Parser Evaluation and the BNC

Jennifer Foster and Josef van Genabith

National Centre for Language Technology
School of Computing
Dublin City University

29th May 2008

What is this work about?

1. Creating a set of gold standard parse trees for 1,000 sentences from the BNC
2. Using these trees as a test set to evaluate various parsers

Outline

The British National Corpus

The BNC is a one hundred million word balanced corpus of British English (Burnard, 2000)

- 90% of the BNC is written text
  - 75% factual
  - 25% fiction
- The 10% spoken component consists of
  - informal dialogue
  - business meetings
  - speeches

The British National Corpus

The BNC is a one hundred million word balanced corpus of British English (Burnard, 2000)

- 90% of the BNC is written text
  - 75% factual
  - 25% fiction
- The 10% spoken component consists of
  - informal dialogue
  - business meetings
  - speeches

The British National Corpus

The BNC is a one hundred million word balanced corpus of British English (Burnard, 2000)

- 90% of the BNC is written text
  - 75% factual
  - 25% fiction
- The 10% spoken component consists of
  - informal dialogue
  - business meetings
  - speeches

BNC Test Set: Choosing the sentences

1,000 sentences in test set

- Not chosen completely at random
- They are *different* from WSJ training data:
  - Contain a verb in BNC but not in WSJ2-21
    - 25,874 verb lemmas in BNC but not in WSJ2-21
    - 14,787 occur only once in BNC (e.g. *jitter, unfade, transpersonalize, kerplonk*)
    - 537 occur greater than 100 times (e.g. *murmur, frown, damn*)
  - Likely to represent a difficult test for WSJ-trained parsers

BNC Test Set: Choosing the sentences

1,000 sentences in test set

- Not chosen completely at random
- They are *different* from WSJ training data:
  - Contain a verb in BNC but not in WSJ2-21
    - 25,874 verb lemmas in BNC but not in WSJ2-21
    - 14,787 occur only once in BNC (e.g. *jitter, unfade, transpersonalize, kerplonk*)
    - 537 occur greater than 100 times (e.g. *murmur, frown, damn*)
  - Likely to represent a difficult test for WSJ-trained parsers

BNC Test Set: Choosing the sentences

1,000 sentences in test set

- Not chosen completely at random
- They are *different* from WSJ training data:
  - Contain a verb in BNC but not in WSJ2-21
    - 25,874 verb lemmas in BNC but not in WSJ2-21
    - 14,787 occur only once in BNC (e.g. *jitter*, *unfade*, *transpersonalize*, *kerplonk*)
    - 537 occur greater than 100 times (e.g. *murmur*, *frown*, *damn*)
  - Likely to represent a difficult test for WSJ-trained parsers

BNC Test Set: Choosing the sentences

1,000 sentences in test set

- Not chosen completely at random
- They are *different* from WSJ training data:
  - Contain a verb in BNC but not in WSJ2-21
    - 25,874 verb lemmas in BNC but not in WSJ2-21
    - 14,787 occur only once in BNC (e.g. *jitter, unfade, transpersonalize, kerplonk*)
    - 537 occur greater than 100 times (e.g. *murmur, frown, damn*)
  - Likely to represent a difficult test for WSJ-trained parsers

BNC Test Set: Choosing the sentences

1,000 sentences in test set

- Not chosen completely at random
- They are *different* from WSJ training data:
  - Contain a verb in BNC but not in WSJ2-21
    - 25,874 verb lemmas in BNC but not in WSJ2-21
    - 14,787 occur only once in BNC (e.g. *jitter*, *unfade*, *transpersonalize*, *kerplonk*)
    - 537 occur greater than 100 times (e.g. *murmur*, *frown*, *damn*)
  - Likely to represent a difficult test for WSJ-trained parsers

BNC Test Set: Choosing the sentences

1,000 sentences in test set

- Not chosen completely at random
- They are *different* from WSJ training data:
  - Contain a verb in BNC but not in WSJ2-21
    - 25,874 verb lemmas in BNC but not in WSJ2-21
    - 14,787 occur only once in BNC (e.g. *jitter, unfade, transpersonalize, kerplonk*)
    - 537 occur greater than 100 times (e.g. *murmur, frown, damn*)
  - Likely to represent a difficult test for WSJ-trained parsers

BNC Test Set: Some examples

Text Type	**#**	**Example**
Spoken	10	The seconder of formally seconded
Poem	9	Groggily somersaulting to get airborne
Caption	4	Community Personified
Headline	2	Drunk priest is nicked driving to a funeral

Average sentence length: 28 words

BNC Test Set: Some examples

Text Type	#	Example
Spoken	10	The seconder of formally seconded
Poem	9	Groggily somersaulting to get airborne
Caption	4	Community Personified
Headline	2	Drunk priest is nicked driving to a funeral

Average sentence length: 28 words

BNC Test Set: Annotation Process

- One annotator
- Two passes through the data
- Approximately 100 hours
- As references, the annotator used
  1. Penn Treebank bracketing guidelines (Bies *et al* 1995)
  2. Penn Treebank itself
- Functional tags and traces not annotated

BNC Test Set: Annotation Difficulties

What happens when the references clash?

- The noun phrase *almost certain death* occurs in BNC gold standard sentence
- According to the guidelines, it should be annotated as *(NP (ADJP almost certain) death)*
- A search for *almost* in the Penn Treebank yields the following example *(NP almost unimaginable speed)*
- In such cases, annotator chose the analysis set out in the guidelines

BNC Test Set: Annotation Difficulties

What happens when the references clash?

- The noun phrase *almost certain death* occurs in BNC gold standard sentence
- According to the guidelines, it should be annotated as *(NP (ADJP almost certain) death)*
- A search for *almost* in the Penn Treebank yields the following example *(NP almost unimaginable speed)*
- In such cases, annotator chose the analysis set out in the guidelines

BNC Test Set: Annotation Difficulties

What happens when the references clash?

- The noun phrase *almost certain death* occurs in BNC gold standard sentence
- According to the guidelines, it should be annotated as
  (NP (ADJP almost certain) death)
- A search for *almost* in the Penn Treebank yields the following example
  (NP almost unimaginable speed)
- In such cases, annotator chose the analysis set out in the guidelines

BNC Test Set: Annotation Difficulties

What happens when the references clash?

- The noun phrase *almost certain death* occurs in BNC gold standard sentence
- According to the guidelines, it should be annotated as
  (NP (ADJP almost certain) death)
- A search for *almost* in the Penn Treebank yields the following example
  (NP almost unimaginable speed)
- In such cases, annotator chose the analysis set out in the guidelines

BNC Test Set: Annotation Difficulties

What happens when the references clash?

- The noun phrase *almost certain death* occurs in BNC gold standard sentence
- According to the guidelines, it should be annotated as
  (NP (ADJP almost certain) death)
- A search for *almost* in the Penn Treebank yields the following example
  (NP almost unimaginable speed)
- In such cases, annotator chose the analysis set out in the guidelines

BNC Test Set: Annotation Difficulties

69 sentences marked as difficult

- Attachment ambiguities
  He has had to come to terms with the tragic loss of friends ***from the very start of his climbing career****.*
- Adverbials
  a few seats ***down*** *from them*

BNC Test Set: Annotation Difficulties

69 sentences marked as difficult

- Attachment ambiguities
  He has had to come to terms with the tragic loss of friends ***from the very start of his climbing career***.
- Adverbials
  a few seats ***down*** *from them*

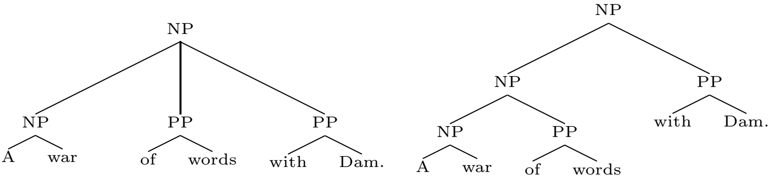
BNC Test Set: Annotation Difficulties

69 sentences marked as difficult

- Attachment ambiguities
  He has had to come to terms with the tragic loss of friends ***from the very start of his climbing career****.*
- Adverbials
  a few seats ***down*** *from them*

BNC Test Set: Annotation Difficulties

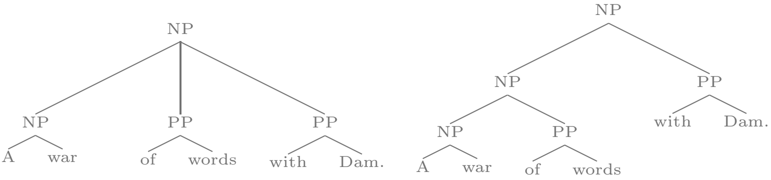
- Noun phrase structure
  a war of words with Damascus



- Miscellaneous
  - *As likely to be queuing at a supermarket checkout as at a communion rail*
  - *day in day out*
  - ***Other than that** he showed up Giggs...*

BNC Test Set: Annotation Difficulties

- Noun phrase structure
  a war of words with Damascus



- Miscellaneous
  - *As likely to be queuing at a supermarket checkout as at a communion rail*
  - *day in day out*
  - ***Other than that** he showed up Giggs...*

Outline

Which Parsers?

Various versions of the Charniak parser

- History-based generative statistical parser (Charniak, 2000)
- Reranking parser (Charniak and Johnson, 2005)
  - First-stage generative parser
  - Discriminative reranker re-orders n-best list returned by first-stage parser

Which Parsers?

Various versions of the Charniak parser

- History-based generative statistical parser (Charniak, 2000)
- Reranking parser (Charniak and Johnson, 2005)
  - First-stage generative parser
  - Discriminative reranker re-orders n-best list returned by first-stage parser

Which Parsers?

Various versions of the Charniak parser

- NANC self-trained parser (McClosky et al, 2006)
  - Reranking parser parses NANC sentences
  - First-stage parser is retrained with NANC trees plus WSJ gold standard trees
- BNC self-trained parser (Foster et al, 2007)

Which Parsers?

Various versions of the Charniak parser

- NANC self-trained parser (McClosky et al, 2006)
  - Reranking parser parses NANC sentences
  - First-stage parser is retrained with NANC trees plus WSJ gold standard trees
- BNC self-trained parser (Foster et al, 2007)

Any other parsers?

Berkeley parser (Petrov et al, 2006)

- Unlexicalised PCFG parser
- To learn PCFG:
  1. Start with x-bar grammar read from Penn Treebank
  2. Split each nonterminal category into two subcategories
  3. Train a grammar (using Expectation Maximisation learning)
  4. For each pair of subcategories
     - Merge the subcategories
     - Measure the information loss after the merge
     - If loss is small, keep the merge
  5. Repeat steps 2-4
- We use PCFG obtained using 5 split/merge iterations

Any other parsers?

Berkeley parser (Petrov et al, 2006)

- Unlexicalised PCFG parser
- To learn PCFG:
  1. Start with x-bar grammar read from Penn Treebank
  2. Split each nonterminal category into two subcategories
  3. Train a grammar (using Expectation Maximisation learning)
  4. For each pair of subcategories
     - Merge the subcategories
     - Measure the information loss after the merge
     - If loss is small, keep the merge
  5. Repeat steps 2-4
- We use PCFG obtained using 5 split/merge iterations

Any other parsers?

Berkeley parser (Petrov et al, 2006)

- Unlexicalised PCFG parser
- To learn PCFG:
  1. Start with x-bar grammar read from Penn Treebank
  2. Split each nonterminal category into two subcategories
  3. Train a grammar (using Expectation Maximisation learning)
  4. For each pair of subcategories
     - Merge the subcategories
     - Measure the information loss after the merge
     - If loss is small, keep the merge
  5. Repeat steps 2-4
- We use PCFG obtained using 5 split/merge iterations

Any other parsers?

Berkeley parser (Petrov et al, 2006)

- Unlexicalised PCFG parser
- To learn PCFG:
  1. Start with x-bar grammar read from Penn Treebank
  2. Split each nonterminal category into two subcategories
  3. Train a grammar (using Expectation Maximisation learning)
  4. For each pair of subcategories
     - Merge the subcategories
     - Measure the information loss after the merge
     - If loss is small, keep the merge
  5. Repeat steps 2-4
- We use PCFG obtained using 5 split/merge iterations

Evaluation Metrics

Phrase-structure evaluation

1. Parseval (evalb implementation) (Black et al, 1991)
2. Leaf Ancestor (Sampson and Barbarczy, 2002)
3. Tree Distance (Emms, 2008)

Dependency evaluation

Relies on constituent to dependency conversion

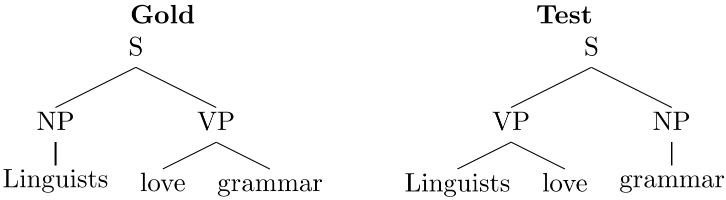
Parseval

- Tree as a set of *labelled spans*
- Precision, recall and f-score over gold and test sets

Gold: { *(S Linguists love grammar), (NP Linguists), (VP love grammar)* }
Test: { *(S Linguists love grammar), (VP Linguists love), (NP grammar)* }

Parseval

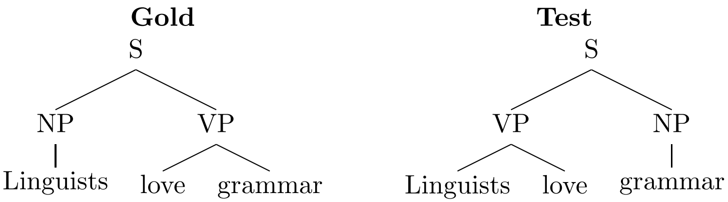
- Tree as a set of *labelled spans*
- Precision, recall and f-score over gold and test sets



Gold: { *(S Linguists love grammar), (NP Linguists), (VP love grammar)* }
Test: { *(S Linguists love grammar), (VP Linguists love), (NP grammar)* }

Parseval

- Tree as a set of *labelled spans*
- Precision, recall and f-score over gold and test sets



Gold: { *(S Linguists love grammar), (NP Linguists), (VP love grammar)* }
Test: { *(S Linguists love grammar), (VP Linguists love), (NP grammar)* }

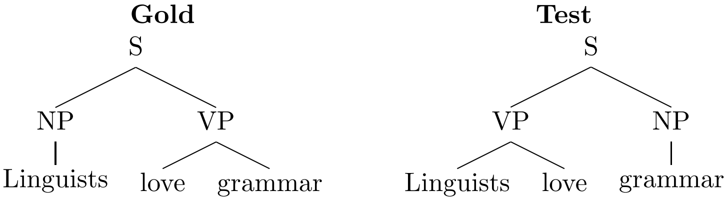
Leaf Ancestor

- Tree as a set of *lineages*
- Each lineage in test set assigned a score
- Score based on edit distance from gold lineage
- Lineage scores are averaged

Gold: { <*Linguists NP] [S*>, <*love [VP S*>, <*grammar VP S]*> }
Test: { <*Linguists VP [S*>, <*love VP] S*>, <*grammar [NP S]*> }

Leaf Ancestor

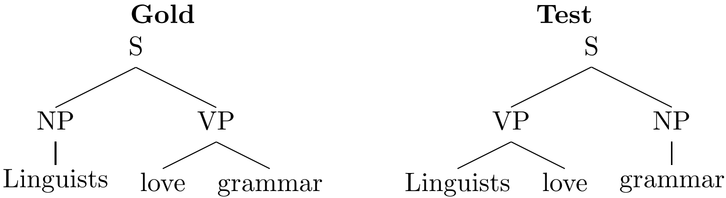
- Tree as a set of *lineages*
- Each lineage in test set assigned a score
- Score based on edit distance from gold lineage
- Lineage scores are averaged



Gold: { <*Linguists NP] [S*>, <*love [VP S*>, <*grammar VP S]*> }
Test: { <*Linguists VP [S*>, <*love VP] S*>, <*grammar [NP S]*> }

Leaf Ancestor

- Tree as a set of *lineages*
- Each lineage in test set assigned a score
- Score based on edit distance from gold lineage
- Lineage scores are averaged



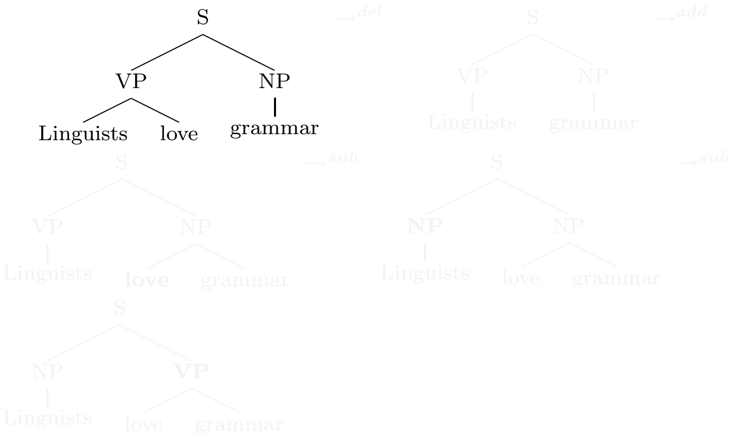
Gold: { <*Linguists NP] [S*>, <*love [VP S*>, <*grammar VP S]*> }
Test: { <*Linguists VP [S*>, <*love VP] S*>, <*grammar [NP S]*> }

Tree Distance

- Edit distance on actual trees
- Calculate the minimum cost of transforming test tree to gold tree

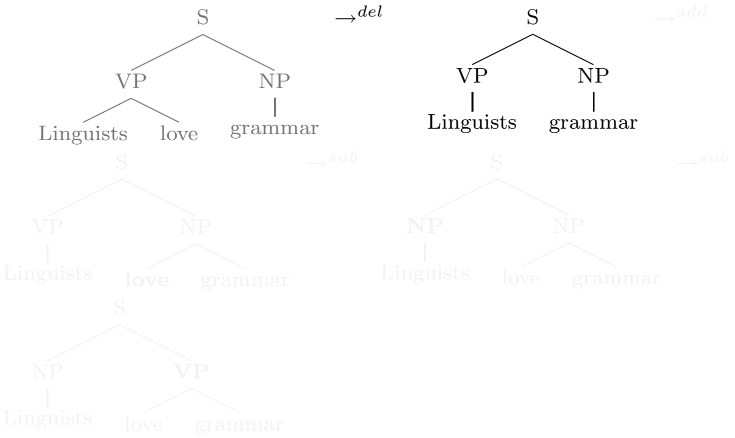
Tree Distance

- Edit distance on actual trees
- Calculate the minimum cost of transforming test tree to gold tree



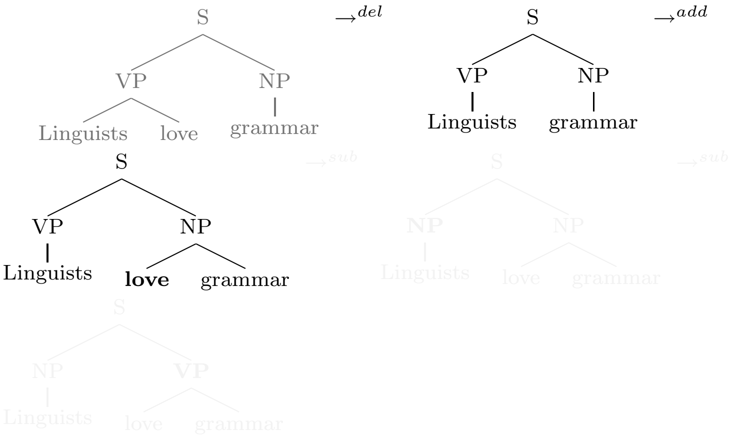
Tree Distance

- Edit distance on actual trees
- Calculate the minimum cost of transforming test tree to gold tree



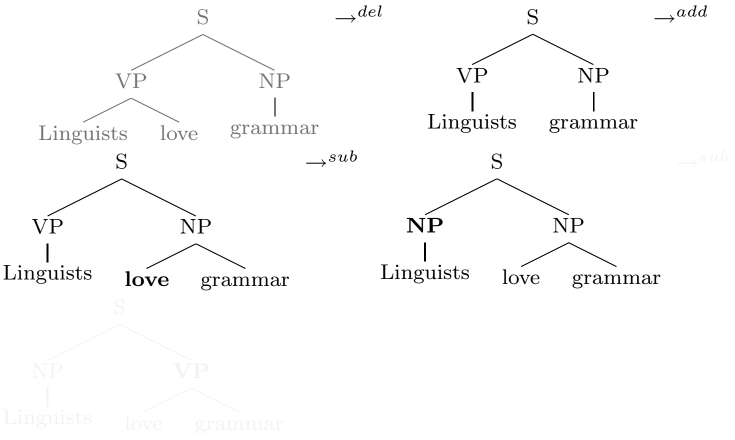
Tree Distance

- Edit distance on actual trees
- Calculate the minimum cost of transforming test tree to gold tree



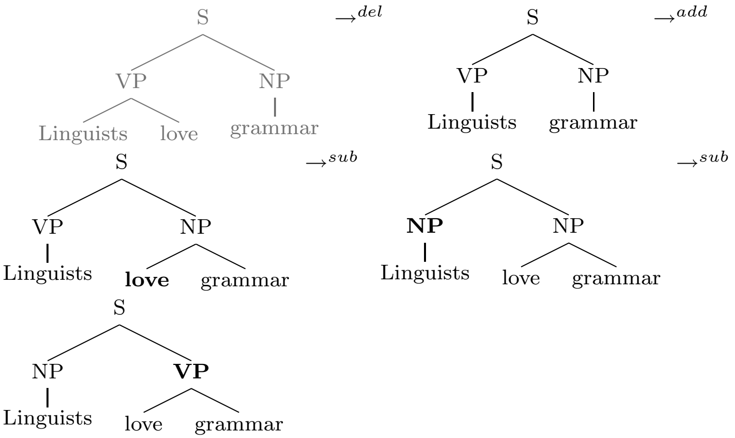
Tree Distance

- Edit distance on actual trees
- Calculate the minimum cost of transforming test tree to gold tree



Tree Distance

- Edit distance on actual trees
- Calculate the minimum cost of transforming test tree to gold tree



Dependency Evaluation

- Tree as a set of word-word dependency tuples <`word`,`head`,`label`>
  (Linguists,love,subj), (grammar,love,obj)
- Automatic conversion procedure (Johansson and Nugues, 2007)
- Works better when Penn-II functional tags are available
- Use automatic functional tag labeller of Chrupala et al, 2007

Evaluation Results

Parser	Parseval	TreeDist	LA	Dep
Berkeley	82.0	89.8	91.1	81.6
Charniak	82.5	90.0	91.6	82.5
C&J Rerank	83.4↑	90.3	91.8	82.8
C&J NANC	83.9	90.6	91.7	83.0
C&J BNC	85.4↑	91.3	92.6	84.2

- Approx 7% drop moving from WSJ23 to BNC
- Evaluation metrics tell roughly the same story
- Reranking improves performance
- Best parser on the BNC test data is BNC self-trained parser
- Using *in*-domain data for self-training appears to be more effective

Evaluation Results

Parser	Parseval	TreeDist	LA	Dep
Berkeley	82.0	89.8	91.1	81.6
Charniak	82.5	90.0	91.6	82.5
C&J Rerank	83.4↑	90.3	91.8	82.8
C&J NANC	83.9	90.6	91.7	83.0
C&J BNC	85.4↑	91.3	92.6	84.2

- Approx 7% drop moving from WSJ23 to BNC
- Evaluation metrics tell roughly the same story
- Reranking improves performance
- Best parser on the BNC test data is BNC self-trained parser
- Using *in*-domain data for self-training appears to be more effective

Evaluation Results

Parser	**Parseval**	**TreeDist**	**LA**	**Dep**
Berkeley	82.0	89.8	91.1	81.6
Charniak	82.5	90.0	91.6	82.5
C&J Rerank	83.4$\uparrow$	90.3	91.8	82.8
C&J NANC	83.9	90.6	91.7	83.0
C&J BNC	85.4$\uparrow$	91.3	92.6	84.2

- Approx 7% drop moving from WSJ23 to BNC
- Evaluation metrics tell roughly the same story
- Reranking improves performance
- Best parser on the BNC test data is BNC self-trained parser
- Using *in*-domain data for self-training appears to be more effective

Evaluation Results

Parser	Parseval	TreeDist	LA	Dep
Berkeley	82.0	89.8	91.1	81.6
Charniak	82.5	90.0	91.6	82.5
C&J Rerank	83.4↑	90.3	91.8	82.8
C&J NANC	83.9	90.6	91.7	83.0
C&J BNC	85.4↑	91.3	92.6	84.2

- Approx 7% drop moving from WSJ23 to BNC
- Evaluation metrics tell roughly the same story
- Reranking improves performance
- Best parser on the BNC test data is BNC self-trained parser
- Using *in*-domain data for self-training appears to be more effective

Evaluation Results

Parser	Parseval	TreeDist	LA	Dep
Berkeley	82.0	89.8	91.1	81.6
Charniak	82.5	90.0	91.6	82.5
C&J Rerank	83.4↑	90.3	91.8	82.8
C&J NANC	83.9	90.6	91.7	83.0
C&J BNC	85.4↑	91.3	92.6	84.2

- Approx 7% drop moving from WSJ23 to BNC
- Evaluation metrics tell roughly the same story
- Reranking improves performance
- Best parser on the BNC test data is BNC self-trained parser
- Using *in*-domain data for self-training appears to be more effective

Evaluation Results

Parser	Parseval	TreeDist	LA	Dep
Berkeley	82.0	89.8	91.1	81.6
Charniak	82.5	90.0	91.6	82.5
C&J Rerank	83.4↑	90.3	91.8	82.8
C&J NANC	83.9	90.6	91.7	83.0
C&J BNC	85.4↑	91.3	92.6	84.2

- Approx 7% drop moving from WSJ23 to BNC
- Evaluation metrics tell roughly the same story
- Reranking improves performance
- Best parser on the BNC test data is BNC self-trained parser
- Using *in*-domain data for self-training appears to be more effective

Evaluation Results

Parser	Parseval	TreeDist	LA	Dep
Berkeley	82.0	89.8	91.1	81.6
Charniak	82.5	90.0	91.6	82.5
C&J Rerank	83.4↑	90.3	91.8	82.8
C&J NANC	83.9	90.6	91.7	83.0
C&J BNC	85.4↑	91.3	92.6	84.2

- Approx 7% drop moving from WSJ23 to BNC
- Evaluation metrics tell roughly the same story
- Reranking improves performance
- Best parser on the BNC test data is BNC self-trained parser
- Using *in*-domain data for self-training appears to be more effective

Error Analysis

Problematic areas for all parsers

- Coordination
  The pistol had been a prop in the film ***in which my father had starred*** *and* ***after filming was over he forgot to return it.***
- Adverbs
  Incidentally Ciccolini *also plays several works for piano 4 hands*
- Noun/verb confusions
  In winter that ***walk*** *back home must have been hell.*
  This faithful rig has been served to ***ground-run*** *engines.*

Error Analysis

Problematic areas for all parsers

- Coordination
  The pistol had been a prop in the film ***in which my father had starred*** *and* ***after filming was over he forgot to return it.***
- Adverbs
  Incidentally Ciccolini *also plays several works for piano 4 hands*
- Noun/verb confusions
  In winter that ***walk*** *back home must have been hell.*
  This faithful rig has been served to ***ground-run*** *engines.*

Error Analysis

Problematic areas for all parsers

- Coordination
  The pistol had been a prop in the film ***in which my father had starred*** *and* ***after filming was over he forgot to return it.***
- Adverbs
  Incidentally Ciccolini *also plays several works for piano 4 hands*
- Noun/verb confusions
  In winter that ***walk*** *back home must have been hell.*
  This faithful rig has been served to ***ground-run*** *engines.*

Error Analysis

Problematic areas for all parsers

- Coordination
  *The pistol had been a prop in the film **in which my father had starred** and **after filming was over he forgot to return it.***
- Adverbs
  ***Incidentally Ciccolini** also plays several works for piano 4 hands*
- Noun/verb confusions
  *In winter that **walk** back home must have been hell.*
  *This faithful rig has been served to **ground-run** engines.*

Error Analysis

More problematic areas

- Fragments
  Moles burrowing away underground out of sight of each other but with a common purpose.
- Parentheticals
  *...but **and here's the rub-a-dub**, it was at least three pits further out.*

Self-training on BNC data gives modest improvements in most areas

Error Analysis

More problematic areas

- Fragments
  Moles burrowing away underground out of sight of each other but with a common purpose.
- Parentheticals
  *...but **and here's the rub-a-dub**, it was at least three pits further out.*

Self-training on BNC data gives modest improvements in most areas

Error Analysis

More problematic areas

- Fragments
  Moles burrowing away underground out of sight of each other but with a common purpose.
- Parentheticals
  ...but ***and here's the rub-a-dub****, it was at least three pits further out.*

Self-training on BNC data gives modest improvements in most areas

Error Analysis

More problematic areas

- Fragments
  Moles burrowing away underground out of sight of each other but with a common purpose.
- Parentheticals
  ...but ***and here's the rub-a-dub****, it was at least three pits further out.*

Self-training on BNC data gives modest improvements in most areas

Concluding remarks

Future Work

- More detailed error analysis
  - Behaviour of Leaf Ancestor metric
  - Difference between Berkeley and Charniak parser
  - Difference between BNC Test Set and WSJ23
- Annotation of traces and functional tags

Another parsing test set for English - 1,000 BNC sentences

Available at http://nclt.computing.dcu.ie/~jfoster/resources

Thank you for listening

Concluding remarks

Future Work

- More detailed error analysis
  - Behaviour of Leaf Ancestor metric
  - Difference between Berkeley and Charniak parser
  - Difference between BNC Test Set and WSJ23
- Annotation of traces and functional tags

Another parsing test set for English - 1,000 BNC sentences

Available at
http://nclt.computing.dcu.ie/~jfoster/resources

Thank you for listening

Concluding remarks

Future Work

- More detailed error analysis
  - Behaviour of Leaf Ancestor metric
  - Difference between Berkeley and Charniak parser
  - Difference between BNC Test Set and WSJ23
- Annotation of traces and functional tags

Another parsing test set for English - 1,000 BNC sentences

Available at
http://nclt.computing.dcu.ie/~jfoster/resources

Thank you for listening

Ann Bies, Mark Ferguson, Karen Katz, and Robert MacIntyre.
1995.
Bracketing guidelines for treebank ii style, penn treebank project.
Technical Report Tech Report MS-CIS-95-06, University of Pennsylvania, Philadelphia, PA.

E. Black, Steve Abney, Dan Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, Fred Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski.
1991.
A procedure for quantitatively comparing the syntactic coverage of english grammars.
In *Proceedings of the 1991 DARPA Speech and Natural Language Workshop*, pages 306–311.

Eugene Charniak and Mark Johnson.
2005.
Course-to-fine n-best-parsing and maxent discriminative reranking.
In *Proceedings of the 43rd Annual Meeting of the ACL (ACL-05)*, pages 173–180, Ann Arbor, Michigan, June.

Eugene Charniak.
2000.
A maximum-entropy-inspired parser.
In *Proceedings of the Annual Meeting of the North American Association for Computational Linguistics (NAACL-00)*, pages 132–139, Seattle, Washington.

Grzegorz Chrupała, Nicolas Stroppa, Josef van Genabith, and Georgiana Dinu.
2007.
Better training for function labeling.
In *Proceedings of the Recent Advances in Natural Language Processing Conference (RANLP-07)*, pages 133–138, Borovets, Bulgaria.

Jennifer Foster, Joachim Wagner, Djamé Seddah, and Josef van Genabith.
2007.
Adapting wsj-trained parsers to the british national corpus using in-domain self-training.
In *Proceedings of the Tenth International Workshop on Parsing Technologies (IWPT-07)*, pages 33–35, Prague, Czech Republic.

David McClosky, Eugene Charniak, and Mark Johnson.
2006.
Reranking and self-training for parser adaptation.
In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL (COLING-ACL-06)*, pages 337–344, Sydney, Australia, July.

Geoffrey Sampson and Anna Babarczy.
2002.
A test of the leaf-ancestor metric for parse accuracy.
In John Carroll, Anette Frank, Dekang Lin, Detlef Prescher, and Hans Uszkoreit, editors, *Proceedings of the "Beyond Parseval - Towards Improved Evaluation Measures for Parsing Systems"*

Workshop at the 3rd International Conference on Linguistic Resources and Evaluation (LREC), Las Palmas, Gran Canaria.