Using a Corpus of Sentence Orderings Defined by Many Experts to Evaluate Metrics of Coherence for Text Structuring

Nikiforos Karamanis
Computational Linguistics Research Group
University of Wolverhampton, UK
N.Karamanis@wlv.ac.uk

Chris Mellish
Department of Computing Science
University of Aberdeen, UK
cmellish@csd.abdn.ac.uk

Abstract

This paper addresses two previously unresolved issues in the automatic evaluation of Text Structuring (TS) in Natural Language Generation (NLG). First, we describe how to verify the generality of an existing collection of sentence orderings defined by one domain expert using data provided by additional experts. Second, a general evaluation methodology is outlined which investigates the previously unaddressed possibility that there may exist many optimal solutions for TS in the employed domain. This methodology is implemented in a set of experiments which identify the most promising candidate for TS among several metrics of coherence previously suggested in the literature.[1]

1 Introduction

Research in NLG focused on problems related to TS from very early on, [McKeown, 1985] being a classic example. Nowadays, TS continues to be an extremely fruitful field of diverse active research. In this paper, we assume the so-called search-based approach to TS [Karamanis *et al.*, 2004] which employs a *metric* of coherence to select a text structure among various alternatives. The TS module is hypothesised to simply *order* a preselected set of information-bearing items such as sentences [Barzilay *et al.*, 2002; Lapata, 2003; Barzilay and Lee, 2004] or database facts [Dimitromanolaki and Androutsopoulos, 2003; Karamanis *et al.*, 2004].

Empirical work on the evaluation of TS has become increasingly automatic and corpus-based. As pointed out by [Karamanis, 2003; Barzilay and Lee, 2004] inter alia, using corpora for automatic evaluation is motivated by the fact that employing human informants in extended psycholinguistic experiments is often simply unfeasible. By contrast, large-scale automatic corpus-based experimentation takes place much more easily.

[Lapata, 2003] was the first to present an experimental setting which employs the *distance between two orderings* to estimate automatically how close a sentence ordering produced by her probabilistic TS model stands in comparison to orderings provided by several human judges.

[Dimitromanolaki and Androutsopoulos, 2003] derived sets of facts from the database of MPIRO, an NLG system that generates short descriptions of museum artefacts [Isard *et al.*, 2003]. Each set consists of 6 facts each of which corresponds to a sentence as shown in Figure 1. The facts in each set were manually assigned an order to reflect what a domain expert, i.e. an archaeologist trained in museum labelling, considered to be the most natural ordering of the corresponding sentences. Patterns of ordering facts were automatically learned from the corpus created by the expert. Then, a classification-based TS approach was implemented and evaluated in comparison to the expert's orderings.

Database fact		**Sentence**
`subclass(ex1, amph)`	→	This exhibit is an amphora.
`painted-by(ex1, p-Kleo)`	→	This exhibit was decorated by the Painter of Kleofrades.
`painter-story(p-Kleo, en4049)`	→	The Painter of Kleofrades used to decorate big vases.
`exhibit-depicts(ex1, en914)`	→	This exhibit depicts a warrior performing splachnoscopy before leaving for the battle.
`current-location(ex1, wag-mus)`	→	This exhibit is currently displayed in the Martin von Wagner Museum.
`museum-country(wag-mus, ger)`	→	The Martin von Wagner Museum is in Germany.

Figure 1: MPIRO database facts corresponding to sentences

A subset of the corpus created by the expert in the previous study (to whom we will henceforth refer as E0) is employed by [Karamanis *et al.*, 2004] who attempt to distinguish between many metrics of coherence with respect to their usefulness for TS in the same domain. Each human ordering of facts in the corpus is scored by each of these metrics which are then penalised proportionally to the amount of alternative orderings of the same material that are found to score equally to or better than the human ordering. The few metrics which manage to outperform two simple baselines in their overall performance across the corpus emerge as the most suitable candidates for TS in the investigated domain. This methodology is very similar to the way [Barzilay and Lee, 2004] evaluate their probabilistic TS model in comparison to the approach of [Lapata, 2003].

Because the data used in the studies of [Dimitromanolaki

[1] Chapter 9 of [Karamanis, 2003] reports the study in more detail.

and Androutsopoulos, 2003] and [Karamanis *et al.*, 2004] are based on the insights of just one expert, an obvious unresolved question is whether they reflect general strategies for ordering facts in the domain of interest. This paper addresses this issue by enhancing the dataset used in the two studies with orderings provided by three additional experts. These orderings are then compared with the orders of E0 using the methodology of [Lapata, 2003]. Since E0 is found to share a lot of common ground with two of her colleagues in the ordering task, her reliability is verified, while a fourth "stand-alone" expert who uses strategies not shared by any other expert is identified as well.

As in [Lapata, 2003], the same dependent variable which allows us to estimate how different the orders of E0 are from the orders of her colleagues is used to evaluate some of the metrics which perform best in [Karamanis *et al.*, 2004]. As explained in the next section, in this way we investigate the previously unaddressed possibility that there may exist many optimal solutions for TS in our domain. The results of this additional evaluation experiment are presented and emphasis is laid on their relation with the previous findings.

Overall, this paper addresses two general issues: a) how to verify the generality of a dataset defined by one expert using sentence orderings provided by other experts and b) how to employ these data for the automatic evaluation of a TS approach. Given that the methodology discussed in this paper does not rely on the employed metrics of coherence or the assumed TS approach, our work can be of interest to any NLG researcher facing these questions.

The next section discusses how the methodology implemented in this study complements the methods of [Karamanis *et al.*, 2004]. After briefly introducing the employed metrics of coherence, we describe the data collected for our experiments. Then, we present the employed dependent variable and formulate our predictions. In the results section, we state which of these predictions were verified. The paper is concluded with a discussion of the main findings.

2 An additional evaluation test

As [Barzilay *et al.*, 2002] report, different humans often order sentences in distinct ways. Thus, there might exist more than one equally good solution for TS, a view shared by almost all TS researchers, but which has not been accounted for in the evaluation methodologies of [Karamanis *et al.*, 2004] and [Barzilay and Lee, 2004].[2]

Collecting sentence orderings defined by many experts in our domain enables us to investigate the possibility that there might exist many good solutions for TS. Then, the measure of [Lapata, 2003], which estimates *how close* two orderings stand, can be employed not only to verify the reliability of E0 but also to compare the orderings preferred by the assumed TS approach with the orderings of the experts.

However, this evaluation methodology has its limitations as well. Being engaged in other obligations, the experts normally have just a limited amount of time to devote to the NLG researcher. Similarly to standard psycholinguistic experiments, consulting these informants is difficult to extend to a larger corpus like the one used e.g. by [Karamanis *et al.*, 2004] (122 sets of facts).

In this paper, we reach a reasonable compromise by showing how the methodology of [Lapata, 2003] supplements the evaluation efforts of [Karamanis *et al.*, 2004] using a similar (yet by necessity smaller) dataset. Clearly, a metric of coherence that has already done well in the previous study, gains extra bonus by passing this additional test.

3 Metrics of coherence

[Karamanis, 2003] discusses how a few basic notions of coherence captured by Centering Theory (CT) can be used to define a large range of metrics which might be useful for TS in our domain of interest.[3] The metrics employed in the experiments of [Karamanis *et al.*, 2004] include:

> M.NOCB which penalises NOCBs, i.e. pairs of adjacent facts without any arguments in common [Karamanis and Manurung, 2002]. Because of its simplicity M.NOCB serves as the first baseline in the experiments of [Karamanis *et al.*, 2004].
>
> PF.NOCB, a second baseline, which enhances M.NOCB with a global constraint on coherence that [Karamanis, 2003] calls the PageFocus (PF).
>
> PF.BFP which is based on PF as well as the original formulation of CT in [Brennan *et al.*, 1987].
>
> PF.KP which makes use of PF as well as the recent reformulation of CT in [Kibble and Power, 2000].

[Karamanis *et al.*, 2004] report that PF.NOCB outperformed M.NOCB but was overtaken by PF.BFP and PF.KP. The two metrics beating PF.NOCB were not found to differ significantly from each other.

This study employs PF.BFP and PF.KP, i.e. two of the best performing metrics of the experiments in [Karamanis *et al.*, 2004], as well as M.NOCB and PF.NOCB, the two previously used baselines. An additional random baseline is also defined following [Lapata, 2003].

4 Data collection

16 sets of facts were randomly selected from the corpus of [Dimitromanolaki and Androutsopoulos, 2003].[4] The sentences that each fact corresponds to and the order defined by E0 was made available to us as well. We will subsequently refer to an unordered set of facts (or sentences that the facts correspond to) as a *Testitem*.

4.1 Generating the BestOrders for each metric

Following [Karamanis *et al.*, 2004], we envisage a TS approach in which a metric of coherence M assigns a score to

[2] A more detailed discussion of existing corpus-based methods for evaluating TS appears in [Karamanis and Mellish, 2005].

[3] Since discussing the metrics in detail is well beyond the scope of this paper, the reader is referred to Chapter 3 of [Karamanis, 2003] for more information on this issue.

[4] These are distinct from, yet very similar to, the sets of facts used in [Karamanis *et al.*, 2004].

each possible ordering of the input set of facts and selects the best scoring ordering as the output. When many orderings score best, M chooses randomly between them. Crucially, our hypothetical TS component only considers orderings starting with the `subclass` fact (e.g. `subclass(ex1, amph)` in Figure 1) following the suggestion of [Dimitromanolaki and Androutsopoulos, 2003]. This gives rise to $5! = 120$ orderings to be scored by M for each Testitem.

For the purposes of this experiment, a simple algorithm was implemented that first produces the 120 possible orderings of facts in a Testitem and subsequently ranks them according to the scores given by M. The algorithm outputs the set of BestOrders for the Testitem, i.e. the orderings which score best according to M. This procedure was repeated for each metric and all Testitems employed in the experiment.

4.2 Random baseline

Following [Lapata, 2003], a random baseline (RB) was implemented as the lower bound of the analysis. The random baseline consists of 10 randomly selected orderings for each Testitem. The orderings are selected irrespective of their scores for the various metrics.

4.3 Consulting domain experts

Three archaeologists (E1, E2, E3), one male and two females, between 28 and 45 years of age, all trained in cataloguing and museum labelling, were recruited from the Department of Classics at the University of Edinburgh.

Each expert was consulted by the first author in a separate interview. First, she was presented with a set of six sentences, each of which corresponded to a database fact and was printed on a different filecard, as well as with written instructions describing the ordering task.[5] The instructions mention that the sentences come from a computer program that generates descriptions of artefacts in a virtual museum. The first sentence for each set was given by the experimenter.[6] Then, the expert was asked to order the remaining five sentences in a coherent text.

When ordering the sentences, the expert was instructed to consider which ones should be together and which should come before another in the text without using hints other than the sentences themselves. She could revise her ordering at any time by moving the sentences around. When she was satisfied with the ordering she produced, she was asked to write next to each sentence its position, and give them to the experimenter in order to perform the same task with the next randomly selected set of sentences. The expert was encouraged to comment on the difficulty of the task, the strategies she followed, etc.

5 Dependent variable

Given an unordered set of sentences and two possible orderings, a number of measures can be employed to calculate the distance between them. Based on the argumentation in [Howell, 2002], [Lapata, 2003] selects Kendall's τ as the most appropriate measure and this was what we used for our analysis as well. Kendall's τ is based on the number of *inversions* between the two orderings and is calculated as follows:

$$\tau = 1 - \frac{2I}{P_N} = 1 - \frac{2I}{N(N-1)/2} \quad (1)$$

P_N stands for the number of pairs of sentences and N is the number of sentences to be ordered.[7] I stands for the number of inversions, that is, the number of adjacent transpositions necessary to bring one ordering to another. Kendall's τ ranges from -1 (inverse ranks) to 1 (identical ranks). The **higher** the τ value, the **smaller** the distance between the two orderings.

Following [Lapata, 2003], the Tukey test is employed to investigate significant differences between *average* τ *scores*.[8] First, the average distance between (the orderings of)[9] two experts e.g. E0 and E1, denoted as $T(E0_{E1})$, is calculated as the mean τ value between the ordering of E0 and the ordering of E1 taken across all 16 Testitems. Then, we compute $T(EXP_{EXP})$ which expresses the overall average distance between all expert pairs and serves as the upper bound for the evaluation of the metrics. Since a total of E experts gives rise to $P_E = \frac{E(E-1)}{2}$ expert pairs, $T(EXP_{EXP})$, is computed by summing up the average distances between all expert pairs and dividing the sum by P_E.

While [Lapata, 2003] always appears to single out a unique best scoring ordering, we often have to deal with many best scoring orderings. To account for this, we first compute the average distance between e.g. the ordering of an expert E0 and the BestOrders of a metric M for a given Testitem. In this way, M is rewarded for a BestOrder that is close to the expert's ordering, but penalised for every BestOrder that is not. Then, the average $T(E0_M)$ between the expert E0 and the metric M is calculated as their mean distance across all 16 Testitems. Finally, yet most importantly, $T(EXP_M)$ is the average distance between all experts and M. It is calculated by summing up the average distances between each expert and M and dividing the sum by the number of experts. As the next section explains in more detail, $T(EXP_M)$ is compared with the upper bound of the evaluation $T(EXP_{EXP})$ to estimate the performance of M in our experiments.

RB is evaluated in a similar way as M using the 10 randomly selected orderings instead of the BestOrders for each Testitem. $T(EXP_{RB})$ is the average distance between all experts and RB and is used as the lower bound of the evaluation.

[5] The instructions are given in Appendix D of [Karamanis, 2003] and are adapted from the ones used in [Barzilay *et al.*, 2002].

[6] This is the sentence corresponding to the `subclass` fact.

[7] In our data, N is always equal to 6.

[8] Provided that an omnibus ANOVA is significant, the Tukey test can be used to specify which of the conditions $c_1, ..., c_n$ measured by the dependent variable differ significantly. It uses the set of means $m_1, ..., m_n$ (corresponding to conditions $c_1, ..., c_n$) and the mean square error of the scores that contribute to these means to calculate a critical difference between any two means. An observed difference between any two means is significant if it exceeds the critical difference.

[9] Throughout the paper we often refer to e.g. "the distance between the orderings of the experts" with the phrase "the distance between the experts" for the sake of brevity.

$E0_{E1}$: 0.692			**	**	**
	$E0_{E2}$: 0.717		**	**	**
		$E1_{E2}$: 0.758	**	**	**
			$E0_{E3}$: 0.258		
				$E1_{E3}$: 0.300	
					$E2_{E3}$: 0.192

CD at 0.01: 0.338
CD at 0.05: 0.282
F(5,75)=14.931, p<0.000

Table 1: Comparison of distances between the expert pairs

$E0_{E1}$: 0.692			**	**	**
	$E0_{E2}$: 0.717		**	**	**
		$E1_{E2}$: 0.758	**	**	**
			$E0_{RB}$: 0.323		
				$E1_{RB}$: 0.347	
					$E2_{RB}$: 0.352

CD at 0.01: 0.242
CD at 0.05: 0.202
F(5,75)=18.762, p<0.000

$E0_{E3}$: 0.258			
	$E1_{E3}$: 0.300		
		$E2_{E3}$: 0.192	
			$E3_{RB}$: 0.302

CD at 0.01: 0.219
CD at 0.05: 0.177
F(3,45)=1.223, p=0.312

Table 2: Comparison of distances between the experts (E0, E1, E2, E3) and the random baseline (RB)

6 Predictions

Despite any potential differences between the experts, one expects them to share some common ground in the way they order sentences. In this sense, a particularly welcome result for our purposes is to show that the average distances between E0 and most of her colleagues are short and not significantly different from the distances between the other expert pairs, which in turn indicates that she is not a "stand-alone" expert.

Moreover, we expect the average distance between the expert pairs to be significantly smaller than the average distance between the experts and RB. This is again based on the assumption that even though the experts might not follow completely identical strategies, they do not operate with absolute diversity either. Hence, we predict that $T(EXP_{EXP})$ will be significantly greater than $T(EXP_{RB})$.

Due to the small number of Testitems employed in this study, it is likely that the metrics do not differ significantly from each other with respect to their average distance from the experts. Rather than comparing the metrics *directly* with each other (as [Karamanis *et al.*, 2004] do), this study compares them *indirectly* by examining their behaviour with respect to the upper and the lower bound. For instance, although $T(EXP_{PF.KP})$ and $T(EXP_{PF.BFP})$ might not be significantly different from each other, one score could be significantly different from $T(EXP_{EXP})$ (upper bound) and/or $T(EXP_{RB})$ (lower bound) while the other is not.

We identify the best metrics in this study as the ones whose average distance from the experts (i) is significantly greater from the lower bound **and** (ii) does not differ significantly from the upper bound.[10]

7 Results

7.1 Distances between the expert pairs

On the first step in our analysis, we computed the T score for each expert pair, namely $T(E0_{E1})$, $T(E0_{E2})$, $T(E0_{E3})$, $T(E1_{E2})$, $T(E1_{E3})$ and $T(E2_{E3})$. Then we performed all 15 pairwise comparisons between them using the Tukey test, the results of which are summarised in Table 1.[11]

The cells in the Table report the level of significance returned by the Tukey test when the difference between two distances exceeds the critical difference (CD). Significance beyond the 0.05 threshold is reported with one asterisk (*), while significance beyond the 0.01 threshold is reported with two asterisks (**). A cell remains empty when the difference between two distances does not exceed the critical difference. For example, the value of $T(E0_{E1})$ is 0.692 and the value of $T(E0_{E3})$ is 0.258. Since their difference exceeds the CD at the 0.01 threshold, it is reported to be significant beyond that level by the Tukey test, as shown in the top cell of the third column in Table 1.

As the Table shows, the T scores for the distance between E0 and E1 or E2, i.e. $T(E0_{E1})$ and $T(E0_{E2})$, as well as the T for the distance between E1 and E2, i.e. $T(E1_{E2})$, are quite high which indicates that on average the orderings of the three experts are quite close to each other. Moreover, these T scores are not significantly different from each other which suggests that E0, E1 and E2 share quite a lot of common ground in the ordering task. Hence, E0 is found to give rise to similar orderings to the ones of E1 and E2.

However, when any of the previous distances is compared with a distance that involves the orderings of E3 the difference is significant, as shown by the cells containing two asterisks in Table 1. In other words, although the orderings of E1 and E2 seem to deviate from each other and the orderings of E0 to more or less the same extent, the orderings of E3 stand much further away from all of them. Hence, there exists a "stand-alone" expert among the ones consulted in our studies, yet this is not E0 but E3.

This finding can be easily explained by the fact that by contrast to the other three experts, E3 followed a very schematic way for ordering sentences. Because the orderings of E3 manifest rather peculiar strategies, at least compared to the orderings of E0, E1 and E2, the upper bound of the analysis, i.e. the average distance between the expert pairs $T(EXP_{EXP})$, is computed without taking into account these orderings:

(2) $T(EXP_{EXP}) = 0.722 = \frac{T(E0_{E1})+T(E0_{E2})+T(E1_{E2})}{3}$

7.2 Distances between the experts and RB

As the upper part of Table 2 shows, the T score between any two experts other than E3 is significantly greater than their distance from RB beyond the 0.01 threshold. Only the dis-

[10] Criterion (ii) can only be applied provided that the average distance between the experts and *at least one* metric M_x is found to be significantly lower than $T(EXP_{EXP})$. Then, if the average distance between the experts and another metric M_y does not differ significantly from $T(EXP_{EXP})$, M_y performs better than M_x.

[11] The Table also reports the result of the omnibus ANOVA, which is significant: F(5,75)=14.931, p<0.000.

tances between E3 and another expert, shown in the lower section of Table 2, are not significantly different from the distance between E3 and RB.

Although this result does not mean that the orders of E3 are similar to the orders of RB,[12] it shows that E3 is roughly as far away from e.g. E0 as she is from RB. By contrast, E0 stands significantly closer to E1 than to RB, and the same holds for the other distances in the upper part of the Table. In accordance with the discussion in the previous section, the lower bound, i.e. the overall average distance between the experts (excluding E3) and RB $T(EXP_{RB})$, is computed as shown in (3):

(3) $T(EXP_{RB}) = 0.341 = \frac{T(E0_{RB})+T(E1_{RB})+T(E2_{RB})}{3}$

7.3 Distances between the experts and each metric

So far, E3 was identified as an "stand-alone" expert standing further away from the other three experts than they stand from each other. We also identified the distance between E3 and each expert as similar to her distance from RB.

Similarly, E3 was found to stand further away from the metrics compared to their distance from the other three experts.[13] This result, gives rise to the set of formulas in (4) for calculating the overall average distance between the experts (excluding E3) and each metric.

(4) (4.1): $T(EXP_{PF.BFP}) = 0.629 = \frac{T(E0_{PF.BFP})+T(E1_{PF.BFP})+T(E2_{PF.BFP})}{3}$

(4.2): $T(EXP_{PF.KP}) = 0.571 = \frac{T(E0_{PF.KP})+T(E1_{PF.KP})+T(E2_{PF.KP})}{3}$

(4.3): $T(EXP_{PF.NOCB}) = 0.606 = \frac{T(E0_{PF.NOCB})+T(E1_{PF.NOCB})+T(E2_{PF.NOCB})}{3}$

(4.4): $T(EXP_{M.NOCB}) = 0.487 = \frac{T(E0_{M.NOCB})+T(E1_{M.NOCB})+T(E2_{M.NOCB})}{3}$

In the next section, we present the concluding analysis for this study which compares the overall distances in formulas (2), (3) and (4) with each other. As we have already mentioned, $T(EXP_{EXP})$ serves as the upper bound of the analysis whereas $T(EXP_{RB})$ is the lower bound. The aim is to specify which scores in (4) are significantly greater than $T(EXP_{RB})$, but not significantly lower than $T(EXP_{EXP})$.

7.4 Concluding analysis

The results of the comparisons of the scores in (2), (3) and (4) are shown in Table 3. As the top cell in the last column of the Table shows, the T score between the experts and RB, $T(EXP_{RB})$, is significantly lower than the average distance between the expert pairs, $T(EXP_{EXP})$ at the 0.01 level. This result verifies one of our main predictions showing that the orderings of the experts (modulo E3) stand much closer to each other compared to their distance from randomly assembled orderings.

As expected, most of the scores that involve the metrics are not significantly different from each other, except for $T(EXP_{PF.BFP})$ which is significantly greater than $T(EXP_{M.NOCB})$ at the 0.05 level. Yet, what we are mainly interested in is how the distance between the experts and each metric compares with $T(EXP_{EXP})$ and $T(EXP_{RB})$. This is shown in the first row and the last column of Table 3.

Crucially, $T(EXP_{RB})$ is significantly lower than $T(EXP_{PF.BFP})$ as well as $T(EXP_{PF.NOCB})$ and $T(EXP_{PF.KP})$ at the 0.01 level. Notably, even the distance of the experts from M.NOCB, $T(EXP_{M.NOCB})$, is significantly greater than $T(EXP_{RB})$, albeit at the 0.05 level. These results show that the distance from the experts is significantly reduced when using the best scoring orderings of any metric, even M.NOCB, instead of the orderings of RB. Hence, all metrics score significantly better than RB in this experiment.

However, simply using M.NOCB to output the best scoring orders is not enough to yield a distance from the experts which is comparable to $T(EXP_{EXP})$. Although the PF constraint appears to help towards this direction, $T(EXP_{PF.KP})$ remains significantly lower than $T(EXP_{EXP})$, whereas $T(EXP_{PF.NOCB})$ falls only 0.009 points short of CD at the 0.05 threshold. Hence, PF.BFP is the most robust metric, as the difference between $T(EXP_{PF.BFP})$ and $T(EXP_{EXP})$ is clearly not significant.

Finally, the difference between $T(EXP_{PF.NOCB})$ and $T(EXP_{M.NOCB})$ is only 0.006 points away from the CD. This result shows that the distance from the experts is reduced to a great extent when the best scoring orderings are computed according to PF.NOCB instead of simply M.NOCB. Hence, this experiment provides additional evidence in favour of enhancing M.NOCB with the PF constraint of coherence, as suggested in [Karamanis, 2003].

8 Discussion

A question not addressed by previous studies making use of a certain collection of orderings of facts is whether the strategies reflected there are specific to E0, the expert who created the dataset. In this paper, we address this question by enhancing E0's dataset with orderings provided by three additional experts. Then, the distance between E0 and her colleagues is computed and compared to the distance between the other expert pairs. The results indicate that E0 shares a lot of common ground with two of her colleagues in the ordering task deviating from them as much as they deviate from each other, while the orderings of a fourth "stand-alone" expert are found to manifest rather individualistic strategies.

The same variable used to investigate the distance between the experts is employed to automatically evaluate the best scoring orderings of some of the best performing metrics in [Karamanis *et al.*, 2004]. Despite its limitations due to the necessarily restricted size of the employed dataset, this eval-

[12] This could have been argued, if the value of $T(E3_{RB})$ had been much closer to 1.

[13] Due to space restrictions, we cannot report the scores for these comparisons here. The reader is referred to Table 9.4 on page 175 of Chapter 9 in [Karamanis, 2003].

EXP_{EXP}: 0.722			**	**	**
	$EXP_{PF.BFP}$: 0.629			*	**
		$EXP_{PF.NOCB}$: 0.606			**
			$EXP_{PF.KP}$: 0.571		**
				$EXP_{M.NOCB}$: 0.487	*
					EXP_{RB}: 0.341

CD at 0.01: 0.150
CD at 0.05: 0.125
F(5,75)=19.111, p<0.000

Table 3: Results of the concluding analysis comparing the distance between the expert pairs (EXP_{EXP}) with the distance between the experts and each metric (PF.BFP, PF.NOCB, PF.KP, M.NOCB) and the random baseline (RB)

uation task allows us to explore the previously unaddressed possibility that there exist many good solutions for TS in the employed domain.

Out of a much larger set of possibilities, 10 metrics were evaluated in [Karamanis *et al.*, 2004], only a handful of which were found to overtake two simple baselines. The additional test in this study carries on the elimination process by pointing out PF.BFP as the single most promising metric to be used for TS in the explored domain, since this is the metric that manages to clearly survive both tests.

Equally crucially, our analysis shows that all employed metrics are superior to a random baseline. Additional evidence in favour of the PF constraint on coherence introduced in [Karamanis, 2003] is provided as well. The general evaluation methodology as well as the specific results of this study will be useful for any subsequent attempt to automatically evaluate a TS approach using a corpus of sentence orderings defined by many experts.

As [Reiter and Sripada, 2002] suggest, the best way to treat the results of a corpus-based study is as hypotheses which eventually need to be integrated with other types of evaluation. Although we followed the ongoing argumentation that using perceptual experiments to choose between many possible metrics is unfeasible, our efforts have resulted into a single preferred candidate which is much easier to evaluate with the help of psycholinguistic techniques (instead of having to deal with a large number of metrics from very early on). This is indeed our main direction for future work in this domain.

Acknowledgments

We are grateful to Aggeliki Dimitromanolaki for entrusting us with her data and for helpful clarifications on their use; to Mirella Lapata for providing us with the scripts for the computation of τ together with her extensive and prompt advice; to Katerina Kolotourou for her invaluable assistance in recruiting the experts; and to the experts for their participation. This work took place while the first author was studying at the University of Edinburgh, supported by the Greek State Scholarship Foundation (IKY).

References

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

[Barzilay *et al.*, 2002] Regina Barzilay, Noemie Elhadad, and Kathleen McKeown. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:35–55, 2002.

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

[Dimitromanolaki and Androutsopoulos, 2003] Aggeliki Dimitromanolaki and Ion Androutsopoulos. Learning to order facts for discourse planning in natural language generation. In *Proceedings of the 9th European Workshop on Natural Language Generation*, Budapest, Hungary, 2003.

[Howell, 2002] David C. Howell. *Statistical Methods for Psychology*. Duxbury, Pacific Grove, CA, 5th edition, 2002.

[Isard *et al.*, 2003] Amy Isard, Jon Oberlander, Ion Androutsopoulos, and Colin Matheson. Speaking the users' languages. *IEEE Intelligent Systems Magazine*, 18(1):40–45, 2003.

[Karamanis and Manurung, 2002] Nikiforos Karamanis and Hisar Maruli Manurung. Stochastic text structuring using the principle of continuity. In *Proceedings of INLG 2002*, pages 81–88, Harriman, NY, USA, July 2002.

[Karamanis and Mellish, 2005] Nikiforos Karamanis and Chris Mellish. A review of recent corpus-based methods for evaluating text structuring in NLG. 2005. Submitted to *Using Corpora for NLG* workshop.

[Karamanis *et al.*, 2004] Nikiforos Karamanis, Chris Mellish, Jon Oberlander, and Massimo Poesio. A corpus-based methodology for evaluating metrics of coherence for text structuring. In *Proceedings of INLG04*, pages 90–99, Brockenhurst, UK, 2004.

[Karamanis, 2003] Nikiforos Karamanis. *Entity Coherence for Descriptive Text Structuring*. PhD thesis, Division of Informatics, University of Edinburgh, 2003.

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

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

[McKeown, 1985] Kathleen McKeown. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Studies in Natural Language Processing. Cambridge University Press, 1985.

[Reiter and Sripada, 2002] Ehud Reiter and Somayajulu Sripada. Should corpora texts be gold standards for NLG? In *Proceedings of INLG 2002*, pages 97–104, Harriman, NY, USA, July 2002.