Automatic Metrics for Genre-specific Text Quality

Annie Louis
University of Pennsylvania
Philadelphia, PA 19103, USA
lannie@seas.upenn.edu

Abstract

To date, researchers have proposed different ways to compute the readability and coherence of a text using a variety of lexical, syntax, entity and discourse properties. But these metrics have not been defined with special relevance to any particular genre but rather proposed as general indicators of writing quality. In this thesis, we propose and evaluate novel text quality metrics that utilize the unique properties of different genres. We focus on three genres: academic publications, news articles about science, and machine generated text, in particular the output from automatic text summarization systems.

1 Introduction

Automatic methods to measure the writing quality of a text can be quite useful for several applications, for example search and recommendation systems, and writing support and grading tools. There are two main categories of prior work on this topic. The first is studies on 'readability' which have proposed metrics to select texts appropriate (easy to read) for an audience of given age and education level (Flesch, 1948; Collins-Thompson and Callan, 2004). These metrics typically classify texts as suitable for adult or child, or into a more fine-grained set of 12 educational grade levels. The second line of work are recent computational metrics to predict coherence. These methods identify regularities in words (Barzilay and Lee, 2004), entity coreference (Barzilay and Lapata, 2008) and discourse relations (Pitler and Nenkova, 2008) from a large collection of articles and use these patterns to predict the coherence. They assume a particular competency level (adult educated readers) and also fix the text (typically news articles, which are appropriate for adult readers). By removing the focus on age/education level, these methods compute textual differences between good and poorly written texts as perceived by a single audience level.

In my thesis, I propose a new definition – text quality: the overall well-written characteristic of an article. It differs from prior work in three respects:

1. We consider a single fixed audience level and the texts that audience is typically exposed to. For example, a college educated reader of a newspaper might find some articles better written than others, even though he understands and can read nearly all of them with ease.

2. It is a holistic property of texts. At a minimum, at least four factors influence quality: the content/topic that is discussed, sentence level-grammaticality, discourse coherence and writing style. Here writing style refers to extra properties introduced into the text by the author but do not necessarily interfere with coherence if not provided. For example, the use of metaphors, examples and humour can have connections with quality. Previous work on coherence metrics do not consider these aspects.

3. Such a property would also have genre-specific dimensions: an academic article should above all be clear and a thriller-story should be fast-paced and interesting. Further even if the same

Proceedings of the NAACL HLT 2012 Student Research Workshop, pages 54–59,
Montréal, Canada, June 3-8, 2012. ©2012 Association for Computational Linguistics

quality aspect is relevant for multiple genres, it has higher weight in one versus another. Prior readability and coherence studies were not proposed with relvance to any particular genre.

These aspects make the investigation of text quality linguistically interesting because by definition the focus is on a wide range of properties of the text itself rather than appropriateness for a reader.

In this thesis, we propose computable measures to capture genre-specific text quality. Our hypothesis is that writing quality is a combination some generic aspects that matter for most texts, such as grammatical sentences, and other unique ones which have high impact in a particular genre.

Specifically, we consider three genres which have high relevance for writing quality research—academic writing, science journalism and output of automatic summarization systems.

Both academic writing and science news articles describe science, but their audience is quite different. Academic writing aims to clearly explain the details of the research to other experts, while science news conveys interesting research findings to lay readers. This fact creates distinctive content and writing style in the two genres. There is also a huge opportunity in these genres for developing applications involving text quality, for example, authoring tools for academic writing and information retrival and recommendation for news articles. We also include a third genre—automatically generated summaries. Here, when systems produce multi-sentence text, they must ensure that the text is readable and coherent. Automatic evaluation of content and linguistic quality is therefore necessary for system development in this genre.

2 Thesis Summary and Contributions

For this thesis, we only consider the discourse and style components of text quality, aspects that have received less focus in prior work. Sentence-level problems have been widely explored and recently, even specifically for academic writing (Dale and Kilgarriff, 2010). We also do not consider content in our work, for example, academic writing quality also depends on the ideas and arguments presented but these aspects are outside the scope of this thesis. As defined previously, we focus on a fixed audience level. We assume a reader at the top level of the competency spectrum: an adult educated reader for science news and automatic summaries, and for academic articles, an expert on the topic. This definition has minimal focus on reader abilities and allows us to analyze textual differences exclusively.

The specific contributions of this thesis are:

1. Defining text quality in terms of linguistic aspects rather than readability: Our work is the first to propose a quality definition where well-written nature is the central focus and including genre-dependent aspects and writing style.

2. Investigating genre-specific metrics: This study is also the first to design and evaluate genre-specific features for text quality prediction. For each genre: academic writing, science journalism and automatic summaries, we develop metrics unique to the genre and evaluate their ability to predict text quality both individually and in combination with generic features put forth in prior work.

3. Proposing new discourse-level features: In prior work, there are discourse-based features based on coreference, discourse relations and word co-occurrence between adjacent sentences. We introduce new features which capture aspects such as organization of communicative goals and general-specific nature of sentences.

Specifically, we introduce the following metrics:

a) Patterns in communicative goals (Section 5): Every text has a purpose and the author uses a sequence of communicative goals realized as sentences to convey that purpose. We introduce a metric that predicts coherence based on the size and sequence of communicative goals for a genre. This aspect is most relevant for research writing: academic and science journalism because there is a clear goal and well-defined purpose for these articles.

b) General-specific nature of sentences (Section 6): Some sentences in a text convey only general content, others provide details and a well-written text would have a certain balance between the two. Particularly, while creating summaries, there is a length contraint, so it cannot include all specific content but some information must be made more general. We introduce a method to predict the specificity for a sentence and examine how specificity and sequence of general-specific sentences is related to the quality of automatic summaries.

c) Information cohesiveness (Section 7): This idea is also proposed for automatic summaries, they must have a focus and present a small set of ideas with easy to understand links between them. We show that cohesiveness properties (computed automatically) of the source text to be summarized can be linked to the expected content quality of summaries that can be generated for that text. This work will be extended to analyze the relationship of cohesiveness with ratings of focus for the summaries.

d) Aspects of style (Section 8): Here we investigate metrics beyond coherence and related to extra features included in the article. We consider the genre of science journalism and investigate whether surprise-invoking sentence construction, visual descriptions and emotional content of the articles are also correlated with perceived quality.

We will evaluate our approaches in two ways:

1. We investigate the extent to which genre-specific metrics are indicative of text quality and whether they complement generic features.

2. We also examine how unique these metrics are for a given genre, for example: are surprising articles always considered well-written even if they are not science-news? For this analysis, we will consider a set of randomly selected news texts (no genre division) with text quality ratings. On this set, we will test the performance of generic and each set of genre-specific metrics. We expect that on this data, the generic features would be best with little improvement from the genre-specific metrics.

So far, we have designed some of the metrics that we described above and have found them to be predictive of writing quality. We will carry out extensive evaluation of these measures in future work.

3 Related work

Early readability metrics used sentence length, number of syllables in words and number of 'easy' words to distinguish texts from different grade levels (Flesch, 1948; Gunning, 1952; Dale and Chall, 1948). Other measures are based on word familiarity (Collins-Thompson and Callan, 2004; Si and Callan, 2001), difficulty of concepts (Zhao and Kan, 2010) and features of sentence syntax (Schwarm and Ostendorf, 2005). There are also readability studies for audience distinctions other than grade levels. Feng et al. (2009) consider adult readers with intellectual disability and therefore introduce features such as the number of entities a person should keep in working memory for that text and how far entity links stretch. Heilman et al. (2007) show that grammatical features make a bigger impact while predicting readability for second language learners in contrast to native speakers.

Newer coherence measures do not focus on reader abilities. They are typically run on news articles and assume an adult audience. They show that word co-occurrence (Soricut and Marcu, 2006), subtopic structure (Barzilay and Lee, 2004), discourse relations (Pitler and Nenkova, 2008; Lin et al., 2011) and coreference patterns (Barzilay and Lapata, 2008) learn from large corpora can be used to predict coherence.

But prior metrics are not proposed as unique to any genre. Some metrics using word patterns (Si and Callan, 2001; Barzilay and Lee, 2004) are domain-dependent in that they require documents from the target domain for training. But they can be trained for any domain in this manner.

However recent work show that genre-specific indicators could be quite useful for applications. McIntyre and Lapata (2009) automatically generate short children's stories using patterns of event and entity co-occurrences. They find that people judge their stories as better when the text is optimized not only for coherence and but also its interesting nature. They use a supervised approach to predict the interest value for a story during the generation process. Burstein et al. (2010) find that for predicting the coherence of student essays, better accuracies can be obtained by augmenting generic coherence metrics with features related to student writing such as word variety and spelling errors.

In my own work on automatic evaluation of summaries (Pitler et al., 2010), I have observed the impact of genre. We consider a corpus of summaries written by people and those produced by automatic systems. Psycholinguistic metrics previously proposed for analyzing coherence of human texts work successfully on human summaries but are less accurate for system summaries. Similarly, metrics which predict the fluency of machine translations accurately, work barely above baseline for judging the grammaticality of sentences from human summaries. But they give high accuracies on machine summary sentences. So for machine and human generated text, clearly different features matter.

4 Corpora for text quality

For the automatic summarization genre, several years of evaluation workshops organized by NIST[1] have created large-scale datasets of automatic summaries rated manually by people for content and linguistic quality. We utilize this data for our experiments but such corpora do not exist for other genres.

For academic writing, we plan to use a collection of biology journal articles marked with the impact factor of the journal. The intuition is that the popular journals are more competitive and so the writing is on average better than less impactful venues. It is however not a direct measure of text quality. For some of our experiments done so far, we have taken an approach that is common with prior studies on coherence (Barzilay and Lee, 2004; Barzilay and Lapata, 2008; Lin et al., 2011). We take an original article and create a random permutation of its sentences, the latter we consider as an incoherent article and the original version as coherent.

For science news, we expect that Amazon Mechanical Turk will be a suitable platform for obtaining ratings of popular and interesting articles from the target audience. We also plan to use proxies such as lists of most emailed/viewed articles from news websites. Here the negative examples would be other articles published during the *same* day/period but not appearing in the popular article list.

5 Patterns in communicative goals

Consider the related work section of a conference paper. One might suppose that a good structure for this section would contain a description of an attribute of the current work, followed by previous work on the topic and then reporting how the current work is different and addresses shortcomings if any of prior work. In fact, this intuition of seeing texts as a sequence of semantic zones is well-understood for the academic writing genre. Prior research has identified that a small set of argumentative zones exist in academic articles such as motivation, results, prior work, speculations and descriptions. They also found that sentences could be manually annotated into zones with high agreement and automatically predicting the zone for a sentence can also be done with high accuracy (Teufel and Moens, 2000; Liakata et al., 2010). We hypothesize that these zones would also have a certain distribution and sequence in well-written articles versus others and propose a metric based on this aspect for the academic writing and science journalism genres.

Rather than using a predefined set of communicative goals, we develop an unsupervised technique to identify analogs to semantic zones and use the patterns in zones to predict coherence (Louis and Nenkova, 2012a). Our key idea is that the syntax of a sentence can be a useful proxy for its communicative goal. For example, questions and definition sentences have unique syntax. We extend this idea to a large scale analysis. Our model represents a sentence either using productions from its constituency parse tree or as a sequence of phrasal nodes. Then we employ two methods that learn patterns in these representations from a collection of articles. The first local method detects patterns in the syntax of adjacent sentences. The second approach is global, where sentences are first grouped into clusters based on syntactic similarity and a Hidden Markov Model is used to record patterns. Each hidden state is assumed to generate the syntax of sentences from a particular zone.

We have evaluated our method on conference publications from the ACL anthology. Our results indicate that we can distinguish an original introduction, abstract or related work section from a corresponding perturbed version (where the sentences are randomly permuted and is therefore incoherent text) with accuracies of 64 to 74% over a 50% baseline.

6 General-specific nature of sentences

In any article, some sentences convey the topic at a high level with other sentences providing details such as justification and examples. The idea is particularly relevant for summaries. Since summaries are much shorter than their source documents, they cannot include all the details from the source. Some details have to be omitted and others made more general. So we explore the preferred degree of general-specific content and its relationship to text quality for summaries.

[1] http://www.nist.gov/tac/

We developed a classifier to distinguish between general and specific sentences from news articles (Louis and Nenkova, 2011a; Louis and Nenkova, 2012b). The classifier uses features such as the word specificity, presence of named entities, word polarity, counts of different phrase types, sentence length, likelihood under language models and the identities of the words themselves. For example, sentences with named entities tended to be specific whereas sentences with shorter verb phrases and more polarity words were general. This classifier was trained on sentences multiply annotated by people as general or specific and produces an accuracy of about 79%. Further the classifier confidence was found to be indicative of the annotator agreement on the sentences; when there was high agreement that a sentence was either general or specific, the classifier also made a very confident prediction for the correct class. So our system also provides a graded score for specificity rather than binary predictions.

Using the classifier we analyzed a large corpus of news summaries created by people and by automatic systems (Louis and Nenkova, 2011b). We found that summaries written by people have more general content than automatic summaries. Similarly, when people were asked to rate automatic summaries for content quality, they gave higher scores to general summaries than specific. On the linguistic quality side an opposite trend was found. Summaries that were more specific had higher scores. Our examinations revealed that general sentences, since they are topic oriented and high level, need to be followed by proper substantiation and details. But automatic systems are rather poor at achieving such ordering. So even though more general content is preferred in summaries, proper ordering of general-specific sentences is needed to create the right effect.

7 Information cohesiveness

If an article has too many ideas it would be difficult to read. Also if the ideas were not closely related in the article that would create additional difficulty. This aspect is important for machine generated text: an automatic summary should focus on a few main aspects rather than present a bag of many unrelated facts. In fact, in large scale evaluation workshops, automatic summaries are also manually graded for a ‘focus’ aspect. For this purpose, we want to identify metrics which can indicate cohesiveness and focus of an article. In our studies so far, we have have developed cohesiveness metrics for clusters of articles (Nenkova and Louis, 2008; Louis and Nenkova, 2009). In future work, we will explore how these metrics work for individual articles.

Information quality also arises in the context of source documents given for automatic summarization. Particularly for systems which summarize online news, the input is created by clustering together news on the same topic from different sources. For example, a cluster may be created for the Japanese earthquake and aftermath. When the period covered is too large or when the documents discuss many different opinions and ideas it becomes hard for a system to point out the most relevant facts. So one proxy for cohesiveness of the input cluster is the average quality of a number of automatic summaries produced for it by different methods. If most of these methods fail to produce a good summary, then that input can be deemed as difficult and incohesive.

We used a large collection of inputs, their automatic summaries and summary scores from the DUC workshops. We computed the average content quality score given by people to each summary and computed the average performance on summaries created for the same input. This value represents the expected system performance for that input and we develop features to predict the same. We simplify the task as binary prediction, average system performance above mean value – low difficulty, and high difficulty otherwise.

One indicative feature was the entropy of the distribution of words in the input. When the entropy was low, the difficulty was less since there are few main ideas to summarize. Another useful feature was the divergence computed between the word distribution in an input and that of a random collection of documents not on any topic. If the input distribution was closer to random documents it indicates the lack of a coherent topic for the source cluster and such inputs were under the hard category. We envision that similar features might help to predict judgements of focus for automatic summaries.

8 Current and future work

For future work, we want to focus on metrics related to style of writing. We will do this analysis for science news articles since journalists employ creative ways to convey technical research content to non-experts readers. For example, authors use analogies and visual language and incorporate a story line. We also noticed that some of the most emailed articles are entertaining and even contain humor. Two example snippets from such articles are provided below to demonstrate some of our intuitions about text quality in this genre. Our aim is to obtain lexical and syntactic correlates that capture some of these unique factors for this domain.

[1]... caused by defects in the cilia—solitary slivers that poke out of almost every cell in the body. They are not the wisps that wave Rockette-like in our airways.

[2] News flash: we're boring. New research that makes creative use of sensitive location-tracking data from cellphones in Europe suggests that most people can be found in one of just a few locations at any time.

Future work will also include extensive evaluation of our proposed models.

References

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

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

J. Burstein, J. Tetreault, and S. Andreyev. 2010. Using entity-based features to model coherence in student essays. In *Proceedings of HLT-NAACL*, pages 681–684.

K. Collins-Thompson and J. Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of HLT-NAACL*, pages 193–200.

E. Dale and J. S. Chall. 1948. A formula for predicting readability. *Edu. Research Bulletin*, 27(1):11–28.

R. Dale and A. Kilgarriff. 2010. Helping our own: text massaging for computational linguistics as a new shared task. In *Proceedings of INLG*, pages 263–267.

L. Feng, N. Elhadad, and M. Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of EACL*, pages 229–237.

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

R. Gunning. 1952. *The technique of clear writing*. McGraw-Hill; Fouth Printing edition.

M. Heilman, K. Collins-Thompson, J. Callan, and M. Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of HLT-NAACL*, pages 460–467.

M. Liakata, S. Teufel, A. Siddharthan, and C. Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of LREC*.

Z. Lin, H. Ng, and M. Kan. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of ACL-HLT*, pages 997–1006.

A. Louis and A. Nenkova. 2009. Performance confidence estimation for automatic summarization. In *Proceedings of EACL*, pages 541–548.

A. Louis and A. Nenkova. 2011a. Automatic identification of general and specific sentences by leveraging discourse annotations. In *Proceedings of IJCNLP*, pages 605–613.

A. Louis and A. Nenkova. 2011b. Text specificity and impact on quality of news summaries. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 34–42.

A. Louis and A. Nenkova. 2012a. A coherence model based on syntactic patterns. *Technical Report, University of Pennsylvania*.

A. Louis and A. Nenkova. 2012b. A corpus of general and specific sentences from news. In *Proceedings of LREC*.

N. McIntyre and M. Lapata. 2009. Learning to tell tales: A data-driven approach to story generation. In *Proceedings of ACL-IJCNLP*, pages 217–225.

A. Nenkova and A. Louis. 2008. Can you summarize this? identifying correlates of input difficulty for multi-document summarization. In *Proceedings of ACL-HLT*, pages 825–833.

E. Pitler and A. Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of EMNLP*, pages 186–195.

E. Pitler, A. Louis, and A. Nenkova. 2010. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of ACL*.

S. Schwarm and M. Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of ACL*, pages 523–530.

L. Si and J. Callan. 2001. A statistical model for scientific readability. In *Proceedings of CIKM*, pages 574–576.

R. Soricut and D. Marcu. 2006. Discourse generation using utility-trained coherence models. In *Proceedings of COLING-ACL*, pages 803–810.

S. Teufel and M. Moens. 2000. What's yours and what's mine: determining intellectual attribution in scientific text. In *Proceedings of EMNLP*, pages 9–17.

J. Zhao and M. Kan. 2010. Domain-specific iterative readability computation. In *Proceedings of JDCL*, pages 205–214.