LINGUISTICS 7800

Statistical Dependency Parsing Korean:

From Corpus Generation To Automatic Parsing

By Jinho D. Choi and Martha Palmer

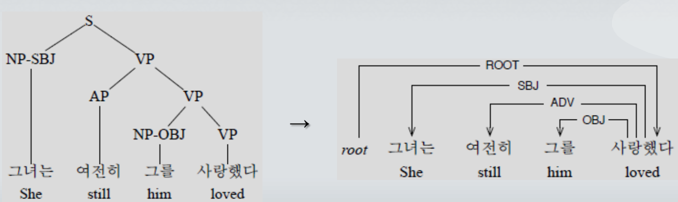
DEPARTMENT OF LINGUISTICS
April 4, 2013
SEUNG HAN LEE

Issues in this paper

- Less work done on **dependency parsing** in Korean because of the lack of training data in dependency structure.
- **Convert constituent** Treebank in Korean to **dependency** Treebank
  - by applying head-percolation rules and heuristics
- How to **extract** useful **features** from morphologically rich Korean
  - ∠∴™ □ ∠/NNG + ∴/XSV + ™/EF
    talk (verb) talk(noun) do ending marker
- Parsing **evaluation**
  - three different genres with gold-standard & automatic morphological analysis
  - impact of fine vs. coarse-grained morphologies on dependency parsing

Dependency Tree

- **Dependency tree** from constituent trees



In Dependency Treebank and Parsing

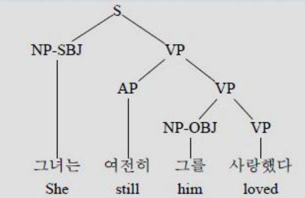
- **No restriction** on word-order unlike phrase structure
- Suitable for **flexible** word-order and **morphologically** rich languages
  - **Korean** (SVO, but free order with case particles)
- For Korean dependency parsing, use **Sejong** constituent Treebank (60K sentences)

Related Work

- **Stanford dependencies**
  - system for extracting dependencies from Penn Treebank style constituent trees (Marnefee et al. 2006)
- Penn Korean Treebank
  - constituent trees for newswire and military corpora (Han et al., 2002)
- KAIST tree-annotated corpus (Lee 1998)

Dependency Conversion: Sejong Treebank

- **Constituent** tree and **morphological** analysis for *'she still loved him'* in Korea



그녀는 → 그녀(she)/NP+는/JX
여전히 → 여전히(still)/MAG
그를 → 그(he)/NP+를/JKO
사랑했다 → 사랑(love)/NNG+하/XSV+았/EP+다/EF

Dependency Conversion: Sejong Treebank

- **POS tags** are used in **morphemes** within tokens

NNG	General noun	MM	Adnoun	EP	Prefinal EM	JX	Auxiliary PR
NNP	Proper noun	MAG	General adverb	EF	Final EM	JC	Conjunctive PR
NNB	Bound noun	MAJ	Conjunctive adverb	EC	Conjunctive EM	IC	Interjection
NP	Pronoun	JKS	Subjective CP	ETN	Nominalizing EM	SN	Number
NR	Numeral	JKC	Complemental CP	ETM	Adnominalizing EM	SL	Foreign word
VV	Verb	JKG	Adnomial CP	XPN	Noun prefix	SH	Chinese word
VA	Adjective	JKO	Objective CP	XSN	Noun DS	NF	Noun-like word
VX	Auxiliary predicate	JKB	Adverbial CP	XSV	Verb DS	NV	Predicate-like word
VCP	Copula	JKV	Vocative CP	XSA	Adjective DS	NA	Unknown word
VCN	Negation adjective	JKQ	Quotative CP	XR	Base morpheme	SF, SP, SS, SE, SO, SW	

Table 1: POS tags in the Sejong Treebank (PM: predicate marker, CP: case particle, EM: ending marker, DS: derivational suffix, PR: particle, SF SP SS SE SO: different types of punctuation).

Dependency Conversion: Sejong Treebank

- Tree consists of various **phrasal nodes** and **function tags**.
  - **each token** is annotated with a **phrasal-level** tag.
  - **function tags,** relations between phrases and siblings, can be used **as dependency labels**.

Phrase-level tags		Function tags	
S	Sentence	SBJ	Subject
Q	Quotative clause	OBJ	Object
NP	Noun phrase	CMP	Complement
VP	Verb phrase	MOD	Noun modifier
VNP	Copula phrase	AJT	Predicate modifier
AP	Adverb phrase	CNJ	Conjunctive
DP	Adnoun phrase	INT	Vocative
IP	Interjection phrase	PRN	Parenthetical

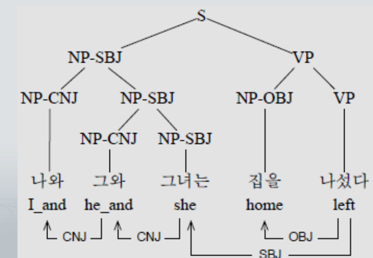
Dependency Conversion: head-percolation

- **Head-percolation rules** in Sejong Treebank
  - **find** the **head** of the phrase and make its **dependent**
  - **generate dependency trees** from constituent trees and **guarantee** dependency trees **well-formed** (root, head, connected, acyclic)

S	r	VP;VNP;S;NP|AP;Q;*
Q	l	S|VP|VNP|NP;Q;*
NP	r	NP;S;VP;VNP;AP;*
VP	r	VP;VNP;NP;S;IP;*
VNP	r	VNP;NP;S;*
AP	r	AP;VP;NP;S;*
DP	r	DP;VP;*
IP	r	IP;VNP;*
X|L|R	r	*

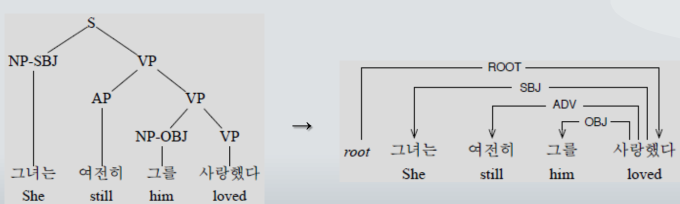
Dependency Conversion: heuristics

- **Heuristics**
  - resolve some **special** cases (e.g., coordination)
  - constituent and dependency trees for '*I and he and she left home*'
  - *she* is the **head** of both *I* and *he*.



Dependency Conversion: dependency labels

- **Dependency labels** from constituent trees
  - **function tag becomes the dependency label** to it head.



Dependency Conversion: dependency labels

- **Algorithm 1** shows how to **infer the other labels.**
  - ROOT is the dependency label of the root node.
  - ADV is adverbials.
  - (A|D|N|V) MOD are (adverb, adnoun, noun, verb) modifiers.

input : (c, p), where c is a dependent of p.
output: A dependency label l as $c \xleftarrow{l} p$.
begin
  if $p = root$ **then** ROOT → l
  elif c.pos = AP **then** ADV → l
  elif p.pos = AP **then** AMOD → l
  elif p.pos = DP **then** DMOD → l
  elif p.pos = NP **then** NMOD → l
  elif p.pos = VP | VNP | IP **then** VMOD → l
  else DEP → l
end

Algorithm 1: Getting inferred labels.

Morphological Analyzers: IMA and Mach

- **Two systems** to generate automatic morphemes and POS tags
  - Intelligent Morphological Analyzer (**IMA**): fine-grained & rich POS tag
  - **Mach** (Shim & Yang 2002): coarse-grained POS tag
  - **mapping between POS tags** generated by two systems for comparing the impact of fine vs. coarse grained morphologies

NN	→	NNG | NNP | SL | SH	VX	→	VX (verb)	SN	→	XSN
NX	→	NNB	AX	→	VX (adjective)	SV	→	XSV
NP	→	NP	DT	→	MM	SJ	→	XSA
NU	→	NR | SN	AD	→	MA⋆	IJ	→	IC
VI	→	VV (intransitive)	JO	→	J⋆	NR	→	NF
VT	→	VV (transitive)	EP	→	EP	UK	→	NA | NV | XR
AJ	→	VA | VCN	EM	→	EF | EC | ET⋆	SY	→	SF | SP | SS | SE | SO | SW
CP	→	VCP	PF	→	XPN			

Table 5: Mappings between POS tags generated by Mach and IMA. In each column, the left-hand and right-hand sides show POS tags generated by Mach and IMA, respectively.

Dependency Parsing

- Parsing algorithm
  - Transition-based dependency parsing approach (Choi and Palmer 2011)
- Machine learning alogrithm
  - Liblinear L2-regularized L1-loss SVM

Dependency Parsing

- Feature extraction
  - extract **features** from **POS tags**
  - **some types of morphemes** used to extract features for dependency parsing models

FS	The first morpheme
LS	The last morpheme before JO | DS | EM
JK	Particles (J⋆ in Table 1)
DS	Derivational suffixes (XS⋆ in Table 1)
EM	Ending markers (E⋆ in Table 1)
PY	The last punctuation, only if there is no other morpheme followed by the punctuation

Dependency Parsing

- Feature extraction example
  - the types of morphemes extracted from the tokens

낙랑공주는 → 낙랑/NNP+공주/NNG+는/JX
Nakrang + Princess + JX
호동왕자를 → 호동/NNP+왕자/NNG+를/JKO
Hodong + Prince + JKO
사랑했다. → 사랑/NNG+하/XSV+았/EP+다/EF+./SF
Love + XSV + EP + EF + .

→

FS	LS	JK	DS	EM	PY
낙랑/NNP	공주/NNG	는/JX	-	-	-
호동/NNP	왕자/NNG	를/JKO	-	-	-
사랑/NNG	-	-	하/XSV	다/EF	./SF

Experiments: corpora

- Grouping Sejong corpora into **6 genres**
  - Newspaper (NP), Magazine (MZ), Fiction (FI), Memoir (ME), Informative Book(IB), Educational Cartoon (EC)
  - These corpora are divided into training(T), development(D), and evaluation sets(E) which **ensures** the robustness of **parsing model**.

	NP	MZ	FI	ME	IB	EC
T	8,060	6,713	15,646	5,053	7,983	1,548
D	2,048	-	2,174	-	1,307	-
E	2,048	-	2,175	-	1,308	-

Table 10: Number of sentences in training (T), development (D), and evaluation (E) sets for each genre.

Experiments: evaluations

- Parsing model evaluation based on **gold-standard** morphology[gold, fine-grained], **IMA** [auto, fine-grained], and **Mach** [auto, coarse-grained]

	Gold, fine-grained			Auto, fine-grained			Auto, coarse-grained		
	LAS	UAS	LS	LAS	UAS	LS	LAS	UAS	LS
NP	82.58	84.32	94.05	79.61	82.35	91.49	79.00	81.68	91.50
FI	84.78	87.04	93.70	81.54	85.04	90.95	80.11	83.96	90.24
IB	84.21	85.50	95.82	80.45	82.14	92.73	81.43	83.38	93.89
Avg.	**83.74**	**85.47**	**94.57**	**80.43**	**83.01**	**91.77**	**80.14**	**82.89**	**91.99**

Table 11: Parsing accuracies achieved by three models (in %). LAS - labeled attachment score, UAS - unlabeled attachment score, LS - label accuracy score

- On the average LAS, [gold, fine-grained] better than [auto, fine-grained]
- [auto, fine-grained] has a POS tagging accuracy of 94.66% on correctly segmented morphemes.

Experiments: evaluations

	Gold, fine-grained			Auto, fine-grained			Auto, coarse-grained		
	LAS	UAS	LS	LAS	UAS	LS	LAS	UAS	LS
NP	82.58	84.32	94.05	79.61	82.35	91.49	79.00	81.68	91.50
FI	84.78	87.04	93.70	81.54	85.04	90.95	80.11	83.96	90.24
IB	84.21	85.50	95.82	80.45	82.14	92.73	81.43	83.38	93.89
Avg.	**83.74**	**85.47**	**94.57**	**80.43**	**83.01**	**91.77**	**80.14**	**82.89**	**91.99**

Table 11: Parsing accuracies achieved by three models (in %). LAS - labeled attachment score, UAS - unlabeled attachment score, LS - label accuracy score

- The difference between [auto, fine-grained] and [auto, coarse-grained] models are small; 'a more fine-grained morphology is not necessarily a better morphology for dependency parsing'.
- High LS implies that models successfully learn labeling information from morphemes.
- Models perform worse on NP genre, and this needs to improve accuracy.

Project

- Based on Sejong Treebank
  - Make the dependency labels in Korean **more rich**
  - **Compare** dependency labels in English with ones in Korean
    AGENT, CSUBJ, CSUBJPASS, EXPL, NSUBJ, NSUBJPASS, ATTR, DOBJ, IOBJ, OPRD, AUX, AUXPASS, HMOD, HYPH, ACOMP, CCOMP, XCOMP, COMPLM, ADVCL, ADVMOD, etc.
  - Find the **rules and morphemes** for generating the dependency labels in Korean