Extending the Entity-based Coherence Model with Multiple Ranks

Vanessa Wei Feng
Department of Computer Science
University of Toronto
Toronto, ON, M5S 3G4, Canada
weifeng@cs.toronto.edu

Graeme Hirst
Department of Computer Science
University of Toronto
Toronto, ON, M5S 3G4, Canada
gh@cs.toronto.edu

Abstract

We extend the original entity-based coherence model (Barzilay and Lapata, 2008) by learning from more fine-grained coherence preferences in training data. We associate multiple ranks with the set of permutations originating from the same source document, as opposed to the original pairwise rankings. We also study the effect of the permutations used in training, and the effect of the coreference component used in entity extraction. With no additional manual annotations required, our extended model is able to outperform the original model on two tasks: *sentence ordering* and *summary coherence rating*.

1 Introduction

Coherence is important in a well-written document; it helps make the text semantically meaningful and interpretable. Automatic evaluation of coherence is an essential component of various natural language applications. Therefore, the study of coherence models has recently become an active research area. A particularly popular coherence model is the entity-based local coherence model of Barzilay and Lapata (B&L) (2005; 2008). This model represents local coherence by transitions, from one sentence to the next, in the grammatical role of references to entities. It learns a pairwise ranking preference between alternative renderings of a document based on the probability distribution of those transitions. In particular, B&L associated a lower rank with automatically created permutations of a source document, and learned a model to discriminate an original text from its permutations (see Section 3.1 below). However, coherence is matter of degree rather than a binary distinction, so a model based only on such pairwise rankings is insufficiently fine-grained and cannot capture the subtle differences in coherence between the permuted documents.

Since the first appearance of B&L's model, several extensions have been proposed (see Section 2.3 below), primarily focusing on modifying or enriching the original feature set by incorporating other document information. By contrast, we wish to refine the learning procedure in a way such that the resulting model will be able to evaluate coherence on a more fine-grained level. Specifically, we propose a concise extension to the standard entity-based coherence model by learning not only from the original document and its corresponding permutations but also from ranking preferences among the permutations themselves.

We show that this can be done by assigning a suitable objective score for each permutation indicating its dissimilarity from the original one. We call this a *multiple-rank model* since we train our model on a multiple-rank basis, rather than taking the original pairwise ranking approach. This extension can also be easily combined with other extensions by incorporating their enriched feature sets. We show that our multiple-rank model outperforms B&L's basic model on two tasks, *sentence ordering* and *summary coherence rating*, evaluated on the same datasets as in Barzilay and Lapata (2008).

In *sentence ordering*, we experiment with different approaches to assigning dissimilarity scores and ranks (Section 5.1.1). We also experiment with different entity extraction approaches

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 315–324,
Avignon, France, April 23 - 27 2012.

	Manila	Miles	Island	Quake	Baco
1	–	–	X	X	–
2	S	–	O	–	–
3	X	X	X	X	X

Table 1: A fragment of an entity grid for five entities across three sentences.

(Section 5.1.2) and different distributions of permutations used in training (Section 5.1.3). We show that these two aspects are crucial, depending on the characteristics of the dataset.

2 Entity-based Coherence Model

2.1 Document Representation

The original entity-based coherence model is based on the assumption that a document makes repeated reference to elements of a set of entities that are central to its topic. For a document d, an entity grid is constructed, in which the columns represent the entities referred to in d, and rows represent the sentences. Each cell corresponds to the grammatical role of an entity in the corresponding sentence: subject (S), object (O), neither (X), or nothing (−). An example fragment of an entity grid is shown in Table 1; it shows the representation of three sentences from a text on a Philippine earthquake. B&L define a local transition as a sequence $\{S, O, X, -\}^n$, representing the occurrence and grammatical roles of an entity in n adjacent sentences. Such transition sequences can be extracted from the entity grid as continuous subsequences in each column. For example, the entity "Manila" in Table 1 has a bigram transition $\{S, X\}$ from sentence 2 to 3. The entity grid is then encoded as a feature vector $\mathbf{\Phi}(d) = (p_1(d), p_2(d), \ldots, p_m(d))$, where $p_t(d)$ is the probability of the transition t in the entity grid, and m is the number of transitions with length no more than a predefined optimal transition length k. $p_t(d)$ is computed as the number of occurrences of t in the entity grid of document d, divided by the total number of transitions of the same length in the entity grid.

For entity extraction, Barzilay and Lapata (2008) had two conditions: **Coreference**+ and **Coreference**−. In **Coreference**+, entity coreference relations in the document were resolved by an automatic coreference resolution tool (Ng and Cardie, 2002), whereas in **Coreference**−, nouns are simply clustered by string matching.

2.2 Evaluation Tasks

Two evaluation tasks for Barzilay and Lapata (2008)'s entity-based model are *sentence ordering* and *summary coherence rating*.

In sentence ordering, a set of random permutations is created for each source document, and the learning procedure is conducted on this synthetic mixture of coherent and incoherent documents. Barzilay and Lapata (2008) experimented on two datasets: news articles on the topic of earthquakes (*Earthquakes*) and narratives on the topic of aviation accidents (*Accidents*). A training data instance is constructed as a pair consisting of a source document and one of its random permutations, and the permuted document is always considered to be less coherent than the source document. The entity transition features are then used to train a support vector machine ranker (Joachims, 2002) to rank the source documents higher than the permutations. The model is tested on a different set of source documents and their permutations, and the performance is evaluated as the fraction of correct pairwise rankings in the test set.

In summary coherence rating, a similar experimental framework is adopted. However, in this task, rather than training and evaluating on a set of synthetic data, system-generated summaries and human-composed reference summaries from the Document Understanding Conference (DUC 2003) were used. Human annotators were asked to give a coherence score on a seven-point scale for each item. The pairwise ranking preferences between summaries generated from the same input document cluster (excluding the pairs consisting of two human-written summaries) are used by a support vector machine ranker to learn a discriminant function to rank each pair according to their coherence scores.

2.3 Extended Models

Filippova and Strube (2007) applied Barzilay and Lapata's model on a German corpus of newspaper articles with manual syntactic, morphological, and NP coreference annotations provided. They further clustered entities by semantic relatedness as computed by the WikiRelated! API (Strube and Ponzetto, 2006). Though the improvement was not significant, interestingly, a short subsection in

their paper described their approach to extending pairwise rankings to longer rankings, by supplying the learner with rankings of all renderings as computed by Kendall's τ, which is one of our extensions considered in this paper. Although Filippova and Strube simply discarded this idea because it hurt accuracies when tested on their data, we found it a promising direction for further exploration. Cheung and Penn (2010) adapted the standard entity-based coherence model to the same German corpus, but replaced the original linguistic dimension used by Barzilay and Lapata (2008) — grammatical role — with topological field information, and showed that for German text, such a modification improves accuracy.

For English text, two extensions have been proposed recently. Elsner and Charniak (2011) augmented the original features used in the standard entity-based coherence model with a large number of entity-specific features, and their extension significantly outperformed the standard model on two tasks: *document discrimination* (another name for *sentence ordering*), and *sentence insertion*. Lin et al. (2011) adapted the entity grid representation in the standard model into a discourse role matrix, where additional discourse information about the document was encoded. Their extended model significantly improved ranking accuracies on the same two datasets used by Barzilay and Lapata (2008) as well as on the *Wall Street Journal* corpus.

However, while enriching or modifying the original features used in the standard model is certainly a direction for refinement of the model, it usually requires more training data or a more sophisticated feature representation. In this paper, we instead modify the learning approach and propose a concise and highly adaptive extension that can be easily combined with other extended features or applied to different languages.

3 Experimental Design

Following Barzilay and Lapata (2008), we wish to train a discriminative model to give the correct ranking preference between two documents in terms of their degree of coherence. We experiment on the same two tasks as in their work: *sentence ordering* and *summary coherence rating*.

3.1 Sentence Ordering

In the standard entity-based model, a discriminative system is trained on the pairwise rankings between source documents and their permutations (see Section 2.2). However, a model learned from these pairwise rankings is not sufficiently fine-grained, since the subtle differences between the permutations are not learned. Our major contribution is to further differentiate among the permutations generated from the same source documents, rather than simply treating them all as being of the same degree of coherence.

Our fundamental assumption is that there exists a canonical ordering for the sentences of a document; therefore we can approximate the degree of coherence of a document by the similarity between its actual sentence ordering and that canonical sentence ordering. Practically, we automatically assign an objective score for each permutation to estimate its dissimilarity from the source document (see Section 4). By learning from all the pairs across a source document and its permutations, the effective size of the training data is increased while no further manual annotation is required, which is favorable in real applications when available samples with manually annotated coherence scores are usually limited. For r source documents each with m random permutations, the number of training instances in the standard entity-based model is therefore $r \times m$, while in our multiple-rank model learning process, it is $r \times \binom{m+1}{2} \approx \frac{1}{2} r \times m^2 > r \times m$, when $m > 2$.

3.2 Summary Coherence Rating

Compared to the standard entity-based coherence model, our major contribution in this task is to show that by automatically assigning an objective score for each machine-generated summary to estimate its dissimilarity from the human-generated summary from the same input document cluster, we are able to achieve performance competitive with, or even superior to, that of B&L's model without knowing the true coherence score given by human judges.

Evaluating our multiple-rank model in this task is crucial, since in summary coherence rating, the coherence violations that the reader might encounter in real machine-generated texts can be more precisely approximated, while the sentence ordering task is only partially capable of doing so.

4 Dissimilarity Metrics

As mentioned previously, the subtle differences among the permutations of the same source document can be used to refine the model learning process. Considering an original document $\boldsymbol{d}$ and one of its permutations, we call $\boldsymbol{\sigma} = (1, 2, \ldots, N)$ the *reference ordering*, which is the sentence ordering in $\boldsymbol{d}$, and $\boldsymbol{\pi} = (o_1, o_2, \ldots, o_N)$ the *test ordering*, which is the sentence ordering in that permutation, where N is the number of sentences being rendered in both documents.

In order to approximate different degrees of coherence among the set of permutations which bear the same content, we need a suitable metric to quantify the dissimilarity between the test ordering $\boldsymbol{\pi}$ and the reference ordering $\boldsymbol{\sigma}$. Such a metric needs to satisfy the following criteria: (1) It can be automatically computed while being highly correlated with human judgments of coherence, since additional manual annotation is certainly undesirable. (2) It depends on the particular sentence ordering in a permutation while remaining independent of the entities within the sentences; otherwise our multiple-rank model might be trained to fit particular probability distributions of entity transitions rather than true coherence preferences.

In our work we use three different metrics: *Kendall's τ distance*, *average continuity*, and *edit distance*.

Kendall's τ distance: This metric has been widely used in evaluation of sentence ordering (Lapata, 2003; Lapata, 2006; Bollegala et al., 2006; Madnani et al., 2007)[1]. It measures the disagreement between two orderings $\boldsymbol{\sigma}$ and $\boldsymbol{\pi}$ in terms of the number of inversions of adjacent sentences necessary to convert one ordering into another. Kendall's τ distance is defined as

$$\tau = \frac{2m}{N(N-1)},$$

where m is the number of sentence inversions necessary to convert $\boldsymbol{\sigma}$ to $\boldsymbol{\pi}$.

Average continuity (AC): Following Zhang (2011), we use average continuity as the second dissimilarity metric. It was first proposed by Bollegala et al. (2006). This metric estimates the quality of a particular sentence ordering by the number of correctly arranged continuous sentences, compared to the reference ordering. For example, if $\boldsymbol{\pi} = (\ldots, 3, 4, 5, 7, \ldots, o_N)$, then $\{3, 4, 5\}$ is considered as continuous while $\{3, 4, 5, 7\}$ is not. Average continuity is calculated as

$$AC = \exp\left(\frac{1}{n-1}\sum_{i=2}^{n}\log\left(P_i + \alpha\right)\right),$$

where $n = \min(4, N)$ is the maximum number of continuous sentences to be considered, and $\alpha = 0.01$. P_i is the proportion of continuous sentences of length i in $\boldsymbol{\pi}$ that are also continuous in the reference ordering $\boldsymbol{\sigma}$. To represent the *dis*similarity between the two orderings $\boldsymbol{\pi}$ and $\boldsymbol{\sigma}$, we use its complement $AC' = 1 - AC$, such that the larger AC' is, the more dissimilar two orderings are[2].

Edit distance (ED): Edit distance is a commonly used metric in information theory to measure the difference between two sequences. Given a test ordering $\boldsymbol{\pi}$, its edit distance is defined as the minimum number of edits (i.e., insertions, deletions, and substitutions) needed to transform it into the reference ordering $\boldsymbol{\sigma}$. For permutations, the edits are essentially movements, which can be considered as equal numbers of insertions and deletions.

5 Experiments

5.1 Sentence Ordering

Our first set of experiments is on sentence ordering. Following Barzilay and Lapata (2008), we use all transitions of length ≤ 3 for feature extraction. In addition, we explore three specific aspects in our experiments: rank assignment, entity extraction, and permutation generation.

5.1.1 Rank Assignment

In our multiple-rank model, pairwise rankings between a source document and its permutations are extended into a longer ranking with multiple ranks. We assign a rank to a particular permutation, based on the result of applying a chosen dissimilarity metric from Section 4 (τ, AC, or ED) to the sentence ordering in that permutation.

We experiment with two different approaches to assigning ranks to permutations, while each

[1] Filippova and Strube (2007) found that their performance dropped when using this metric for longer rankings; but they were using data in a different language and with manual annotations, so its effect on our datasets is worth trying nonetheless.

[2] We will refer to AC' as AC from now on.

source document is always assigned a zero (the highest) rank.

In the **raw** option, we rank the permutations directly by their dissimilarity scores to form a full ranking for the set of permutations generated from the same source document.

Since a full ranking might be too sensitive to noise in training, we also experiment with the **stratified** option, in which C ranks are assigned to the permutations generated from the same source document. The permutation with the smallest dissimilarity score is assigned the same (zero, the highest) rank as the source document, and the one with the largest score is assigned the lowest $(C-1)$ rank; then ranks of other permutations are uniformly distributed in this range according to their raw dissimilarity scores. We experiment with 3 to 6 ranks (the case where $C = 2$ reduces to the standard entity-based model).

5.1.2 Entity Extraction

Barzilay and Lapata (2008)'s best results were achieved by employing an automatic coreference resolution tool (Ng and Cardie, 2002) for extracting entities from a source document, and the permutations were generated only afterwards — entity extraction from a permuted document depends on knowing the correct sentence order and the oracular entity information from the source document — since resolving coreference relations in permuted documents is too unreliable for an automatic tool.

We implement our multiple-rank model with full coreference resolution using Ng and Cardie's coreference resolution system, and entity extraction approach as described above — the **Coreference**+ condition. However, as argued by Elsner and Charniak (2011), to better simulate the real situations that human readers might encounter in machine-generated documents, such oracular information should not be taken into account. Therefore we also employ two alternative approaches for entity extraction: (1) use the same automatic coreference resolution tool on permuted documents — we call it the **Coreference**± condition; (2) use no coreference resolution, i.e., group head noun clusters by simple string matching — B&L's **Coreference**− condition.

5.1.3 Permutation Generation

The quality of the model learned depends on the set of permutations used in training. We are not aware of how B&L's permutations were generated, but we assume they are generated in a perfectly random fashion.

However, in reality, the probabilities of seeing documents with different degrees of coherence are not equal. For example, in an essay scoring task, if the target group is (near-) native speakers with sufficient education, we should expect their essays to be less incoherent — most of the essays will be coherent in most parts, with only a few minor problems regarding discourse coherence. In such a setting, the performance of a model trained from permutations generated from a uniform distribution may suffer some accuracy loss.

Therefore, in addition to the set of permutations used by Barzilay and Lapata (2008) (PS_{BL}), we create another set of permutations for each source document (PS_M) by assigning most of the probability mass to permutations which are mostly similar to the original source document. Besides its capability of better approximating real-life situations, training our model on permutations generated in this way has another benefit: in the standard entity-based model, all permuted documents are treated as incoherent; thus there are many more incoherent training instances than coherent ones (typically the proportion is 20:1). In contrast, in our multiple-rank model, permuted documents are assigned different ranks to further differentiate the different degrees of coherence within them. By doing so, our model will be able to learn the characteristics of a coherent document from those near-coherent documents as well, and therefore the problem of lacking coherent instances can be mitigated.

Our permutation generation algorithm is shown in Algorithm 1, where $\alpha = 0.05$, $\beta = 5.0$, $MAX\_NUM = 50$, and K and K' are two normalization factors to make $p(swap\_num)$ and $p(i, j)$ proper probability distributions. For each source document, we create the same number of permutations as PS_{BL}.

5.2 Summary Coherence Rating

In the summary coherence rating task, we are dealing with a mixture of multi-document summaries generated by systems and written by humans. Barzilay and Lapata (2008) did not assume

Algorithm 1 Permutation Generation.

Input: $S_1, S_2, \ldots, S_N$; $\boldsymbol{\sigma} = (1, 2, \ldots, N)$
  Choose a number of sentence swaps
    swap_num with probability $e^{-\alpha \times swap\_num} / K$
  for $i = 1 \rightarrow swap\_num$ **do**
    Swap a pair of sentence (S_i, S_j)
      with probability $p(i, j) = e^{-\beta \times |i-j|} / K'$
  end for
Output: $\boldsymbol{\pi} = (o_1, o_2, \ldots, o_N)$

a simple binary distinction among the summaries generated from the same input document cluster; rather, they had human judges give scores for each summary based on its degree of coherence (see Section 3.2). Therefore, it seems that the subtle differences among incoherent documents (system-generated summaries in this case) have already been learned by their model.

But we wish to see if we can replace human judgments by our computed dissimilarity scores so that the original supervised learning is converted into unsupervised learning and yet retain competitive performance. However, given a summary, computing its dissimilarity score is a bit involved, due to the fact that we do not know its correct sentence order. To tackle this problem, we employ a simple sentence alignment between a system-generated summary and a human-written summary originating from the same input document cluster. Given a system-generated summary $D_s = (S_{s1}, S_{s2}, \ldots, S_{sn})$ and its corresponding human-written summary $D_h = (S_{h1}, S_{h2}, \ldots, S_{hN})$ (here it is possible that $n \neq N$), we treat the sentence ordering $(1, 2, \ldots, N)$ in D_h as $\boldsymbol{\sigma}$ (the original sentence ordering), and compute $\boldsymbol{\pi} = (o_1, o_2, \ldots, o_n)$ based on D_s. To compute each o_i in $\boldsymbol{\pi}$, we find the most similar sentence S_{hj}, $j \in [1, N]$ in D_h by computing their cosine similarity over all tokens in S_{hj} and S_{si}; if all sentences in D_h have zero cosine similarity with S_{si}, we assign -1 to o_i.

Once $\boldsymbol{\pi}$ is known, we can compute its "dissimilarity" from $\boldsymbol{\sigma}$ using a chosen metric. But because now $\boldsymbol{\pi}$ is not guaranteed to be a permutation of $\boldsymbol{\sigma}$ (there may be repetition or missing values, i.e., -1, in $\boldsymbol{\pi}$), Kendall's τ cannot be used, and we use only *average continuity* and *edit distance* as dissimilarity metrics in this experiment.

The remaining experimental configuration is the same as that of Barzilay and Lapata (2008), with the optimal transition length set to ≤ 2.

6 Results

6.1 Sentence Ordering

In this task, we use the same two sets of source documents (*Earthquakes* and *Accidents*, see Section 3.1) as Barzilay and Lapata (2008). Each contains 200 source documents, equally divided between training and test sets, with up to 20 permutations per document. We conduct experiments on these two domains separately. For each domain, we accompany each source document with two different sets of permutations: the one used by B&L (PS_{BL}), and the one generated from our model described in Section 5.1.3 (PS_M). We train our multiple-rank model and B&L's standard two-rank model on each set of permutations using the SVM^{rank} package (Joachims, 2006), and evaluate both systems on their test sets. Accuracy is measured as the fraction of correct pairwise rankings for the test set.

6.1.1 Full Coreference Resolution with Oracular Information

In this experiment, we implement B&L's fully-fledged standard entity-based coherence model, and extract entities from permuted documents using oracular information from the source documents (see Section 5.1.2).

Results are shown in Table 2. For each test situation, we list the best accuracy (in *Acc* columns) for each chosen dissimilarity metric, with the corresponding rank assignment approach. C represents the number of ranks used in stratifying raw scores ("N" if using **raw** configuration, see Section 5.1.1 for details). Baselines are accuracies trained using the standard entity-based coherence model[3].

Our model outperforms the standard entity-based model on both permutation sets for both datasets. The improvement is not significant when trained on the permutation set PS_{BL}, and is achieved only with one of the three metrics;

[3]There are discrepancies between our reported accuracies and those of Barzilay and Lapata (2008). The differences are due to the fact that we use a different parser: the Stanford dependency parser (de Marneffe et al., 2006), and might have extracted entities in a slightly different way than theirs, although we keep other experimental configurations as close as possible to theirs. But when comparing our model with theirs, we always use the exact same set of features, so the absolute accuracies do not matter.

Condition: Coreference+					
Perms	Metric	*Earthquakes*		*Accidents*	
		C	Acc	*C*	Acc
PS_{BL}	τ	3	79.5	3	82.0
	AC	4	85.2	3	**83.3**
	ED	3	**86.8**	6	82.2
	Baseline		85.3		83.2
PS_M	τ	3	**86.8**	3	**85.2***
	AC	3	**85.6**	1	**85.4***
	ED	*N*	**87.9***	4	**86.3***
	Baseline		85.3		81.7

Table 2: Accuracies (%) of extending the standard entity-based coherence model with multiple-rank learning in sentence ordering using **Coreference+** option. Accuracies which are significantly better than the baseline ($p < .05$) are indicated by *.

Condition: Coreference±					
Perms	Metric	*Earthquakes*		*Accidents*	
		C	Acc	*C*	Acc
PS_{BL}	τ	3	71.0	3	73.3
	AC	3	***76.8**	3	**74.5**
	ED	4	***77.4**	6	**74.4**
	Baseline		71.7		73.8
PS_M	τ	3	**55.9**	3	51.5
	AC	4	**53.9**	6	49.0
	ED	4	**53.9**	5	52.3
	Baseline		49.2		**53.2**

Table 3: Accuracies (%) of extending the standard entity-based coherence model with multiple-rank learning in sentence ordering using **Coreference±** option. Accuracies which are significantly better than the baseline ($p < .05$) are indicated by *.

but when trained on PS_M (the set of permutations generated from our biased model), our model's performance significantly exceeds B&L's[4] for all three metrics, especially as their model's performance drops for dataset *Accidents*.

From these results, we see that in the ideal situation where we extract entities and resolve their coreference relations based on the oracular information from the source document, our model is effective in terms of improving ranking accuracies, especially when trained on our more realistic permutation sets PS_M.

6.1.2 Full Coreference Resolution without Oracular Information

In this experiment, we apply the same automatic coreference resolution tool (Ng and Cardie, 2002) on not only the source documents but also their permutations. We want to see how removing the oracular component in the original model affects the performance of our multiple-rank model and the standard model. Results are shown in Table 3.

First we can see when trained on PS_M, running full coreference resolution significantly hurts performance for both models. This suggests that, in real-life applications, where the distribution of training instances with different degrees of coherence is skewed (as in the set of permutations generated from our model), running full coreference resolution is not a good option, since it almost makes the accuracies no better than random guessing (50%).

Moreover, considering training using PS_{BL}, running full coreference resolution has a different influence for the two datasets. For *Earthquakes*, our model significantly outperforms B&L's while the improvement is insignificant for *Accidents*. This is most probably due to the different way that entities are realized in these two datasets. As analyzed by Barzilay and Lapata (2008), in dataset *Earthquakes*, entities tend to be referred to by pronouns in subsequent mentions, while in dataset *Accidents*, literal string repetition is more common.

Given a balanced permutation distribution as we assumed in PS_{BL}, switching distant sentence pairs in *Accidents* may result in very similar entity distribution with the situation of switching closer sentence pairs, as recognized by the automatic tool. Therefore, compared to *Earthquakes*, our multiple-rank model may be less powerful in indicating the dissimilarity between the sentence orderings in a permutation and its source document, and therefore can improve on the baseline only by a small margin.

6.1.3 No Coreference Resolution

In this experiment, we do not employ any coreference resolution tool, and simply cluster head

[4]Following Elsner and Charniak (2011), we use the Wilcoxon Sign-rank test for significance.

Condition: Coreference–					
Perms	Metric	*Earthquakes*		*Accidents*	
		C	Acc	*C*	Acc
PS_{BL}	τ	4	82.8	*N*	**82.0**
	AC	3	78.0	3	**84.2**
	ED	*N*	78.2	3	*82.7
	Baseline		**83.7**		80.1
PS_M	τ	3	**86.4	*N*	**85.7
	AC	4	*84.4	*N*	**86.6
	ED	5	**86.7	*N*	**84.6
	Baseline		82.6		77.5

Table 4: Accuracies (%) of extending the standard entity-based coherence model with multiple-rank learning in sentence ordering using **Coreference**– option. Accuracies which are significantly better than the baseline are indicated by * ($p < .05$) and ** ($p < .01$).

nouns by string matching. Results are shown in Table 4.

Even with such a coarse approximation of coreference resolution, our model is able to achieve around 85% accuracy in most test cases, except for dataset *Earthquakes*, training on PS_{BL} gives poorer performance than the standard model by a small margin. But such inferior performance should be expected, because as explained above, coreference resolution is crucial to this dataset, since entities tend to be realized through pronouns; simple string matching introduces too much noise into training, especially when our model wants to train a more fine-grained discriminative system than B&L's. However, we can see from the result of training on PS_M, if the permutations used in training do not involve swapping sentences which are too far away, the resulting noise is reduced, and our model outperforms theirs. And for dataset *Accidents*, our model consistently outperforms the baseline model by a large margin (with significance test at $p < .01$).

6.1.4 Conclusions for Sentence Ordering

Considering the particular dissimilarity metric used in training, we find that *edit distance* usually stands out from the other two metrics. *Kendall's τ distance* proves to be a fairly weak metric, which is consistent with the findings of Filippova and Strube (2007) (see Section 2.3). Figure 1 plots the testing accuracies as a function of different

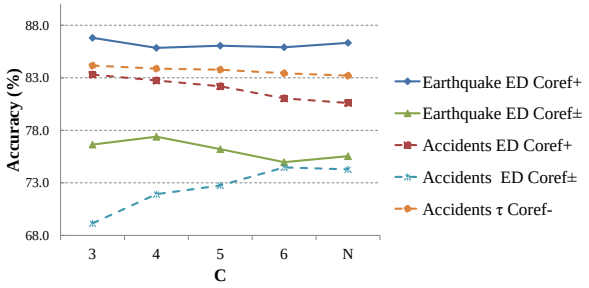

Figure 1: Effect of *C* on testing accuracies in selected *sentence ordering* experimental configurations.

choices of *C*'s with the configurations where our model outperforms the baseline model. In each configuration, we choose the dissimilarity metric which achieves the best accuracy reported in Tables 2 to 4 and the PS_{BL} permutation set. We can see that the dependency of accuracies on the particular choice of *C* is not consistent across all experimental configurations, which suggests that this free parameter *C* needs careful tuning in different experimental setups.

Combining our multiple-rank model with simple string matching for entity extraction is a robust option for coherence evaluation, regardless of the particular distribution of permutations used in training, and it significantly outperforms the baseline in most conditions.

6.2 Summary Coherence Rating

As explained in Section 3.2, we employ a simple sentence alignment between a system-generated summary and its corresponding human-written summary to construct a test ordering $\boldsymbol{\pi}$ and calculate its dissimilarity between the reference ordering $\boldsymbol{\sigma}$ from the human-written summary. In this way, we convert B&L's supervised learning model into a fully unsupervised model, since human annotations for coherence scores are not required. We use the same dataset as Barzilay and Lapata (2008), which includes multi-document summaries from 16 input document clusters generated by five systems, along with reference summaries composed by humans.

In this experiment, we consider only *average continuity (AC)* and *edit distance (ED)* as dissimilarity metrics, with **raw** configuration for rank assignment, and compare our multiple-rank model with the standard entity-based model using either full coreference resolution[5] or no resolution

[5]We run the coreference resolution tool on all documents.

Entities	Metric	Same	Full
Coreference+	*AC*	**82.5**	***72.6**
	ED	**81.3**	****73.0**
	Baseline	78.8	70.9
Coreference−	*AC*	76.3	72.0
	ED	78.8	71.7
	Baseline	**80.0**	**72.3**

Table 5: Accuracies (%) of extending the standard entity-based coherence model with multiple-rank learning in summary rating. Baselines are results of standard entity-based coherence model. Accuracies which are significantly better than the corresponding baseline are indicated by * ($p < .05$) and ** ($p < .01$).

for entity extraction. We train both models on the ranking preferences (144 in all) among summaries originating from the same input document cluster using the SVM^{rank} package (Joachims, 2006), and test on two different test sets: *same-cluster test* and *full test*. *Same-cluster* test is the one used by Barzilay and Lapata (2008), in which only the pairwise rankings (80 in all) between summaries originating from the same input document cluster are tested; we also experiment with *full test*, in which pairwise rankings (1520 in all) between all summary pairs excluding two human-written summaries are tested.

Results are shown in Table 5. **Coreference+** and **Coreference−** denote the configuration of using full coreference resolution or no resolution separately. First, clearly for both models, performance on *full test* is inferior to that on *same-cluster* test, but our model is still able to achieve performance competitive with the standard model, even if our fundamental assumption about the existence of canonical sentence ordering in documents with same content may break down on those test pairs not originating from the same input document cluster. Secondly, for the baseline model, using the **Coreference−** configuration yields better accuracy in this task (80.0% vs. 78.8% on *same-cluster test*, and 72.3% vs. 70.9% on *full test*), which is consistent with the findings of Barzilay and Lapata (2008). But our multiple-rank model seems to favor the **Coreference+** configuration, and our best accuracy even exceeds B&L's best when tested on the same set: 82.5% vs. 80.0% on *same-cluster test*, and 73.0% vs. 72.3% on *full test*.

When our model performs poorer than the baseline (using **Coreference−** configuration), the difference is not significant, which suggests that our multiple-rank model with unsupervised score assignment via simple cosine matching can remain competitive with the standard model, which requires human annotations to obtain a more fine-grained coherence spectrum. This observation is consistent with Banko and Vanderwende (2004)'s discovery that human-generated summaries look quite extractive.

7 Conclusions

In this paper, we have extended the popular coherence model of Barzilay and Lapata (2008) by adopting a multiple-rank learning approach. This is inherently different from other extensions to this model, in which the focus is on enriching the set of features for entity-grid construction, whereas we simply keep their original feature set intact, and manipulate only their learning methodology. We show that this concise extension is effective and able to outperform B&L's standard model in various experimental setups, especially when experimental configurations are most suitable considering certain dataset properties (see discussion in Section 6.1.4).

We experimented with two tasks: *sentence ordering* and *summary coherence rating*, following B&L's original framework. In sentence ordering, we also explored the influence of removing the oracular component in their original model and dealing with permutations generated from different distributions, showing that our model is robust for different experimental situations. In summary coherence rating, we further extended their model such that their original supervised learning is converted into unsupervised learning with competitive or even superior performance.

Our multiple-rank learning model can be easily adapted into other extended entity-based coherence models with their enriched feature sets, and further improvement in ranking accuracies should be expected.

Acknowledgments

This work was financially supported by the Natural Sciences and Engineering Research Council of Canada and by the University of Toronto.

References

Michele Banko and Lucy Vanderwende. 2004. Using n-grams to understand the nature of summaries. In *Proceedings of Human Language Technologies and North American Association for Computational Linguistics 2004: Short Papers*, pages 1–4.

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 42rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 141–148.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: an entity-based approach. *Computational Linguistics*, 34(1):1–34.

Danushka Bollegala, Naoaki Okazaki, and Mitsuru Ishizuka. 2006. A bottom-up approach to sentence ordering for multi-document summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 385–392.

Jackie Chi Kit Cheung and Gerald Penn. 2010. Entity-based local coherence modelling using topological fields. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 186–195.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Micha Elsner and Eugene Charniak. 2011. Extending the entity grid with entity-specific features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, pages 125–129.

Katja Filippova and Michael Strube. 2007. Extending the entity-grid coherence model to semantically related entities. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 2007)*, pages 139–142.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002)*, pages 133–142.

Thorsten Joachims. 2006. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, pages 217–226.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 545–552.

Mirella Lapata. 2006. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):471–484.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, pages 997–1006.

Nitin Madnani, Rebecca Passonneau, Necip Fazil Ayan, John M. Conroy, Bonnie J. Dorr, Judith L. Klavans, Dianne P. O'Leary, and Judith D. Schlesinger. 2007. Measuring variability in sentence ordering for news summarization. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 2007)*, pages 81–88.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002)*, pages 104–111.

Michael Strube and Simone Paolo Ponzetto. 2006. Wikirelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 1219–1224.

Renxian Zhang. 2011. Sentence ordering driven by local and global coherence for summary generation. In *Proceedings of the ACL 2011 Student Session*, pages 6–11.