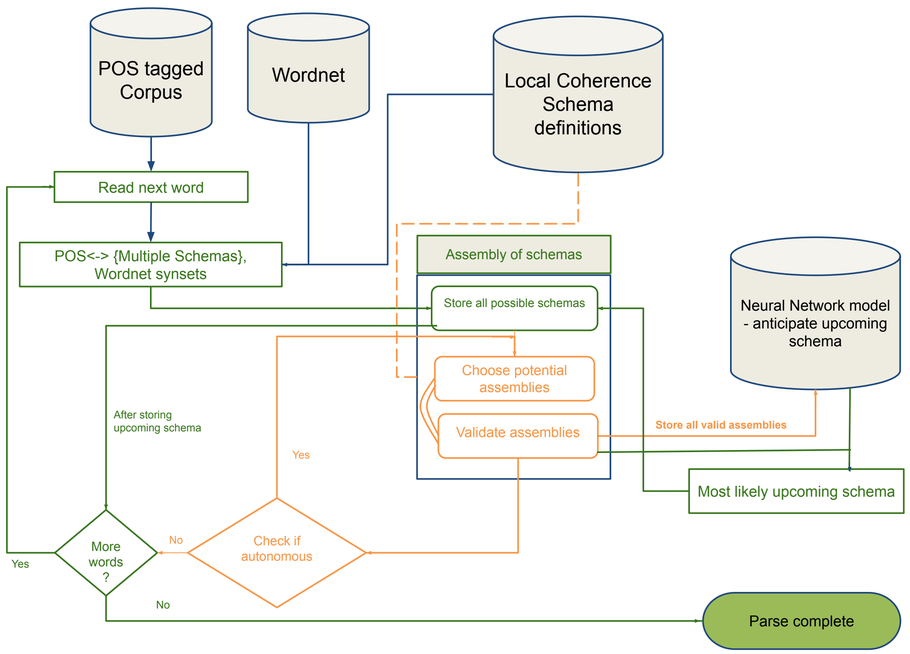
Automatic Grammar Induction from Free Text Using Insights from Cognitive Grammar

Introduction

- Natural Language Parsing – The task of learning the grammatical structure of natural language sentences
- Parsing Approaches – Rule-based, Supervised and Unsupervised statistical learning
- Theory of Grammar - Formal / Generative school, Functional / Cognitive school
- We present a grammar induction system. It learns grammatical structure from raw text with little annotation.
- Parsing - *expect and assemble* mechanism applied on the running text. Schematised usage patterns are recognised, upcoming schemas are expected, running schemas are assembled in meaningful ways.
- Implementation: schema definitions for local coherence - rule-based, anticipation of upcoming schemas - neural network, incremental assembly of partial structures results in final parse

Implementation Overview



Local Coherence - Construction schemas

- Grammatical structure: Meaningful. Sentence: A smaller version of text. Local coherence plays meaningful role.
- Every expression: 3 axes. 1st axis - patterns with what construction: composition; 2nd - how it elaborates or associates with other constructions: interaction; 3rd - whether it is autonomous or dependent
- Most abstract: 7 composition schemas, 6 interaction schemas. THING, RELATION, PRONOUN, PROCESS, STATUS, OPERATOR, EVENT - composition
- CONTINUATIVE, COMBINATIVE, PARTICIPANT, DESCRIPTIVE, QUALIFIER, CLOSED
- Complex grammatical structures arise from locally coherent assembly of these patterns.

Implementation

1. Three components: Schema definition, Schema assembly and Schema prediction.
2. Rules - Schema definition and schema assembly.
3. Neural network - Learn patterns of valid assembly
4. Wordnet information - To improve the learning. Additional to POS tags.
5. POS tags mapped to possible basic schemas - just a practical decision for bootstrapping. In principle, It is possible to implement a fully supervised / unsupervised basic schema tagger.
6. A sentence is passed through all the three components for many iterations until all the words are exhausted -> results in one final construction schema.
7. The order in which all intermediate schemas are assembled to form the final schema can be called as the parse of the sentence.
8. Multiple valid ways to arrive at final schema - No one gold standard to compare against.

Evaluation methodology

- Implementation done for English and Welsh. Project was mainly motivated towards developing a Welsh parser.
- Training - BNC corpus 4500 sentences were taken. Testing: 100 sentences
- No gold standard - So evaluation was done by generating new sentences using the intermediate schemas as templates and human volunteers rated them on a scale of 1 to 5.
- For English, there are constituency treebanks. Comparison of intermediate schema structures against treebank phrase structures and reported the accuracy.

Results

Condition	No. of sentences	Welsh (average score)	English (average score)
Without wordnet	100	To be done	3.5
With wordnet	100	To be done	4

For English alone

Condition	Accuracy
Without wordnet	78.19%
With wordnet	83.26%

References

1. Langacker, R. W. (2008). *Cognitive grammar: A basic introduction*. Oxford University Press.
2. Bod, R. (2009). From exemplar to grammar: A probabilis- tic analogy-based model of language learning. *Cognitive Science*, 33(5):752–793.

Vigneshwaran M Dawn Knight Irena Spasic

Cardiff University