Adapting Graph Summaries to the Users' Reading Levels

Priscilla Moraes, Kathleen McCoy and Sandra Carberry
Department of Computer and Information Sciences
University of Delaware, Newark, Delaware, USA
[pmoraes | mccoy | carberry]@udel.edu

Abstract

Deciding on the complexity of a generated text in NLG systems is a contentious task. Some systems propose the generation of simple text for low-skilled readers; some choose what they anticipate to be a "good measure" of complexity by balancing sentence length and number of sentences (using scales such as the D-level sentence complexity) for the text; while others target high-skilled readers. In this work, we discuss an approach that aims to leverage the experience of the reader when reading generated text by matching the syntactic complexity of the generated text to the reading level of the surrounding text. We propose an approach for sentence aggregation and lexical choice that allows generated summaries of line graphs in multimodal articles available online to match the reading level of the text of the article in which the graphs appear. The technique is developed in the context of the SIGHT (Summarizing Information Graphics Textually) system. This paper tackles the micro planning phase of sentence generation discussing additionally the steps of lexical choice, and pronominalization.

1 Introduction

Multimodal documents from online popular media often contain information graphics that augment the information found in the text. These graphics, however, are inaccessible to blind users. The SIGHT system is an ongoing project that proposes methods of making this information accessible to visually impaired users by generating a textual summary capturing the high-level message of the graphic along with visually distinctive features. Figure 1 shows an example of an information graphic found in popular media. This graphic ostensibly conveys that there was a change in the trend of ocean levels, which is first stable until about 1940 and then rising through 2003. Earlier work on the system (Wu, Carberry, Elzer, & Chester, 2010) is able to infer this high-level message given a representation of the graphic.

Nevertheless, a generated summary should convey more than just the intended message. It should provide important visual features that jump out at a person who views the graphic (such as the fluctuation in the data values as seen in the graph in Figure 1). The set of remarkable features is different for different graphics. Previous work of ours (Moraes, Carberry, & McCoy, 2013) presents methods that capture these most important features and allow the composition of customized summaries for each graph. Thus, given a graphic, our previous work has resulted in a system that can produce a set of propositions to include in its summary. In this paper, we turn to the subsequent phases of generation: given a set of propositions, how these propositions should be realized such that the resultant text is adapted to the user's reading level and thus is coherent and understandable.

Therefore, this work presents novel strategies that have been deployed in the text generation phase of the SIGHT system applied to line graphs. It describes the micro planning phase, emphasizing sentence aggregation, lexical choice and pronominalization. The contribution of this work is the provision of coherent and concise textual summaries that narrate line graphs' high-level content to visually impaired users through approaches that rely on 1) making the right wording choices and 2) making appropriate syntactical decisions in order to achieve a desired reading level for the generated text.

Previous work in generation assumes a particular level of complexity for all texts created. Our hypothesis is that the graph's summary should vary depending on the user's reading level. Although one could explicitly inquire about the user's reading level, this would be intrusive and

Proceedings of the 8th International Natural Language Generation Conference, pages 64–73,
Philadelphia, Pennsylvania, 19-21 June 2014.

would detract from the overall experience. Thus we hypothesize that the level of complexity of the article in which the graph appears roughly equates with the user's reading level --- that is, users generally choose articles that are at their own reading comfort level. Therefore, our approach is to generate summaries that reflect the reading level of the accompanying article. Not only will such summaries be coherent and understandable to the user, but also the summary should fit seamlessly into the user's reading of the article.

The decision to match the text complexity of the generated text to that of the article's text was inspired by results of an experiment performed with college students aiming to evaluate the content determination output. In the experiment, sentences were generated for each proposition selected by the system. Comments made by the subjects revealed that the simplest possible text was not easier to understand. Rather, it caused them confusion and discomfort when reading it. Based on these results, we decided to tackle the problem of deciding on the text complexity of automatically generated text by following the same syntactical complexity of the surrounding text, by reading level. In addition, we use word frequencies to select more common lexical items to compose summaries of lower reading levels.

The next section presents the background and motivation for our work. Section 3 discusses some related work concerned with text generation and simplification. Section 4 presents our proposed approach to text generation that adapts the output to the reading level of the surrounding text. Section 5 shows some examples of text generated in different grade level groups. Section 6 shows our preliminary evaluation and it is followed by some conclusions and ideas for future work in Section 7 and 8, respectively.

2 Background

The approaches presented in this work are deployed in the context of the SIGHT system. The system is concerned with providing access to information graphics present in multimodal documents from popular media such as the graphic in Figure 1. For this graphic, the content selection module[1] (Moraes et al., 2013) chooses the following propositions for inclusion in the initial summary:

- graph type (line graph);
- entity being measured (annual difference from Seattle's 1899 sea level, in inches);
- the intended message of the graphic (changing trend: stable then rising);
- the high fluctuation of the data values;
- the description of the individual segments of the graphic;
- the initial value (annotated end point);
- the ending value (annotated end point).

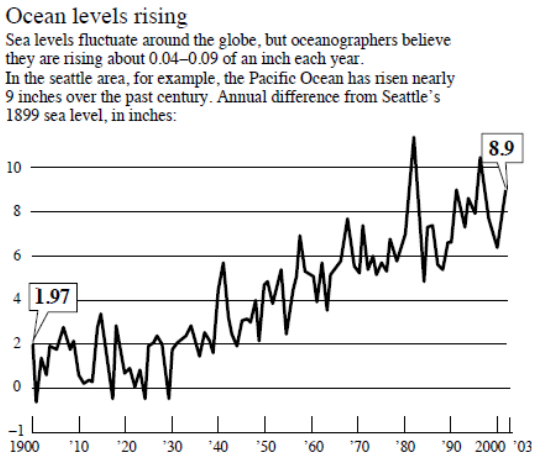

Figure 1: Example of a graphic that has a Changing Trend as its intended message and presents out-standing visual features (volatility and annotations on end points).

These propositions are not necessarily selected in this listed order, nor in the order they will be mentioned in the summary. They are selected based on their overall importance in the context of the graphic since the content selection framework is based on an adapted version of a centrality-based algorithm. Once these propositions are selected, an overarching organizational strategy must be chosen to decide on the most appropriate ordering. Our system gives most importance to the overall intended message of the graphic and thus this will be mentioned first. Next, a description of the features of the individual trend(s) will be provided. Finally, summary information about the whole graph will be given. The system must make further decisions when the graph conveys more than one trend (such as the graph in Figure 1). For such cases, the system must further decide whether to organize the description of the trends (1) by the trends themselves – e.g. either in left to right order - when no trend is considered more important than the others; or (2) by importance – when a trend has a

[1] The content selection module has been presented in a previous paper and is outside the scope of this paper.

greater set of features selected for the discourse or it composes a candidate intended message, which augments the intended message (Moraes et al., 2013). In the latter case, if a piece of the graphic (trend) has significantly more features selected, meaning that it possesses a higher number of visually outstanding features, it will be described first, then followed by the other trends. The organization of the sentences is a separate step that happens prior to the realization phase, which is the focus here, and will not be discussed further in this paper.

Having the set of ordered propositions selected, the question that arises is how to realize this information to the user. The most straightforward way of realizing the summary would be to realize each proposition as a single sentence. This strategy was applied in an evaluation experiment (briefly described next) that aimed to test the preciseness of the content selection framework. The experiment presented the subjects with line graphs and their correspondent generated initial summaries (the propositions were properly ordered for this experiment). Subjects were asked whether or not the most important information about the graphic was part of the summary and whether the summary presented unnecessary or redundant information. They were also offered the opportunity to provide additional comments.

For the experiment, the initial summary for the graphic in Figure 1 was the following:

> The image shows a line graph. The line graph is highly volatile. The line graph presents the number of annual difference from Seattle's 1899 sea level, in inches. The line graph shows a trend that changes. The changing trend consists of a stable trend from 1900 to 1928 followed by a rising trend through 2003. The first segment is the stable trend. The stable trend has a starting value of 1.97 inches. The second segment is the rising trend. The rising trend has an ending value of 8.9 inches.

Although the experiment was intended to evaluate the content present in the summaries, various comments addressed the syntactical construction of the text. These comments highlighted the lack of aggregation and pronominalization. For instance, a common theme of the comments was that some of the information could be "combined" and presented more succinctly.

All the participants of the experiment were graduate students. These results showed that more sophisticated readers *prefer* text that is more sophisticated. This finding pointed to the necessity of an aggregation step before the delivery of the summaries. However, questions arose concerning how much aggregation to do, how to measure aggregation to choose one strategy over another, or to decide on a desired level of aggregation.

To answer the above questions, we decided to examine the text complexity of the text surrounding the graphic --- that is, the text from the article in which the graph appears. We presume that this text complexity equates with the user's reading level and thus summaries at this level of complexity will be understandable and coherent to the users. This approach seemed to be the best way of customizing the text complexity of the summaries in order to tailor summaries to individual users.

3 Related Work

Research on generating text concerned with low-skilled users has been conducted by (Williams & Reiter, 2004, 2005a, 2005b, 2008; Williams, Reiter, & Osman, 2003). As stated by (Williams & Reiter, 2005b), most NLG systems generate text for readers with good reading ability. Thus, they developed a system called *SkillSum* which adapts its output for readers with poor literacy after assessing their reading and numeracy skills. Their results show that, for these target readers, the micro planning choices made by *SkillSum* enhanced readability. (Siddharthan, 2003) proposes a regeneration phase for syntactical text simplification in order to preserve discourse structure "aiming to make the text easier to read for some target group (like aphasics and people with low reading ages) or easier to process by some program (like a parser or machine translation system). (J. Carroll et al., 1999) presents a text simplification methodology to help language-impaired users. (Rello & Baeza-Yates, 2012) investigates dyslexic errors on the Web and (Rello, Baeza-Yates, Bott, & Saggion, 2013) propose a system that uses lexical simplification to enhance readability and understandability of text for people with dyslexia. They help users to understand the text by offering as options the replacement of more complicated lexical items

by simpler vocabulary. They performed experiments with people with no visual impairments and with people with dyslexia and concluded that the system improved readability for the users with dyslexia and improved comprehensibility for users with no visual impairments. Experiments performed with blind users and the usability of a system that provides access to charts and graphs is presented by (Ferres, Lindgaard, Sumegi, & Tsuji, 2013).

Other NLG systems make decisions on text complexity based on available scales such as the D-level sentence complexity (Covington, He, Brown, Naci, & Brown, 2006). One example is presented in (Demir et al., 2010) where tree structures are built representing all the possible ways sentences can be aggregated and the choice of the tree tries to balance the number of sentences, their D-level complexity, and the types of relative clauses.

Although text simplification is crucial to target low-skilled readers and users with language disabilities, our experiment with college students showed that the simplest text was rather unpleasant to read for them. We therefore propose a technique that focuses on adjusting the generated text to the reading level of the surrounding text. Thus, our system should satisfy both high-level and low-level readers.

4 Aggregation and Text Complexity

The initial summaries generated by the system are composed of individual sentences that were realized from atomic concept units. Since we use a bottom-up approach when selecting content, in order to achieve different text complexity levels, a sentence aggregation step is needed. The aggregation module is in charge of merging propositions that describe an entity, creating a more complex sentence that will encompass the information selected that describes the referring expression.

The approach proposed by (Wilkinson, 1995) presents the aggregation process divided in two major steps: semantic grouping and sentence structuring. Although they are interdependent, both are needed in order to achieve aggregation in a text. Initiatives on automatic aggregation (or only semantic grouping) of text using learning techniques also exist. (Barzilay, 2006), (Bayyarapu, 2011), (Walker, Rambow, & Rogati, 2001) are some examples of learning aggregation rules and grouping constrains in order to aggregate text. (Demir, 2010) presents a mechanism in which each proposition is a single node tree which can be realized as a sentence and attempts to form more complex trees by combining trees in such a way so that the more complex tree (containing multiple propositions) can still be realized as a single sentence. In order to decide which tree is the best one to be realized, Demir's work applies the revised D-level sentence complexity scale, which measures the syntactic complexity of a sentence according to its syntactic structure.

Although learning methodologies are innovative, they strive to train the algorithms in order to choose the best text plan based in a specific task or environment (defined by the training data and the decision of which plan is the "best" given the human subjects' judgments). Our contention is that a given sentence plan can be perfectly suitable in one context and, at the same time, be ineffective in another one, making the choice of the best text plan a variable. For this reason, we decided to take into consideration the article reading level when choosing the text plan that will be used to design the aggregation of summaries generated by our system. This approach allows the summary of the line graph to fit coherently within the article's text. Text plans, in the context of this work, refer to the different set of rules that are followed in order to aggregate propositions before the realization phase. Each text plan decides how propositions related to a given entity should be combined in order to produce sentences.

4.1 Reading Level Assessment

Much effort has been devoted to developing automated approaches for assessing text complexity. Some examples are the use of support vector machines (Schwarm & Ostendorf, 2005) in order to find topical texts at a given reading level. Another approach is the use of statistical language models (Collins-Thompson & Callan, 2005; Collins-Thompson & Callan, 2004) for predicting reading difficulty. The combination of vocabulary and grammatical features in order to predict reading difficulty for first and second language texts is the object of study in (Heilman, Collins-Thompson, Callan, & Eskenazi, 2007).

(Sheehan, Kostin, Futagi, & Flor, 2010) developed a system called *SourceRater* (now named *TextEvaluator*), which considers features of text that go beyond syntactical features. The authors list a set of dimensions of text that influences in a text reading complexity. These dimensions are: Spoken vs. Written Language, Aca-

demic Orientation, Syntactic Complexity, Narrative Style, Overt Expression of Persuasion, Vocabulary Difficulty, and Negation. They divide texts into literary and informational in order to assess these features and their impact in reading difficulty after finding that these styles have substantial differences. They evaluate their technique by comparing their results with assessments done using Flesh-Kincaid reading level assessment (Kincaid, Fishburne, Rogers, & Chissom, 1975) applied to text categorized into grade levels by the Common Core Standards ("Common Core State Standards Initiative," 2014).

Another tool, Coh-Metrix (Graesser et al., 2004), was designed to analyze text on measures of cohesion, language and readability. This evaluator also categorizes the input text into one of Scientific, Narrative or Informational and it considers features such as cohesion relations, user world knowledge, language, and discourse characteristics besides syntactical features such as word and sentence length when assessing the text complexity.

To generate text that complies with a given reading level, we consider that a common, well-know, widely-used metric such as Flesch-Kincaid or SMOG (Laughlin, 1969) will suffice for providing input to the text planning phase of our system. To assure the usefulness of this metric in our context, we evaluated the similarity between assessments done by Flesch-Kincaid and SMOG and assessments made by *TextEvaluator*. For this comparison, we used 55 articles from our corpus[2]. The results showed that for only 20 percent of the articles was the reading level assessment provided by Flesch-Kincaid and SMOG different from the text complexity classification done by *TextEvaluator*. From these results, we concluded that simple reading assessments such as Flesch-Kincaid and SMOG would suffice for guiding the choice of syntactical text complexity in our generated summaries.

4.2 Generating Summaries for Different Reading Levels

When generating the initial summaries of line graphs, our system creates different text plans for each group of grade levels (each group comprises two or more grade levels starting at the 5th grade) and applies the appropriate one depending upon the assessed reading level of the text in the article containing the graphic.

Because the summary is not long enough to be exact when determining its reading level (since longer texts result in more accurate assessment of their reading level), we decided not to create one text plan for each grade level. Instead, we have created five grade level groups and each one comprises two or more grades. For each group of grade levels, we define a text plan that increases a sentence syntactic structure complexity as the grade gets higher. We define a text plan for summaries that can range between grades 5 (inclusive) and 7 (exclusive), another text plan for grades between 7 (inclusive) and 9 (excusive). A third text plan is defined for grades 9 inclusive and 11 (exclusive), one for 11 (inclusive) and 13 (exclusive) and, finally, another one for grades greater than or equal to 13 (college level).

The content selection framework, as mentioned earlier, defines the content of a given summary dynamically. Due to this fact, the amount of information (or the number of propositions) selected for inclusion in a summary varies per graphic. Our intention is to make sure that the reading level of the summaries generated by our system do not exceed the reading level of their respective article's text. It is admissible, however, for the summary to have a slightly lower reading level than the one from the text.

The organization phase, which is a previous step, divides the set of propositions produced by the content selection module into three groups: 1) propositions that comprise an introduction containing the high-level message of the graphic, 2) propositions that detail the individual trends of the graph, and 3) propositions that convey computational information about the overall graph. Thus, from the set of selected propositions, the text plan of a given group defines rules on Noun Phrase (NP) density and lexical choice. When describing an entity, attributes of this entity can be added to the NP as modifiers using either adjectives e.g. "*a steep rising trend*", conjunctions e.g., "*the rising trend is steep and volatile*" or relative clauses e.g. "*a rising trend, which is steep*". When the modifier of an NP is a Verb Phrase (VP), it is combined using a relative clause e.g., "*the line graph, which presents the number of jackets sold in 2013...*" VPs can be modified by adverbs e.g., "*the falling trend is very steep*". The text plans applies rules within sets of propositions that are grouped hierarchically. Within these major groups, propositions can only be aggregated if they belong to the same

[2] Our Digital Library contains multimodal articles collected from popular media. It is available at http://ir.cis.udel.edu/~moraes/udgraphs

entity. The decision of using one syntactic structure over the other is currently based on discourse strategies. The complexity added by a relative clause over the one added by an adjective, for example, is the focus of current investigation (more details in Section 8) and will be considered when choosing one construction over another.

4.3 Lexical Choice

Most of the work on text simplification and readability assessment considers lexicalization a crucial aspect for readability and comprehensibility. (Rello, Baeza-Yates, Bott, & Saggion, 2013) presents a system that increases the understandability and readability of text by helping users understand the text by replacing complex words with more common ones in the lexicon. (Laughlin, 1969) states that longer and more precise words are usually harder to understand.

This led us to use more common words at lower grade levels to increase the chance of the text being easily understood by the reader. For this, we use the Word Frequency Data from the Corpus of Contemporary American English (Davies, 2008). Precise and specific words (which are less frequently used) that describe visual features of line graphs such as *volatility* and *steepness* are replaced by other words or expressions that are more commonly used but still carry the same meaning, such as *"peaks and valleys"* or *"ups and downs"*. The experiment presented in Section 6 corroborates this claim, showing that college level students were comfortable with the use of such lexical items whereas fifth graders complained about them and asserted they did not know their meanings. Future work concerns the use of lexical items categorized by reading levels (details in Section 8).

4.4 Pronominalization

Another important feature is the pronominalization of referring expressions. This technique avoids reintroduction of entities every time they are mentioned. The experiment mentioned in Section 2 showed that the reintroduction of entities or the repetition of referring expressions (when a pronoun could be used) in fact jeopardized the understanding of some passages in the summaries. The participants would usually complain that a given summary was confusing because it could be "better presented" and they would additionally provide us with comments regarding the reintroduction of the referring expressions. From these results, we concluded that it would be valuable to include a pronominalization step in the aggregation phase so that even the summaries that are at a lower grade level would not repeat the referring expression when using multiple non aggregated sentences.

The propositions chosen by the content selection framework contain the information about their memberships (features such as volatility and steepness point to the segment of the graphic they belong to). This membership information is the clue used to define discourse focus. Our work follows the approach applied in the TEXT system (McKeown, 1992), in which pronouns are used in order to refer to the entity being focused in subsequent sentences. Also inspired by the work presented by (McCoy & Strube, 1999) our system makes use of other anaphoric expressions besides pronouns, such as "the trend" or "the graph". These alternative anaphoric expressions are used to reintroduce entities when the discourse focus changes. The following example shows the use of pronouns and the reintroduction of the entity in the last set of propositions. The entities that are in focus in each sentence are underlined and the referring expressions are bolded.

> The image shows a line graph. The line graph presents the number of cumulative, global unredeemed frequent-flier miles. **It** conveys a rising trend from 1999 to 2005. **It** has a starting value of 5.5. **It** has an ending value of 14.2. **The graph** shows an overall increase of 8.7.

The last sentence changes the focus back to the overall graph. Even though the entity *line graph* was already mentioned, the focus had changed to the entity *rising trend*, so when the focus returns to the entity *line graph*, the system makes use of a definite reference to reintroduce it.

5 Examples of Summaries Generated for Different Reading Levels

Below are examples of some of the summaries that our system generates for the graph in Figure 1 at different reading levels. Their assessed reading levels provided by SMOG are also shown[3]. The summaries in these examples are also pro-

[3] These results were obtained from using a tool available in the GNU project Style and Diction (FSF, 2005).

nominalized. The pronominalization phase is described in Section 4.4.

Summary for Grades > 5 and <= 7

The image shows a line graph. The line graph has ups and downs. It presents the number of annual difference from Seattle's 1899 sea level, in inches. It conveys a changing trend. It consists of a stable trend from 1900 to 1928 followed by a rising trend through 2003. The first segment is the stable trend. It has a starting value of 1.97 inches. The second segment is the rising trend. It has an ending value of 8.9 inches.
(SMOG 4.8)

Summary for Grades > 11 and <= 13

The image shows a highly volatile line graph, which presents the number of annual difference from Seattle's 1899 sea level, in inches, in addition to conveying a changing trend that consists of a stable trend from 1900 to 1928 followed by a rising trend through 2003. The first segment is the stable trend that has starting value of 1.97 inches. The second segment is the rising trend that has ending value of 8.9 inches.
(SMOG 10.0)

The assessed reading level of these passages are below the maximum threshold due to the limited number of propositions selected by the content determination algorithm.

6 Evaluation

This work on aggregation was motivated by the evaluation described in Section 2, which was intended to evaluate the content selection phase of the system. Much to our surprise, many of the comments indicated that the summaries were difficult to read because they lacked aggregation! This result caused us to implement the work presented here. Our first evaluation therefore replicated our first experiment where, instead of using a simple sentence for each proposition, sentences were aggregated to reflect a 7^{th} – 9^{th} grade reading level (the level slightly lower than the median of the articles collected for our corpus).

Table 1 compares the results of these two initial experiments. The results[4] show a dramatic drop in the comments related to issues with aggregation. From this preliminary experiment results, we felt encouraged to pursue the generation of summaries suited to grade levels.

	Number of Subjects	Number of Responses	Number of complaints
Experiment 1	16	201	22
Experiment 2	29	331	4

Table 1. Comparison of results from preliminary experiment.

Our second experiment targeted our generation of grade-level appropriate text. In this experiment, we wished to judge whether readers at different reading levels would prefer texts generated by our system aimed at their reading level. We therefore recruited two groups of participants: (1) students from a fifth grade elementary school in the area and (2) undergraduate students in an introductory CS course at a university.

Participants were presented with 2 summaries from each of 5 different graphs. One of the summaries was generated to be at a 5^{th} – 7^{th} grade reading level and the other at a 11^{th} – 13^{th} grade reading level. The participants were asked to select the summary they liked the best and to provide comments on what they did not like in either summary.

Table 2 shows the results of this experiment. Five students from 5^{th} grade and thirty-four freshmen college students were recruited to participate. From these results we can see that, in fact, the majority in both groups preferred the grade-level appropriate summary. For the freshmen college students, the fact that the subjects were almost evenly split on their choices, even though they are at the same grade level, was expected. This shows that reading preferences may vary even among people from same age/grade level. Since there were subjects who preferred simple to complex text, we can assume that reading skills can vary even within a grade level group. Our contention is that readers who prefer simple text would read venues that use simple text structure and syntax. That is where our ap-

[4] The number of complaints presented in Table 1 are concerned only with syntactical issues.

proach plays an even better role when looking into the surrounding text the user is reading. Following this approach, instead of assessing or asking the user which level they are in, gives us more chances of being successful at producing text that will be more appropriate to each user.

Analyzing the results on the choices of the opposite summary to their target group, we noticed that there was an agreement amongst subjects regarding the type of the graph. Kids who showed a preference for the complex text, for example, did so only for graphics describing a simple trend, therefore having a small amount of information an making it easy for them to follow.

Some college students who chose the simpler summary provided comments that showed to be independent of the reading level decisions of the system. Some subjects pointed that a default connective applied by the realizer ("in addition to") was making the summary complicated to read. That can actually be the cause of the choice for the simple summary, and not necessarily the amount of aggregation. To address this, we consider that changing the connective to a more common one (e.g. "and") would make the text more fluid.

From these results, we conclude that, indeed, adapting the generated text to the complexity of text commonly read by a user is a promising path to follow. An experiment where we provide the subjects with the article accompanying the graph and ask them to choose the summary that they believe fits the text complexity of the summary is intended and planned as future work. We have initiated investigation in some automated ways of generating text within these different grade level groups and we discuss it further in Section 8.

	Chose Summaries for 5th – 7th Grades (%)	**Chose Summaries for 11th - 13th Grades (%)**
5th grade	80	20
Freshmen students	47	53

Table 2. Results from experiment measuring choices of summaries in different reading levels.

7 Conclusion

Most NLG systems available today generate text that focus on specific target readers. Some of them focus on text generation for low-skilled readers, while others generate text for high-skilled readers. In this work, we presented an approach that offers a solution that attends to the needs of readers at different grade levels.

Our system generates initial summaries of line graphs available in popular media, so visually impaired users can have access to the high-level message these resources carry. Our contention is that users read articles from venues that they feel comfortable with reading. Therefore, we assert that generating summaries that fit the text complexity of the overall article leverages the quality of the generated text. We showed an approach that uses Flesch-Kincaid and SMOG reading assessments in order to determine the syntactical complexity of the generated text. From the experiments performed, we conclude that pursuing the generation of natural language text that fits the reading level of the surrounding text is promising.

8 Path Forward

Investigation on more automated ways of deciding on how to aggregate propositions is the next step to take. Our current aggregation method relies on templates for each group. We anticipate some techniques to learn how different text constructions can affect reading measures and then using them when choosing an adjective over a relative clause for increasing the NP density and use of passive voice, for example. This would allow the aggregation phase to be easily applied to NLG systems in different contexts.

Another important point is the choice of lexical items by reading level or age. We plan on investigating how the usage of word frequency by age/grade level (Carroll, 1972) might help achieving a more appropriate summary for a given grade level. Then, the lexical items that are listed as common to the target grade reading level would be applied in their respective context.

Some comments provided on the second experiment described in Section 6 were that it was not so easy to understand long sentences on which values and dates were also present. This aspect deserves investigation on acquiring numeracy skills along with reading skills as clues to assess the best text complexity to present. Research that assess numeracy and literacy skills of users is presented by (Williams & Reiter, 2008).

From the accessibility prospective, an experiment with blind users is anticipated. We intend to evaluate the effect of generating text in different reading levels for people with visual and/or reading impairments.

References

Barzilay, R. (2006). *Aggregation via set partitioning for natural language generation.* Paper presented at the In HLT-NAACL.

Bayyarapu, H. S. (2011). *Efficient algorithm for Context Sensitive Aggregation in Natural Language generation.* Paper presented at the RANLP.

Carroll, J. B. (1972). A New Word Frequency Book. *Elementary English, 49*(7), pp. 1070-1074.

Collins-Thompson, K., & Callan, J. (2005). Predicting Reading Difficulty with Statistical Language Models. *J. Am. Soc. Inf. Sci. Technol., 56*(13), 1448-1462.

Collins-Thompson, K., & Callan, J. P. (2004). *A Language Modeling Approach to Predicting Reading Difficulty.* Paper presented at the HLT-NAACL.

Common Core State Standards Initiative. (2014). Retrieved 2014-01-09, from http://www.corestandards.org/

Covington, M., He, C., Brown, C., Naci, L., & Brown, J. (2006). *How Complex is that Sentence? A Proposed Revision of the Rosenberg and Abbeduto D-Level Scale.* Paper presented at the Research Report, Artificial Intelligence Center, University of Georgia.

Davies, M. (2008). Word frequency data: Corpus of Contemporary American English.

Demir, S. (2010). *Sight for visually impaired users: Summarizing information graphics textually*. University of Delaware.

Demir, S., Oliver, D., Schwartz, E., Elzer, S., Carberry, S., McCoy, K. F., & Chester, D. (2010). Interactive SIGHT: textual access to simple bar charts. *New Rev. Hypermedia Multimedia, 16*, 245-279.

Ferres, L., Lindgaard, G., Sumegi, L., & Tsuji, B. (2013). Evaluating a Tool for Improving Accessibility to Charts and Graphs. *ACM Trans. Comput.-Hum. Interact., 20*(5), 28:21-28:32.

FSF. (2005). Style and Diction GNU project. from www.gnu.org/software/diction

Graesser, A. C., McNamara, D. S., Louwerse, M. M., Cai, Z., Dempsey, K., Floyd, Y., . . . Correspondence, F. Y. (2004). *Coh-Metrix: Analysis of text on cohesion and language.* Paper presented at the M. Louwerse Topics in Cognitive Science.

Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2007). *Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts.* Paper presented at the HLT-NAACL.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.

Laughlin, G. H. M. (1969). SMOG Grading-a New Readability Formula. *Journal of Reading, 12*(8), pp. 639-646.

McCoy, K., & Strube, M. (1999). *Generating Anaphoric Expressions: Pronoun or Definite Description?* Paper presented at the ACL WORKSHOP ON DISCOURSE AND REFERENCE STRUCTURE.

McKeown, K. (1992). *Text Generation*: Cambridge University Press.

Moraes, P. S., Carberry, S., & McCoy, K. (2013). *Providing access to the high-level content of line graphs from online popular media.* Paper presented at the Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility, Rio de Janeiro, Brazil.

Rello, L., Baeza-Yates, R., Bott, S., & Saggion, H. (2013). *Simplify or Help?: Text Simplification Strategies for People with Dyslexia.* Paper presented at the Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility, New York, NY, USA.

Rello, L., & Baeza-Yates, R. A. (2012). The presence of English and Spanish dyslexia in the Web. *The New Review of Hypermedia and Multimedia, 18*(3), 131-158.

Schwarm, S. E., & Ostendorf, M. (2005). *Reading Level Assessment Using Support Vector Machines and Statistical Language Models.* Paper presented at the Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, USA.

Sheehan, K. M., Kostin, I., Futagi, Y., & Flor, M. (2010). Generating automated text complexity classifications that are aligned with targeted text complexity standards.

Walker, M. A., Rambow, O., & Rogati, M. (2001). *SPoT: a trainable sentence planner*. Paper presented at the Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies, Stroudsburg, PA, USA.

Wilkinson, J. (1995). Aggregation in Natural Language Generation: Another Look.

Williams, S., & Reiter, E. (2008). Generating basic skills reports for low-skilled readers*. *Natural Language Engineering, 14*(4), 495-525.

Wu, P., Carberry, S., Elzer, S., & Chester, D. (2010). *Recognizing the intended message of line graphs.* Paper presented at the Proceedings of the 6th international conference on

Diagrammatic representation and inference, Berlin, Heidelberg.