ESSLLI 2006 Workshop on Coherence

Entity versus Rhetorical Coherence for Information Ordering:
Initial Experimentation

Nikiforos Karamanis

Nikiforos.Karamanis@cl.cam.ac.uk

Natural Language and
Information Processing group
Computer Laboratory

UNIVERSITY OF
CAMBRIDGE

Overview

- Background
- Centering theory (applied to GNOME)
- Supplementing Centering
- Evaluating the new model on GNOME: Methodology and Results

Coherence

- A felicitous text has to be coherent: the content has to be organised in a way that is easy to read and comprehend.
- Arbitrary reordering the sentences that an understandable text consists of gives rise to documents that do not make sense.

Entity coherence

- Important aspect of textual felicity.
- Arises from the way NP referents relate subsequent clauses in the text.
- Centering Theory (Grosz et al. 1995): An influential framework for modelling entity coherence.
- Contrasts with models of relational coherence such as Rhetorical Structure Theory (Mann and Thompson 1987).

Information ordering

- Deciding in which sequence to present a set of preselected information-bearing items.
- Important problem in automatic text production.

(Barzilay et al. 2002, Lapata 2003, Barzilay and Lee 2004, Barzilay and Lapata 2005, Bollegala et al. 2006, Ji and Pullman 2006, inter alia)

Centering for information ordering

- Assume a system which receives an unordered set of clauses as its input...

  and uses a metric of coherence to output the highest scoring ordering of these clauses.
- Define metrics of coherence using Centering and compare them with each other to find the most appropriate one for information ordering (Karamanis 2003).

Experiments on GNOME

- 20 descriptions of museum artefacts (museum labels) from the GNOME corpus (GNOME-LAB).
- Simplest metric (and most remote to Centering) sets a baseline which cannot be overtaken by other metrics which utilise additional Centering-specific notions.
- However, the baseline does not perform well enough to be used on its own for information ordering (Karamanis et al. 2004).

This talk

- Can the model of local rhetorical coherence suggested by Knott et al. (2001) boost the performance of the metrics?
- Same metrics and texts as in our previous work.

The GNOME corpus

- Texts from various genres reliably annotated for features relevant to Centering (Poesio et al. 2004):
  - finite units
  - grammatical role of NPs
  - coreferent NPs
- Characteristic museum label:

(1) (a) [Item]$_S$144 is a torc. (b) Its present [arrangement]$_S$, twisted into three rings, may be a modern alteration; (c) [it]$_S$ should probably be a single ring, worn around the neck. (d) The [terminals]$_S$ are in the form of goats' heads.

Computing the CF list

For each finite unit:

- Rank NP referents (aka forward looking centers, CFs) in order of prominence according to the function of Brennan et al. (1987).
- The referent of the NP which is marked as subject is the first member of CF list (aka CP).
- Ranking ties between referents are resolved according to linear order of corresponding NPs.
- Intra-unit coreference: Highest ranked grammatical role is used to place referent in the CF list.

The backward looking center (CB)

- The highest member of the current CF list which also appears in the previous list.

CF(1a): {*Item*:de374, *torc*:de375}

CF(1b): {*arrangement*:de376, *its*:de374, … }

CF(1c): {*it*:de374, *ring*:de379, … }

CF(1d): {*terminals*:de380, *form*:de381, … }

Transitions, Coherence, Salience

Brennan et al. (1987), Kibble and Power (2000, 2004)

	Coherence: CB(Un)=CB(Un-1) or CB(Un-1) undef.	Coherence*: CB(Un)≠CB(Un-1)
Salience: CB(Un)=CP(Un)	Continue	Smooth-Shift
Salience*: CB(Un)≠CP(Un)	Retain	Rough-Shift

NOCB: No referents in common between two subsequent CF lists.

Transition preferences, Cheapness

- Brennan et al. (1987):

  Continue >> Retain >> S-Shift >> R-Shift
- This is the same as ranking Coherence over Salience (Kibble 2001, Beaver 2004).
- Cheapness (Strube and Hahn 1999): CB(Un)=CP(Un-1)

Transitions and Cheapness

CF(1a): {*Item*:de374, *torc*:de375}

CF(1b): {*arrangement*:de376, *its*:de374, … }

Retain

CF(1c): {*it*:de374, *ring*:de379, … }

Continue, Cheapness*

CF(1d): {*terminals*:de380, *form*:de381, … }

NOCB

Supplementing Centering

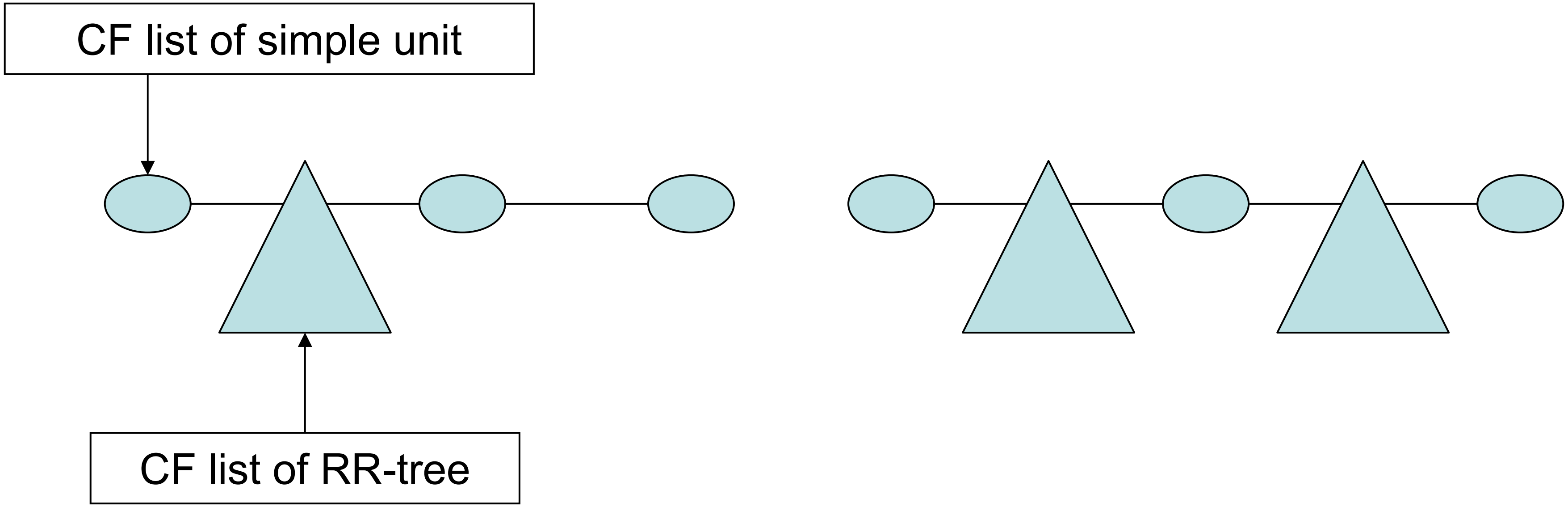
- Most Centering-based studies ignore other coherence-inducing factors (see e.g. the studies in Walker et al. 1998).
- Centering needs to be supplemented with other models of coherence (Kibble 2001):

  The model of Knott et al. (2001) seems to be a good candidate in our domain of interest (Poesio et al. 2004).

Beyond elaboration (Knott et al. 2001)

- Elaboration does not capture entity coherence adequately: Use more appropriate models such as Centering instead.
- Entity coherence in descriptive texts is supplemented by trees of Rhetorical Relations (RR-trees) which apply *locally*, i.e. between adjacent clauses.

A hybrid coherence model



RR-trees in GNOME

- RR-trees identified using cue phrases such as *because*, *but*, *although*, etc (Knott and Dale 1994).
- 19 RR-trees in 12 texts (GNOME-RR).
- In all but one case, the finite units related via a local RR appear within the same sentence.
- 15 sentences consisting of more than one finite unit which are not related to each other with an explicit cue (e.g. units 1b and 1c; C.f. Power et al. 2003).
- Computing CF lists for RR-trees reduces the number of CF lists per text by 1.58 lists on average.

Example of an RR-tree in GNOME

(2a) Access to the cartonnier’s lower half can only be gained by the doors at the sides,

(2b) because the table would have blocked the front.

CF(2a): {*access*:de12, *half*:de13, … }

CF(2b):{*table*:de9, *front*:de18}

Computing the CF list of the RR-tree

Use *sentence* instead of *finite unit* and keep all other Centering parameters such as CF ranking the same:

(3) Access to the cartonnier's lower half can only be gained by the doors at the sides, because the table would have blocked the front.

CF(3): {*access*:de12, *table*:de9 … }

Input to information ordering

- Think of the set of CF lists as the unordered input set of information-bearing items:

CF: {*Item*:de374, *torc*:de375}, CF: {*terminals*:de380, *form*:de381, … }, CF: {*it*:de374, *ring*:de379, … }, CF: {*arrangement*:de376, *its*:de374, … }

Centering for information ordering

- Is it possible to order the CF lists using (combinations of) other Centering-based notions such as the CB, transitions, principles, etc?

1. CF: {*Item*:de374, *torc*:de375}
2. CF: {*arrangement*:de376, *its*:de374, … }
3. CF: {*it*:de374, *ring*:de379, … }
4. CF: {*terminals*:de380, *form*:de381, … }

Assumptions

- Assume an approach to information ordering in which a metric is used to select the best scoring ordering of CF lists among several alternatives (Karamanis and Manurung 2002, Althaus et al. 2004).

(C.f. Mellish et al. 1998, Kibble and Power 2000/2004, Cheng 2000/2002, inter alia).

Select the best scoring ordering

Possible orderings:	**Scores by metric M:**
Ordering1	Score1
…	…
OrderingN	ScoreN

Centering-based metrics

- Centering concepts such as the various types of transitions or the different kinds of principles can be combined with each other in more or less complicated ways to define a large number of metrics (Karamanis 2003, chapter 3).

Possible metrics

- M.NOCB: Simply prefer ordering with fewest NOCBs.
- M.CHEAP: Fewest violations of Cheapness (Strube and Hahn 1999).
- M.BFP: Transition preferences of Brennan et al. (1987):
  Continue > Retain > Smooth-Shift > Rough-Shift
- M.KP: Lowest total cost of NOCBs plus violations of Cheapness, Coherence and Salience (Kibble and Power 2000).

- NOCBs ranked higher than the sum of other costs (Kibble and Power 2004).
- Other possible combinations and rankings:
  - NOCB>Coherence*>Cheapness*>Salience* (Karamanis 2003)
  - Coherence*> Cheapness*+Salience* (Kibble 2001)
  - …
- NOCB+Rough-Shift (Miltsakaki and Kukich 2000/2004).
- Continue > Retain > Shift (Grosz et al. 1995).
- Transitions in Strube and Hahn (1999).
- …

Which metrics?

- Present a general methodology for identifying automatically and prior to the actual task which of the many possible metrics represent the most promising candidates for information ordering.

Corpus-based experiments

- Plausible alternative to (generally more costly) human-based evaluation.
- Needs to be integrated with other types of evaluation for which it provides testable hypotheses (Reiter and Sripada 2002).

Main premise

The ordering of CF lists in a corpus text (GSO) represents a good solution:

- How likely is a metric to prefer the GSO over its alternatives?
- If a metric takes an alternative ordering to score better than the GSO, it has to be penalised.

Experimental question

Possible orderings:	**Scores by metric M:**
Ordering1	Score1
…	…
OrderingC: ordering in the corpus	ScoreC: best score
…	…
OrderingN	ScoreN

Methodology

- GSO: ordering of CF lists observed in the text.
- M: metric of coherence.
- Search through the space of possible orderings and compute:
  - Better(M, GSO): the percentage of orderings that score better than the GSO according to M.
  - Equal(M, GSO): the percentage of orderings that score equal to the GSO according to M.

Classification rate

- Weighted sum of the percentage of alternative orderings that score equally to or better than the GSO:

$$u(M, GSO)=Better(M, GSO) + Equal(M, GSO)/2$$

Smaller classification rates are better:

If $u(M_y, GSO)$ is smaller than $u(M_x, GSO)$, then M_y is more suitable than M_x for producing the GSO.

Experimenting on a corpus

- Calculate how many texts in a corpus C return a lower classification rate for M_y than for M_x (and vise versa).
- Check whether the difference in the number of texts is significant using the Sign Test.
- If there exist significantly more texts with a lower classification rate for M_y than for M_x, then M_y is more suitable than M_x for producing the orderings in C.

Experimental questions

- Which is the best performing metric in GNOME-RR?
- Is it different from the one in GNOME-LAB?
- Does taking RR-trees into account improve the performance of the metrics?

GNOME-RR corpus	M.NOCB			p
	lower	greater	ties	
M.CHEAP	10	2	0	0.038
M.KP	11	1	0	0.006
M.BFP	7	5	0	0.774
N	12			

Table 1: Comparing the baseline M.NOCB with more complicated Centering-based metrics in GNOME-RR.

GNOME-LAB	M.NOCB			p
corpus	lower	greater	ties	
M.CHEAP	18	2	0	<0.000
M.KP	16	2	2	0.002
M.BFP	12	3	5	0.036
N	20			

Table 2: Comparing the baseline M.NOCB with more complicated Centering-based metrics in GNOME-LAB.

	GNOME-RR			p
metrics	lower	greater	ties	
M.NOCB	3	9	0	0.146
M.CHEAP	9	3	0	0.146
M.KP	10	2	0	0.038
M.BFP	5	7	5	0.774
N	12			

Table 3: Changes in classification rate of the metrics in GNOME-RR compared to their performance in GNOME-LAB.

Discussion

- M.NOCB is a better choice for ordering the CF lists of museum labels than M.KP, M.CHEAP and M.BFP.
- This holds irrespective of whether RR-trees are taken into account for the computation of the CF lists.
- About 1/4 of alternative orderings in GNOME-RR (and 1/5 in GNOME-LAB) are taken to be more coherent than GSO.

Related work

- Poesio et al. (2004): evaluate different configurations of Centering (mainly different ways of computing the CF list).
- Prefer configurations with smaller number of disfavoured transitions such as NOCBs.
- But disfavoured transitions are very frequent: 57% of transitions in GNOME-LAB and 53% in GNOME-RR are NOCBs.
- This suggests that:
  - Entity coherence has little to do with our domain.
  - Things (slightly) improve when we use RR-trees.

A different perspective

- Average classification rate Y of M.NOCB is about 20% in GNOME-LAB and 23% in GNOME-RR.
- The GSO has fewest NOCBs than the overwhelming majority (between 77% and 80%) of alternative orderings: so entity coherence appears to be quite relevant!
- Yet, things are getting worse when RR-trees are taken into account!

Supporting results

- Baseline overwhelmingly beating its competitors in three other corpora.
- Trying to enhance the metrics with a global focus feature has the same negative effect as the RR-trees in the GNOME domain.

MPIRO corpus	M.NOCB			p
	lower	greater	ties	
M.CHEAP	110	12	0	<.000
M.KP	103	16	3	<.000
M.BFP	41	31	49	.121
N	122			

Table 4: Comparing the baseline M.NOCB with more complicated Centering-based metrics in the MPIRO corpus

- Y(M.NOCB)= 20% (very similar to GNOME-LAB)

NEWS corpus	M.NOCB			p
	lower	greater	ties	
M.CHEAP	155	44	1	<0.000
M.KP	131	68	1	<0.000
M.BFP	121	71	8	<0.000
N	200			

Table 5: Comparing the baseline M.NOCB with more complicated Centering-based metrics in the NEWS corpus.

- Y(M.NOCB)= 31% (worst performance)

ACCS corpus	M.NOCB			p
	lower	greater	ties	
M.CHEAP	183	17	0	<0.000
M.KP	167	33	0	<0.000
M.BFP	100	100	0	1.000
N	200			

Table 6: Comparing the baseline M.NOCB with more complicated Centering-based metrics in the ACCS corpus.

- Y(M.NOCB)= 16% (best performance)

	M.NOCB			p
	lower	greater	ties	
PF.NOCB (GNOME-LAB)	9	2	9	0.066
PF.NOCB (GNOME-RR)	6	0	6	0.032

Table 7: Comparing the baseline M.NOCB with a version which incorporates a constraint on global focus (PF) in GNOME-LAB and GNOME-RR.

- Y(M.NOCB) = 20% in GNOME-LAB, 23% in GNOME-RR.
- Y(PF.NOCB)= 22% in GNOME-LAB, 27% in GNOME-RR.

Conclusions

- M.NOCB, the metric which is most remotely related to Centering, is the most appropriate for information ordering (in all investigated genres).
- Local RR-trees in the GNOME domain do not help.
- Provide researchers with a simple and easily extendable evaluation framework as well as a robust baseline to deploy for their own meaningful comparisons.

Acknowledgments

Joined work with:

- Chris Mellish, Uni of Aberdeen.
- Massimo Poesio, Uni of Essex / Uni of Trento.
- Jon Oberlander, Uni of Edinburgh.

Special thanks to:

- James Soutter, Uni of Edinburgh.

References

Grosz et al. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Kibble (2001). A Reformulation of Rule 2 of Centering Theory. *Computational Linguistics* 27 (4):579-587.

Poesio et al. (2004). Centering: a parametric theory and its instantiations. *Computational Linguistics* 30 (3):309-363.

Walker et al. (1998). Centering in naturally occurring discourse: An overview. In Walker et al. (eds.) *Centering Theory in Discourse*. Clarendon Press, Oxford: 1-30.

References

Kibble and Power (2004). Optimising Referential Coherence in Text Generation. *Computational Linguistics* 30 (4):401-416.

Knott et al. (2001). Beyond Elaboration: The Interaction of Relations and Focus in Coherent Text. In Sanders et al. (eds.) *Text Representation: Linguistic and Psycholinguistic Aspects.* J. Benjamins, Amsterdam, 181-196.

Karamanis et al. (2004). Evaluating centering-based metrics of coherence using a reliably annotated corpus. *Proceedings of ACL 2004:* 391–398.

Karamanis (2003). *Entity Coherence for Descriptive Text Structuring*. Ph.D. thesis, Division of Informatics, University of Edinburgh.