CSE/LIN 467/567: Extra assignment report

Jui Kate (50288204)
Vakul Bhatia (50290967)

Abstract

The spanish language has different dependency structure, grammar, syntax from English language. The spanish corpus in penn-treebank format was prepared to generate dependency structure. The dependency structure for English language was known and given. We found these structures for Spanish as well. All executable codes are available at https://github.com/JuiKate/Computational-Linguistics/tree/master/extra-assignment

1 Introduction

We have taken Spanish language to work on. The objective of this project is to make required changes in java files of Stanford parser, build the project, create executable jar file, modifying generate-data-corrected.sh file to use the updated jar file to generate dependency structure for the given input Spanish penn-treebank file.

2 Converting rules

We made changes in the following java files of package edu.stanford.nlp.trees

1. **EnglishPatterns.java** – We identified words which are used in Spanish and which will be detected in the dependency conversion. Copular verbs like "ser","estar" were added which are frequently and commonly used in Spanish. Time related words were added like the days of the week, months

of the year, relative time like today/tomorrow/morning/evening, seasons of the year etc.

2. **SemanticHeadFinder.java** – We added header rules so that it identifies the head nodes and then the dependency structure is understood. We added head rules for tokens like CL/CONP/PRED. Without adding these rules the parser gives error since it cannot identify head rules for these tags.
3. **UniversalSemanticHeadFinder.java** – Rules used in SemanticHeadFinder.java were similarly added in this java file.
4. **EnglishGrammaticalRelations.java** – We made few changes in this file since Spanish has a different grammatical structure compared to English so few grammatical rules were updated as per Spanish grammar.

3 Parsing results

Sample dependency parsing results:

1	Zedillo	_	N	N	_	2	nsubj	_	_
2	asegura	_	V	V	_	0	root	_	_
3	en	_	PREP	PREP	_	2	prep	_	_
4	California	_	N	N	_	3	dep	_	_
5	que	_	C	C	_	13	dep	_	_
6	las	_	ART	ART	_	8	dep	_	_
7	reformas	_	N	N	_	8	dep	_	_
8	politicas	_	ADJ	ADJ	_	13	dep	_	_
9	en	_	PREP	PREP	_	8	prep	_	_
10	México	_	N	N	_	9	dep	_	_
11	no	_	ADV	ADV	_	13	advmod	_	_
12	tienen	_	V	V	_	13	dep	_	_
13	marcha_atrás	_	N	N	_	2	dep	_	_
14	.	_	PUNCT	PUNCT	_	2	dep	_	_

1	EEUU	_	N	N	_	5	nsubj	_	_
2	concederá	_	V	V	_	5	dep	_	_
3	la	_	ART	ART	_	5	dep	_	_
4	residencia	_	N	N	_	5	dep	_	_
5	legal	_	ADJ	ADJ	_	0	root	_	_
6	a	_	PREP	PREP	_	5	prep	_	_
7	medio	_	Q	Q	_	6	dep	_	_
8	millón	_	N	N	_	7	dep	_	_
9	de	_	PREP	PREP	_	7	prep	_	_
10	centroamericanos	_	N	N	_	9	dep	_	_
11	.	_	PUNCT	PUNCT	_	5	dep	_	_

4 Discussion and conclusion

Due to the difference between grammatical structure of English and Spanish languages respective changes are required to form dependency relation. We made the changes as per our understanding and more changes can be made for improving the quality of Spanish parser.

References

de Marneffe, M.-C. and Manning, C. D. (2008). The Stanford Typed Dependencies Representation.
In
Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser
Evaluation
, pages 1–8, Manchester, UK. Coling 2008 Organizing Committee.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The
Stanford CoreNLP Natural Language Processing Toolkit. In
Proceedings of 52nd Annual Meeting
of the Association for Computational Linguistics: System Demonstrations
, pages 55–60, Balti-
more, Maryland. Association for Computational Linguistics.