Learning from a Neighbor: Adapting a Japanese Parser for Korean through Feature Transfer Learning

Hiroshi Kanayama
IBM Research - Tokyo
Koto-ku, Tokyo, Japan
hkana@jp.ibm.com

Youngja Park
IBM Research - T.J.Watson Research Center
Yorktown Heights, NY, USA
young_park@us.ibm.com

Yuta Tsuboi
IBM Research - Tokyo
Koto-ku, Tokyo, Japan
yutat@jp.ibm.com

Dongmook Yi
Korea Software Solutions Laboratory, IBM Korea
Gangnam-gu, Seoul, Korea
dmyi@kr.ibm.com

Abstract

We present a new dependency parsing method for Korean applying cross-lingual transfer learning and domain adaptation techniques. Unlike existing transfer learning methods relying on aligned corpora or bilingual lexicons, we propose a *feature transfer* learning method with minimal supervision, which adapts an existing parser to the target language by transferring the features for the source language to the target language. Specifically, we utilize the *Triplet/Quadruplet Model*, a hybrid parsing algorithm for Japanese, and apply a delexicalized feature transfer for Korean. Experiments with *Penn Korean Treebank* show that even using only the transferred features from Japanese achieves a high accuracy (81.6%) for Korean dependency parsing. Further improvements were obtained when a small annotated Korean corpus was combined with the Japanese training corpus, confirming that efficient cross-lingual transfer learning can be achieved without expensive linguistic resources.

1 Introduction

Motivated by increasing demands for advanced natural language processing (NLP) applications such as sentiment analysis (Pang et al., 2002; Nasukawa and Yi, 2003) and question answering (Kwok et al., 2001; Ferrucci et al., 2010), there is a growing need for accurate syntactic parsing and semantic analysis of languages, especially for non-English languages with limited linguistic resources. In this paper, we propose a new dependency parsing method for Korean which requires minimal human supervision. Dependency parsing can handle long-distance relationships and coordination phenomena very well, and has proven to be very effective for parsing free-order languages such as Korean and Japanese (Kübler et al., 2009).

Most statistical parsing methods rely on annotated corpora labeled with phrase structures or dependency relationships, but it is very expensive to create a large number of consistent annotations. Recently, treebanks have become available for many languages such as English, German, Arabic, and Chinese. However, the parsing results on these treebanks vary a lot depending on the size of annotated sentences and the type of annotations (Levy and Manning, 2003; McDonald et al., 2013). Further, many languages lack annotated corpus, or the size of the annotated corpus is too small to develop a reliable statistical method. To address these problems, there have been several attempts at unsupervised parsing (Seginer, 2007; Spitkovsky et al., 2011), grammar induction (Klein and Manning, 2004; Naseem et al., 2010), and cross-lingual transfer learning using annotated corpora of other languages (McDonald et al., 2011). However, the accuracies of unsupervised methods are unacceptably low, and results from cross-lingual transfer learning show different outcomes for different pairs of languages, but, in most cases, the parsing accuracy is still low for practical purposes. A recent study by McDonald *et al.* (2013) concludes that cross-lingual transfer learning is beneficial when the source and target languages were similar. In particular, it reports that Korean is an outlier with the lowest scores (42% or less in UAS) when a model was trained from European languages.

In this paper, we present a new cross-lingual

transfer learning method that learns a new model for the target language by transferring the features for the source language. Unlike other approaches which rely on aligned corpora or a bilingual lexicon, we learn a parsing model for Korean by reusing the features and annotated data used in the Japanese dependency parsing, the *Triplet/Quadruplet Model* (Kanayama et al., 2000), which is a hybrid approach utilizing both grammatical knowledge and statistics.

We exploit many similarities between the two languages, such as the head-final structure, the noun to verb modification via case and topic markers, and the similar word-order constraints. It was reported that the mapping of the grammar formalism in the language pair was relatively easy (Kim et al., 2003b; Kim et al., 2003a). However, as the two languages are classified into independent language families (Gordon and Grimes, 2005), there are many significant differences in their morphology and grammar (especially in the writing systems), so it is not trivial to handle the two languages in a uniform way.

We show the *Triplet/Quadruplet Model* is suitable for bilingual transfer learning, because the grammar rules and heuristics reduce the number of modification candidates and can mitigate the differences between two languages efficiently. In addition, this model can handle the relationships among the candidates as a richer feature space, making the model less dependent upon the lexical features of the content words that are difficult to align between the two languages. Similarly to the delexicalized parsing model in (McDonald et al., 2011), we transfer only part-of-speech information of the features for the content words. We create new mapping rules to extract syntactic features for Korean parsing from the Japanese annotated corpus and refine the grammar rules to get closer modification distributions in two languages.

Our experiments with Penn Korean Treebank (Han et al., 2002) confirm that the *Triplet/Quadruplet Model* adapted for Korean outperforms a distance-based dependency parsing method, achieving 81.6% accuracy when no annotated Korean corpus was used. Further performance improvements were obtained when a small size of annotated Korean corpus was added, confirming that our algorithm can be applied without more expensive linguistic resources such as an aligned corpora or bilingual lexicons. Moreover,

the delexicalized feature transfer method enables the algorithm applicable to any two languages that have similar syntactic structures.

2 Related Work

2.1 Parsing for Korean

Since Korean is a morphologically-rich language, many efforts for Korean parsing have focused on automatically extracting rich lexical information such as the use of case frame patterns for the verbs (Lee et al., 2007), the acquisition of case frames and nominal phrases from raw corpora (Park et al., 2013), and effective features from phrases and their neighboring contexts (Choi and Palmer, 2011). Recently, Choi et al. (2012) discussed the transformation of *eojeol*-based Korean treebank to entity-based treebank to effectively train probabilistic CFG parsers. We apply similar techniques as in (Choi et al., 2012) to mitigate the differences between Korean and Japanese syntactic structures.

Chung and Rim (2003) applied the *Triplet/Quadruplet Model* for Korean parsing as done in our work. They reported that the model performed well for long-distance dependencies, but, in their experiments, the number of modification candidates was not effectively reduced (only 91.5% of phrases were in one of the three positions, while it was 98.6% in Kanayama's work for Japanese). In this paper, we introduce more sophisticated grammatical knowledge and heuristics to have similar dependency distributions in the two languages. Smith and Smith (2004) attempted a bilingual parsing for English and Korean by combining statistical dependency parsers, probabilistic context-free grammars, and word translation models into a unified framework that jointly searches for the best English parse, Korean parse and word alignment. However, we utilize an existing parser and align the features from the source language to the features for the target language, and, thus, our method is applicable to situations where there is no aligned corpora or word translation models.

2.2 Transfer learning and domain adaptation

Recently, transfer learning has attracted much attention, as it can overcome the lack of training data for new languages or new domains for both classification and regression tasks (Pan and Yang, 2010). Transfer learning has also been applied to
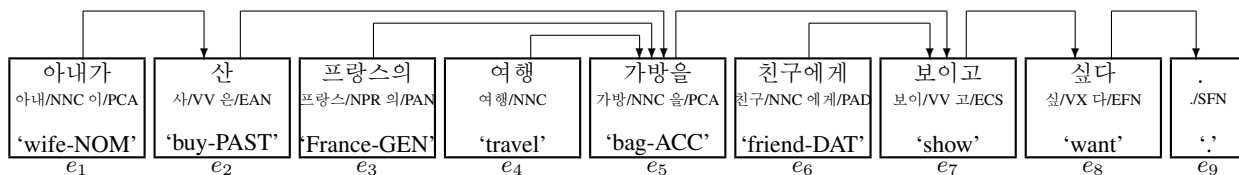
아내가	산	프랑스의	여행	가방을	친구에게	보이고	싶다	.
아내/NNC 이/PCA	사/VV 은/EAN	프랑스/NPR 의/PAN	여행/NNC	가방/NNC 을/PCA	친구/NNC 에게/PAD	보이/VV 고/ECS	싶/VX 다/EFN	./SFN
'wife-NOM'	'buy-PAST'	'France-GEN'	'travel'	'bag-ACC'	'friend-DAT'	'show'	'want'	'.'
e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9

Figure 1: An example of dependency structures of a Korean sentence "아내가 산 프랑스의 여행 가방을 친구에게 보이고 싶다." ('(I) want to show the French travel bag which (my) wife bought to (my) friend'). Each box corresponds to a Korean phrasal unit *eojeol*.

妻が	買った	フランスの	旅行かばんを	友達に	見せたい。
妻/n が/pc	買/v た/aux	フランス/np の/pn	旅行/n かばん/n を/pc	友達/n に/pc	見せ/v たい/aux 。/
'wife-NOM'	'buy-PAST'	'France-GEN'	'travel bag-ACC'	'friend-DAT'	'want to show'
b_1	b_2	b_3	b_4	b_5	b_6

Figure 2: A dependency structure of a Japanese sentence which corresponds to the Korean sentence in Figure 1, "妻が買ったフランスの旅行かばんを友達に見せたい。". Each box corresponds to a Japanese phrasal unit *bunsetsu*.

syntactic parsing, where a parsing model for a target language is learned from linguistic resources in one or more different languages (Hwa et al., 2005; Zeman and Resnik, 2008; McDonald et al., 2011; Durrett et al., 2012; Georgi et al., 2012; Naseem et al., 2012). McDonald et al. (2011) proposed a delexicalized parsing model for cross-lingual dependency parsing and demonstrated that a high accuracy parsing was achieved for Indo-European languages where significant amount of parallel texts exist. However, in more recent work, McDonald et al. (2013) showed that, unlike transfer learning within close language families, building a Korean parser from European languages was not successful with a very low accuracy. Durrett et al. (2012) and Georgi et al. (2012) show that transfer parsing can be improved when additional bilingual resources are available, such as bilingual dictionaries and parallel corpora of glossed texts respectively.

Our method does not require such resources and does not have restrictions on the sentence type that can be parsed. Instead, we use a mixture of a small corpus in a target language (i.e., Korean) and a larger corpus of a source language (Japanese). This task is similar to domain adaptation, and our objective is to outperform the training model built on each language separately. To avoid the loss of accuracy due to the differences between two domains, we apply the domain adaptation technique proposed by Daumé III (2007) which duplicates the feature space into three categories with each

of the features trained by source, by target, and by combined domains.

3 Dependency Structures of Korean and Japanese

A dependency structure in Korean is typically analyzed in terms of *eojeol* units, a basic phrase that consists of a content word agglutinated with optional functional morphemes such as postpositional particles or endings for verbs. Figure 1 shows an example Korean sentence with the dependencies between the *eojeol*s indicated by arrows. Figure 2 illustrates the dependency structure between the *bunsetsu*s in Japanese that corresponds to the Korean structure in Figure 1. As these figures show, the syntactic structures are quite similar in these languages: All of the dependencies are directed from left to right, and the postpositional particles determine if the content word modifies a verb ("가" in e_1 and "が" in b_1) or a noun ("의" in e_3 and "の" in b_3). The *eojeol*s in Korean roughly correspond to the *bunsetsu*s in Japanese. In the remainder of this paper, we denote both an *eojeol* or a *bunsetsu* as a '*PU*' (phrasal unit) when distinction is not needed.

While Korean and Japanese have similar syntactic structures, the two languages have many differences. The *eojeol*s in Korean are separated by white space, while the *bunsetsu*s in Japanese are not. Further, the statistics show several differences in the two languages. Table 1 compares a Korean corpus, *Penn Korean Treebank* (henceforth KTB)

Table 1: Statistics of Korean and Japanese corpora.

	KTB (Korean)	EDR (Japanese)
Average number of characters (except for whitespace) per sentence	73.7	28.0
Average number of PUs per sentence	25.5	8.53
Average number of morphemes per PU	1.83	2.86
Ratio of modification to the next PU	70.0%	61.8%

Table 2: Simplified examples of Japanese grammar rules.

Rightmost morpheme of the modifier PU	Conditions for the modified PUs
postpositional "を" *wo* (accusative)	verb, adjective
postpositional "の" *no* (genitive, nominative)	noun, verb, adjective
postpositional "と" *to* (conjunctive)	noun, verb, adjective, adverb "一緒に" *isshoni* ('together')
adverb	verb, adjective, adverb, copula

(Han et al., 2002), and a Japanese corpus, *EDR Corpus* (EDR, 1996). Both corpora consist of word-level bracketed constituents, so they are converted into PU-level dependency structures using the method described in Choi and Palmer (2011). Though both corpora consist mainly of newspaper or magazine articles, the sentences are not aligned with each other, so the statistics show the comparisons of the two corpora, rather than the theoretical comparisons of the two languages. However, we can see that Korean sentences tend to be longer than Japanese sentences both in terms of the number of characters and PUs. More *eojeol*s modify an adjacent *eojeol* in Korean than in Japanese. For instance, e_1, e_4, e_6, e_7, and e_8 modify the next *eojeol* in Figure 1, but only b_1, b_3, and b_5 modify the next *bunsetsu* in Figure 2. Those differences suggest some of the difficulties in applying the Japanese dependency model to Korean. The Japanese parsing method that will be described in the next section exploits these characteristics, which we apply to Korean parsing.

4 Triplet/Quadruplet Model

This section describes the *Triplet/Quadruplet Model* (Kanayama et al., 2000) which was originally designed for Japanese parsing. First, we review the two main ideas of the model – restriction of modification candidates and feature selection for probability calculation. Then, we describe how we apply the *Triplet/Quadruplet Model* to Korean parsing in Section 4.3.

4.1 Restriction of modification candidates

The *Triplet/Quadruplet Model* utilizes a small number (about 50) of hand-crafted grammar rules

that determine whether a PU can modify each PU to its right in a sentence. The main goal of the grammar rules is to maximize the coverage, and the rules are simple describing high-level syntactic dependencies, and, thus, the rules can be easily created without worrying about the precision or contradictory rules. The statistical information is later used to select the right rules for a given sentence to produce an accurate parsing result. Table 2 shows several grammar rules for Japanese, in which the modified PUs are determined depending on the conditions of the rightmost morpheme in the modifier PU.

An analysis of the EDR corpus shows that 98.6% of the correct dependencies are either the nearest PU, the second nearest PU, or the farthest PU from the modifier (more details in Table 4(a)). Therefore, the model can be simplified by restricting the candidates to these three candidates and by ignoring the other PUs with a small sacrifice (1.4%) of parsing accuracy.

4.2 Calculation of modification probabilities

Let u be a modifier PU in question, c_{un} the u's n-th modification candidate PU, Φ_u and $\Psi_{c_{un}}$ the attributes of u and c_{un}, respectively. Then the probability that u modifies its n-th candidate is calculated by the triplet equation (1) or the quadruplet equation (2) when u has two or three candidates, respectively [1].

$$P(u \rightarrow c_{un}) = P(n \mid \Phi_u, \Psi_{c_{u1}}, \Psi_{c_{u2}}) \qquad (1)$$
$$P(u \rightarrow c_{un}) = P(n \mid \Phi_u, \Psi_{c_{u1}}, \Psi_{c_{u2}}, \Psi_{c_{u3}}) \ (2)$$

---

[1]It is trivial to show that $P(u \rightarrow cu_1) = 1$, when u has only one candidate.

Table 3: Simplified examples of Korean grammar rules.

Rightmost morpheme of the modifier PU	Conditions for the modified PUs
PCA,PCJ,PAD,PAU (postpositional particles)	V* (verb, adjective or auxiliary), CO (copula)
EAN (nominal verb ending *e.g.* "은" *eun*)	N* (noun)
ADV (adverb), ADC (conjunction)	N* (noun), V* (verb, adjective or auxiliary), ADV (adverb), ADC (conjunction)
postpositional "과" *gwa* (conjunctive)	N* (noun), V* (verb, adjective or aux), adverb "함께" *hamkke* ('together')
N* (noun)	N* (noun), V* (verb, adjective or auxiliary)

Table 4: Distribution (percentage) of the position of the correct modified PU among the candidate PUs selected by the initial grammar rules. The column 'Sum' shows the coverage of the 1st, 2nd and last PUs. The EDR corpus was used for Japanese, and the KTB was used for Korean in this analysis.

(a) Japanese						(b) Korean (with the initial grammar)					
# of Candidates	Ratio	1st	2nd	Last	Sum	# of Candidates	Ratio	1st	2nd	Last	Sum
1	32.7	100.0	–	–	100.0	1	10.5	100.0	–	–	100.0
2	28.1	74.3	26.7	–	100.0	2	11.4	85.9	14.1	–	100.0
3	17.5	70.6	12.6	16.8	100.0	3	10.4	76.2	13.4	10.4	100.0
4	9.9	70.4	11.1	13.8	95.3	4	9.3	74.7	11.3	8.0	93.9
≥ 5	11.8	70.2	11.1	10.5	91.9	≥ 5	58.4	75.5	10.0	4.9	90.5
Total	100	–	–	–	98.6	Total	100	–	–	–	93.9

These probabilities are estimated by the maximum entropy method with a feature set to express Φ and Ψ. Assuming the independence of those modifications, the probability of the dependency tree for an entire sentence $P(T)$ is calculated as the product of the probabilities of all of the dependencies in the sentence using beam search to maximize $P(T)$ under the constraints of the projected structure.

$$P(T) \simeq \prod_{u} P(u \to c_{un}) \quad (3)$$

In comparison, a traditional statistical parser (Collins, 1997) uses Equation (4) to calculate the probability of u modifying t.

$$P(u \to t) = P(True \mid \Phi_u, \Psi_t, \Delta_{u,t}) \quad (4)$$

We call the model based on Equation (4) the *Distance Model*, since $\Delta_{u,t}$ (the distance between u and t) is typically used as the key feature. Though other contextual information, in addition to the attributes of u and t, can be added, the model calculates the probabilities of the dependencies between u and t independently and thus often fails to incorporate appropriate contextual information.

Equations (1) and (2) have two major advantages over the *Distance Model*: First, all the attributes of the modifier and its candidates can be handled simultaneously. The combination of those attributes helps the model to express the context of

the modifications. Second, the probability of each modification is calculated based on the relative positions of the candidates, instead of the distance from the modifier PU in the surface sentence, and, thus, the model is more robust.

4.3 Korean dependency parsing with the *Triplet/Quadruplet Model*

We design the Korean parser by adapting the *Triplet/Quadruplet Model* based on the analogous characteristics of Japanese and Korean. First, we created the Korean grammar rules for generating candidate modified PUs by modifying the rules for Japanese shown in Table 2 for Korean. The rule set, containing fewer than 50 rules, is simple enough to be created manually, because the rules simply describe possible dependencies, and Japanese phenomena are good hints for Korean phenomena. Table 3 shows some examples of the rules for Korean based on the POS schema used in the KTB corpus. We did not automatically extract the rules from the annotated corpora so that the rules are general and independent of the training corpus. Nonetheless, 96.6% of the dependencies in KTB are covered by the grammar rules. The remaining dependencies (3.4%) not covered by the rule set are mainly due to rare modifications and may indicate inconsistencies in the annotations, so we do not seek any grammar rules to achieve nearly 100%.
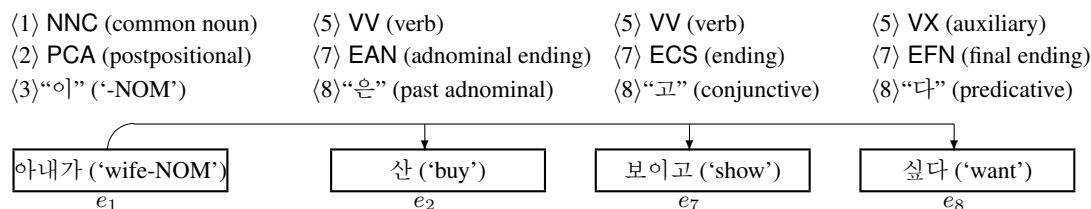
⟨1⟩ NNC (common noun)	⟨5⟩ VV (verb)	⟨5⟩ VV (verb)	⟨5⟩ VX (auxiliary)
⟨2⟩ PCA (postpositional)	⟨7⟩ EAN (adnominal ending)	⟨7⟩ ECS (ending)	⟨7⟩ EFN (final ending)
⟨3⟩"이" ('-NOM')	⟨8⟩"은" (past adnominal)	⟨8⟩"고" (conjunctive)	⟨8⟩"다" (predicative)

아내가 ('wife-NOM')	산 ('buy')	보이고 ('show')	싶다 ('want')
e_1	e_2	e_7	e_8

Figure 3: The features used to select the modified PU of e_1 among its three candidates. The full sentence of this example is shown in Figure 1. The numbers in brackets correspond to the feature IDs in Table 5.

Table 5: The features to express attributes of a modifier and modification candidates.

Feature set ID	Description
⟨1⟩	PoS of the head morpheme of the modifier
⟨2⟩	PoS of the last morpheme of the modifier
⟨3⟩	Lex of the postpositional or endings of the modifier
⟨4⟩	Lex of the adverb of the modifier
⟨5⟩	PoS of the head morpheme of the modification candidate
⟨6⟩	Lex of the head morpheme of the modification candidate
⟨7⟩	PoS of the last morpheme of the modification candidate
⟨8⟩	Lex of the postpositional or endings of the modification candidate
⟨9⟩	Existence of a quotation expression "다고" *dago* or "라고" *rago*
⟨10⟩	Number of "은" *eun* (TOPIC marker) between the modifier and modification candidate
⟨11⟩	Number of commas between the modifier and modification candidate
combination	⟨1⟩ × ⟨5⟩ / ⟨2⟩ × ⟨5⟩ / ⟨2⟩ × ⟨7⟩ / ⟨3⟩ × ⟨5⟩ / ⟨3⟩ × ⟨8⟩

Table 4(a) and (b) show the distribution of the numbers of candidate PUs and the position of the correct modified PUs obtained from the analysis of the EDR corpus and the KTB corpus respectively. As we can see, the first candidate is preferred in both languages, but the preference of the nearer candidate is stronger in Korean. For instance, when there are more than one candidates, the probability that the first candidate is the correct one is 78% for Korean but 71% for Japanese. Further, when there are more than 2 candidates, Japanese prefers the last candidate, while Korean prefers the second candidate. Based on the analysis results, the number of modification candidates is restricted to at most three (the first, second and last candidates) for Korean as well.

The next step is to design Φ_u and $\Psi_{c_{un}}$, which are required in Equations (1) and (2) to choose the correct modified PU. We converted the feature set from the Japanese study to get the Korean features as listed in Table 5. For example, to find the modified PU of e_1 "아내가" *anae-ga* ('wife-NOM') in the sentence shown in Figure 1, the attributes of e_1 and the attributes of the three candidates, e_2, e_7, and e_8, are extracted as shown in Figure 3, and their attributes are used to estimate the probability of each candidate in Equation (2).

5 Adaptation for Bilingual Transfer Learning

In Section 4.3, we explained how the *Triplet/Quadruplet Model* can be used for Korean. In this section, we describe the feature adaption techniques in more detail and investigate if the new model with transferred features works well when a small amount of annotated corpus for the target language is provided. Further, we study if we can leverage the annotated corpus for the source language in addition to the parsing model and train a model for the target language using the training data for the source language.

5.1 Feature Transfer

With the assumption that Korean and Japanese have similar syntactic dependencies, we adopt the delexicalized parsing model presented in McDonald et al. (2011). We transfer the part-of-speech (POS) in the Japanese features to the POS scheme in the KTB corpus, and translate Japanese functional words to the corresponding functional words in Korean. This transfer process is mandatory because we use the language specific POS systems to capture language-specific dependency phenomena, unlike other works using language universal but coarser POS systems.

We do not transfer lexical knowledge on con-

Table 6: Example of mapping rules for parts-of-speech and functional words.

Japanese PoS		Korean PoS
common noun		NNC
verb		VV
adjective		VJ
nominal suffix		XSF
case particle	"で","に","へ","から"	PAD
	others	PCA

Japanese particle		Korean particle
"を" *wo* ('-ACC')		"을" *eul*
"より" *yori* ('from')		"부터" *buteo*
"は" *ha* ('-TOPIC')		"은" *eun*
"も" *mo* ('too')		"도" *do*
"が" *ga*	case particle ('-NOM')	"이" *i*
	conjunctive particle ('but')	"지만" *jiman*

tent words and exceptional cases, so feature sets ⟨4⟩ and ⟨6⟩ are not transferred. Table 6 shows some examples of the feature transfer which handle POS tags and functional words. We note that the Korean features shown in Figure 3 are directly extracted from Japanese corpus using those rules.

5.2 Adaptation of parsing rules

While Japanese and Korean are similar in terms of syntactic dependencies, there are significant differences between the two languages in the distribution of modification as shown in the Table 4(a) and (b): In Korean, more than half of modifiers have 5 or more candidates, while only 12% of Japanese modifiers do. In Japanese, 98.6% of correct modified PUs are located in one of the three positions (1st, 2nd or last), but, in Korean, the ratio falls to 93.9% as shown in Table 4. Another major difference of the two languages is the different average numbers of PUs per sentence as shown in Table 1. Korean has 25.5 PUs per sentence, while the number is only 8.5 in Japanese. This is mainly caused by the difference between *eojeol* in Korean and *bunsetsu* in Japanese. In Japanese, compound nouns and verb phrases with an auxiliary verb are likely to form a single *bunsetsu*, while, in Korean, they are split into multiple *eojeol*s with a whitespace in-between.

These differences significantly reduce the effect of transfer learning. To address these problems, we further refine the grammar rules as in the following. We added heuristic rules for the Korean model to effectively reduce the number of candidates in compound nouns which consist of a noun sequence in multiple *eojeol*s, and verbs or adjectives followed by auxiliary verbs. Figure 4 shows an algorithm to reduce the number of modified PUs considering the structure of compound nouns. In this example, both PUs e_4 ("travel") and e_5 ("bag-ACC") can be candidate PUs for *eojeol* e_3. However, based on the rule in Figure 4, e_4 and e_5 are considered as a compound noun (line 1),

and e_4 is determined to modify e_5 (line 3). Subsequently, e_3's modifiability to e_4 is rejected (line 5), and, thus, the correct modified PU of e_3 is determined as e_5. After refining the rules for compound nouns and auxiliary verbs, the probability of the correct modified PU being the 1st, 2nd or last candidate PU increases from 93.9% to 96.3% as shown in Table 7, and the distribution of the candidate's positions for Korean became closer to the Japanese distribution shown in Table 4(a).

5.3 Learning from heterogeneous bilingual corpora

The feature transfer and rule adaptation methods described in previous sections generate a very accurate Korean parser using only a Japanese corpus as shown in the first row in Table 8. The next question is if we can leverage bilingual corpora to further improve the accuracy, when annotated corpus for the target language (Korean) is available. We note that the two corpora do not need to be aligned and can come from different domains. To mitigate the side effects of merging heterogeneous training data in different languages, we apply the domain adaptation method proposed by Daumé III (2007) and augment the feature set to a source language-specific version, a target-specific version and a general version. Specifically, a feature set x in Table 5 is expanded as follows:

$$x_K = < x, 0, x > \tag{5}$$
$$x_J = < 0, x, x > \tag{6}$$

where x_K and x_J denote the feature sets extracted from the Korean corpus and the Japanese corpus respectively. Then, the features specific to Korean and Japanese get higher weights for the first part or the second part respectively, and the characteristics existing in both languages influence the last part.

| if PoS of u_i's last morpheme is N* **and** PoS of u_{i+1}'s first morpheme is N* |
| **then** |
| u_i must modify u_{i+1} |
| **if** u_{i-1}'s last morpheme is **not** "의" **then** u_{i-1} cannot modify u_{i+1} |
| **else** u_{i-1} cannot modify u_i |
| u_1 to u_{i-2} cannot modify u_i |

프랑스의
프랑스/NPR 의/PAN
'France-GEN'
e_3

여행
여행/NNC
'travel'
e_4

가방을
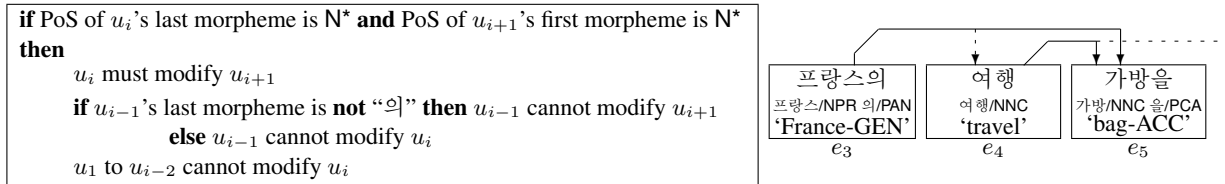가방/NNC 을/PCA
'bag-ACC'
e_5

Figure 4: Heuristic rules to reduce the number of modification candidates surrounding compound nouns in Korean. The example in the right figure shows that candidates in the dotted lines are removed by the heuristics.

Table 7: Distribution of the position of correct modified PU for Korean after the refinement of the Korean grammar rules.

# of candidates	Ratio	1st	2nd	Last	Sum
1	46.4%	100.0%	–	–	100.0%
2	9.8%	79.0%	21.0%	–	100.0%
3	9.2%	75.5%	12.7%	11.8%	100.0%
4	8.0%	71.0%	11.8%	9.6%	92.4%
≥ 5	26.6%	70.4%	10.1%	7.8%	88.3%
Total	100%	–	–	–	96.3%

6 Experiments

In this section, we validate the effectiveness of learning a Korean parser using the feature transfer learning from the Japanese parser and compare the Korean model with other baseline cases. We also compare the parsing results when various sizes of bilingual corpora were used to train the Korean model.

6.1 Korean parsing using the *Triplet/Quadruplet Model*

First, to validate the effectiveness of the *Triplet/Quadruplet Model* for parsing Korean, we built eight Korean dependency parsing models using different numbers of training sentences for Korean. The KTB corpus Version 2.0 (Han et al., 2002) containing 5,010 annotated sentences was used in this study. We first divide the corpus into 5 subsets by putting each sentence into its (sentence ID `mod` 5)-th group. We use sentences from the first subgroup for estimating the parameters, sentences from the second subgroup for testing, and use the remaining three subgroups for training. We built 8 models in total, using from 0 sentence up to 3,006 sentences selected from the training set. The number of training sentences in each model is shown in the first column in Table 8. The parameters were estimated by the maximum entropy method, and the most preferable tree is selected using each dependency probability and the beam search. The test data set contains 1,043 sentences.

We compare the *Triplet/Quadruplet Model*-based models with the Distance Model. For the Distance Model, we used the same feature set as in Table 5, and added the distance feature ($\Delta_{u,t}$) by grouping the distance between two PUs into 3 categories (1, 2 to 5, and 6 or more). The performances are measured by UAS (unlabeled attachment score), and the results of the two methods are shown in the second column, where Japanese Corpus Size=0, in Table 8 (a) and (b) respectively. The top leftmost cells (80.61% and 71.63%) show the parsing accuracies without any training corpora. In these cases the nearest candidate PU is selected as the modified PU. The difference between two models suggests the effect of restriction of modification candidates by the grammar rules. We note that the *Triplet/Quadruplet Model* produces more accurate results and outperforms the Distance Model by more than 2 percentage points in all cases. The results confirm that the method for Japanese parsing is suitable for Korean parsing.

6.2 Results of bilingual transfer learning

Next, we evaluate the transfer learning when annotated sentences for Japanese were also added. Table 8(a) shows the accuracies of our model when various numbers of training sentences from Korean and Japanese are used. The first row shows the accuracies of Korean parsing when the models were trained only with the Japanese corpus, and

Table 8: The accuracy of Korean dependency parsing with various numbers of annotated sentences in the two languages. † denotes that the mixture of bilingual corpora significantly outperformed ($p < .05$) the parser trained with only the Korean corpus without Japanese corpus.

(a) Triplet/Quadruplet Model

		Japanese Corpus Size			
		0	2,500	5,000	10,000
Korean Corpus Size	0	80.61%	80.78%	81.23% †	81.58% †
	50	82.21%	82.32%	82.40% †	82.43% †
	98	82.36%	82.66% †	82.69% †	82.70% †
	197	83.13%	83.18%	83.30% †	83.28%
	383	83.62%	83.92% †	83.94% †	83.91% †
	750	84.03%	84.00%	84.06%	84.06%
	1,502	84.41%	84.34%	84.32%	84.28%
	3,006	84.77%	84.64%	84.64%	84.65%

(b) Distance Model

		Japanese Corpus Size		
		0	2,500	5,000
Korean Corpus Size	0	71.63%	62.42%	54.92%
	50	79.31%	79.55% †	79.54% †
	98	80.53%	80.63% †	80.72% †
	197	80.91%	80.84%	80.85%
	383	81.86%	81.75%	81.76%
	750	82.10%	81.92%	81.94%
	1,502	82.50%	82.48%	82.50%
	3,006	82.66%	82.57%	82.54%

other rows show the results when the Korean and Japanese corpora were mixed using the method described in Section 5.3.

As we can see from the results, the benefit of transfer learning is larger when the size of the annotated corpus for Korean (*i.e.*, target language) is smaller. In our experiments with Triplet/Quadruplet Model, positive results were obtained by the mixture of the two languages when the Korean corpus is less than 500 sentences, that is, the annotations in the source language successfully compensated the small corpus of the target language. When the size of the Korean corpus is relatively large ($\geq 1,500$ sentences), adding the Japanese corpus decreased the accuracy slightly, due to syntactic differences between the two languages. Also the effect of the corpus from the source language tends to saturate as the size of the source corpus, when the target corpus is larger. This is mainly because our mapping rules ignore lexical features, so few new features found in the larger corpus were incorrectly processed.

When merging the corpus in two languages, if we simply concatenate the transferred features from the source language and the features from the target language (instead of using the duplicated features shown in Equations (5) and (6)), the accuracy dropped from 82.70% to 82.26% when the Korean corpus size was 98 and Japanese corpus size was 10,000, and from 83.91% to 83.40% when Korean=383. These results support that there are significant differences in the dependencies between two languages even if we have im-

proved the feature mapping, and our approach with the domain adaptation technique (Daumé III, 2007) successfully solved the difficulty.

Table 8(b) shows the results of the Distance Model. As we can see from the first row, using only the Japanese corpus did not help the Distance Model at all in this case. The Distance Model was not able to mitigate the differences between the two languages, because it does not use any grammatical rules to control the modifiability. This demonstrates that the hybrid parsing method with the grammar rules makes the transfer learning more effective. On the other hand, the domain adaptation method described in (5) and (6) successfully counteracted the contradictory phenomena in the two languages and increased the accuracy when the size of the Korean corpus was small (size=50 and 98). This is because the interactions among multiple candidates which cannot be captured from the small Korean corpus were provided by the Japanese corpus.

Some of previous work reported the parsing accuracy with the same KTB corpus; 81% with trained grammar (Chung et al., 2010) and 83% with Stanford parser after corpus transformation (Choi et al., 2012), but as Choi et al. (2012) noted it is difficult to directly compare the accuracies.

6.3 Discussion

The analysis of e_2's dependency in Figure 1 is a good example to illustrate how the *Triplet/Quadruplet Model* and the Japanese corpus help Korean parsing. *Eojeol* e_2 has three modification candidates, e_3, e_5, and e_6. In the

Distance Model, e_3 is chosen because the distance between the two *eojeol*s (Δ_{e_2,e_3}) was 1, which is a very strong clue for dependency. Also, in the *Triplet/Quadruplet Model* trained only with a small Korean corpus, e_3 received a higher probability than e_5 and e_6. However, when a larger Japanese corpus was combined with the Korean corpus, e_5 was correctly selected as the Japanese corpus provided more samples of the dependency relation of "verb-PAST" (e_2) and "common noun-ACC (을)" (e_5) than that of "verb-PAST" and "proper noun-GEN (의)" (e_3).

As we can notice, larger contextual information is required to make the right decision for this case, which may not exist sufficiently in a small corpus due to data sparseness. The grammar rules in the *Triplet/Quadruplet Model* can effectively capture such contextual knowledge even from a relatively small corpus. Further, since the grammar rules are based only on part-of-speech tags and a small number of functional words, they are similar to the delexicalized parser (McDonald et al., 2011). These delexicalized rules are more robust to linguistic idiosyncrasies, and, thus, are more effective for transfer learning.

7 Conclusion

We presented a new dependency parsing algorithm for Korean by applying transfer learning from an existing parser for Japanese. Unlike other transfer learning methods relying on aligned corpora or bilingual lexical resources, we proposed a *feature transfer* method utilizing a small number of hand-crafted grammar rules that exploit syntactic similarities of the source and target languages. Experimental results confirm that the features learned from the Japanese training corpus were successfully applied for parsing Korean sentences and mitigated the data sparseness problem. The grammar rules are mostly delexicalized comprising only POS tags and a few functional words (*e.g.*, case markers), and some techniques to reduce the syntactic difference between two languages makes the transfer learning more effective. This methodology is expected to be applied to any two languages that have similar syntactic structures, and it is especially useful when the target language is a low-resource language.

References

Jinho D. Choi and Martha Palmer. 2011. Statistical dependency parsing in Korean: From corpus generation to automatic parsing. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages*, pages 1–11.

DongHyun Choi, Jungyeul Park, and Key-Sun Choi. 2012. Korean treebank transformation for parser training. In *Proceedings of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages*, pages 78–88.

Hoojung Chung and Heechang Rim. 2003. A new probabilistic dependency parsing model for head-final, free word order languages. *IEICE TRANSACTIONS on Information and Systems*, E86-1(11):2490–2493.

Tagyoung Chung, Matt Post, and Daniel Gildea. 2010. Factors affecting the accuracy of korean parsing. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 49–57.

Michael Collins. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics*.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263.

Greg Durrett, Adam Pauls, and Dan Klein. 2012. Syntactic transfer using a bilingual lexicon. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1–11.

EDR. 1996. EDR (Japan Electronic Dictionary Research Institute, Ltd.) electronic dictionary version 1.5 technical guide.

David A. Ferrucci, Eric W. Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John M. Prager, Nico Schlaefer, and Christopher A. Welty. 2010. Building Watson: An overview of the DeepQA project. *AI Magazine*, 31(3):59–79.

Ryan Georgi, Fei Xia, and William D Lewis. 2012. Improving dependency parsing with interlinear glossed text and syntactic projection. In *Proceedings of COLING 2012*, pages 371–380.

Raymond G Gordon and Barbara F Grimes. 2005. *Ethnologue: Languages of the world*, volume 15. SIL international Dallas, TX.

Chung-hye Han, Na-Rae Han, Eon-Suk Ko, Heejong Yi, and Martha Palmer. 2002. Penn Korean treebank: Development and evaluation. In *Proc. Pacific Asian Conf. Language and Comp.*

Rebecca Hwa, Philip Resnik, and Amy Weinberg. 2005. Breaking the resource bottleneck for multilingual parsing. Technical report, DTIC Document.

Hiroshi Kanayama, Kentaro Torisawa, Yutaka Mitsuishi, and Jun'ichi Tsujii. 2000. A hybrid Japanese parser with hand-crafted grammar and statistics. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 411–417.

Roger Kim, Mary Dalrymple, Ronald M Kaplan, and Tracy Holloway King. 2003a. Porting grammars between typologically similar languages: Japanese to korean. In *Proceedings of the 17th Pacific Asia Conference on Language, Information.*

Roger Kim, Mary Dalrymple, Ronald M Kaplan, Tracy Holloway King, Hiroshi Masuichi, and Tomoko Ohkuma. 2003b. Multilingual grammar development via grammar porting. In *ESSLLI 2003 Workshop on Ideas and Strategies for Multilingual Grammar Development*, pages 49–56.

Dan Klein and Christopher D Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 478–487.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. Dependency parsing. *Synthesis Lectures on Human Language Technologies*, 1(1):1–127.

Cody Kwok, Oren Etzioni, and Daniel S Weld. 2001. Scaling question answering to the web. *ACM Transactions on Information Systems (TOIS)*, 19(3):242–262.

Hyeon-Yeong Lee, Yi-Gyu Hwang, and Yong-Seok Lee. 2007. Parsing of Korean based on CFG using sentence pattern information. *International Journal of Computer Science and Network Security*, 7(7).

Roger Levy and Christopher Manning. 2003. Is it harder to parse chinese, or the chinese treebank? In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, ACL, pages 439–446.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 62–72.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of ACL 2013*.

Tahira Naseem, Harr Chen, Regina Barzilay, and Mark Johnson. 2010. Using universal linguistic knowledge to guide grammar induction. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1234–1244.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 629–637.

Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the Second International Conferences on Knowledge Capture*, pages 70–77.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing (EMNLP)*, pages 79–86, Philadelphia, Pennsylvania.

Jungyeul Park, Daisuke Kawahara, Sadao Kurohashi, and Key-Sun Choi. 2013. Towards fully lexicalized dependency parsing for Korean. In *Proceedings of The 13th International Conference on Parsing Technologies*.

Yoav Seginer. 2007. Fast unsupervised incremental parsing. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 384–391.

David A. Smith and Noah A. Smith. 2004. Bilingual parsing with factored estimation: Using English to parse Korean. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 49–56.

Valentin I Spitkovsky, Hiyan Alshawi, Angel X Chang, and Daniel Jurafsky. 2011. Unsupervised dependency parsing without gold part-of-speech tags. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1281–1290.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *IJCNLP*, pages 35–42.