Detecting grammatical errors using probabilistic parsing

Joachim Wagner, Jennifer Foster and Josef van Genabith

National Centre for Language Technology

Dublin City University

http://nclt.computing.dcu.ie

Research Question

- Can the output of existing probabilistic, treebank-trained parsers be exploited to judge grammaticality of sentences?

Two stage probabilistic parsing

- First stage – trained on grammatical data
- Second stage – trained on ungrammatical data
- Second stage is needed because state-of-the-art treebank-trained probabilistic parsers are robust BUT
  - Not necessarily accurate
  - *(NP (NP The closure) (PP in (NP computed breadth-first)))*
- Error detection is needed to decide when to apply the second ungrammatical stage

Error Detection Method 1

- Investigate whether the probability of a sentence's most likely parse can be predicted such that the deviation between the estimated parse probability (EPP) and the actual parse probability (APP) reflects the sentence's grammaticality
- Predict the EPP for a test sentence
  - based on APP of *similar* sentences from training set of grammatical sentences
- If EPP is some factor greater than the APP, then the sentence is classified as ungrammatical.

Error Detection Method 1
Do your circles overlap ? (-61.3)

Distance	Sentence	Log. Prob.
0.24	Is Mr Fatuzzo there ?	-60.3
0.42	Is Burma really isolated ?	-62.0
0.68	Should embryos be cloned ?	-57.5
0.73	(Mr Crowley refused)	-59.1
0.74	Should we reprimand ministers ?	-59.8
0.76	Subject : Phare - Poland	-71.7
0.77	Subject : ASEAN and Burma	-70.4
0.80	Is Mr Duisenberg present ?	-64.0
0.81	Structural Funds (continuation)	-57.1
0.81	Have I understood correctly ?	-49.6
		-61.2

Error Detection Method 1

Does your circles overlap ? (-63.3)

Distance	Sentence	Log. Prob.
0.05	Is Mr Fatuzzo there ?	-60.3
0.44	Is Burma really isolated ?	-62.0
0.45	(Mr Crowley refused)	-59.1
0.57	Structural Funds (continuation)	-57.1
0.60	Euro-Mediterranean cooperation (continuation)	-67.1
0.60	Have I understood correctly ?	-49.6
0.60	Have I understood correctly ?	-49.6
0.60	Have I understood correctly ?	-49.6
0.60	Should we reprimand ministers ?	-59.8
0.63	(Loud sustained applause)	-57.7
		-57.2

Error Detection Method 2

- Use machine learning to classify a sentence as grammatical or ungrammatical
- Training data:
  - parsed grammatical sentences
  - parsed ungrammatical sentences

Automatic Error Creation (1)

- Why is it useful to have a large amount of ungrammatical data?
  - Test data
  - Training data
- Why is it necessary to do this automatically?
  - Finding and annotating errors is so time-consuming
- Empirically motivated method
  - Tagged corpus of grammatical language
  - Attempt to introduce an error into each sentence
  - Based on error analysis of hand-crafted corpus

Hand-crafted corpus (1)

- A small corpus of ungrammatical written language
  - Grammaticality judgements on English sentences *in context*
  - Definition of “ungrammatical”
    - A sentence is ungrammatical if it contains an error and all words in the sentence are well-formed.
    - *The theory in empirical* is included. *The theory is empirrical* is not.
  - Applied above definition to reading material over 18 months

Hand-crafted corpus (2)

- A corpus of ungrammatical language
  - Each sentence is corrected > parallel corpus
  - 925 ungrammatical sentences, 1117 grammatical sentences
  - Some sentences had more than one correction with the same meaning (given the context):
    - *The longest **journey begin** with one step.*
    - *The longest journey **begins** with one step.*
    - *The longest **journeys** begin with one step.*
  - Error analysis based on the correction which was applied to make the ungrammatical sentence well-formed.

Hand-crafted corpus (3)

Replace a word, **48%**

- *I didn't **wanted** to delete the track → I didn't **want** to delete the track*
- *It was the **fist** signs of the façade beginning to fracture → It was the **first** signs of the façade beginning to fracture*

Add a word, **24%**

- *Will be declaring their undying love for each other? →Will **they** be declaring their undying love for each other?*

Hand-crafted corpus (4)

- Delete a word, **17%**
  - *A joint development which will **the** provide 10 new apartments → A joint development which will provide 10 new apartments*
- Combination of above (composite errors), **11%**
  - *What does a single line **yellow** mean? → What does a single **yellow** line mean?*
  - *This means **to allow** structure sharing → This means structure sharing **is allowed***

Automatic Error Creation (2)

- Deletion errors
  - repeated word errors

    *I think I'll get Fred **to to** wash his own overalls*
  - double syntactic function errors

    *Do you ever go and visit **the any** of them?*
  - random extra word errors

    *It'd be one thing less for Neil to worry **and** about*

Automatic Error Creation (3)

- Insertion errors
  - *He does not mind being butt of his colleagues' jokes*
- Context-sensitive spelling errors
  - *I came **too** the mountain very casually*
- Agreement errors
  - ***The contrasts** was startling*
  - *The first of **these** visiting **scientist** begin in January*

Automatic Error Creation (4)

- Limitations
  - Some ungrammatical constructions not covered
    - wrong verb form

    *Brent would often **became** stunned by resentment.*
  - Only one error per sentence
  - Only simple errors (involving one correction operation)
- Applied to the sentences in the British National Corpus – 9 million ungrammatical sentences

Error Detection Details (1)

- Training data:
  - Method 1:
    - 400,000 parsed grammatical BNC sentences
  - Method 2:
    - 200,000 parsed grammatical BNC sentences
    - 200,000 parsed ungrammatical BNC sentences
- Parser used: Charniak's parser (August 2005)
- Sentence length: 10-20 words

Error Detection Details (2)

- Learning algorithms
  - Method 1: Own implementation of k-nn
  - Method 2: Weka implementation of support vector machines
- Evaluation carried out using 10-fold cross validation on training data

Error Detection Details (3)

- Training features
  - #words
  - Height,#nodes of most probable parse tree
  - POS counts, e.g. #IN, #TO,#DT, etc.
  - ratio of closed class to open class words in sentence
  - language model probabilities (unigram token, pcfg terminal rules)
  - probability of most probable parse tree (Method 2 only)
  - probability of 2nd most probable parse tree (Method 2 only)

Preliminary Results

First Column: Method 1, Second Column: Method 2

Error Type	Precision		Recall		F-Score	
Extra Word	63.9	70.7	64.5	66.7	64.2	68.6
Missing Word	58.8	61.4	58.4	59.8	58.6	60.6
CS Spelling	41.5	70.6	35.8	68.9	38.4	69.7
Agreement	58.4	62.1	57.3	63.2	57.8	62.6

Future Work

- Method 1:
  - Surface similarity measures such as bleu score
- Method 2:
  - Distribution of first *fifty* parse probabilities
- Vary sentence length range of training data
- PCFG parser instead of history-based
- Influence of corpus size
- Moving from sentences to phrases
- Other languages, e.g. German
- Other learning algorithms

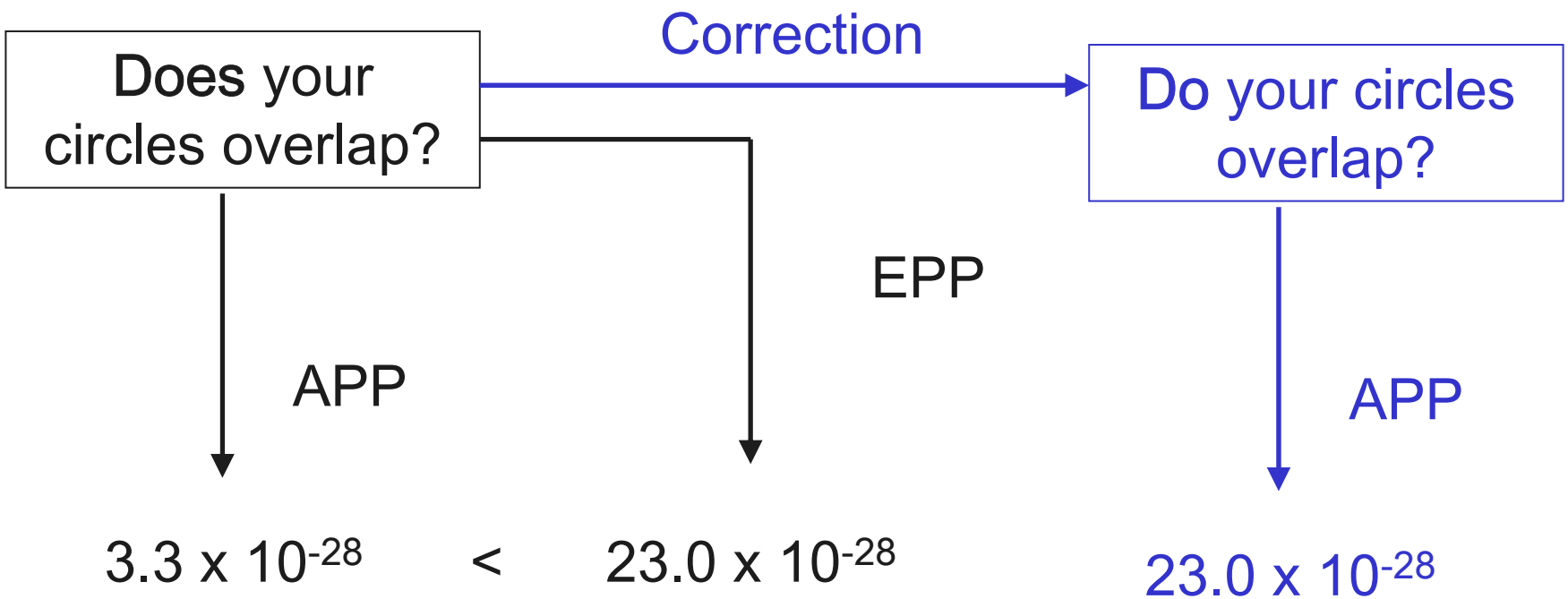
References

- Foster 2000. A Unification Strategy for Parsing Agreement Errors. In Proceedings of the Student Session of the European Summer School for Logic, Language and Information (ESSLLI-00), pp.77-87, Birmingham.
- Foster and Vogel 2004. Good Reasons for Noting Bad Grammar: Constructing a Corpus of Ungrammatical Language. In Stephan Kepser and Marga Reis (eds.) *Pre-proceedings of the International Conference on Linguistic Evidence, pp.151-152,* Tuebingen.
- Foster 2004. Parsing Ungrammatical Input: An Evaluation Procedure. *Proceedings of the 4th International Conference on Linguistic Resources and Evaluation, Vol. 6, pp.2039-2042,* Lisbon.
- Foster and Vogel 2004. Parsing Ill-formed Text using an Error Grammar. *AIReview, Special AICS2003 Issue, 21, pp.269-291.*
- Foster 2005. Good Reasons for Noting Bad Grammar: Empirical Investigations into the Parsing of Ungrammatical Written English, Phd Thesis, University of Dublin, Trinity College, Dublin.

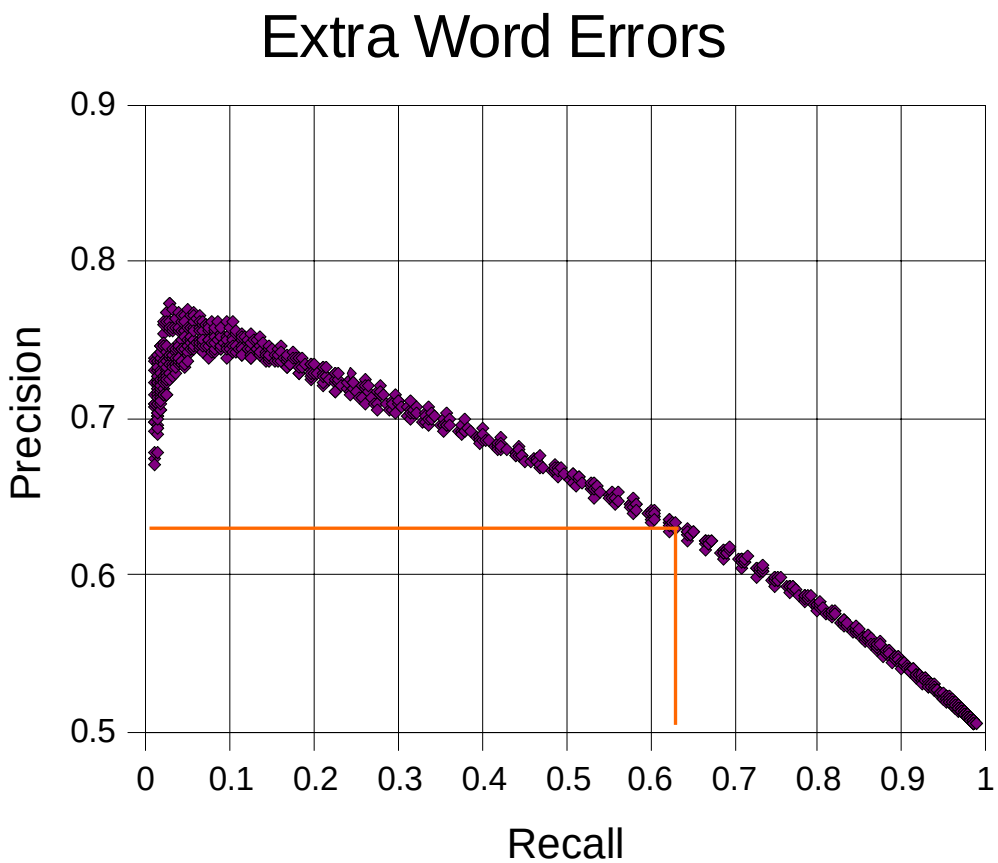
Thank you!

- Any questions?

Error Detection Method 1



Precision-Recall Graphs



Future Work: Comparison

- Discriminative grammars
  - XLE
  - RASP
- Thresholded corpus-induced PCFG
  - Low frequency rules
  - Rules more frequently used in parsing error corpus
- POS n-gram statistics