名詞→助詞
에	γ	○	×	に、で、へ、や	要、で→に

誤訳の例の中で a は名詞の誤訳であり、助詞の翻訳ではない。ここでは、次の助詞の誤訳は b、c、d、e のようなものを対象とする。

表３　　助詞誤訳の原因別頻度（数字は％を表す）

	은	는	이	가	을	를	으로	로	에게	에서	의	과	와	도	에
助詞(認識率)	78	65	68	74	79	70	100	90	100	100	92	96	98	86	100
名詞→助詞	8	0	4	8	4	0	0	6	0	0	4	10	2	8	0
助詞→名詞	8	0	16	12	2	0	0	4	0	0	4	4	0	6	0
助詞→用活形	6	25	0	0	5	16	0	0	0	0	0	0	0	0	0
用活形→助詞	8	10	0	0	6	6	0	0	0	0	0	0	0	0	0
多議語の選択	0	0	14	6	4	8	35	18	20	35	0	10	0	0	20
助詞＋用言	0	0	2	0	16	18	10	8	12	18	0	0	0	0	15
未登録（連）	5	8	6	4	12	18	12	14	10	2	4	2	0	0	0
生成型	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0

2.6　問題点と誤訳の分析

韓国の助詞と日本語の助詞間に統語的に使い方で大きな差があるものがある。韓国語では属格の「의」は省略可能だが、日本語では省略できない場合（姜１９９６）がある。しかし、zero 格の処理の問題は今後の課題にする。また、助詞を含んでいる慣用句の誤訳が多かったが、例えば「意識を失う」の韓国語を「意識をなくす」というふうに翻訳している。助詞のみの翻訳には成功しているが、文全体の面で

見ると誤訳になってしまう。助詞と用言との関連性をもってプログラムを作る必要性があると考えられる。この点も今後の課題である。全般的に助詞の誤訳の環境として一番多かったのは外来語、人名、固有名詞、地名の後に来る助詞であった。登録されてない情報を外来語なり、地名として判断できるようなプログラムを作る必要がある。この点は助詞に限る問題ではないが、翻訳率を上げる上での今後の課題である。

3．結論

助詞に関わるcorpusをHITACHIのHICOM/MT機械翻訳にかけた結果、生じる誤訳には2つのタイプがあることが分かった。

一つは形態分析のレベルにおける同音異義の形式への誤訳である。用言の活用型と助詞の同音異義の例として、1．動詞の現在連体形(語幹＋은、는）と、名詞＋助詞은、는(日本語の「は」に相当）、2．動詞の未来連体形(語幹＋을，를）と、名詞＋助詞을，를(日本語の「を」に相当）がある。もう一つは意味分析のレベルにおける、多義語の助詞から多義語の助詞への翻訳に際する誤訳である。韓国語の1つの助詞に対して、訳語となりうる日本語の助詞が2つ以上存在する場合があり、文の中でその選択の間違いによる誤訳が起こる。つまり、日本語を生成する過程において多義語のなかで選択が難しい助詞がある。そこで、助詞によって誤訳のパターンが異なることに注目し、別の誤訳の防止プログラムを考える必要がある。「は」に当たる韓国語「은 」は名詞の「銀」と同音異義の環境なので、「銀」を名詞単語句形ではなく、「銀」＋助詞形即ち、「은은」、「은을」、「은을」の形態で入力することは助詞の翻訳率を挙げる1つの提案である。また、辞書の登録単位として、β typeの誤訳については名詞＋助詞、代名詞＋助詞を設ける必要がある。特に、誤訳例 eのように接続 TABLEでは充分条件の訳である名詞＋用言の連体形はその順序通りに訳すことも可能であるが、形態素の成立必要条件であるBLANKの後のTABLEの認識がされればこの類の誤訳が避けられると考えられる。（用言の連体形の結果を出す際の条件として必ず前後にBLANKの存在の有無を義務条件とする)そして、多義語のγ typeは助詞＋用言を1つの辞書の単位として入力するのが適当な方法である。なぜならば「으로、에서」の前の名詞、固有名詞が道具、原因、出発点、到達点になりうる情報を全部入力することは不可能だからである。

助詞のそれぞれの誤訳のtypeは次の通りである。

同音異義型：는、를、과、로　　多義語型：으로、에、에서
α type　：가 、이、 로　　β type：　은、는、을、 를
γ type　：으로、에서、에

「参考文献」
崔杞鮮・金泰完（１９９６）「日韓機械翻訳システムの現状および分析」（言語処理学会、言語処理学会第２回年次大会発表論文集、pp.433-443）
野間秀樹（１９９３）「現代朝鮮語の対格と動詞の統辞論」（語研資料１５ 言語研究Ⅲ 東京外国語大学　語学研究所、PP.77-168）
新田春夫（１９９５）「結合価理論から見た日本語文型」(言語・情報・テキスト VOL.2 1994-1995, 東京大学大学院総合文化研究科、言語情報科学専攻、PP.67-88）
金泰錫・浦昭二（１９９２）日韓機械翻訳における意味接続関係を用いた韓国語の生成方法（情報処理学会 Vol ３３、pp.1578-1588）
金泰錫（１９９３）「日韓機械翻訳における対立語の処理」（東義大学校、産業技術研究誌、第９券、pp.217-221）
金泰錫・浦昭二（１９９３）「日韓機械翻訳における否定の処理」（情報処理学会 Vol 34、pp.892-904）
金泰錫・金政仁（１９９５）「拡張翻訳テーブルを用いた日韓機械翻訳」（情報処理学会５１回全大集、Vol.3、pp.89-90）
金政仁（１９９６）「日韓機械翻訳における活用語処理のための拡張翻訳テーブルの改善」（言語処理学会第２回年次大会、pp.9-12）
松田純一・河野勝也（１９９３）「構文ダイレクト方式による日韓機械翻訳システム」（情報処理学会全国大会論文集（３）、pp.139-140）
姜龍熙（１９９６）「韓・日機械翻訳における誤訳及び考察」（ハングル及び韓国語情報処理、第８回ハングル及び韓国語情報処理学術大会、人間と機械と言語、PP.351-366）