Evaluating Centering for Sentence Ordering in Two New Domains

Nikiforos Karamanis
Natural Language and Information Processing Group
Computer Laboratory
University of Cambridge
Nikiforos.Karamanis@cl.cam.ac.uk

Abstract

This paper builds on recent research investigating sentence ordering in text production by evaluating the Centering-based metrics of coherence employed by Karamanis et al. (2004) using the data of Barzilay and Lapata (2005). This is the first time that Centering is evaluated empirically as a sentence ordering constraint in several domains, verifying the results reported in Karamanis et al.

1 Introduction

As most literature in text linguistics argues, a felicitous text should be *coherent* which means that the content has to be organised in a way that makes the text easy to read and comprehend. The easiest way to demonstrate this claim is by arbitrarily reordering the sentences that an understandable text consists of. This process very often gives rise to documents that do not make sense although the information content remains the same. Hence, deciding in which sequence to present a set of preselected information-bearing items is an important problem in automatic text production.

Entity coherence, which arises from the way NP referents relate subsequent sentences in the text, is an important aspect of textual felicity. *Centering Theory* (Grosz et al., 1995) has been an influential framework for modelling entity coherence in computational linguistics in the last two decades. Karamanis et al. (2004) were the first to evaluate Centering-based metrics of coherence for ordering clauses in a subset of the GNOME corpus (Poesio et al., 2004) consisting of 20 artefact descriptions. They introduced a novel experimental methodology that treats the observed ordering of clauses in a text as the gold standard, which is scored by each metric. Then, the metric is penalised proportionally to the amount of alternative orderings of the same material that score equally to or better than the gold standard.

This methodology is very similar to the way Barzilay and Lapata (2005) evaluate automatically another model of coherence called the entity grid using a larger collection of 200 articles from the North American News Corpus (NEWS) and 200 accident narratives from the National Transportation Safety Board database (ACCS). The same data and similar methods were used by Barzilay and Lee (2004) to compare their probabilistic approach for ordering sentences with that of Lapata (2003).

This paper discusses how the Centering-based metrics of coherence employed by Karamanis et al. can be evaluated on the data prepared by Barzilay and Lapata. This is the first time that Centering is evaluated empirically as a sentence ordering constraint in more than one domain, verifying the results reported in Karamanis et al.

The paper also contributes by emphasising the following methodological point: To conduct our experiments, we need to produce several alternative orderings of sentences and compare them with the gold standard. As the number of possible orderings grows factorially, enumerating them exhaustively (as Barzilay and Lee do) becomes impractical. In this paper, we make use of the methods of Karamanis (2003) which allow us to explore a

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 65–68,
New York, June 2006.

Table 1A	NP referents						
Sentences	department	trial	microsoft	...	products	brands	...
(a)	S	O	S	...	–	–	...
(b)	–	–	O	...	S	O	...

Table 1B	CF list:			CHEAPNESS
Sentences	{CP, next two referents}	CB	Transition	CB_n=CP_{n-1}
(a)	{department, microsoft, trial, ...}	n.a.	n.a.	n.a.
(b)	{products, microsoft, brands, ...}	microsoft	RETAIN	*

Table 1: (A) Fragment of the entity grid for example (1); (B) CP (i.e. first member of the CF list), next two referents, CB, transition and violations of CHEAPNESS (denoted with a $*$) for the same example.

sufficient number of alternative orderings and return more reliable results than Barzilay and Lapata, who used a sample of just 20 randomly produced orderings (often out of several millions).

2 Materials and methods

2.1 Centering data structures

Example (1) presents the first two sentences of a text in NEWS (Barzilay and Lapata, Table 2):

(1) (a) [The Justice Department]$_S$ is conducting [an anti-trust trial]$_O$ against [Microsoft Corp.]$_X$ with [evidence]$_X$ that [the company]$_S$ is increasingly attempting to crush [competitors]$_O$. (b) [Microsoft]$_O$ is accused of trying to forcefully buy into [markets]$_X$ where [its own products]$_S$ are not competitive enough to unseat [established brands]$_O$. (...)

Barzilay and Lapata automatically annotated their corpora for the grammatical role of the NPs in each sentence (denoted in the example by the subscripts S, O and X for subject, object and other respectively)[1] as well as their coreferential relations. This information is used as the basis for the computation of the entity grid: a two-dimensional array that captures the distribution of NP referents across sentences in the text using the aforementioned symbols for their grammatical role and "–" for a referent that does not occur in a sentence. Table 1A illustrates a fragment of the grid for the sentences in example (1).[2]

Our data transformation script computes the basic structure of Centering (known as CF list) for each row of the grid using the referents with the symbols S, O and X (Table 1B). The members of the CF list are ranked according to their grammatical role (Brennan et al., 1987) and their position in the grid.[3] The derived sequence of CF lists can then be used to compute other important Centering concepts:

- The CB, i.e. the referent that links the current CF list with the previous one such as microsoft in (b).
- Transitions (Brennan et al., 1987) and NOCBs, that is, cases in which two subsequent CF lists do not have any referent in common.
- Violations of CHEAPNESS (Strube and Hahn, 1999), COHERENCE and SALIENCE (Kibble and Power, 2000).

2.2 Metrics of coherence

Karamanis (2003) assumes a system which receives an unordered set of CF lists as its input and uses a metric to output the highest scoring ordering. He discusses how Centering can be used to define many different metrics of coherence which might be useful for this task. In our experiments we made use of the four metrics employed in Karamanis et al. (2004):

- The baseline metric M.NOCB which simply prefers the ordering with the fewest NOCBs.
- M.CHEAP which selects the ordering with the fewest violations of CHEAPNESS.
- M.KP, introduced by Kibble and Power, which sums up the NOCBs as well as the violations of CHEAPNESS, COHERENCE and SALIENCE, preferring the ordering with the lowest total cost.
- M.BFP which employs the transition preferences of Brennan et al.

[1] Subjects in passive constructions such as "Microsoft" in (1b) are marked with O.

[2] If a referent such as microsoft is attested by several NPs, e.g. "Microsoft Corp." and "the company" in (1a), the role with the highest priority (in this case S) is used.

[3] The referent department appears in an earlier grid column than microsoft because "the Justice Department" is mentioned before "Microsoft Corp." in the text. Since grid position corresponds to order of mention, the former can be used to resolve ties between referents with the same grammatical role in the CF list similarly to the use of the latter e.g. by Strube and Hahn.

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

Table 2: Comparing M.NOCB with M.CHEAP, M.KP and M.BFP in the NEWS corpus.

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

Table 3: Comparing M.NOCB with M.CHEAP, M.KP and M.BFP in the ACCS corpus.

2.3 Experimental methodology

As already mentioned, previous work assumes that the gold standard ordering (GSO) observed in a text is more coherent than any other ordering of the sentences (or the corresponding CF lists) it consists of. If a metric takes a randomly produced ordering to be more coherent than the GSO, it has to be penalised.

Karamanis et al. (2004) introduce a measure called the *classification rate* which estimates this penalty as the weighted sum of the percentage of alternative orderings that score equally to or better than the GSO.[4] When comparing several metrics with each other, the one with the lowest classification rate is the most appropriate for sentence ordering.

Karamanis (2003) argues that computing the classification rate using a random sample of one million orderings provides reliable results for the entire population of orderings. In our experiments, we used a random sample of that size for GSOs which consisted of more than 10 sentences. This allows us to explore a sufficient portion of possible orderings (without having to exhaustively enumerate every ordering as Barzilay and Lee do). Arguably, our experiments also return more reliable results than those of Barzilay and Lapata who used a sample of just a few randomly produced orderings.

Since the Centering-based metrics can be directly deployed on unseen texts without any training, we treated all texts in NEWS and ACCS as testing data.[5]

[4]The classification rate is computed according to the formula Better(M,GSO) + Equal(M,GSO)/2. Better(M,GSO) stands for the percentage of orderings that score better than the GSO according to a metric M, whilst Equal(M,GSO) is the percentage of orderings that score equal to the GSO.

[5]By contrast, Barzilay and Lapata used 100 texts in each domain to train their probabilistic model and 100 to test it. Note that although they experiment with quite large corpora their reported results are not verified by statistical tests.

3 Results

The experimental results of the comparisons of the metrics from section 2.2 are reported in Table 2 for the NEWS corpus and in Table 3 for ACCS. Following Karamanis et al., the tables compare the baseline metric M.NOCB with each of M.CHEAP, M.KP and M.BFP. The exact number of GSOs for which the classification rate of M.NOCB is lower than its competitor for each comparison is reported in the second column of the Table. For example, M.NOCB has a lower classification rate than M.CHEAP for 155 (out of 200) GSOs from NEWS. M.CHEAP achieves a lower classification rate for just 44 GSOs, while there is a single tie in which the classification rate of the two metrics is the same. The p value returned by the two-tailed sign test for the difference in the number of GSOs, rounded to the third decimal place, is reported in the fifth column of Table 2.[6]

Overall, the Table shows that M.NOCB does significantly better in NEWS than the other three metrics which employ additional Centering concepts. Similarly, M.CHEAP and M.KP are overwhelmingly beaten by the baseline in ACCS. Also note that since M.BFP fails to significantly overtake M.NOCB in ACCS, the baseline can be considered the most promising solution in that case too by applying Occam's razor.

Table 4 shows the results of the evaluation of the metrics in GNOME from Karamanis et al. These results are strikingly similar to ours despite the much smaller size of their sample. Hence, M.NOCB is the most suitable among the investigated metrics for ordering the CF lists in both NEWS and ACCS in addition to GNOME.

[6]The sign test was chosen by Karamanis et al. to test significance because it does not carry specific assumptions about population distributions and variance.

GNOME corpus	M.NOCB			p
	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 of texts	20			

Table 4: Comparing M.NOCB with M.CHEAP, M.KP and M.BFP in the GNOME corpus.

4 Discussion

Our experiments have shown that the baseline M.NOCB performs better than its competitors. This in turn indicates that simply avoiding NOCB transitions is more relevant to sentence ordering than the additional Centering concepts employed by the other metrics.

But how likely is M.NOCB to come up with the GSO if it is actually used to guide an algorithm which orders the CF lists in our corpora? The *average classification rate* of M.NOCB is an estimate of exactly this variable.

The average classification rate for M.NOCB is 30.90% in NEWS and 15.51% in ACCS. The previously reported value for GNOME is 19.95%.[7] This means that on average M.NOCB takes approximately 1 out of 3 alternative orderings in NEWS and 1 out of 6 in ACCS to be more coherent that the GSO. As already observed by Karamanis et al., there results suggest that M.NOCB cannot be put in practical use.

However, the fact that M.NOCB is shown to overtake its Centering-based competitors across several corpora means that it is a simple, yet robust, baseline against which other similar metrics can be tested. For instance, Barzilay and Lapata report a ranking accuracy of around 90% for their best grid-based sentence ordering method, which we take to correspond to a classification rate of approximately 10% (assuming that there do not exist any equally scoring alternative orderings). This amounts to an improvement over M.NOCB of almost 5% in ACCS and 20% in NEWS.

Given the deficiencies of the evaluation in Barzilay and Lapata, this comparison can only be provisional. In our future work, we intend to directly evaluate their method using a substantially large number of alternative orderings and M.NOCB as the baseline. We will also try to supplement M.NOCB with other features of coherence to improve its performance.

[7] The variability is presumably due to the different characteristics of each corpus (which do not prevent M.NOCB from always beating its competitors).

Acknowledgments

Many thanks to Regina Barzilay and Mirella Lapata for their data, to Le An Ha for the data transformation script and to Chris Mellish, Massimo Poesio and three anonymous reviewers for comments. Support from the Rapid Item Generation project (Wolverhampton University) and the BBSRC-funded Flyslip grant (No 16291) is also acknowledged.

References

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of ACL 2005*, pages 141–148.

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models with applications to generation and summarization. In *Proceedings of HLT-NAACL 2004*, pages 113–120.

Susan E. Brennan, Marilyn A. Friedman [Walker], and Carl J. Pollard. 1987. A centering approach to pronouns. In *Proceedings of ACL 1987*, pages 155–162, Stanford, California.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Nikiforos Karamanis, Massimo Poesio, Chris Mellish, and Jon Oberlander. 2004. Evaluating centering-based metrics of coherence using a reliably annotated corpus. In *Proceedings of ACL 2004*, pages 391–398, Barcelona, Spain.

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

Rodger Kibble and Richard Power. 2000. An integrated framework for text planning and pronominalisation. In *Proceedings of INLG 2000*, pages 77–84, Israel.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of ACL 2003*, pages 545–552, Saporo, Japan, July.

Massimo Poesio, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. 2004. Centering: a parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363.

Michael Strube and Udo Hahn. 1999. Functional centering: Grounding referential coherence in information structure. *Computational Linguistics*, 25(3):309–344.