Graph-based Local Coherence Modeling

Camille Guinaudeau and **Michael Strube**
Heidelberg Institute for Theoretical Studies gGmbH
Schloss-Wolfsbrunnenweg 35
69118 Heidelberg, Germany
(camille.guinaudeau|michael.strube)@h-its.org

Abstract

We propose a computationally efficient graph-based approach for local coherence modeling. We evaluate our system on three tasks: sentence ordering, summary coherence rating and readability assessment. The performance is comparable to entity grid based approaches though these rely on a computationally expensive training phase and face data sparsity problems.

1 Introduction

Many NLP applications which process or generate texts rely on information about local coherence, i.e. information about which entities occur in which sentence and how the entities are distributed in the text. This led to the development of many theories and models accounting for local coherence. One popular model, the centering model (Grosz et al., 1995), uses a ranking of discourse entities realized in particular sentences and computes transitions between adjacent sentences to provide insight in the felicity of texts. Centering models local coherence rather generally and has been applied to the generation of referring expressions (Kibble and Power, 2004), to resolve pronouns (Brennan et al., 1987, inter alia), to score essays (Miltsakaki and Kukich, 2004), to arrange sentences in the correct order (Karamanis et al., 2009), and to many other tasks. Poesio et al. (2004) observe that it is not clear how to set parameters in the centering model so that optimal performance in different tasks and languages can be achieved. Barzilay and Lapata (2008) criticize research on centering to be too dependent on manually annotated input. This led them to propose a local coherence model relying on a more parsimonious representation, the entity grid model.

The entity grid is a two dimensional array where the rows represent sentences and the columns discourse entities. From this grid Barzilay and Lapata (2008) derive probabilities of transitions between adjacent sentences which are used as features for machine learning algorithms. They evaluate this approach successfully on sentence ordering, summary coherence rating, and readability assessment. However, their approach has some disadvantages which they point out themselves: data sparsity, domain dependence and computational complexity, especially in terms of feature space issues while building their model (Barzilay and Lapata (2008, p.8, p.10, p.30), Elsner and Charniak (2011, p.126, p.127)).

In order to overcome these problems we propose to represent entities in a graph and then model local coherence by applying centrality measures to the nodes in the graph (Section 3). We claim that a graph is a more powerful representation for local coherence than the entity grid (Barzilay and Lapata, 2008) which is restricted to transitions between adjacent sentences. The graph can easily span the entire text without leading to computational complexity and data sparsity problems. Similar to the application of graph-based methods in other areas of NLP (e.g. work on word sense disambiguation by Navigli and Lapata (2010); for an overview over graph-based methods in NLP see Mihalcea and Radev (2011)) we model local coherence by relying only on centrality measures applied to the nodes in the graph. We apply our graph-based model to the three tasks handled by Barzilay and Lapata (2008) to show that it provides the same flexibility over disparate tasks as the entity grid model: sentence ordering (Section 4.1), summary coherence ranking (Section 4.2), and readability assessment (Section 4.3). In the

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 93–103,
Sofia, Bulgaria, August 4-9 2013.

The Turkish government fell after mob-tie allegations.

Turkey's constitution mandates a secular republic despite its Muslim majority.

Military and secular leaders pressured President Demirel to keep the Islamic-oriented Virtue Party on the fringe.

Business leaders feared Virtue would alienate the EU.

Table 1: Excerpt of a manual summary M from DUC2003

	GOVERNMENT	ALLEGATION	TURKEY	CONSTITUTION	SECULAR	REPUBLIC	MAJORITY	MILITARY	LEADER	PRESIDENT	DEMIREL	VIRTUE	PARTY	FRINGE	BUSINESS	EU
s_1	S	X	–	–	–	–	–	–	–	–	–	–	–	–	–	–
s_2	–	–	X	S	X	O	X	–	–	–	–	–	–	–	–	–
s_3	–	–	–	–	X	–	–	X	S	X	S	X	O	X	–	–
s_4	–	–	–	–	–	–	–	–	S	–	–	S	–	–	X	O

Table 2: Entity Grid representation of summary M

experiments sections, we discuss the impact of genre and stylistic properties of documents on the local coherence computation. We also show that, though we do not need a computationally expensive learning phase, our model achieves state-of-the-art performance. From this we conclude that a graph is an alternative to the entity grid model: it is computationally more tractable for modeling local coherence and does not suffer from data sparsity problems (Section 5).

2 The Entity Grid Model

Barzilay and Lapata (2005; 2008) introduced the entity grid, a method for local coherence modeling that captures the distribution of discourse entities across sentences in a text.

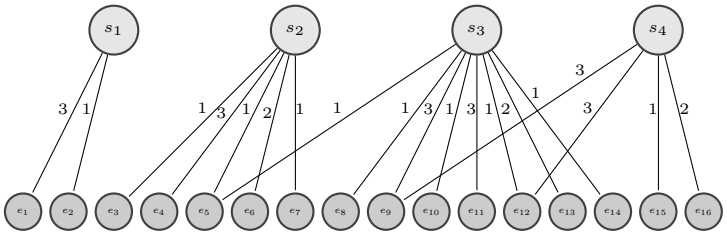
An entity grid is a two dimensional array, where rows correspond to sentences and columns to discourse entities. For each discourse entity e_j and each sentence s_i in the text, the corresponding grid cell c_{ij} contains information about the presence or absence of the entity in the sentence. If the entity does not appear in the sentence, the corresponding grid cell contains an absence marker "$-$". If the entity is present in the sentence, the cell contains a representation of the entity's syntactic role: "S" if the entity is a subject, "O" if it is an object and "X" for all other syntactic roles (cf. Table 2). When a noun is attested more than once with a different grammatical role in the same sentence, the role with the highest grammatical ranking is chosen to represent the entity (a subject is ranked higher than an object, which is ranked higher than other syntactic roles).

Barzilay and Lapata (2008) capture local coherence by means of local entity transitions, i.e. sequences of grid cells $(c_{1j} \ldots c_{ij} \ldots c_{nj})$ representing the syntactic function or absence of an entity in adjacent sentences[1]. The coherence of a sentence in relation to its local context is determined by the local entity transitions of the entities present or absent in the sentence. To make this representation accessible to machine learning algorithms, Barzilay and Lapata (2008) compute for each document the probability of each transition and generate feature vectors representing the sentences. Coherence assessment is then formulated as a ranking learning problem where the ranking function is learned with SVM^{light} (Joachims, 2002).

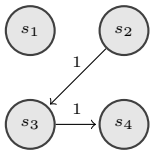
The entity grid approach has already been applied to many applications relying on local coherence estimation: summary rating (Barzilay and Lapata, 2005), essay scoring (Burstein et al., 2010) or story generation (McIntyre and Lapata, 2010). It was also used successfully in combination with other systems or features. Soricut and Marcu (2006) show that the entity grid model is a critical component in their sentence ordering model for discourse generation. Barzilay and Lapata (2008) combine the entity grid with readability-related features to discriminate documents between easy- and difficult-to-read categories. Lin et al. (2011) use discourse relations to transform the entity grid representation into a discourse role matrix that is used to generate feature vectors for machine learning algorithms similarly to Barzilay and Lapata (2008).

Several studies propose to extend the entity grid model using different strategies for entity selection. Filippova and Strube (2007) aim to improve the entity grid model performance by grouping entities by means of semantic relatedness. In their studies, Elsner and Charniak extend the number and type of entities selected and consider that each entity has to be dealt with accordingly with its information status (Elsner et al., 2007) or its named-entity category (Elsner and Charniak, 2011). Finally, they include a heuristic coreference resolution component by linking mentions which share a

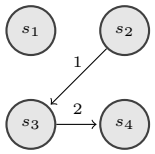
[1]For complexity reasons, Barzilay and Lapata consider only transitions between at most three sentences.



(a) Bipartite Graph



(b) Unweighted One-mode Projection



(c) Weighted One-mode Projection

	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}	e_{11}	e_{12}	e_{13}	e_{14}	e_{15}	e_{16}
s_1	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s_2	0	0	1	3	1	2	1	0	0	0	0	0	0	0	0	0
s_3	0	0	0	0	1	0	0	1	3	1	3	1	2	1	0	0
s_4	0	0	0	0	0	0	0	0	3	0	0	3	0	0	1	2

(d) Incidence Matrix

	s_1	s_2	s_3	s_4
s_1	0	0	0	0
s_2	0	0	1	0
s_3	0	0	0	1
s_4	0	0	0	0

(e) Unweighted Adjacency Matrix

	s_1	s_2	s_3	s_4
s_1	0	0	0	0
s_2	0	0	1	0
s_3	0	0	0	2
s_4	0	0	0	0

(f) Weighted Adjacency Matrix

Figure 1: Bipartite graph for summary M from Table 1, one-mode projections and associated incidence and adjacency matrices. Weights in Figure 1(a) are assigned as follows: "S" = 3, "O" = 2, "X" = 1, "–" = 0 (no edge).

head noun. These extensions led to the best results reported so far for the sentence ordering task.

3 Method

Our model is based on the insight that the entity grid (Barzilay and Lapata, 2008) corresponds to the incidence matrix of a bipartite graph representing the text (see Newman (2010) for more details on graph representation). A fundamental assumption underlying our model is that this bipartite graph contains the entity transition information needed for local coherence computation, rendering feature vectors and learning phase unnecessary. The bipartite graph $G = (V_s, V_e, L, w)$ is defined by two independent sets of nodes – that correspond to the set of sentences V_s and the set of entities V_e of the text – and a set of edges L associated with weights w. An edge between a sentence node s_i and an entity node e_j is created in the bipartite graph if the corresponding cell c_{ij} in the entity grid is not equal to "–". Each edge is associated with a weight $w(e_j, s_i)$ that depends on the grammatical role of the entity e_j in the sentence s_i[2]. In contrast to Barzilay and Lapata's entity grid that contains information about absent entities, our graph-based representation only contains "positive" information. Figure 1(a) shows an example of the bipartite graph that corresponds to the grid in Table 2. The incidence matrix of this graph (Figure 1(d)) is very similar to the entity grid.

[2]The assignment of weights is described in Section 4.

By modeling entity transitions, Barzilay and Lapata rely on links that exist between sentences to model local coherence. In the same spirit, we apply different kinds of one-mode projections to the sentence node set V_s of the bipartite graph to represent the connections that exist between – potentially non adjacent – sentences in the graph. These projections result in graphs where nodes correspond to sentences. An edge is created between two nodes if the corresponding sentences have a least one entity in common. Contrary to the bipartite graph, one-mode projections are directed as they follow the text order. Therefore, in projection graphs an edge can exist between the first and the second sentence while the inverse is not possible. In our model, we define three kinds of projection graphs, P_U, P_W and P_{Acc}, depending on the weighting scheme associated with their edges. In P_U, weights are binary and equal 1 when two sentences have a least one entity in common (Figure 1(b)). In P_W, edges are weighted according to the number of entities "shared" by two sentences (Figure 1(c)). In P_{Acc} syntactic information is accounted for by integrating the edge weights in the bipartite graph. In this case, weights are equal to

$$W_{ik} = \sum_{e \in E_{ik}} w(e, s_i) \cdot w(e, s_k) \,,$$

where E_{ik} is the set of entities shared by s_i and s_k. Distance between sentences s_i and s_k can also be integrated in the weight of one-mode projections to decrease the importance of links that ex-

ists between non adjacent sentences. In this case, the weights of the projection graphs are divided by $k - i$.

From this graph-based representation, the local coherence of a text T can be measured by computing the average outdegree of a projection graph P. This centrality measure was chosen for two main reasons. First, it allows us to evaluate to which extent a sentence is connected, in terms of discourse entities, with the other sentences of the text. Second, compared to other centrality measures, the computational complexity of the average outdegree is low ($\mathcal{O}(\frac{N*(N-1)}{2})$ for a document composed by N sentences), keeping the local coherence estimation feasible on large documents and on large corpora. Formally, the local coherence of a text T is equal to

$$\begin{aligned} \textit{LocalCoherence}(T) &= \textit{AvgOutDegree}(P) \\ &= \frac{1}{N} \sum_{i=1..N} \textit{OutDegree}(s_i) \, , \end{aligned}$$

where $\textit{OutDegree}(s_i)$ is the sum of the weights associated to edges that leave s_i and N is the number of sentences in the text. This value can also be seen as the sum of the values of the adjacency matrix of the projection graph (Figures 1(e) and 1(f)) divided by the number of sentences.

4 Experiments

We compare our model with the entity grid approach and evaluate the influence of the different weighting schemes used in the projection graphs, either P_W or P_{Acc}, where weights are potentially decreased by distance information $Dist$. Our baseline corresponds to local coherence computation based on the unweighted projection graph P_U.

For graph construction, all nouns in a document are considered as discourse entities, even those which do not head NPs as this is beneficial for the entity grid model as described in Elsner and Charniak (2011). We also propose to use a coreference resolution system and consider coreferent entities to be the same discourse entity. To do so, we use one of the top performing systems from the CoNLL 2012 shared task (Martschat et al., 2012). As the coreference resolution system is trained on well-formed textual documents and expects a correct sentence ordering, we use in all our experiments only features that do not rely on sentence order (e.g. alias relations, string matching, etc.).

Grammatical information associated with each entity is extracted automatically thanks to the Stanford parser using dependency conversion (de Marneffe et al., 2006). Syntactic weights in the bipartite graph are defined following the linguistic intuition that subjects are more important than objects, which are themselves more important than other syntactic roles. Preliminary experiments show that as long as weight assignment follows the scheme S $>$ O $>$ X, then more coherent documents are associated with a higher local coherence value than less coherent document in 90% of cases (while this value equals 49% when no restriction is given on syntactic weights order). Moreover, as the local coherence computation is a linear combination of the syntactic weights, the function is smooth and no large variations of the local coherence values are observed for small changes of weights' values. For these reasons, weights $w(e, s_i)$ are set as follows: 3 if e is subject in s_i, 2 if e is an object and 1 otherwise.

We evaluate the ability of our graph-based model to estimate the local coherence of a textual document with three different experiments. First, we perfom a sentence ordering task (Section 4.1) as proposed in Barzilay and Lapata (2008). Then, as the first task uses "artificial" documents, we also work on two other tasks that involve "real" documents: summary coherence rating (Section 4.2), and readability assessment (Section 4.3). In these experiments, distance computation and syntactic weights are the same for all tasks and all corpora. However, the model is also flexible and can be adaptated to the different tasks by optimizing the parameters on a development data set, which may give better results.

4.1 Sentence Ordering

The first experiment consists in ranking alternative sentence orderings of a document, as proposed by Barzilay and Lapata (2008) and Elsner and Charniak (2011).

4.1.1 Experimental Settings

The sentence ordering task can be performed in two ways: discrimination and insertion. Discrimination consists in comparing a document to a random permutation of its sentences. For this, our system associates local coherence values with the original document and its permutation, the output of our system being considered as correct if the score for the original document is higher than the

score of its permutation. In the insertion task, proposed by Elsner and Charniak (2011), we evaluate the ability of our system to retrieve the original position of a sentence previously removed from a document. For this, each sentence is removed in turn and a local coherence score is computed for every possible reinsertion position. The system output is considered as correct if the document associated with the highest local coherence score is the one in which the sentence is reinserted in the correct position.

These two tasks were performed on documents extracted from the English test part of the CoNLL 2012 shared task (Pradhan et al., 2012). This corpus, composed by documents of multiple news sources – spoken or written – was preferred to the ACCIDENTS and EARTHQUAKES corpora used by Barzilay and Lapata (2008) for two reasons. First, as mentioned by Elsner and Charniak (2008), these corpora use a very constrained style and are not typical of normal informative documents[3]. Second, we want to evaluate the influence of automatically performed coreference resolution in a controlled fashion. The coreference resolution system used performs well on the CoNLL 2012 data. In this dataset, documents composed by the concatenation of differents news articles or too short to have at least 20 permutations were discarded from the corpus. This filtering results in 61 documents composed of 36.1 sentences or 2064 word tokens on average. In both discrimination and insertion, we compare our system against a random baseline where random values are associated with the different orderings.

4.1.2 Discrimination

Accuracy is used to evaluate the ability of our system to discriminate a document from 20 different permutations. It equals the number of times our system gives the highest score to the original document, divided by the number of comparisons. Since the model can give the same score for a permutation and the original document, we also compute F-measure where recall is *correct/total* and precision equals *correct/decisions*. We test significance using the Student's t-test that can detect significant differences between paired samples. Moreover, as increasing the number of hypotheses in a test can also increase the likelihood of witnessing a rare event, and therefore, the chance to reject the null hypothesis when it is true, we use the Bonferroni correction to adjust the increased random likelihood of apparent significance.

	Acc	F	Acc	F
Random	0.496	0.496		
B&L	0.877	0.877		
E&C	0.915	0.915		
	wo coref		w coref	
P_U, *Dist*	0.830	0.830	0.833	0.833
P_W, *Dist*	0.871	0.871	0.849	0.849
P_{Acc}, *Dist*	0.889	0.889	0.852	0.852

Table 3: Discrimination, reproduced baselines (B&L: Barzilay and Lapata (2008); E&C Elsner and Charniak (2011)) vs. graph-based

Table 3 presents the values obtained by three baseline systems when applied to our corpus. Results for the entity grid models described by Barzilay and Lapata (2008) and Elsner and Charniak (2011) are obtained by using Micha Elsner's reimplementation in the Brown Coherence Toolkit[4]. The system was trained on the English training part of the CoNLL 2012 shared task filtered in the same way as the test part.

Table 3 also displays the results for our model. These values show that our system performs comparable to the state-of-the-art. Indeed, the difference between our best results and those of Elsner and Charniak are not statistically significant.

In this experiment, distance information is critical. Without it, it is not possible to distinguish between an original document and one of its permutation as both contain the same number and kind of entities. Distance however can detect changes in the distribution of entities within the document as space between entities is significantly modified when sentence order is permuted. When the number of entities "shared" by two sentences is taken into account (P_W), the accuracy of our system grows (from 0.830 to 0.871). Table 3 finally shows that syntactic information improves the performance of our system (yet not significantly) and gives the best results (P_{Acc}).

We also evaluated the influence of coreference resolution on the performance of our system. Us-

[3]Our graph-based model obtains for the discrimination task an accuracy of 0.846 and 0.635 on the ACCIDENTS and EARTHQUAKES datasets, respectively, compared to 0.904 and 0.872 as reported by Barzilay and Lapata (2008).

[4]`https://bitbucket.org/melsner/browncoherence`; B&L is Elsner's "baseline entity grid" (command line option '-n'), E&C is Elsner's "extended entity grid" ('-f')

	Acc.	Ins.	Acc.	Ins.
Random	0.028	0.071		
E&C	0.068	0.167		
	wo coref		w coref	
P_U, *Dist*	0.062	0.101	0.068	0.120
P_W, *Dist*	0.075	0.114	0.070	0.138
P_{Acc}, *Dist*	0.071	0.102	0.067	0.097

Table 4: Insertion, reproduced baselines vs. graph-based

ing coreference resolution improves the performance of the system when distance information is used alone in the system (Table 3). However, this improvement is not statistically significant.

4.1.3 Insertion

Sentence insertion is much more difficult than discrimination for two reasons. First, in insertion, permutations only differ by one sentence. Second, a document is compared to many more permutations in insertion task than in discrimination.

In complement to accuracy, we use the insertion score introduced by Elsner and Charniak (2011) for evaluation. This score – the higher, the better – computes the proximity between the initial and the proposed position of a sentence, averaged by the number of sentences.

Table 4 shows that, as expected, results for this task are much lower than those obtained for discrimination. However they are still comparable with the results of Elsner and Charniak (2011)[5].

As previously and for the same reasons, distance information is critical for this task. The best results, that present a statistically significant improvement when compared to the random baseline, are obtained when distance information and the number of entities "shared" by two sentences are taken into account (P_W). We can see that the accuracy value obtained with our system is higher than the one provided with the entity grid model. However, the entity grid model reaches a significantly higher insertion score. This means that, if it makes more mistakes than our system, the position chosen by the entity grid model is usually closer to the correct position. Finally, contrary to the discrimination task, syntactic information (P_{Acc}) does not improve the performance of our system.

[5]Their results are slightly lower than those presented in their paper, probably because our corpus is composed by documents that can be longer than the ones used in their experiments (Wall Street Journal articles).

When the coreference resolution system is used, the best accuracy value decreases while the insertion score increases from 0.114 to 0.138 (Table 4). Therefore, coreference resolution tends to associate positions that are closer to the original ones.

4.2 Summary Coherence Rating

To reconfirm the hypothesis that our model can estimate the local coherence of a textual document, we perform a second experiment, summary coherence rating. To this end, we apply our model on the corpus used and proposed by Barzilay and Lapata (2008). As the objective of our model is to estimate the *coherence* of a summary, we prefer this dataset to other summarization evaluation task corpora, as these account for other dimensions of the summaries: content selection, fluency, etc. Starting with a pair of summaries, one slightly more coherent than the other, the objective of the task is to order the two summaries according to local coherence.

4.2.1 Experimental Settings

For the summary coherence rating experiment, pairs to be ordered are composed of summaries extracted from the Document Understanding Conference (DUC 2003). Summaries, provided either by humans or by automatic systems, were judged by seven humans annotators and associated with a coherence score (for more details on this score see Barzilay and Lapata (2008)). 80 pairs were then created, each of these being composed by two summaries of a same document where the score of one of the summaries is significantly higher than the score of the second one. Even though all summaries are of approximately the same length (114.2 words on average), their sentence length can vary considerably. Indeed, more coherent summaries tend to have more sentences and contain less entities.

For evaluation purposes, the accuracy still corresponds to the number of correct ratings divided by the number of comparisons, while the F-measure combines recall and precision measures. As before, significance is tested with the Student's t-test accounting for the Bonferroni correction.

4.2.2 Results

Table 5 compares the results reported by Barzilay and Lapata (2008) on the exact same corpus with the results obtained with our system. It shows that

	Acc.	F	Acc.	F
B&L	0.833			
	wo coref		w coref	
P_U	0.800	0.815	0.700	0.718
P_W	0.613	0.613	0.538	0.548
P_{Acc}	0.700	0.704	0.638	0.638
P_U, *Dist*	0.650	0.658	0.550	0.557
P_W, *Dist*	0.525	0.525	0.513	0.513
P_{Acc}, *Dist*	0.700	0.700	0.588	0.588

Table 5: Summary Coherence Rating, reported results from Barzilay and Lapata (2008) vs. graph-based

our system gives results comparable to those obtained by Barzilay and Lapata (2008).

This table also shows that, contrary to sentence ordering task, accounting for the distance between two sentences (*Dist*) tends to decrease the results. This difference is explained by the fact that a manual summary, usually considered as more coherent by humans annotators, tends to contain more (and shorter) sentences than an automatic one. As adding distance information decreases the value of our local coherence score, our graph-based model gives better results without it.

Moreover, in contrast to the first experiment, when accounting for the number of entities "shared" by two sentences (P_W), values of accuracy and F-measure are lower. We explain this behaviour by the number of sentences contained in the less coherent documents. Indeed, they are composed by a smaller number of sentences but contain more entities on average. This means that, in these documents, two sentences tend to share a larger number of entities and therefore have a higher local coherence score when the P_W projection graph is used.

When combined with distance information, syntactic information still improves the results (P_{Acc}), though not significantly, but does not lead to the best results for this task.

Finally, Table 5 also shows that using a coreference resolution system for document representation does not improve the performance of our system. We believe that, as mentioned by Barzilay and Lapata (2008), this degradation is related to the fact that automatic summarization systems do not use anaphoric expressions which makes the coreference resolution system useless in this case.

With our graph-based model, the best results are obtained by the baseline (P_U), and experiments show that adding information about distance or syntax does not help in this context. It seems therefore necessary to integrate information that is more appropriate to summaries. Although making the model more appropriate for a specific task is out of the scope of this paper, our model is flexible and accounting for information about genre differences or sentence length, by adding weights in the graph-based representation of the document, is feasible without any modification of the model.

4.3 Readability Assessment

Barzilay and Lapata (2008) argue that grid models are domain and style dependent. Therefore they proposed a readability assessment task to test if the entity grid model can be used for style classification. They combined their model with Schwarm and Ostendorf's (2005) readability features and use Support Vector Machines to classify documents in two categories. With the same intention, we evaluate the ability of our model to differentiate "easy to read" documents from difficult ones.

4.3.1 Experimental Settings

The objective of the readability assessment task is to evaluate how difficult to read a document is. We perform this task on the data used by Barzilay and Lapata (2008), a corpus collected originally by Barzilay and Elhadad (2003) from the *Encyclopedia Britannica* and its version for children, the *Britannica Elementary*. Both versions contain 107 articles. In *Encyclopedia Britannica*, documents are composed by an average of 83.1 sentences while they contain 36.6 sentences in *Britannica Elementary*. Although these texts are not explicitly annotated with grade levels, they represent two broad readability categories.

In order to estimate the complexity of a document, our model computes the local coherence score for each article in the two categories. The article associated with the higher score is considered to be the more readable as it is more coherent, needing less interpretation from the reader than a document associated with a lower local coherence score. Values presented in the following section correspond to accuracy, where the system is correct if it assigns the higher local coherence score to the most "easy to read" document, and F-measure.

	Acc.	F	Acc.	F
S&O	0.786			
B&L	0.509			
B&L + S&O	0.888			
	wo coref		w coref	
P_U	0.589	0.589	0.374	0.374
P_W	0.579	0.579	0.383	0.383
P_{Acc}	0.645	0.645	0.421	0.421
P_U, *Dist*	0.589	0.589	0.280	0.280
P_W, *Dist*	0.570	0.570	0.290	0.290
P_{Acc}, *Dist*	0.766	0.766	0.308	0.308

Table 6: Readability, reported results from Barzilay and Lapata (2008) vs. graph-based (S&O: Schwarm and Ostendorf (2005))

4.3.2 Results

In order to compare our results with those reported by Barzilay and Lapata (2008), entities used for the graph-based representation are discourse entities that head NPs.

Table 6 shows that, for this task, syntactic information plays a dominant role (P_{Acc}). A statistically significant improvement is provided by including syntactic information. It gives more weight to subject entities that are more numerous in the *Britannica Elementary* documents which are composed by simpler and shorter sentences. Finally, when distance is accounted for together with syntactic information, the accuracy is significantly improved ($p < 0.01$) with regard to the results obtained with syntactic information only.

Table 6 also shows that when the number of entities "shared" by two sentences is accounted for (P_W), the results are lower. Indeed, *Encyclopedia Britannica* documents are composed by longer sentences, that contain a higher number of entities. This increases the local coherence value of difficult documents more than the value of "easy to read" documents, that contain less entities.

When our graph-based representation used the coreference resolution system, unlike the observation of Barzilay and Lapata (2008), the results of our model decrease significantly. The poor performance of our system in this case can be explained by the fact that the coreference resolution system regroups more entities in *Encyclopedia Britannica* documents than in *Britannica Elementary* ones. Therefore, the number of entities that are "shared" by two sentences increases more importantly in the *Encyclopedia Britannica* corpus, while the distance between two occurrences of one entity decreases in a more significant manner. For these reasons, the coherence scores associated with "difficult to read" documents tend to be higher when coreference resolution is performed on our data, which reduces the performance of our system. As before, syntactic information leads to the best results, but does not allow the accuracy to be higher than random anymore.

Compared to the results provided by Barzilay and Lapata (2008) with the entity grid model alone, our representation outperforms their model significantly. We believe that this difference is caused by how syntactic information is introduced in the document representation and by the fact that our system can deal with entities that appear throughout the whole document while the entity grid model only looks at entities within a three sentences windows. Our model which captures exclusively local coherence is almost on par with the results reported for Schwarm & Ostendorf's (2005) system which relies on a wide range of lexical, syntactic and semantic features. Only when Barzilay and Lapata (2008) combine the entity grid with Schwarm & Ostendorf's features they reach performance considerably better than ours.

In addition to the experiments proposed by Barzilay and Lapata (2008), we used a third readability category, the *Britannica Student*, that contains articles targeted for youths (from 11 to 14 years old). These documents, which are quite similar to the *Encyclopedia Britannica* ones, are composed by an average of 44.1 sentences. As we were only able to find 99 articles out of the 107 original ones in this category, sub corpora of the three categories were used for the comparison with the *Britannica Student* articles.

Table 7 shows the results obtained for the comparisons between the two first categories and the *Britannica Student* articles. As previously, coreference resolution tends to lower the results, therefore only values obtained without coreference resolution are reported in the table.

When articles from *Britannica Student* are compared to articles extracted from *Encyclopedia Britannica*, Table 7 shows that the different parameters have the same influence as for comparing between *Encyclopedia Britannica* and *Britannica Elementary*: statistically significant improvement with syntactic information, higher values when distance is taken into account, etc. However, it

	Brit. vs. Stud.		*Stud. vs. Elem.*	
	Acc.	F	Acc.	F
P_U	0.444	0.444	0.667	0.667
P_W	0.434	0.434	0.636	0.636
P_{Acc}	0.465	0.465	0.707	0.707
P_U, *Dist*	0.475	0.475	0.646	0.646
P_W, *Dist*	0.485	0.485	0.616	0.616
P_{Acc}, *Dist*	0.556	0.556	0.657	0.657

Table 7: Readability, comparison between *Encyclopedia Britannica*, *Britannica Elementary* and *Britannica Student*

can also be seen that accuracy and F-measure are lower for comparing these two corpora. This is probably due to the stylistic difference between these two kinds of articles, which is less significant than the difference between articles from *Encyclopedia Britannica* and *Britannica Elementary*.

Concerning the comparison between *Britannica Student* and *Britannica Elementary* articles, Table 7 shows that integrating distance information gives slightly different results and tends to decrease the values of accuracy and F-measure. This is explained by the fact that *Britannica Elementary* documents contain fewer entities than *Britannica Student* articles. As the length of the two kinds of articles is similar, distance between entities in *Britannica Elementary* documents is more important. As a result, accounting for distance information lowers the local coherence values for the more coherent document, which reduces the performance of our model. As previously, syntactic information improves the results and, for this comparison, the best result is obtained when syntactic information alone is accounted for. This leads to an accuracy which is almost equal to the one when comparing *Encyclopedia Britannica* and *Britannica Elementary* (0.707 against 0.766).

These two additional experiments show that our model is style dependent. It obtains better results when it has to distinguish between *Encyclopedia Britannica* and *Britannica Elementary* or *Britannica Student* and *Britannica Elementary* articles which present a more important difference from a stylictic point of view than articles from *Encyclopedia Britannica* and *Britannica Elementary*.

5 Conclusions

In this paper, we proposed an unsupervised and computationally efficient graph-based local coherence model. Experiments show that our model is robust among tasks and domains, and reaches reasonable results for three tasks with the same parameter values and settings (i.e. accuracy values of 0.889, 0.70 and 0.766 for sentence ordering, summary coherence rating and readability assessment tasks respectively (P_{Acc}, *Dist*)). Moreover, our model can be optimized and obtains results comparable with entity grid based methods when proper settings are used for each task.

Our model has two main advantages over the entity grid model. First, as the graph used for document representation contains information about entity transitions, our model does not need a learning phase. Second, as it relies only on graph centrality, our model does not suffer from the computational complexity and data sparsity problems mentioned by Barzilay and Lapata (2008).

Our current model leaves space for improvement. Future work should first investigate the integration of information about entities. Indeed, our model only uses entities as indications of sentence connection although it has been shown that distinguishing important from unimportant entities, according to their named-entity category, has a positive impact on local coherence computation (Elsner and Charniak, 2011). Moreover, future work should also examine the use of discourse relation information, as proposed in (Lin et al., 2011). This can be easily done by adding edges in the projection graphs when sentences contain entities related from a discourse point of view while Lin et al.'s approach suffers from complexity and data sparsity problems similar to the entity grid model.

Finally, these promising results on local coherence modeling make us believe that our graph-based representation can be used without much modification for other tasks, e.g. extractive summarization or topic segmentation. This could be achieved with link analysis algorithms such as PageRank, that decide on the importance of a (sentence) node within a graph based on global information recursively drawn from the entire graph.

Acknowledgments. This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The first author has been supported by a HITS postdoctoral scholarship. We would like to thank Mirella Lapata and Regina Barzilay for making their data available and Micha Elsner for providing his toolkit.

References

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, Sapporo, Japan, 11–12 July 2003, pages 25–32.

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Mich., 25–30 June 2005, pages 141–148.

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

Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, Stanford, Cal., 6–9 July 1987, pages 155–162.

Jill Burstein, Joel Tetreault, and Slava Andreyev. 2010. Using entity-based features to model coherence in student essays. In *Proceedings of Human Language Technologies 2010: The Conference of the North American Chapter of the Association for Computational Linguistics*, Los Angeles, Cal., 2–4 June 2010, pages 681–684.

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*, Genoa, Italy, 22–28 May 2006, pages 449–454.

Micha Elsner and Eugene Charniak. 2008. Coreference-inspired coherence modeling. In *Proceedings ACL-HLT 2008 Conference Short Papers*, Columbus, Ohio, 15–20 June 2008, pages 41–44.

Micha Elsner and Eugene Charniak. 2011. Extending the entity grid with entity-specific features. In *Proceedings of the ACL 2011 Conference Short Papers*, Portland, Oreg., 19–24 June 2011, pages 125–129.

Micha Elsner, Joseph Austerweil, and Eugene Charniak. 2007. A unified local and global model for discourse coherence. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, N.Y., 22–27 April 2007, pages 436–443. Read this version: http://www.cs.brown.edu/ melsner/order.pdf.

Katja Filippova and Michael Strube. 2007. Extending the entity-grid coherence model to semantically related entities. In *Proceedings of the 11th European Workshop on Natural Language Generation*, Schloss Dagstuhl, Germany, 17–20 June 2007, pages 139–142.

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.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, 23–26 July 2002, pages 133–142.

Nikiforos Karamanis, Chris Mellish, Massimo Poesio, and Jon Oberlander. 2009. Evaluating centering for information ordering using corpora. *Computational Linguistics*, 35(1):29–46.

Rodger Kibble and Richard Power. 2004. Optimizing referential coherence in text generation. *Computational Linguistics*, 30(4):401–416.

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*, Portland, Oreg., 19–24 June 2011, pages 997–1006.

Sebastian Martschat, Jie Cai, Samuel Broscheit, Éva Mújdricza-Maydt, and Michael Strube. 2012. A multigraph model for coreference resolution. In *Proceedings of the Shared Task of the 16th Conference on Computational Natural Language Learning*, Jeju Island, Korea, 12–14 July 2012, pages 100–106.

Neil McIntyre and Mirella Lapata. 2010. Plot induction and evolutionary search for story generation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 11–16 July 2010, pages 1562–1572.

Rada Mihalcea and Dragomir Radev. 2011. *Graph-based Natural Language Processing and Information Retrieval*. Cambridge Univ. Press, Cambridge, U.K.

Eleni Miltsakaki and Karen Kukich. 2004. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 10(1):25–55.

Roberto Navigli and Mirella Lapata. 2010. An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):678–692.

Mark E.J. Newman. 2010. *Networks: An Introduction*. Oxford University Press, New York, N.Y.

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

Sameer Pradhan, Alessandro Moschitti, and Nianwen Xue. 2012. CoNLL-2012 Shared Task: Modeling multilingual unrestricted coreference in OntoNotes.

In *Proceedings of the Shared Task of the 16th Conference on Computational Natural Language Learning*, Jeju Island, Korea, 12–14 July 2012, pages 1–40.

Sarah E. Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Mich., 25–30 June 2005, pages 523–530.

Radu Soricut and Daniel Marcu. 2006. Discourse generation using utility-trained coherence models. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, 17–21 July 2006, pages 1105–1112.