Resolving Attachment and Clause Boundary Ambiguities
for Simplifying Relative Clause Constructs

Advaith Siddharthan

Natural Language and Information Processing Group

Computer Laboratory, Cambridge

as372@cl.cam.ac.uk

Structure of Talk

- Motivation
- Relative Clause Attachment
  - Methodology
  - Evaluation
- Clause Boundaries
  - Methodology
  - Evaluation
- Conclusions
- Future Work

Motivation

Dis-embedding relative clauses is an important aspect of text simplification, an NLP task that:

- Restructures sentences, making text easier to read (or process)
- Preserves meaning and information content.
- Reduces grammatical (or lexical) complexity

Applications of Text Simplification:

- People with Language Disabilities like Aphasia (Carroll et al., 1998; Carroll et al., 1999)
- Preprocessing before parsing (Chandrasekar et al., 1996; Chandrasekar and Srinivas, 1997)
- Displaying text on Limited Channel Devices

Dis-Embedding Clauses

Previously published work on dis-embedding relative clauses make use of simplification rules that act on:

- Linear text (Chandrasekar et al., 1996)
- Some form of parse tree (Chandrasekar and Srinivas, 1997; Carroll et al., 1998)

A Hand-Crafted Rule:

`W X:NP, Y:Rel_Pr Z.` $\longrightarrow$ `W X. X Z.`

'[The pace] of [life] was slower in [those days],' says [51-year-old Cathy Tinsall] from [South London], *who had [five children]*.

$$\downarrow$$

'[The pace] of [life] was slower in [those days],' says [51-year-old Cathy Tinsall] from [South London]. [51-year-old Cathy Tinsall] from [South London] had [five children].

Defining the Problem

I focus on deciding local vs wide attachment when the noun phrase preceding the relative clause has the structure:

`NP1 Prep NP2`

‘[The pace] of [life] was slower in [those days],’ says [51-year-old Cathy Tinsall] from [South London], *who had [five children]*.

[The suicide note] included [lurid references] to [the economy] run under [the influence] of [Herr Pohl], *who might stop [a British government] from running [its own economic policy]*.

How Important is this Problem?

In the Penn Wall Street Journal Treebank (Marcus et al., 1993):

- 19% of *who* relative clauses
- 24% of *which* relative clauses

are preceded by complex noun phrases having the structure: `NP1 Prep NP2`

Relative Clause Attachment

Why not use a parser?

- The sentences in need of simplification don't come through a parser very well.
- Applications require speed
- Non-restrictive relative clauses are increasingly being treated by parsers as sentential adjuncts (Nunberg, 1990), leaving the attachment decisions to anaphora resolution algorithms

An Example

The board is dominated by *the heirs of* ***the late John T. Dorrance Jr.***, who controlled about 58% of Campbell's stock.

```
(T/leta_s
 (S/np_vp (NP/det_n The_AT (N1/n board_NN1))
  (V/be_ppart/- be+s_VBZ (V/pp dominate+ed_VVN
   (PP/p1 (P1/p_np by_II
    (NP/n2_name (NP/det_n the_AT (N1/n_of heir+s_NN2
     (PP/p1 (P1/p_np of_IO
      (NP/det_n the_AT (N1/ap_n1/- (AP/a1 (A1/a late_JJ))
       (N1/n John_NP1))))))) (NP/n1_posttit
        (N1/name+ T._NP1 (N1/n Dorrance_NP1)) Jr._NNSA1)))))))
 (Tacl/comma-e ,_, (S/whnp_vp who_PNQS
  (V/pp_pp control+ed_VVD
   (PP/p1 (P1/p_np about_II
    (NP/plu3 (N1/num2_nms (NP/num (N1/n 58_MC))
     (N1/n_of %_NNU (PP/p1 (P1/p_np of_IO
      (NP/n2_poss (NP/n1_name/- (N1/n Campbell_NP1)) 's+_POSS)
       (N1/n stock_NN1))))))))))
```

Relative Clause Attachment

An Anaphora Resolution Problem?

On the other hand, anaphora resolution algorithms based on:

- discourse oriented approaches (Hobbs, 1986; Reichman, 1978; Grosz, 1978)
- global focus (Grosz and Sidner, 1986)
- local focus (Carter, 1987; Webber, 1978; Sidner, 1981)

do not deal with relative clause attachment directly.

Agency and WordNet Classes

- It is useful to differentiate between *who* and *which* clauses
- Quirk (Quirk et al., 1985):

  > The relative pronoun *who* is used to refer to something with personality and *which* to something without.

- In terms of the WordNet hierarchy (Miller et al., 1993), *who* can only refer to hyponyms of
  - *humans*
  - *groups (organizations)*
  - *animals*
- *which* cannot refer to *humans*

Prepositional Preferences

- Lexicalization over prepositions is impractical
- We assume that prepositions only influence attachment indirectly, through their preferences for the agency of their arguments
- We classified the subject and object of 15000 prep occurrences

Prepositional Preferences

Prep	P_{who}	P_{which}	Prep	P_{who}	P_{which}
about	0.58	0.43	against	0.53	0.47
among	0.62	0.42	as	0.57	0.43
at	0.46	0.57	before	0.51	0.52
between	0.75	0.41	by	0.63	0.43
during	0.44	0.56	from	0.62	0.53
for	0.55	0.50	in	0.52	0.52
into	0.51	0.51	like	0.39	0.54
near	0.50	0.50	of	0.52	0.52
on	0.62	0.49	over	0.61	0.50
to	0.66	0.61	under	0.34	0.54
with	0.52	0.51	without	0.37	0.54

Probability of the preposition selecting for local attachment (for *who* and *which* clauses)

Boolean Features — 1

0: Target (wide attachment)
1: Target (local attachment)
2: Restrictive Clause (defined by absence of comma)
3: `NP1` is a *person*
4: `NP1` is a *group*
5: `NP1` is an *animal*
6: `NP1` is a *possession*
7: `NP1` is an *entity*
8: `NP1` is an *act*
9: `NP1` is an *abstraction*
10: `NP1` has no WordNet class
11: `NP1` is a proper noun
12: `NP1` is a definite NP (presence of definite determiner)
13: `NP1` has no determiner
14: `Prep` favours local attachment
15: `Prep` favours wide attachment

Boolean Features — 2

16: `NP2` is a *person*
17: `NP2` is a *group*
18: `NP2` is an *animal*
19: `NP2` is a *possession*
20: `NP2` is an *entity*
21: `NP2` is an *act*
22: `NP2` is an *abstraction*
23: `NP2` has no WordNet class
24: `NP2` is a proper noun
25: `NP2` is a definite NP (presence definite determiner)
26: `NP2` has no determiner
27: `Verb` selects for singular subject
28: `Verb` selects for plural subject
29: `NP1` is singular
30: `NP2` is singular

Machine Learning

Examples

An example is a list of the indexes of the features that are present in any particular sentence.

> **Ex:** 0,2,3,4,7,13,33,19,20,21,24,29,34:

Algorithm

SNoW machine learning package (Carlson et al., 1999) using the WINNOW algorithm.

Data

Parse trees from the Penn Wall Street Journal Treebank (Marcus et al., 1993)

Results (*who* clauses)

Data Set	Size	Baseline1	Baseline2	Winnow
Training Set	~200	66.5%	73.3%	91.6%
Test Set	~50	66.5%	73.3%	91.1%

Baseline1: Always attach locally.

Baseline2: Attach according to the preposition's preferences

Another baseline:

I converted the first 100 of these sentences to plain text and parsed them with the Briscoe and Carroll (1995) parser. An analysis of the parse trees gave:

- Recall = 62%
- Precision = 69.35%
- Local Attachment Baseline: Precision = 68%

Error Analysis

41% of the errors came from partitives like:

1. {hundreds|thousands|dozens} of investors
2. a {number|lot} of people
3. the percentage of Americans
4. the kind of guy

where `NP1` does not have a WordNet semantic class.

Example

A Fannie Mae seminar this week promises to attract hundreds of investors, who can be expected to channel tens of billions of dollars into the market.

can be simplified to either of:

A Fannie Mae seminar this week promises to attract hundreds of investors. These investors can be expected to channel tens of billions of dollars into the market.

and:

A Fannie Mae seminar this week promises to attract hundreds of investors. These hundreds of investors can be expected to channel tens of billions of dollars into the market.

Error Analysis

21% of errors arose because the network didn't learn the rule that a restrictive clause cannot attach to a proper noun (in instances like *A former backup singer for Ms Midler who had...*). If this is enforced as a hard rule, the precision figures go up by almost 2%. It is possible that this rule would have been learnt by the network had it been presented with more training data.

Many of the other errors arose because the network had genuinely little information to go on; for example:

- *Some 3.8 million of the 5 million who will...*
- *A major piece of Hollywood manpower who has...*

Results (*which* clauses)

Data Set	Size	Baseline1	Baseline2	Winnow
Training Set	~400	69.7%	62.7%	79.6%
Test Set	~50	69.7%	62.7%	76.5%

Baseline1: Always attach locally.

Baseline2: Attach according to the preposition's preferences

Handling Appositives

“She was an inspirational lady,” says Laura Dobson, a freshman at the University of South Carolina, *who had Mrs. Yeargin in the teacher-cadet class last year*.

One man who is likely to reap the benefits is Vaino Heikkinen, aged 67, a farmer in Lieksa, 10km from the Soviet border, *who claims a Finnish record for shooting 36 bears since 1948*.

The Algorithm

Given the pattern: ...NP_1, NP_2,...,NP_n, *who*..., the task is to select the NP_i that the relative clause refers to.

1. `IF` only one of the NPs is a Proper Noun `THEN` select it
2. `ELSE IF` more than one NP is a Proper Noun `THEN`
   (a) `IF` any of NP_2 through NP_n are a single word proper noun, `THEN` reject them
   (b) Of the rest, `IF` exactly one of them does not contain a preposition `THEN` select it
   (c) `ELSE IF` exactly one of NP_1 and NP_n is a Proper Noun `THEN` select it
3. `ELSE` By default, select NP_1

Examples

- ...<u>Joe Watson</u>, **the prosecutor in the case**, who is...
- ...**the smallest in terms of production**, <u>Chateau Petrus</u>, which costs...
- ...at least one member of the court, <u>Judge Robert Mayer</u>, **a former civil litigator**, who served...
- ...the microphone invented by my grandfather, <u>Emile Berliner</u>, which had...
- ...<u>**its main product**</u>, bleached paperboard, which goes...

blue: correct <u>underlined</u>:Selected by algorithm

bold: acceptable

Evaluation

Data Set	Size	Baseline* %		Algorithm 1 %	
		C	C or A	C	C or A
Training Set[1]	121	88.4	95.0	100.0	100.0
Test Set 1[2]	101	90.1	97.0	99.0	99.0
Test Set 2[3]	58	65.6	87.9	93.1	98.3

* Baseline: Always NP_1.

[1] Only *who* clauses.

[2] Only *who* clauses.

[3] Only *which* clauses.

C = Correct

A = Acceptable

Deciding Clause Boundaries

Determining where a relative clause ends is not always trivial. Non-restrictive relative clauses can extend to the end of the sentence or end with a comma. However, there might be commas internal to the clause so that at each comma after the clause starts, a decision needs to be made on whether the clause ends or not. We devised a set of heuristics for making this decision based on as manual examination of 290 non-restrictive *who* clauses and 846 non-restrictive *which* clauses in our training set derived from the Penn WSJ Treebank. These heuristics are encoded in algorithm 2 below.

Algorithm

1. LET n be the number of commas between *", {who|which}"* and the end of the sentence.
2. IF $n = 0$ THEN clause extends till the end of sentence
3. IF $n > 0$ THEN a decision needs to be made as follows
4. FOR each comma (scanning from left to right) DO
   (a) IF followed by an Appositive THEN INTERNAL comma
   (b) IF followed by a Verb Group THEN
       IF the verb has POS "VB{N|G}" THEN INTERNAL comma
       ELSE END of clause
   (c) IF an implicit conjunction of adjectives or adverbs like "JJ, JJ" or "RB, RB" THEN INTERNAL clause
   (d) IF its a *who* clause THEN
       IF ", CC who" THEN END of clause
       IF ", {which|when|where}" THEN INTERNAL comma
   (e) IF its a *which* clause THEN
       IF ", CC which" THEN END of clause
       IF ", {who|when|where}" THEN INTERNAL comma
5. ELSE by default end clause on first comma

Evaluation — 1

Data Set	Size	Accuracy[1]	Accuracy[2]
Training (*who*)	290	98.97%	97.84%
Training (*which*)	846	98.34%	96.80%
Test (*who*)	236	98.31%	96.75%
Test (*which*)	696	96.70%	94.20%

[1] For all clauses. [2] For only ambiguous clauses.

Evaluation — 2

The second evaluation was against the six systems that participated in the CoNLL-2001 clause identification workshop at ACL-2001 (Daelemans and Zajac, 2001).

Data Set	Average	Best[1]	Our Algo
Who Clauses	28.84%	80.77%	96.15%
Which Clauses	29.65%	80.52%	96.10%

[1] (Carreras and Màrquez, 2001)

The ANLT parser (Briscoe and Carroll, 1995) gave a recall of 77% and precision of 75% on the 26 ambiguous *who* clauses and a recall of 87% and precision of 78% on the 77 ambiguous *which* clauses.

Conclusions

- Local context plays an important role in making disambiguation decisions
- So do knowledge sources like WordNet
- These techniques require only POS tagged text and can be used to aid parse selection
- Machine learning techniques useful to determine the relative importance of features
- Sequential decision based algorithms enough when features consistently select one way or the other

Future Work

- The results presented in this paper pertain to the *analysis* and *transformation* stages suggested by Chandrasekar et al. (1996)
- In practice, we require a third stage—*generation*
- Issues in generation:
  - Generating referring expression (Siddharthan and Copestake, 2002)
  - Preserving discourse structure

References

Ted Briscoe and John Carroll. 1995. Developing and evaluating a probabilistic LR parser of part-of-speech and punctuation labels. In *Proceedings of the ACL/SIGPARSE 4th International Workshop on Parsing Technologies, Prague / Karlovy Vary, Czech Republic*, pages 48–58.

Andrew J. Carlson, Chad M. Cumby, Jeff L. Rosen, and Dan Roth. 1999. The SNoW learning architecture. Technical report, Tech. Report UIUCDCS-R-99-2101, UIUC Computer Science Department, May.

Xavier Carreras and Luís Màrquez. 2001. Boosting trees for clause splitting. In Walter Daelemans and Rémi Zajac, editors, *Proceedings of CoNLL-2001*, pages 73–75. Toulouse, France.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of English newspaper text to assist aphasic readers. In *Proceedings of AAAI98 Workshop on Integrating Artificial Intelligence and Assistive Technology, Madison, Wisconsin*.

John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying English text for language impaired readers. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Bergen, Norway.*

David Carter. 1987. *Interpreting Anaphors in Natural Language Texts*. Ellis Horwood, Chichester.

Raman Chandrasekar and Bangalore Srinivas. 1997. Automatic induction of rules for text simplification. *Knowledge-Based Systems*, 10:183–190.

Raman Chandrasekar, Christine Doran, and Bangalore Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the Sixteenth International Conference on Computational Linguistics (COLING '96)*, Copenhagen, Denmark.

Walter Daelemans and Rémi Zajac, editors. 2001. *Proceedings of CoNLL-2001*. Toulouse, France.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Barbara J. Grosz. 1978. Discourse analysis. In D.E. Walker, editor, *Understanding Spoken Language*, pages 235–268. North-Holland, New York.

Jerry R. Hobbs. 1986. Resolving pronoun references. In Barbara J. Grosz, Karen Sparck-Jones, and Bonnie L. Webber, editors, *Readings in Natural Language Processing*, pages 339–352. Morgan Kaufmann, Los Altos, California.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large natural language corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330.

George A. Miller, Richard Beckwith, Christiane D. Fellbaum, Derek Gross, and Katherine Miller. 1993. Five Papers on WordNet. Technical report, Princeton University, Princeton, N.J.

Geoffrey Nunberg. 1990. *The Linguistics of Punctuation*. Number 18 in CSLI Lecture Notes. Stanford University Press.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman.

Rachel Reichman. 1978. Conversational coherency. *Cognitive Science*, 2:283–327.

Advaith Siddharthan and Ann Copestake. 2002. Generating anaphora for simplifying text. In *4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2002)*.

Candace L. Sidner. 1981. Focusing for interpretation of pronouns. *American Journal of Computational Linguistics*, 7(4):217–231.

Bonnie L. Webber. 1978. A formal approach to discourse anaphora. Technical Report 3761, Cambridge, Massachusetts.