Graph-based Coherence Modeling For Assessing Readability

Mohsen Mesgar and **Michael Strube**
Heidelberg Institute for Theoretical Studies gGmbH
Schloss-Wolfsbrunnenweg 35
69118 Heidelberg, Germany
(mohsen.mesgar|michael.strube)@h-its.org

Abstract

Readability depends on many factors ranging from shallow features like word length to semantic ones like coherence. We introduce novel graph-based coherence features based on frequent subgraphs and compare their ability to assess the readability of Wall Street Journal articles. In contrast to Pitler and Nenkova (2008) some of our graph-based features are significantly correlated with human judgments. We outperform Pitler and Nenkova (2008) in the readability ranking task by more than 5% accuracy thus establishing a new state-of-the-art on this dataset.

1 Introduction

Readability depends on many factors which enable readers to process a text. These factors can be used by readability assessment methods to quantify the difficulty of text understanding. Possible applications of readability assessment are automatic text summarization and simplification systems. Measuring readability can also be used in question answering and knowledge extraction systems to prune texts with low readability (Kate et al., 2010).

Many different text features have been used to assess readability. They include shallow features (Flesch, 1948; Kincaid et al., 1975), language modeling features (Si and Callan, 2001; Collins-Thompson and Callan, 2004), syntactic features (Schwarm and Ostendorf, 2005) and text flow or coherence (Barzilay and Lapata, 2008; Pitler and Nenkova, 2008). In a coherent text each sentence has some connections with other sentences. Although these local connections make the text more readable, the corresponding coherence features used in Pitler and Nenkova (2008) (Section 2) are not strongly correlated with human judgments.

The main goal of this paper is to introduce novel graph-based coherence features for assessing readability. To achieve this goal, we use the entity graph coherence model by Guinaudeau and Strube (2013) (Section 3.1.1) and follow two ideas. The first main idea is to use a graph representation of rhetorical relations between sentences of a text (Section 3.1.2) and to merge the entity graph and the rhetorical graph (Section 3.1.3). Hence we enrich the entity graph and consequently consider the distribution of two aspects of coherence (i.e. entities and discourse relations) simultaneously. The second main idea is to apply subgraph mining algorithms to find frequent subgraphs (i.e. patterns) in texts (Section 3.2). Subgraph mining has been successfully applied to other tasks, e.g. image processing (Nowozin et al., 2007) and language modeling (Biemann et al., 2012). We hypothesize that text coherence correlates with frequent subgraphs (vaguely reminding us of coherence patterns (Daneš, 1974)) and that the mined patterns are good predictors for readability ratings.

Our study is novel in introducing new and informative graph-based coherence features. We examine the predictive power of these feature in two experiments: first, readability rating prediction, and second, ranking texts according to the readability (Section 5).

*Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)*, pages 309–318,
Denver, Colorado, June 4–5, 2015.

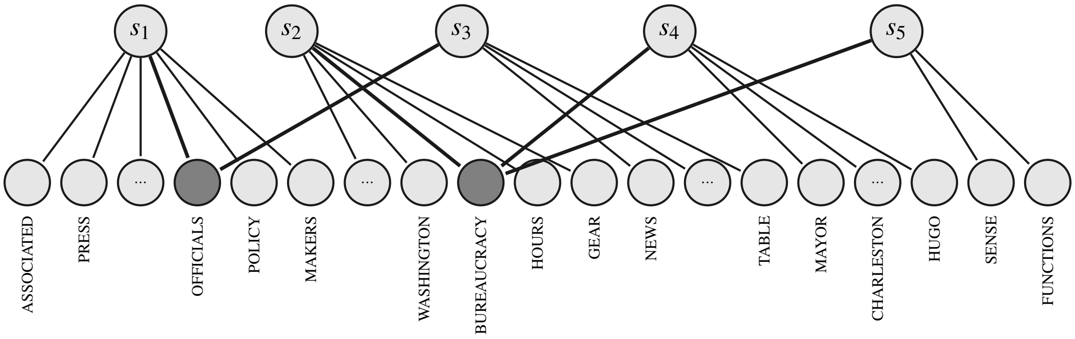

Figure 1: The entity graph representation of the text in Table 1. Dark entities are shared by the sentences.

2 Readability Assessment

The quality of a text depends on different factors which make the text easier to read. These factors range from shallow features like word length to semantic features like coherence. Readability assessment leads to two problems: distinguishing and recognizing readability levels of texts and predicting human readability ratings.

Pitler and Nenkova (2008) use all entity transitions of the entity grid model (Barzilay and Lapata, 2008) as coherence features. They compute the correlation between them and readability ratings and show that none of them is significantly correlated with human readability judgments. Indeed, none of these features on its own is a good predictor to measure coherence and to predict readability as well.

3 Method

We introduce the graph representation of a text and propose to use these graphs to model coherence.

3.1 Graphs

3.1.1 Entity Graph

Guinaudeau and Strube (2013) describe a graph-based version of the entity grid (Barzilay and Lapata, 2008) which models the interaction between entities and sentences as a bipartite graph. This graph contains two sets of nodes: sentences and entities. Sentence and entity nodes are connected if and only if the entity is mentioned in the sentence (Figure 1). Edges are weighted according to the grammatical role of the entity mentioned in the sentence.

Guinaudeau and Strube (2013) model entity transitions between sentences via a one-mode projection of the entity graph. The one-mode projection is a graph consisting of sentence nodes that are connected if and only if they have at least one entity in common in the entity graph. One-mode projections are directed as they follow the text order. Hence, backward edges never occur. Guinaudeau and Strube (2013) introduce three kinds of projections. The unweighted projection P_u^{ER} models the existence of the entity connections between sentences. The weighted projection P_w^{ER} uses the number of shared entities by sentences as a weight for the corresponding edge (Figure 2). P_{acc}^{ER} takes the grammatical function of entities in sentences into account as edge weights. Guinaudeau and Strube (2013) show that P_{acc}^{ER} does not perform well for readability assessment. It does not outperform P_w^{ER} in our

S1: The *[Associated] [Press]*'s *[earthquake] [coverage]* drew *[attention]* to a *[phenomenon]* that deserves some *[thought]* by public *[officials]* and other *[policy] [makers]*.

S2: Private *[relief] [agencies]*, such as the *[Salvation] [Army]* and *[Red] [Cross]*, mobilized almost instantly to help *[people]*, while the *[Washington] [bureaucracy]* "took *[hours]* getting into *[gear]*."

S3: One *[news]* show we saw *[yesterday]* even displayed 25 federal *[officials]* meeting around a *[table]*.

S4: We recall that the *[mayor]* of *[Charleston]* complained bitterly about the federal *[bureaucracy]*'s response to *[Hurricane Hugo]*.

S5: The *[sense]* grows that modern public *[bureaucracies]* simply don't perform their assigned *[functions]* well.

Table 1: A sample text from the Wall Street Journal dataset (Pitler and Nenkova, 2008).

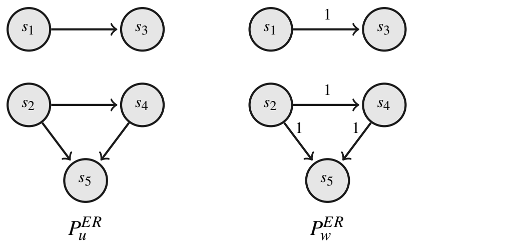

Figure 2: P_u^{ER}: unweighted, and P_w^{ER}: weighted projection graphs. In the weighted projection all edge weights are equal to one, because all sentences share one entity.

experiments as well. Thus, we do explain further details of P_w^{ER} here.

3.1.2 Discourse Relation Graph

Lin et al. (2011) and Lin (2011) use Rhetorical Structure Theory (RST) to describe and model coherence by considering the transitions between discourse relations. Inspired by the entity grid they expand the relation sequence into a two-dimensional matrix whose rows and columns are sentences and entities, respectively. The cell $\langle s_i, e_j \rangle$ corresponds to the set of discourse relations entity e_j is involved with in sentence s_i. These methods are based on entity transitions which, however, are intuitively implausible, because discourse relations connect sentences (or elementary discourse units).

Since discourse relations capture interactions between sentences (Table 2), we model these relations with a graph.

Relation	Arg1	Arg2
Implicit_Expansion	S1	S2
Explicit_Comparison	S2	S2
Implicit_Expansion	S2	S3
Implicit_Temporal	S3	S4
Implicit_Contingency	S4	S5

Table 2: PDTB-style discourse relations (Prasad et al., 2008) of the sample text in Table 1

A discourse relation graph is $P_u^{DR} = (V, R)$, where V is the set of sentence nodes and R is the edge set which represents all discourse relations in the text. Two sentence nodes are adjacent if and only if they are connected by at least one discourse relation. Intra-sentential discourse relations are represented as self-edges. We define P_w^{DR} as a weighted discourse relation graph whose edge weights are the number of discourse relations between sentence nodes (Figure 3).

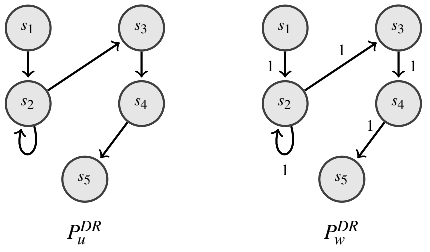

Figure 3: P_u^{DR}: unweighted, and P_w^{DR}: weighted discourse relation graphs.

3.1.3 Combined Entity and Discourse Relation Graphs

Both projection and discourse relation graphs represent different types of connections. These graphs can be merged by employing basic operators.

We use the $\vee$ operator (logical OR) to combine the projection graph P_u^{ER} with the P_u^{DR} graph. The $\vee$ operator takes two sentence nodes and creates an edge between them if they are connected at least by one connection, whether entity transition (P_u^{ER}) or discourse relations (P_u^{DR}). The other basic logical operators (e.g. $\wedge$ or $\oplus$) lose connections. Hence we do not report on their performance. Inspired by linear regression models we combine the weighted graphs by adding ($+$) the edge weights in P_w^{ER} and P_w^{DR} (Figure 4).

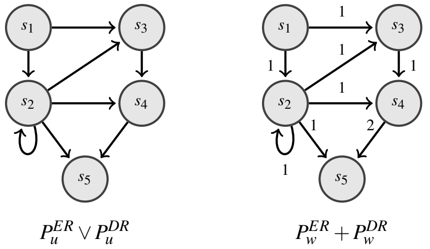

Figure 4: Combined entity and discourse relation graphs.

3.2 Coherence Features

We use the proposed graphs to introduce novel coherence features.

Average outdegree. Measures to which extent a sentence is connected with other sentences (Guinaudeau and Strube, 2013):

$$AvgOutDegree(P) = \frac{\sum_{s \in S} OutDegree(s)}{\|S\|}$$

where $OutDegree(s)$ is the sum of the weights associated with edges that leave node s and $\|S\|$ is the number of sentences in the text.

Number of components. The projection graph can be disconnected. A graph is *disconnected* if there are at least two nodes which are not reachable from each other (like s_1 and s_2 in Figure 2). A maximal non-empty connected subgraph in a graph is called *component*. Each projection graph in Figure 2 contains two components. Intuitively, projection graphs of a more coherent text should contain fewer number of components. The outdegree does not capture this type of connectivity. E.g., in Figure 5 the average outdegree of the two graphs is equal, while the left graph contains more components and should be less coherent.

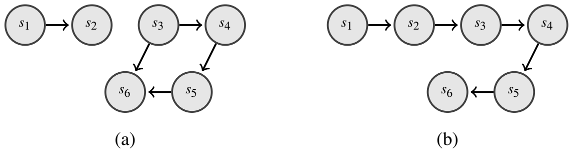

Figure 5: Two graphs with the same outdegree value. Graph (a) has two components. It is less coherent.

Frequent subgraphs. We hypothesize that particular coherence patterns show a correlation with readability. These patterns are encoded as subgraphs in graphs. An advantage is that coherence can be measured beyond simple sentence or node connectivity. We first define the graph concepts employed.

Isomorphic. Two graphs G and G' are *isomorphic*, if they fulfill two conditions: there should be a one-to-one association between nodes of G' and those of G, and two nodes of G' should be connected, if and only if their associated nodes in G are connected.

Subgraph. Graph G' is a *subgraph* of graph G, if G' is isomorphic to a graph whose nodes and edges are in G.

k-node subgraph. A subgraph with k nodes is called *k-node subgraph*.

Induced subgraph. The graph G' is an *induced subgraph* of graph G, if G' is a subgraph of G whose nodes are connected by all edges which connect the corresponding nodes in G (Figure 6). We always mean induced subgraphs when using the term subgraph.

Frequent subgraph & minimum support. Let $\zeta = \{G_1, G_2, \cdots, G_n\}$ be a database of n graphs. For each subgraph sg, $support(sg)$ denotes the number of graphs (in ζ) which contain sg as a subgraph. A subgraph sg is a *frequent subgraph* if and only if $support(sg) > \lambda$, where λ is called *minimum support*.

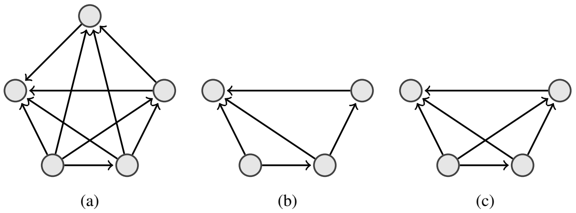

Figure 6: Both graphs (b) and (c) are subgraphs of (a). Only (c) is an induced subgraph of (a).

Graph signature. Given a set of frequent subgraphs $\{sg_1, sg_2, ..., sg_m\}$, a graph signature for $G \in \zeta$ is the vector $\Phi(G) = (\varphi(sg_1, G), \varphi(sg_2, G), ..., \varphi(sg_m, G))$, where

$$\varphi(sg_i, G) = \frac{count(sg_i, G)}{\sum_{sg_j \in (sg_1, sg_2, ..., sg_m)} count(sg_j, G)}$$

Here $count(sg_i, G)$ is the number of occurrences of sg_i in graph G. We use the relative frequency $\varphi(sg_i, G)$ because it compares graphs with different numbers of nodes and different numbers of edges.

Subgraph features are divided into two categories: basic subgraphs and frequent large subgraphs.

Basic subgraphs. Instead of frequent subgraphs all possible 3-node subgraphs (Figure 7) are used as basic subgraphs because they are the smallest meaningful subgraphs that can model coherence patterns.

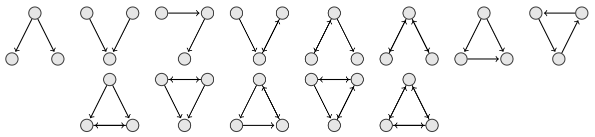

Figure 7: All possible directed 3-node subgraphs.

Because backward edges never occur in one-mode projections, only four subgraphs are feasible (Figure 8).

We interpret these subgraphs as follows:

- sg_1: The connection between a sentence and subsequent ones. In other words, at least two entities are mentioned in one sentence and the subsequent ones are about these entities.

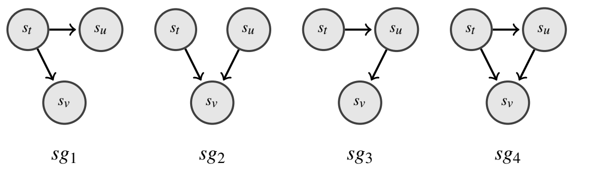

Figure 8: Feasible 3-node subgraph coherence features. Node labels illustrate the order of sentences. Sentence s_t occurs before sentence s_u, and sentence s_u occurs before sentence s_v (i.e. $t < u < v$).

- sg_2: Indicates that entities in s_t and s_u get connected to each other in s_v.
- sg_3: Each sentence tends to refer to the most prominent entity (focus of attention) in preceding sentences (Sidner, 1983; Grosz et al., 1995). The absence of a connection between s_t and s_v indicates that the entity connecting s_t and s_u is different from the entity connecting s_u and s_v. Therefore this subgraph approximately corresponds to the shift of the focus of attention.
- sg_4: Merges sg_1 and sg_3 and represents all connections of these two subgraphs.

We use these feasible 3-node subgraphs and compute the graph signature, Φ, of each $G \in \zeta$. We propose each $\varphi \in \Phi$ (i.e. relative frequency of each subgraph in G) as a connectivity feature of graph G to measure text coherence.

Frequent large subgraphs. Since we observe a strong correlation between basic subgraphs and human readability ratings (Table 4), we mine frequent large subgraphs of projection graphs. Our intuition is that larger subgraphs are more informative coherence patterns. Hence, we extend the coherence features from all feasible 3-node subgraphs to frequent k-node subgraphs. We first use an efficient subgraph mining algorithm to extract all subgraphs with size k and then compute the count of each subgraph as an induced subgraph in each graph $G \in \zeta$. We retain a subgraph sg, if it is frequent (i.e. $support(sg) > \lambda$). The result of these steps is a two-dimensional matrix whose rows represent graphs in ζ and columns represent frequent subgraphs with size k. The cell $\langle G_i, sg_j \rangle$ shows the count of sg_j in graph G_i. Given this matrix, we compute the graph signature of each $G \in \zeta$ and take each element of the graph signature as a coherence feature.

4 Experiments

4.1 Data

We use the dataset created by Pitler and Nenkova (2008) which consists of randomly selected articles from the Wall Street Journal corpus. The articles were rated by three humans on a scale from 1 to 5 for readability based on quality measures that are designed to estimate the coherence of articles. The final readability score of each article is the average of these three ratings.

We exclude three files from this dataset: `wsj--0382` does not exist in the Penn Treebank (Marcus et al., 1994)[1]. `wsj-2090` does not exist in the Penn Discource Treebank (Prasad et al., 2008). `wsj-1398` is a poem.

4.2 Settings

Entity graph. We use the gold parse trees in the Penn Treebank (Marcus et al., 1994) to extract all nouns in a document as mentions. We consider nouns with identical stem[2] as coreferent. We divide the edge weight between two sentence nodes s_i and s_j by their distance $j - i$ to decrease the importance of links that exist between non-adjacent sentences.

Discourse relation graph. We use gold PDTB-style discourse relations (Prasad et al., 2008). We filter out EntRel and NoRel relations.

Number of components. For counting the number of components in each projection graph, the SageMath[3] package is used. This feature is computed on unweighted projections (i.e. P_u^{ER}).

Frequent subgraphs. Since subgraph mining is an NP-complete problem, different algorithms have been introduced to improve the performance of subgraph mining. We use the gSpan[4] algorithm (Yan and Han, 2002) to mine subgraphs of a graph database which contains P_u^{ER} projections. An advantage of using efficient subgraph mining algorithms is that we can exhaustively search very large subgraph spaces. A graph with $\|E\|$ edges, however, potentially has $\mathcal{O}(2^{\|E\|})$ subgraphs. Having sparse graphs

[1] Pitler and Nenkova (2008) also remove one file from their experiments. We assume that it is `wsj-0382`.

[2] We use Stanford CoreNLP (`http://nlp.stanford.edu/software/corenlp.shtml`)

[3] `http://sagemath.org/download-linux.html`

[4] We use the Java package: `http://www.cs.ucsb.edu/~xyan/software/gSpan.htm`

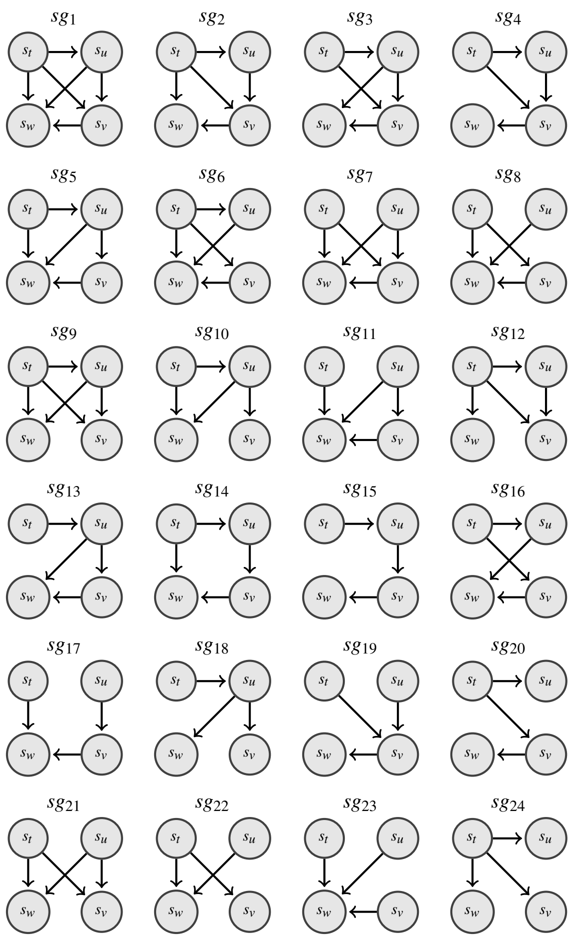

Figure 9: Frequent subgraphs with four nodes where $t < u < v < w$.

and using efficient subgraph mining algorithm lets us to search trough this space. We mine subgraphs with $k = 4$ and $\lambda = 0$ (Figure 9).

4.3 Evaluation

We evaluate on the following benchmark tasks.

Readability assessment. We use the Pearson correlation coefficient to find features correlated with readability scores. It takes feature values and readability scores of all articles and returns $-1 \leq \rho \leq +1$. A high value of $|\rho|$ shows a strong correlation. We report statistical significance on the 0.05-level[5].

Readability as ranking. We rank texts pairwise with respect to their readability. We define a classification problem with a set of text pairs and a label, which indicates whether the first text in a pair is more readable. We use every two texts whose human readability scores differ by at least 0.5. Each text is represented with its graph-based coherence features. We employ WEKA's linear support vector implementation (SMO) to classify the pairs. Performance is evaluated using 10-fold cross-validation.

	ρ	p_value
Entity Graph		
P_u^{ER}	-0.013	0.949
P_w^{ER}	0.151	0.452
P_{acc}^{ER}	0.150	0.455
Discourse Relation Graph		
P_u^{DR}	0.150	0.455
P_w^{DR}	0.155	0.440
Combination of Entity and Discourse Relation		
$P_u^{ER} \vee P_u^{DR}$	0.083	0.681
$P_w^{ER} + P_u^{DR}$	0.185	0.356
$P_w^{ER} + P_w^{DR}$	0.187	0.350

Table 3: The correlation of the average outdegree of different graphs with human readability ratings.

5 Results

Readability assessment. We report the correlation of our coherence models encoded in graph features and compare them with Guinaudeau and Strube's (2013) entity graph as the state-of-the-art coherence model. Pitler and Nenkova (2008) show that the entity transition features extracted from the entity grid model (Barzilay and Lapata, 2008) on its own do not significantly predict human readability ratings. So we do not describe their results here.

The results for the outdegree feature is shown in Table 3. The average outdegree of P_w^{ER} is highly correlated with human readability ratings. This confirms the readability results of Guinaudeau and Strube (2013) on the Encyclopedia Britannica dataset. The outdegrees of discourse relation graphs are more strongly correlated with human readability ratings than the outdegree of the projections in the entity graph, suggesting that efficient graph-based encoding of discourse relations can measure readability well. The outdegree of the combined graph $P_w^{ER} + P_w^{DR}$ is highly correlated, showing that the interaction of entity connections and discourse relations is important for text coherence. However, none of the outdegree measures in this table are significantly correlated with human readability rat-

[5] The results written in bold face (Section 5).

ings, confirming the intuition that outdegree only measures node connectivity in graphs and it is not enough to measure readability.

	ρ	p_value
Number of Components	**−0.391**	**0.044**
Relative frequency of 3-node Subgraphs		
sg_1	0.310	0.116
sg_2	−0.325	0.098
$\mathbf{sg_3}$	**−0.384**	**0.048**
sg_4	0.108	0.592

Table 4: Number of components and subgraph sg_3 are significantly correlated to readability.

Table 4 shows the correlation of two features of projections[6]: The number of components has a strong and significant negative correlation with human readability ratings[7], suggesting that simple properties of graphs measure text coherence. The lower part of Table 4 shows the correlation of the relative frequency of 3-node subgraphs (see Figure 8). More readable articles have many sg_1 and few number of sg_2 patterns. Pattern sg_3 is significantly and negatively correlated with human readability judgments, confirming the intuition that many shifts in focus of attention make texts difficult to read.

Table 5 shows the correlation between the relative frequency of 4-node subgraphs and readability ratings. First, most subgraphs with less than four edges are negatively correlated with readability, except sg_{20} and sg_{24} which are weakly correlated with readability. Few connections between sentences make the text difficult to read.

	number of edges	ρ	p_value
sg_1	6	0.103	0.609
sg_2	5	−0.212	0.288
sg_3	5	−0.176	0.380
sg_4	4	−0.257	0.196
sg_5	5	−0.140	0. 486
sg_6	5	0.200	0.317
$\mathbf{sg_7}$	**5**	**−0.402**	**0.038**
sg_8	4	−0.317	0.107
sg_9	5	0.153	0.446
sg_{10}	4	−0.238	0.232
$\mathbf{sg_{11}}$	**4**	**−0.509**	**0.007**
$\mathbf{sg_{12}}$	**4**	**0.449**	**0.019**
sg_{13}	4	−0.045	0.824
sg_{14}	4	−0.033	0.870
sg_{15}	3	−0.358	0.067
sg_{16}	4	−0.068	0.736
sg_{17}	3	−0.308	0.118
$\mathbf{sg_{18}}$	**3**	**−0.546**	**0.003**
$\mathbf{sg_{19}}$	**3**	**−0.601**	**0.001**
sg_{20}	3	0.094	0.641
sg_{21}	4	0.068	0.736
sg_{22}	3	−0.374	0.055
sg_{23}	3	−0.314	0.111
sg_{24}	3	0.100	0.620

Table 5: The correlation between the relative frequency of 4-node subgraphs and readability ratings.

Second, the highest positive and significant correlation of sg_{12} and the most negatively correlated subgraph sg_{11} show that different patterns of edges in subgraphs capture readability judgments. Stoddard (1991, p.29) explains this by the *ambiguity node* phenomenon: “[...] in some cases, there may be more than one logical, possible node for a given cohesive element in a text, in which case, a reader may see the resulting ambiguity but not be able to decide between the choices”. E.g., in sg_{11} a reader may make a decision about the focus of attention in s_w, while in sg_{12} the focus of attention of s_w is the same as the focus of attention of s_t. This phenomenon can also be observed in all positively correlated subgraphs. If readers have to return to one point in the text, they prefer to return to a sentence which is the core of the preceding sentences. However, we should refrain of interpreting too much into these patterns.

Finally, we conclude that in all strongly negative correlated subgraphs, a subgraph suffers either from edge shortage or the *ambiguity node* phenomenon like sg_7.

Considering the correlation of 3-node subgraphs in Table 4 and 4-node subgraphs in Table 5, two results are noticeable. First, in large subgraphs there are more strongly correlated subgraphs than 3-node subgraphs, confirming our hypothesis that larger subgraphs convey coherence patterns with higher quality. Second, sg_{12} in 4-node subgraphs is more strongly and positively correlated than sg_4 in

[6] Although, the proposed features can be applied on all kind of presented graphs, we evaluate them (except outdegree) only on projections of the entity graph model. We leave the application to the other graph representations for future work.

[7] This supports Karamanis et al. (2009) who report that NOCB transitions in the centering model can be used for the sentence ordering task.

3-node subgraphs, because sg_{12} captures more circumstances about s_t. The relative frequency of sg_{12} is more informative than sg_4's relative frequency.

Readability as ranking. Results of the readability ranking problem are shown in Table 6. Baseline features are entity transition features which are used as coherence features by Pitler and Nenkova (2008)[8].

Features	Accuracy
Baselines	
None (Majority class)	47.85%
Baseline features (Pitler and Nenkova, 2008)	83.25%
Graph-based Features	
Number of components	61.72%
Basic subgraphs (3-node)	79.43%
Frequent large subgraphs (4-node)	89.00%
Frequent basic + large subgraphs	88.52%
Baseline features + frequent large subgraphs	93.30%

Table 6: SVM prediction accuracy.

When classifying with graph signatures based on basic subgraphs, accuracy is lower than with the baseline coherence features. This is probably related to the entity grid features which represent grammatical role transitions of entities, while the basic subgraphs only models the occurrence of entities across sentences. Graph signatures based on large subgraphs improve the performance of basic subgraphs by around 10%. This high accuracy verifies that larger subgraphs capture coherence patterns with high quality. Combining basic (3-node) and large subgraphs (4-node) cannot improve the performance of the large subgraphs features. This probably is because basic subgraphs are implicitly included in larger subgraphs. The combination of coherence baseline features and frequent large subgraphs improves the accuracy.

6 Related Work

There is a research tradition developing metrics for readability and using these metrics to quantify how difficult it is to understand a document. Shallow features such as word, sentence and text length, which only capture superficial properties of a text, have been used traditionally (Flesch, 1948; Kincaid et al., 1975). De Clercq et al. (2014) use traditional shallow features and apply these to a new corpus annotated with two different methodologies. However, some studies indicate that shallow features do not precisely predict the readability of a text (Feng et al., 2009; Petersen and Ostendorf, 2009). Later studies introduce deeper (more semantic) features such as those obtained by language models (Si and Callan, 2001; Collins-Thompson and Callan, 2004) and syntactic features like the number of NPs in sentences or the height of the sentence's parse tree (Schwarm and Ostendorf, 2005; Heilman et al., 2007). Barzilay and Lapata (2008) propose an entity-based coherence model which operationalizes some of the intuitions behind the centering model (Grosz et al., 1995). Although this model works well on the sentence ordering and summary coherence rating tasks, it does not work well for readability assessment. Only when combining the entity grid with features taken from Schwarm and Ostendorf (2005) the entity grid performs competitively.

While most of these studies predict the readability level of documents, Pitler and Nenkova (2008) present a new readability dataset with *Wall Street Journal* articles, where each article is assigned human readability ratings. They analyze the correlation between different readability features and human readability scores. They show no correlation between entity-transition features and readability scores. In contrast to them we are able to report a statistically significant correlation between some entity-based features and human readability ratings.

7 Conclusions

We proposed graph-based coherence features based on the notion of frequent subgraphs. We analyzed these features on the dataset created by Pitler and Nenkova (2008) which associates human readability ratings with each document. We have shown that frequent subgraphs represent coherence patterns in a text. Larger subgraphs obtain a high and statistically significant correlation with human readability ratings.

Pitler and Nenkova (2008) did not achieve statistically significant (positive or negative) correlations between their features derived from the entity grid and human readability ratings. In contrast, some of

[8] The accuracy reported in their paper is 79.42%. Our reimplementation achieves higher accuracy, because our dataset has three articles less.

our automatically induced subgraphs have a strong statistically significant correlation. We also outperform Pitler and Nenkova (2008) in the readability ranking task by more than 5% accuracy thus establishing a new state-of-the-art on this dataset. We conclude that the graph-based representation (Guinaudeau and Strube, 2013) is a better and more informative starting point for assessing readability.

In future work, we plan to induce common subgraphs and apply our method to different datasets (e.g. the dataset created by De Clercq et al. (2014)) combined with other readability features (Schwarm and Ostendorf, 2005).

Acknowledgments

This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The first author has been supported by a HITS Ph.D. scholarship. We would like to thank our colleagues Benjamin Heinzerling, Yufang Hou, Sebastian Martschat, Nafise Moosavi, and Daraksha Parveen who commented on earlier drafts of this paper.

References

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

Chris Biemann, Stefanie Roos, and Karsten Weihe. 2012. Quantifying semantics using complex network analysis. In *Proceedings of the 24th International Conference on Computational Linguistics,* Mumbai, India, 8–15 December 2012, pages 263–278.

Kevyn Collins-Thompson and James P. Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics,* Boston, Mass., 2–7 May 2004, pages 193–200.

František Daneš. 1974. Functional sentence perspective and the organization of the text. In F. Daneš, editor, *Papers on Functional Sentence Perspective*, pages 106–128. Prague: Academia.

Orphée De Clercq, Véronique Hoste, Bart Desmet, Philip Van Oosten, Martine De Cock, and Lieve Macken. 2014. Using the crowd for readability prediction. *Natural Language Engineering*, 20(3):293–325.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics,* Athens, Greece, 30 March – 3 April 2009, pages 229–237.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychlogy*, 32:221–233.

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.

Camille Guinaudeau and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Sofia, Bulgaria, 4–9 August 2013, pages 93–103.

Michael J. Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. 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 460–467.

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

Rohit Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond Mooney, Salim Roukos, and Chris Welty. 2010. Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd International Conference on Computational Linguistics,* Beijing, China, 23–27 August 2010, pages 546–554.

J. Peter Kincaid, Robert P. Jr. Fishburne, Richard L. Rogers, and Brad S. Chisson. 1975. Derivation of new readability formulas (automated readability index, Fog count and Flesch reading ease formula) for navy enlisted personnel. Technical Report 8-75, Naval Technical Training Command, Naval Air Station Memphis-Millington, Tenn., February.

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 (Volume 1: Long Papers),* Portland, Oreg., 19–24 June 2011, pages 997–1006.

Ziheng Lin. 2011. *Discourse parsing: Inferring discourse structure, modeling coherence, and its applications*. Ph.D. thesis, Dept. of Computer Science, School of Computing, National University of Singapore.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen

Katz, and Britta Schasberger. 1994. The Penn treebank: Annotating predicate argument structure. In *Proceedings of ARPA Speech and Natural Language Workshop*.

Sebastian Nowozin, Koji Tsuda, Takeaki Uno, Taku Kudo, and Gokhan BakIr. 2007. Weighted substructure mining for image analysis. In *Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* Minneapolis, Minn., 18-23 June 2007, pages 1–8.

Sarah E. Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer Speech and Language*, 23(1):89–106.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processsing,* Waikiki, Honolulu, Hawaii, 25–27 October 2008, pages 186–195.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation,* Marrakech, Morocco, 26 May – 1 June 2008.

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.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the ACM 10th Conference on Information and Knowledge Management,* Atlanta, Georgia, 5–10 November 2001, pages 574–576.

Candace L. Sidner. 1983. Focusing in the comprehension of definite anaphora. In M. Brady and R.C. Berwick, editors, *Computational Models of Discourse*, pages 267–330. Cambridge, Mass.: MIT Press. Reprinted in: Grosz, Barbara J. et al. (Eds.) (1986). Readings in Natural Language Processing. Morgan Kaufman: Los Altos, Cal., pp.363-394.

Sally Stoddard. 1991. *Text and Texture: Patterns of Cohesion*. Ablex, Norwood, N.J.

Xifeng Yan and Jiawei Han. 2002. gSpan: Graph-based substructure pattern mining. In *Proceedings of the International Conference on Data Mining,* Maebashi City, Japan, 9–12 December 2002, pages 721–724.