Parse Tree Fragmentation of Ungrammatical Sentences

Homa B. Hashemi, Rebecca Hwa

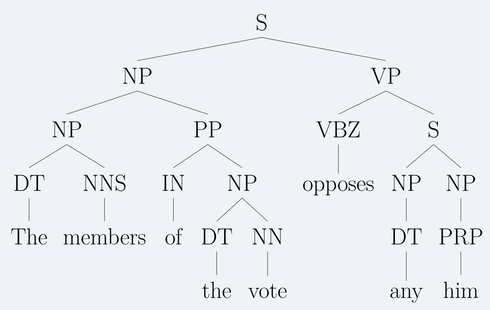
Intelligent Systems Program, Computer Science Department, University of Pittsburgh

Parsing ungrammatical sentences

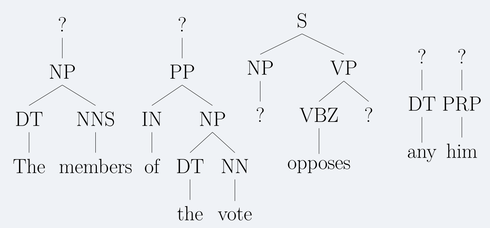
- Some example domains of ungrammatical sentences:
  - Writings of ESL learners
  - Machine translation outputs
- Parsers produce full, syntactically well-formed trees that are not appropriate for ungrammatical sentences

Our proposed approach: Parse Tree Fragmentation

- Identify well-formed syntactic structures for the parts that make sense
- *Parse tree fragmentation* is the process of breaking up the tree
- *Fragments* are reasonable isolated parts of parse trees



Stanford Parse Tree



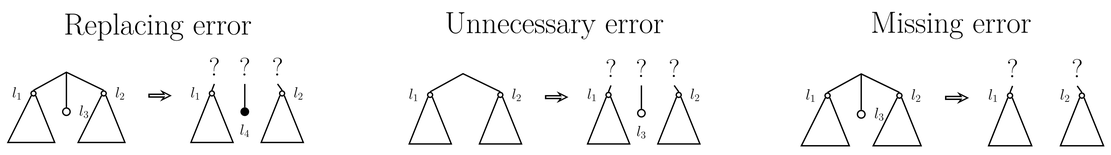
Coherent fragments

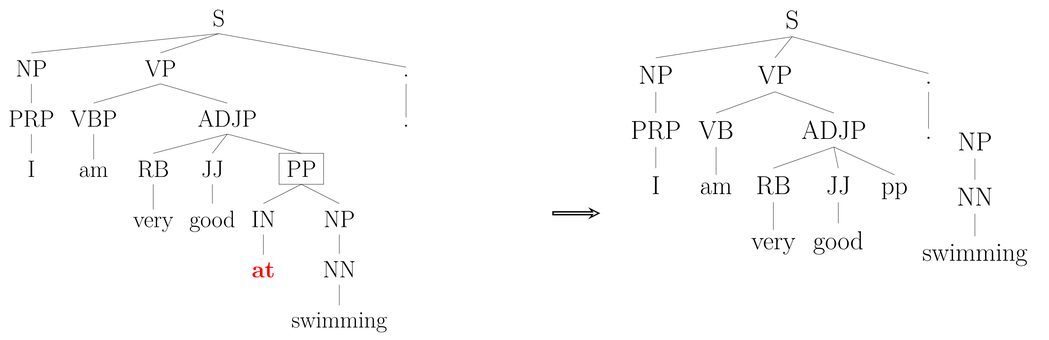
Developing a Fragmentation Corpus

1) Pseudo Gold Fragmentation (PGold)

Given an ungrammatical sentence and its error corrections:

- ESL sentence: *I am very good swimming.*
- Teacher corrections: *I am very good **at** swimming.*





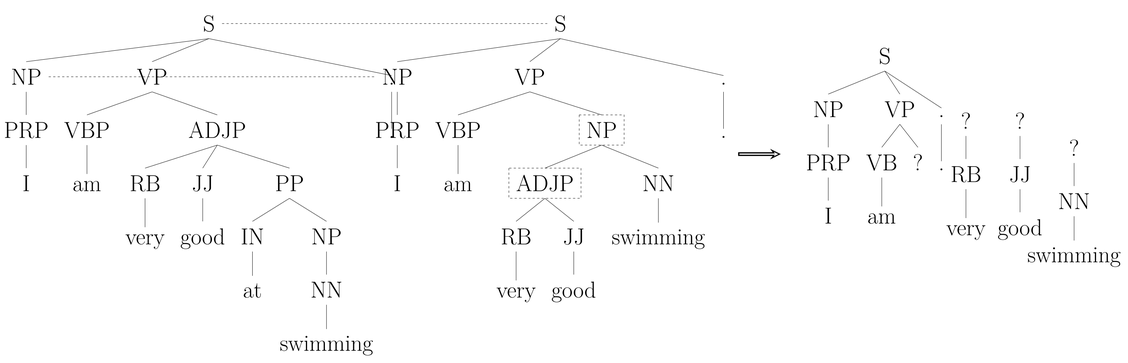
Parse tree of the grammatical sentence:
I am very good at swimming.

PGold fragments of the ungrammatical sentence:
I am very good swimming.

2) REference Fragmentation (REF)

Given an ungrammatical sentence and a grammatical version of the same sentence:

1. Automatically find alignments between two trees
   - Because we don't necessarily know what the error is without some detailed human correction annotations
2. Assign fragments to aligned nodes



Parse tree of grammatical sent. | Parse tree of ungrammatical sent. | REF fragments of ungrammatical sent.

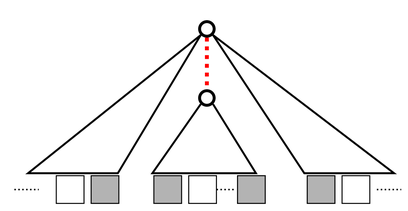
Fragmentation methods

1) Classification-based Parse Tree Fragmentation (CLF)

- **Binary classification**: Each edge is kept or cut
- **Training data**: Parse trees fragments by Reference method

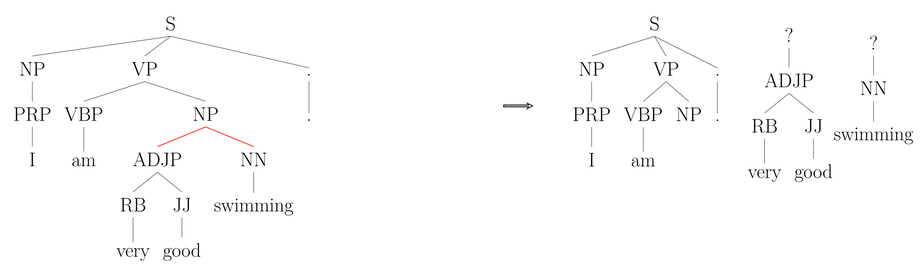
Features:

1. Labels of parent, child, grandparent
2. Depth & height of parent, child
3. Word bigrams and trigrams
4. CFG rule frequencies in Treebank



2) Treebank-based Parse Tree Fragmentation (TBF)

- For domain that do not have parallel corpora, we back off to available resources
- Use context free grammar rule frequencies in treebank to keep or cut an edge



Parse tree of ungrammatical sentence | TBF fragments

Data

1 English as a Second Language corpus (ESL)

- Fluency score is number of errors
- 5000 sentences with 1+ errors
- 7000 sentences with 0+ errors

2 Machine Translation outputs (MT)

- Fluency score calculated by edit rates (HTER)
- 4000 sentences with HTER score > 0.1
- 6000 sentences with HTER scores $\geqslant 0$

Experiments

Extrinsic Evaluation: Fluency Judgment

Binary classification: a sentence has virtually no error or many errors

Regression: Predict number of errors in ESL dataset or HTER in MT dataset

Our feature set: number, avg. size, min size, max size of fragments

	ESL			MT		
	Classification		Regression	Classification		Regression
feature set	Acc.(%)	AUC	Pearson's r	Acc.(%)	AUC	Pearson's r
LM	76.7	0.73	0.279	74.4	0.71	0.307
C&J	76.3	0.74	0.318	68.3	0.6	0.136
TSG	77.3	0.74	0.285	69.8	0.59	0.105
PGold	100	1	0.928	-	-	-
REF	99.8	1	0.84	94.4	0.99	0.782
CLF	79.9	0.81	0.377	73	0.66	0.205
TBF	77.2	0.74	0.298	71.8	0.51	0.04
CLF + LM	**82.2**	**0.86**	**0.462**	74.7	**0.73**	**0.324**

Experiments using 10-fold cross validation with Gradient Boosting Classifier
C&J: Charniak&Johnson,"Coarse-to-fine n-best parsing and MaxEnt discriminative reranking", ACL 2005.
TSG: Post,"Judging grammaticality with tree substitution grammar derivations", ACL 2011.

Conclusion

- Introducing the new task of **parse tree fragmentation**
- Extracting gold fragments using existing corpora for other NLP applications
- Proposing two practical fragmentation methods