On Automated Evaluation of Readability of Summaries: Capturing Grammaticality, Focus, Structure and Coherence

Ravikiran Vadlapudi
Language Technologies Research Center
IIIT Hyderabad
ravikiranv@research.iiit.ac.in

Rahul Katragadda
Language Technologies Research Center
IIIT Hyderabad
rahul_k@research.iiit.ac.in

Abstract

Readability of a summary is usually graded manually on five aspects of readability: *grammaticality*, *coherence and structure*, *focus*, *referential clarity* and *non-redundancy*. In the context of automated metrics for evaluation of summary quality, content evaluations have been presented through the last decade and continue to evolve, however a careful examination of readability aspects of summary quality has not been as exhaustive. In this paper we explore alternative evaluation metrics for '*grammaticality*' and '*coherence and structure*' that are able to strongly correlate with manual ratings. Our results establish that our methods are able to perform pair-wise ranking of summaries based on grammaticality, as strongly as ROUGE is able to distinguish for content evaluations. We observed that none of the five aspects of readability are independent of each other, and hence by addressing the individual criterion of evaluation we aim to achieve automated appreciation of readability of summaries.

1 Introduction

Automated text summarization deals with both the problem of identifying relevant snippets of information and presenting it in a pertinent format. Automated evaluation is crucial to automatic text summarization to be used both to rank multiple participant systems in shared tasks[1], and to developers whose goal is to improve the summarization systems. Summarization evaluations help in the creation of reusable resources and infrastructure; it sets up the stage for comparison and replication of results by introducing an element of competition to produce better results (Hirschman and Mani, 2001).

Readability or Fluency of a summary is categorically measured based on a set of linguistic quality questions that manual assessors answer for each summary. The linguistic quality markers are: *grammaticality*, *Non-Redundancy*, *Referential Clarity*, *Focus* and *Structure and Coherence*. Hence *readability assessment* is a manual method where expert assessors give a rating for each summary on the Likert Scale for each of the linguistic quality markers. Manual evaluation being time-consuming and expensive doesn't help system developers — who appreciate fast, reliable and most importantly *automated* evaluation metric. So despite having a sound manual evaluation methodology for readability, there is an need for reliable automatic metrics.

All the early approaches like Flesch Reading Ease (Flesch, 1948) were developed for general texts and none of these techniques have tried to characterize themselves as approximations to grammaticality or structure or coherence. In assessing readability of summaries, there hasn't been much of dedicated analysis with text summaries, except in (Barzilay and Lapata, 2005) where local coherence was modeled for text summaries and in (Vadlapudi and Katragadda, 2010) where grammaticality of text summaries were explored. In a marginally related work in Natural Language Generation, (Mutton et al., 2007) addresses sentence level fluency regardless of content, while recent work in (Chae and Nenkova, 2009) gives a systematic study on how syntactic features were able to distinguish machine generated translations from human translations. In another related work, (Pitler and Nenkova, 2008) investigated the impact of certain *surface linguistic features*, *syntactic*, *entity coherence* and *discourse* features on the readability of Wall Street Journal (WSJ) Corpus. We use the *syntactic* features used in (Pitler and Nenkova, 2008) as baselines for our experiments on grammaticality in this paper.

[1] The summarization tracks at Text Analysis Conference (TAC) 2009, 2008 and its predecessors at Document Understanding Conferences (DUC).

Proceedings of the NAACL HLT 2010 Student Research Workshop, pages 7–12,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

While studying the coherence patterns in student essays, (Higgins et al., 2004) identified that grammatical errors affect the overall expressive quality of the essays. In this paper, due to the lack of an appropriate baseline and due to the interesting-ness of the above observation we use metrics for grammaticality as a baseline measure for *structure and coherence*. Focus of a summary, is the only aspect of readability that relies to a larger extent on the content of the summary. In this paper, we use Recall Oriented Understudy of Gisting Evaluation (ROUGE) (Lin, 2004) based metrics as one of the baselines to capture *focus* in a summary.

2 Summary Grammaticality

Grammaticality of summaries, in this paper, is defined based on the grammaticality of its sentences, since it is more a sentence level syntactic property. A sentence can either be grammatically correct or grammatically incorrect. The problem of grammatical incorrectness should not occur in summaries being evaluated because they are generated mostly by extract based summarization systems.

But as the distribution of grammaticality scores in Table 1 shows, there are a lot of summaries that obtain very low scores. Hence, We model the problem of grammaticality as "*how suitable or acceptable are the sentence structures to be a part of a summary?*".

The acceptance or non acceptance of sentence structures varies across reviewers because of various factors like usage, style and dialects. Hence, we define a degree to which a sentence structure is acceptable to the reviewers, this is called the *degree of acceptance* throughout this paper.

Grammaticality Score	1	2	3	4	5
Percentage Distribution (in %)	10	13	15	37	25

Table 1: Percentage distribution of grammaticality scores in system summaries

In this paper, the *degree of acceptance* of sentence structures is estimated using language models trained on a corpus of human written summaries. Considering the sentence structures in reference summaries as the best accepted ones (with highest *degree of acceptance*), we estimate the *degree of acceptance* of sentences in system generated summaries by quantifying the amount of similarity/digression from the references using the language models.

The structure of the sentences can be represented by sequences of parts-of-speech (POS) tags and chunk tags. Our previous observations (Vadlapudi and Katragadda, 2010) show that the tagset size plays an important role in determining the *degree of acceptance*. In this paper, we combine the two features of a sentence — the POS-tag sequence and chunk-tag sequence — to generate the POS-Chunk-tag training corpus.

Some aspects of grammatical structure are well identifiable at the level of POS tags, while some other aspects (such as distinguishing between appositives and lists for eg.) need the power of chunk tags, the combination of these two tag-sequences provides the power of both.

Hence, the following approaches use probabilistic models, learned on POS tag corpus and POS-Chunk tag corpus, in 3 different ways to determine the grammaticality of a sentence.

2.1 Enhanced Ngram model

As described in our previous work, the Ngram model estimates the probability of a sentence to be grammatically acceptable with respect to the corpus using language models. Sentences constructed using frequent grammar rules would have higher probability and are said to have a well accepted sentence structure. The grammaticality of a summary is computed as

$$G(Sum) = AVG(P(Seq_i))\ ;\ P(Seq_i) = \log(\sqrt[n]{\prod_{j=1}^{n} P(K_j)})$$

$$P(K_j) = P(t_{j-2}t_{j-1}t_j)$$

$$P(t_1t_2t_3) = \lambda_1 * P(t_3|t_1t_2) + \lambda_2 * P(t_3|t_2) + \lambda_3 * P(t_3)$$

where $G(Sum)$ is grammaticality score of a summary Sum and $G(S_i)$ is grammaticality of sentence S_i which is estimated by the probability ($P(Seq_i)$) of its POS-tag sequence (Seq_i). $P(K_j)$ is probability of POS-tag trigram K_j which is $t_{j-2}t_{j-1}t_j$ and $\forall t_j,\ t_j \in POS\ tags$. The additional tags t_{-1}, t_0 and t_{n+1} are the beginning-of-sequence and end-of-sequence markers. The average AVG of the grammaticality scores of sentences $P(Seq_i)$ in a summary gives the final grammaticality score of the summary. In the prior work, arithmetic mean was used as the averaging technique, which performs consistently well. However, here two other averaging techniques namely *geometric mean* and *harmonic mean* are experimented and based on our experiments, we found *geometric mean* performing better than the other two averaging techniques. All the results reported in this paper are based on *geometric mean*. The above procedure estimates grammaticality of sentence using its POS tags and we call this run '*Ngram (POS)*'. A similar procedure is followed to estimate grammaticality using its POS-Chunk tags (language models trained on POS-chunk-tag training corpus). The corresponding run is called '*Ngram (POS-Chunk)*' in the results.

2.2 Multi-Level Class model

In this model, we view the task of scoring grammaticality as a n-level classification problem. Grammaticality of summaries is manually scored on a scale of 1 to 5, which means the summaries are classified into 5 classes. We assume that each sentence of the summary is also rated on a similar scale which cumulatively decides to which class the summary must belong. In our approach, sentences are classified into 5 classes on the basis of frequencies of underlying grammar rules (trigram) by defining class boundaries on frequencies. Hence, the cumulative score of the rules estimate the score of grammaticality of a sentence and inturn the summary.

Similar to (Vadlapudi and Katragadda, 2010), trigrams are classified into 5 classes C_1, C_2, C_3, C_4 and C_5 and each class is assigned a score on a similar scale ($\forall_j score(C_j) = j$) and class boundaries are estimated using the frequencies of trigrams in the training corpus. The most frequent trigram, for example, would fall into class C_5. POS-Class sequences are generated from POS-tag sequences using class boundaries as shown in Figure 1. This is the first level of classification.

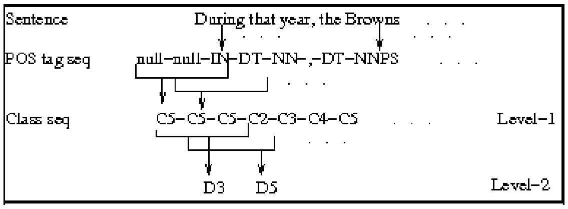
Sentence
During that year, the Browns . . .
POS tag seq
null–null–IN–DT–NN–,–DT–NNPS . . .
Class seq
C5–C5–C5–C2–C3–C4–C5 . . .
Level–1
D3
D5
Level–2

Figure 1: Two-level class model

Like the first level of classification, a series of classifications are performed upto 'k' levels. At each level we apply the scoring method described below to evaluate the grammaticality of summaries. We observed that from 3rd level onwards the structural dissimilarity disappears and the ability to distinguish different structures is lost. Hence, we report on the second level of classification, that captures the grammatical acceptability of summaries very well, and Figure 1 explains the two level classification.

$$G(S_i) = AVG(H(C_{w1}), H(C_{w2}), ..., H(C_{wn})) \quad (1)$$

AVG is the average of $H(C_{wi})$, where $w1$, $w2$, $\ldots wn$ are class trigrams, C_{wi} is the class into which class trigram wi falls into and $H(C_{wi})$ is score assigned to the class C_{wi}. The AVG is computed using geometric mean and this run is referred as '*Class (POS 2 level)*' in the results.

Similar to above approach, the grammaticality of a sentence can also be estimated using POS-Chunk tag sequence and POS-Chunk Class training data, and the corresponding run is referred as '*Class (POS-Chunk 2 level)*'.

2.3 Hybrid Model

As would be later seen in Table 2, the *Ngram (POS)* and *Class (POS 2 level)* runs are able to distinguish various systems based on grammaticality. We also note that these runs are able to very finely distinguish the degree of grammaticality at summary level. This is a very positive result, one that shows the applicability of applying these methods to any test summaries in this genre. To fully utilize these methods we combine the two methods by a linear combination of their scores to form a '*hybrid model*'. As seen with earlier approaches, both the POS-tag sequences and POS-Chunk-tag sequences could be used to estimate the grammaticality of a sentence, and hence the summary. These two runs are called '*Hybrid (POS)*' and '*Hybrid (POS-Chunk)*', respectively.

3 Structure and Coherence

Most automated systems generate summaries from multiple documents by extracting relevant sentences and concatenating them. For these summaries to be comprehensible they must also be *cohesive* and *coherent*, apart from being *content bearing* and *grammatical*. Lexical cohesion is a type of cohesion that arises from links between words in a text (Halliday and Hasan, 1976).A Lexical chain is a sequence of such related words spanning a unit of text. Lexical cohesion along with presuppositions and implications with world knowledge achieves coherence in texts. Hence, *coherence* is what makes text semantically meaningful, and in this paper, we also attempt to automate the evaluation of the "*structure and coherence*" of summaries.

We capture the structure or lexical cohesion of a summary by constructing a lexical chain that spans the summary. The relation between entities (noun phrases) in adjacent sentences could be of type center-reference (pronoun reference or reiteration), or based on semantic relatedness (Morris and Hirst, 1991). A center-reference relation exists if an entity in a sentence is a reference to center in adjacent sentence. Identifying centers of reference expressions can be done using a co-reference resolution tool. Performance of co-reference resolution tools in summaries, being evaluated, is not as good as their performance on generic texts. Semantic relatedness relation cannot be captured by using tools like Wordnet because they are not very exhaustive and hence are not effective. We use a much richer knowledge base to define this relation – Wikipedia.

Coherence of a summary is modelled by its structure and content together. Structure is captured by lexical chains which also give information about focus of each sentence which inturn contribute to the topic focus of the summary. Content presented in the summary must be semantically relevant to the topic focus of the summary. If the content presented by each sentence is semantically relevant to the focus of the sentence, then it would be semantically relevant to the topic focus of the summary. As the foci of sentences are closely related, a prerequisite for being a part of a lexical chain, the summary is said to be coherent. In this paper, the semantic relatedness of topic focus and content is captured using Wikipedia as elaborated in Section 3.1 of this paper.

3.1 Construction of lexical chains

In this approach, we identify the strongest lexical chain possible which would capture the structure of the summary. We define this problem of finding the strongest possible lexical chain as that of finding the best possible parts-of-speech tag sequence for a sentence using the Viterbi algorithm shown in (Brants, 2000). The entities (noun phrases) of each sentence are the nodes and transition probabilities are defined as relatedness score (Figure 2). The strongest lexical chain would have the highest score than other possible lexical chains obtained.

Consider sentence S_k with entity set (e_{11}, e_{12}, $e_{13}, \dots e_{1n}$) and sentence S_{k+1} with entity set (e_{21}, e_{22}, e_{23}, $\dots e_{2m}$). Sentences S_k and S_{k+1} are said to be strongly connected if there exists entities $e_{1i} \in S_k$ and $e_{2j} \in S_{k+1}$ that are closely related. e_{1i} and e_2j are considered closely related if

- e_{2j} is a pronoun reference of the center e_{1i}
- e_{2j} is a reiteration of e_{1i}
- e_{2j} and e_{1i} are semantically related

Pronoun reference In this approach, we resolve the reference automatically by finding more than one possible center for the reference expression using Wikipedia. Since the summaries are generated from news articles, we make a fair assumption that related articles are present in Wikipedia. We ensure that the correct center is one among the possible centers through which S_{k+1} and S_{k+2} might be strongly connected. Entities with *query hits ratio* $\geq \lambda$ are considered as possible centers and entity e_{2j} is replaced by entities that act as the possible centers. Since the chain with the identified correct center is likely to have the highest score, our final lexical chain would contain the correct center.

$$Query\ hit\ ratio = \frac{Query\ hits(e_{1i}\ and\ e_{2j})}{Query\ hits(e_{1i})}$$

Reiteration Generally, an entity with a determiner can be treated as reiteration expression but not vice versa. Therefore, we check whether e_{2j} is actually a reiteration expression or not, using query hits on Wikipedia. If *Query hits* (e_{2j}) $\geq \ \beta$ then we consider it to be a reiteration expression. A reiterating expression of a *named entity* is generally a common noun that occurs in many documents. After identifying a reiteration expression we estimate relatedness using semantic relatedness approach.

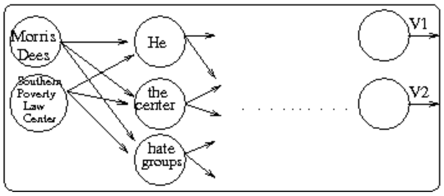

Figure 2: Viterbi trace for identifying lexical chain

Semantic relatedness By using *query hits* over Wikipedia we estimate the *semantic relatedness* of two entities. Such an approach has been previously attempted in (Strube and Ponzetto, 2006). Based on our experiments on grammaticality 2.2, classifying into 5 classes is better suited for evaluation tasks, hence we follow suit and classify *semantic relatedness* into 5 classes. These classes indicate how semantically related the entities are. Each class is assigned a value that is given to the hits which fall into the class. For example, if *query hits* lie in the range (γ_1, γ_2) or if query hit ratio is $\geq \xi$ then it falls into class k and is assigned a score equal to k.

Now that we have computed semantic connectedness between adjacent sentences using the methods explained above, we identify the output node with maximum score (node V2 in Figure 2). This node with best score is selected and by backtacking the Viterbi path we generate the lexical chain for the summary. The constants $\lambda, \gamma_1, \gamma_2and\xi$ are determined based on empirical tuning.

3.2 Coherence

We estimate coherence of the summary by estimating how the sentences stick together and the semantic relevance of their collocation. In a sentence, the semantic relatedness of entities with the focus estimates score for the meaningfulness of the sentence, and the average score of all the sentences estimates the coherence of the summary.

$$C(Summary) = \frac{\Sigma_{i=1}^{N} G(s_i)}{N}$$

$$G(s_i) = \frac{\Sigma_{j=1}^{k-1} H(Q(F\ and\ e_{ij}))}{k}$$

Where $C(Summary)$ is the coherence of summary $Summary$, and $G(s_i)$ is the semantic relatedness of a sentence s_i in $Summary$, while $Q(q)$ denotes the number of query hits of query q. F is the focus of s_i and e_{ij} is an entity in s_i, and $H(Q)$ is the score of class into which query falls.

4 Evaluation

This paper deals with methods that imitate manual evaluation metric for *grammaticality* and *structure and coherence* by producing a score for each summary. An evaluation of these new *summarization evaluation metrics* is based on how well the system rankings produced by them correlate with manual evaluations. We use 3 types of correlation evaluations — Spearman's Rank Correlation, Pearson's Correlation and Kendall's Tau — each describing some aspect of ordering problems.

We used reference summaries from TAC 2008, 2009 for the reference corpus and the experiments described were tested on DUC 2007 query-focused multi-document summarization datasets which have 45 topics and 32 system summaries for each topic apart from 4 human reference summaries.

Table 2 shows the system level correlations of our approaches to *grammaticality* assessment with that of human ratings. We have used four baseline approaches: AverageNPs, AverageVPs, AverageSBARs and AverageParseTreeHeight. Our approaches constitute of the following runs: *Ngram (POS)*, *Ngram (POS-Chunk)*, *Class (POS 2 level)*, *Class (POS-Chunk 2 level)*, *Hybrid (POS)*, *Hybrid (POS-Chunk)*.

RUN	**Spearman's ρ**	**Pearson's r**	**Kendall's τ**
	Baselines		
AverageNPs	0.1971	0.2378	0.1577
AverageSBARs	0.2923	0.4167	0.2138
AverageVPs	0.3118	0.3267	0.2225
ParseTreeHeight	0.2483	0.3759	0.1922
	Our experiments		
Ngram (POS)	0.7366	0.7411	0.5464
Ngram (POS+Chunk)	0.7247	0.6903	0.5421
Class (POS 2 level)	0.7168	0.7592	0.5464
Class (POS+Chunk 2 level)	0.7061	0.7409	0.5290
Hybrid (POS)	0.7273	**0.7845**	0.5205
Hybrid (POS+Chunk)	**0.7733**	0.7485	**0.5810**

Table 2: System level correlations of automated and manual metrics for grammaticality.

RUN	**Spearman's ρ**	**Pearson's r**	**Kendall's τ**
	Experiments		
Ngram (POS)	**0.4319**	**0.4171**	**0.3165**
Ngram (POS+Chunk)	**0.4132**	**0.4086**	**0.3124**
Class (POS 2 level)	0.3022	0.3036	0.2275
Class (POS+Chunk 2 level)	0.2698	0.2650	0.2015
Hybrid (POS)	**0.3652**	**0.3483**	**0.2747**
Hybrid (POS+Chunk)	0.3351	0.3083	0.2498

Table 3: Summary level correlations of automated and manual metrics for *grammaticality* .

RUN	**Spearman's ρ**	**Pearson's r**	**Kendall's τ**
	Baselines		
Human Grammaticality rating	0.5546	**0.6034**	0.4152
Ngram(POS)	0.3236	0.4765	0.2229
	Experiments		
Our coherence model	**0.7133**	0.5379	**0.5173**

Table 4: System level correlations of automated and manual metrics for *coherence* .

Table 4 shows the system level correlations of our approach to *structure and coherence* assessment with that of human ratings. As mentioned earlier in Section 1, human ratings for grammaticality and our

RUN	Spearman's ρ	Pearson's r	Kendall's τ
Baselines			
Human Grammaticality rating	0.5979	0.6463	0.4360
Human Coherence rating	0.9400	0.9108	0.8196
Ngram(POS)	0.4336	0.6578	0.3175
Our coherence model	0.5900	0.5331	0.4125
ROUGE-2	0.3574	0.4237	0.2681

Table 5: System level correlations of automated and manual metrics for *focus*

best performing system for grammaticality are used as baselines for *structure and coherence* assessment. Again, like we previously mentioned, *focus* can be easily characterized using structure and coherence, and to an extent the grammatical well-formedness. Also the *focus* of a summary is also dependent on content of the summary. Hence, we use ROUGE-2, manual rating for grammaticality, manual rating for coherence, and our approaches to both *grammaticality* and *structure and coherence* as baselines as shown in Table 5.

5 Discussion and Conclusion

In this paper, we addressed the problem of identifying the *degree of acceptance* of grammatical formations at sentence level using surface features like Ngrams probabilities (in Section 2.1), and trigrams based class Ngrams (in Section 2.2) and a hybrid model using both Ngram and Class model (in Section 2.3), on the POS-tag sequences and POS-chunk-tag sequences which have produced impressive results improving upon our previous work.

Our approaches have produced high correlations to human judgment on grammaticality. Results in Table 2 show that the Hybrid approach on the POS-Chunk tag sequences outperforms all the other approaches. Our approaches to grammaticality assessment have performed decently at pair-wise ranking of summaries, shown by correlations of the order of 0.4 for many runs. This correlation is of the same order as that of similar figure for content evaluations using ROUGE and Basic Elements.

Table 4 shows that our approach to the 'structure and coherence' assessment outperforms the baselines set and has an impressive correlation with manual ratings. From Table 5 we found that grammaticality is a good indicator of focus while we also observe that *structure and coherence* forms a strong alternative to *focus*.

The focus of this paper was on providing a complete picture on capturing the grammaticality aspects of readability of a summary using relatively shallow features as POS-tags and POS-Chunk-tags. We used lexical chains to capture *structure and coherence* of summaries, whose performance also correlated with *focus* of summaries. None of the five aspects of readability are completely independent of each other, and by addressing the individual criteria for evaluation we aim to achieve overall appreciation of readability of summary.

References

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *ACL*.

Thorsten Brants. 2000. Tnt: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*, pages 224–231, Morristown, NJ, USA. Association for Computational Linguistics.

Jieun Chae and Ani Nenkova. 2009. Predicting the fluency of text with shallow structural features: Case studies of machine translation and human-written text. In *EACL*, pages 139–147. The Association for Computer Linguistics.

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

M.A.K Halliday and Ruqayia Hasan. 1976. Longman publishers.

Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. 2004. Evaluating multiple aspects of coherence in student essays. In *HLT-NAACL 2004: Main Proceedings*, pages 185–192, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Lynette Hirschman and Inderjeet Mani. 2001. Evaluation.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *the proceedings of ACL Workshop on Text Summarization Branches Out*. ACL.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Comput. Linguist.*, 17(1):21–48.

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. 2007. Gleu: Automatic evaluation of sentence-level fluency. In *ACL*. The Association for Computer Linguistics.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *EMNLP*, pages 186–195. ACL.

Michael Strube and Simone Paolo Ponzetto. 2006. Wikirelate! computing semantic relatedness using wikipedia. In *21. AAAI / 18. IAAI 2006*. AAAI Press, july.

Ravikiran Vadlapudi and Rahul Katragadda. 2010. Quantitative evaluation of grammaticality of summaries. In *CICLing*.