Chinese Parsing and Grammatical Relations

Yuchen Liao
Universität des Saarlandes
WS18-19
Typology of Morphosyntax
Prof. Dr. Avgustinova

Outline

- Introduction
- Grammatical Relations
- Models
- Results and Conclusions
- References

Outline

- **Introduction**
- Grammatical Relations
- Models
- Results and Conclusions
- References

Introduction

- **The Penn Chinese Treebank (CTB)**
- A factored-model statistical parser for it shows the implications of differences between Wall Street Journal (WSJ) and CTB.
- Parse errors
- Difficult ambiguities inhenrent in Chinese Grammer
- Treebank-derived CFG : more linguistic ambiguities, genuine and artificial.
- Corpus-based statistical parsing : leading technique to deal with it, using the WSJ section of the English Penn Treebank (ETB).
- Different in Chinese : linguistic, tree-structure

Introduction

- **Translation difficulty**
- A richer set of Chinese grammatical relations between words
- apply the log probability of the phrase orientation classifier as an extra feature in a phrase-based MT system
- Chinese grammatical relations : useful for other NLP tasks.
- Major factor in the difficulty of MT from Chinese to English : Structural differences including

--the ordering of head nouns and relative clauses

--the ordering of prepositional phrases and the heads they modify.

Outline

- Introduction
- **Grammatical Relations**
- Models
- Results and Conclusions
- References

Grammatical Relations

- **Chinese grammatical relations** : designed to be very similar to the **Stanford English typed dependencies**
- Chinese specific structures

--e.g. the usage of 的(DE) : lead to different English translations

--cpm (DE as complementizer) or assm (DE as associative marker)

- The typed dependencies

--annotate these Chinese specific relations

--not provide a mapping onto how they are translated into English.

Grammatical Relations

- **Comparison with English**
- Chinese has more nn, punct, nsubj, rcmod, dobj, advmod, conj, nummod, attr, tmod, and ccomp
- English uses more pobj, det, prep, amod, cc, cop, and xsubj,
- Due to **grammatical differences** between Chinese and English
- E.g. some determiners in English are not mandatory in Chinese

  进出口/import and export总额/total value

  The total value of imports and exports

abbreviation	short description	Chinese example	typed dependency	counts	percentage
nn	noun compound modifier	服务 中心	nn(中心, 服务)	13278	15.48%
punct	punctuation	海关 统计 表明 ，	punct(表明, ，)	10896	12.71%
nsubj	nominal subject	梅花 盛开	nsubj(盛开, 梅花)	5893	6.87%
conj	conjunct (links two conjuncts)	设备 和 原材料	conj(原材料, 设备)	5438	6.34%
dobj	direct object	浦东 颁布 了 七十一 件 文件	dobj(颁布, 文件)	5221	6.09%
advmod	adverbial modifier	部门 先 送上 文件	advmod(送上, 先)	4231	4.93%
prep	prepositional modifier	在 实践 中 逐步 完善	prep(完善, 在)	3138	3.66%
nummod	number modifier	七十一 件 文件	nummod(件, 七十一)	2885	3.36%
amod	adjectival modifier	跨世纪 工程	amod(工程, 跨世纪)	2691	3.14%
pobj	prepositional object	根据 有关 规定	pobj(根据, 规定)	2417	2.82%
rcmod	relative clause modifier	不 曾 遇到 过 的 情况	rcmod(情况, 遇到)	2348	2.74%
cpm	complementizer	开发 浦东 的 经济 活动	cpm(开发, 的)	2013	2.35%
assm	associative marker	企业 的 商品	assm(企业, 的)	1969	2.30%
assmod	associative modifier	企业 的 商品	assmod(商品, 企业)	1941	2.26%
cc	coordinating conjunction	设备 和 原材料	cc(原材料, 和)	1763	2.06%
clf	classifier modifier	七十一 件 文件	clf(文件, 件)	1558	1.82%
ccomp	clausal complement	银行 决定 先 取得 信用 评级	ccomp(决定, 取得)	1113	1.30%
det	determiner	这些 经济 活动	det(活动, 这些)	1113	1.30%
lobj	localizer object	近年 来	lobj(来, 近年)	1010	1.18%
range	dative object that is a quantifier phrase	成交 药品 一亿多 元	range(成交, 元)	891	1.04%
asp	aspect marker	发挥 了 作用	asp(发挥, 了)	857	1.00%
tmod	temporal modifier	以前 不 曾 遇到 过	tmod(遇到, 以前)	679	0.79%
plmod	localizer modifier of a preposition	在 这 片 热土 上	plmod(在, 上)	630	0.73%
attr	attributive	贸易额 为 二百亿 美元	attr(为, 美元)	534	0.62%
mmod	modal verb modifier	利益 能 得到 保障	mmod(得到, 能)	497	0.58%
loc	localizer	占 九成 以上	loc(占, 以上)	428	0.50%
top	topic	建筑 是 主要 活动	top(是, 建筑)	380	0.44%
pccomp	clausal complement of a preposition	据 有关 部门 介绍	pccomp(据, 介绍)	374	0.44%
etc	etc modifier	科技 、 文教 等 领域	etc(文教, 等)	295	0.34%
lccomp	clausal complement of a localizer	中国 对 外 开放 中 升起 的 明星	lccomp(中, 开放)	207	0.24%
ordmod	ordinal number modifier	第七 个 机构	ordmod(个, 第七)	199	0.23%
xsubj	controlling subject	银行 决定 先 取得 信用 评级	xsubj(取得, 银行)	192	0.22%
neg	negative modifier	以前 不 曾 遇到 过	neg(遇到, 不)	186	0.22%
rcomp	resultative complement	研究 成功	rcomp(研究, 成功)	176	0.21%
comod	coordinated verb compound modifier	颁布 实行	comod(颁布, 实行)	150	0.17%
vmod	verb modifier	其 在 支持 外商 企业 方面 的 作用	vmod(方面, 支持)	133	0.16%
prtmod	particles such as 所, 以, 来, 而	在 产业化 所 取得 的 成就	prtmod(取得, 所)	124	0.14%
ba	"ba" construction	把 注意力 转向 市场	ba(转向, 把)	95	0.11%
dvpm	manner DE(地) modifier	有效 地 防止 流失	dvpm(有效, 地)	73	0.09%
dvpmod	a "XP+DEV(地)" phrase that modifies VP	有效 地 防止 流失	dvpmod(防止, 有效)	69	0.08%
prnmod	parenthetical modifier	八五 期间 （ 1990 – 1995 ）	prnmod(期间, 1995)	67	0.08%
cop	copular	原 是 自给自足 的 经济	cop(自给自足, 是)	59	0.07%
pass	passive marker	被 认定 为 高 技术 产业	pass(认定, 被)	53	0.06%
nsubjpass	nominal passive subject	镍 被 称作 现代 工业 的 维生素	nsubjpass(称作, 镍)	14	0.02%

Table 2: Chinese grammatical relations and examples. The counts are from files 1–325 in CTB6.

Shared relations	Chinese	English
nn	15.48%	6.81%
punct	12.71%	9.64%
nsubj	6.87%	4.46%
rcmod	2.74%	0.44%
dobj	6.09%	3.89%
advmod	4.93%	2.73%
conj	6.34%	4.50%
num/nummod	3.36%	1.65%
attr	0.62%	0.01%
tmod	0.79%	0.25%
ccomp	1.30%	0.84%
xsubj	0.22%	0.34%
cop	0.07%	0.85%
cc	2.06%	3.73%
amod	3.14%	7.83%
prep	3.66%	10.73%
det	1.30%	8.57%
pobj	2.82%	10.49%

Table 1: The percentage of typed dependencies in files 1–325 in Chinese (CTB6) and English (English-Chinese Translation Treebank)

Grammatical Relations

- another difference : e.g.

--English uses adjectives (amod) to modify a noun

--Chinese can use noun compounds

西藏/Tibet 金融/finance体制/system 改革/reform

the reform in Tibet ’s financial system

Grammatical Relations

- More specific examples such as:

--**prep and pobj** : English has much more uses of prep and pobj

--九七/1997 之后/after

after 1997

--**cc and punct** : The Chinese sentences contain more punctuation (punct) while the English translation has more conjunctions

--这些/these 城市/city 社会/social 经济/economic 发展/development 迅速/rapid， 地方/local 经济/economic 实力/strength 明显/clearly 增强/enhance

In these municipalities the social and economic development has been rapid, and the local economic strength has clearly been enhanced。

Grammatical Relations

- 3 salient linhuistic differences between English and Chinese :

--CH. makes less use of function words and morphology than EN.

--EN. is left-headed and right-branching, CH. is more mixed.

--subject pro-drop

Grammatical Relations

- Tree-Structural Differences between English and Chinese Treebanks.
- CTB annotation – Government-Binding (GB) theory
- 2 differences :

--requires phrasal projection of all categories

particularly prominent with NPs: CTB adj.-noun mod.

--disdinguishes between levels of adjunction and complementation

made only for VP

- The CTB has fewer types than ETB of equivalent size and has lower branching factor.

Grammatical Relations

- The Penn Chinese TreeBank : Phrase structure annotation of a large corpus
- to improve speed while ensuring annotation quality
- proven to be a crucial resource in the recent success of English Part-Of-Speech (POS) taggers and parsers
- Data : mostly newswire and magazine articles from Xinhua newswire, Hong Kong news and the Sinorama magazine
- The structure of the original articles : maintained

Table 6. *Functional tags and null categories used in CTB*

Functional tags				*Null Categories*	
ADV	adverbial	MNR	manner	*pro*	dropped argument
APP	appositive	OBJ	direct object	*PRO*	used in non-finite constructions
BNF	beneficiary	PN	proper noun		
CND	condition	PRD	predicate	*T*	trace of A'-movement
DIR	direction	PRP	purpose or reason	*	trace of A-movement
EXT	extent	Q	question	*RNR*	right node raising
FOC	focus	SBJ	subject	*OP*	operator
HLN	headline	SHORT	short form	*?*	other unknown empty categories
IJ	interjective	TMP	temporal		
IMP	imperative	TPC	topic		
IO	indirect object	TTL	title		
LGS	logical subject	VOC	vocative		
LOC	locative	WH	wh-phrase		

Outline

- Introduction
- Grammatical Relations
- **Models**
- Results and Conclusions
- References

Models

- Factored Parsing model
- Combining two independent parses :

--maximum likelihood estimated (MLE) PCFG model

--constituent-free dependency parse

- Offers the prospect of increased flexibility in tuning the individual parse models.
- Focus : to refine the PCFG model via stepwise refinements informed by major observed ambiguity classes.

Models

- **Discriminative Recording Model** in phrase-based systems
- use linear distance as the cost for phrase movements
- Disadvantage : insensitivity to the content of the words or phrases.
- Data sparseness can make estimation less reliable.
- **Phrase Orientation Classifier** : build up the target language (EN) translation from left to right.
- Predicts the start position of the next phrase in the source sentence.
- Path Features Using Typed Dependencies
- Feature : two words at positions p and q in the Chinese sentence ($p < q$), the shortest path concatenate all the relations on the path

Outline

- Introduction
- Grammatical Relations
- Models
- **Results and Conclusions**
- References

Results and conclusions

- Chinese typed dependencies with information about grammatical relations between words

--to build path features

--to improve a phrase orientation classifier.

- apply the log probability as an additional feature in a phrase-based MT system
- typed dependencies on the source side : informative for the reordering component in a phrase-based system

Results and conclusions

- An encouraging for the use of detailed error analysis followed by focused tree structure enhancements to improved parser performance.
- Two limitations :

--error types are rare in Treebank data.

--common error types : not the result of shortcomings

major sources of error for the parse : coordination scoping ambiguity (in ETB) and N/V tag ambiguity (for CH.).

References

- Roger Levy, Christopher Manning, *Is it harder to parse Chinese, or the Chinese Treebank?*
- Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky and Christopher D. Manning, *Discriminative Recording with Chinese Grammatical Relations Features*
- Naiwen Xue, Fei Xia, Fu-Dong Chiou and Marta Palmer, *The penn chinese treebank : phrase structure annotation of a large corpus*

Thank you for your attention !