Diagnosing Causes of Reading Difficulty using Bayesian Networks

Pascual Martínez-Gómez
The University of Tokyo
National Institute of Informatics
pascual@nii.ac.jp

Akiko Aizawa
The University of Tokyo
National Institute of Informatics
aizawa@nii.ac.jp

Abstract

There is a need of matching text difficulty to the expected reading skill of the audience. Readability measures were developed with this objective in mind, first by psycholinguists, and more recently, by practitioners of natural language processing. A common strategy was to extract linguistic features that are good predictors of readability, and then train discriminative classification or regression models that correlate well with human judgment. But correlation does not imply causality, which is a necessary property to explain why documents are not readable. Our objective is to provide mechanisms for text producers to adjust the readability of their content. We propose the use of generative models to diagnose causes of reading difficulty, and bring closer the realization of automatic readability optimization.

1 Introduction

Educational institutions, government agencies and some private companies have a special interest in authoring documents for a certain audience, but it is expensive to involve expert linguists to assess the readability of every document they produce. The first psycholinguistic studies developed readability formulas for grading purposes, based on surface linguistic features. Those formulas, despite of their simplicity, performed well and were widely used by editors to grade reading material for young readers. However, content producers might be tempted to adapt their manuscripts by tweaking the text features present in readability formulas, without gaining (or even degrading) real readability (Davison and Kantor, 1982).

Recently, the application of statistical models to linguistic problems proved successful, and ambitious tasks in automatic document transformation such as text summarization or machine translation became a hot topic in computational linguistics. Readability optimization is one of such document transformation problems. Recent studies on readability embrace machine learning techniques to recognize readability with an even higher accuracy. The common approach consists in extracting as many features as possible, and then training a classifier or a regression model using human annotated texts to predict a readability score given the observation of the linguistic features.

Those discriminative models correlate well with human judgment, but fail at explaining why a document is not readable. We call *readability diagnosis* the automatic discovery of the causes that lead to (un)readability, and we believe it is an essential step for readability optimization.

We propose the use of Bayesian causal networks to perform readability diagnosis. That is, given a document, the objective of our Bayesian network is to recognize the specific parts of the document that are difficult to read. Bayesian networks are a type of generative model, where the joint probability distribution is constructed by making certain independence assumptions. Their main advantage is that they allow to query the model regarding any linguistic variable, generalizing the functionality of traditional models.

In the next section we briefly introduce former work by psycholinguists and recent work by practitioners on natural language processing. We describe our application of Bayesian networks to readability diagnosis in Section 3 and summarize the capabilities of the model. Corpora, baseline system description and results are presented in Section 4, where we assess to what extent our generative model predicts cognitive evidence. Pointers to future work and applications that would benefit from our results can be found in Section 5, followed by our conclusions in Section 6.

International Joint Conference on Natural Language Processing, pages 1383–1391,
Nagoya, Japan, 14-18 October 2013.

2 Related Work

Readability formulas have been the subject of investigation long before the existence of current natural language processing techniques. Although sophisticated methods could have been developed, there was an emphasis on easy-to-compute formulas, where the readability score of a text is computed as a function of its linguistic features.

The Flesch-Kincaid formula (Flesch, 1948; Kincaid et al., 1975) was probably the first in gaining wide recognition among publishers. This formula is a linear combination of two variables[1], as:

$$r_{\text{score}} = 0.39 \times \text{ASL} + 11.8 \times \text{ASW} - 15.59 \quad (1)$$

where ASL is the average number of words per sentence, and ASW is the average number of syllables per word. Despite of its simplicity, ASL and ASW are very discriminative linguistic features when assessing readability and this formula correlates surprisingly well with human judgment.

The search for more discriminative linguistic features continued, and Mc Laughlin (1969) found that the number of polysyllabic words in a certain amount of text is also a good predictor of reading difficulty. The rationale behind such a linguistic feature could be that the required lexical processing is higher when the word is longer, or that long words tend to be more infrequent and difficult to read (Rayner and Duffy, 1986). This work was followed by others (Fry, 1990) that counted the number of words in the text that were contained in the vocabulary of specific word lists. The use of word lists introduced a new dimension in readability, since it was possible to design hand-crafted lists that could not only account for lexical frequency, but also for semantic complexity.

Building on the idea of lexical frequency and counting on large amounts of text data, the use of word lists was generalized into unigram language models (Si and Callan, 2001), which increased the correlation with human judgment on readability.

Linguistic features of different nature were also explored, and grammatical features are an example of them (Heilman et al., 2007). Those features alone were found not to be as discriminant as the lexical ones, but performed well in combination with them. However, the effects were not additive, which suggests that variables correlated with each other to a certain extent. This was also noted in some other works (Petersen and Ostendorf, 2009), where syntactic features were also used, in combination with higher order n-gram language models.

Automatic text transformation for readability optimization is a task that naturally follows former readability studies and the large scale need of producing content for specific audiences. Authors in (Carroll et al., 1999; Devlin et al., 1999; Siddharthan, 2003) approached the problem using rules for syntactic transformation, anaphora substitutions and vocabulary simplifications, but those rules were not experimentally tested for their target readers. Williams and Reiter (2005) did test their transformation rules, but they were limited to assess the effects of their set of rules, which had a low coverage. Other authors (Aluísio et al., 2010; François and Fairon, 2012) integrated readability scores in authoring systems, to assist text simplification rather than fully automating it.

Previous work has concentrated on finding linguistic features that are good predictors of readability, and building discriminative models that best correlate with human judgment. But those models can only indicate whether a piece of text is readable or not, and fail in explaining the causes. In view of (semi-)automatic text simplification and readability optimization, we propose Bayesian causal networks as a generative model for readability. In this approach, readability is modeled as a factored joint probability distribution over lexical, part of speech, syntactic, semantic and discourse features. This provides an interpretable model to gain linguistic insight about what features impact most on readability in a *specific document* and to understand how that text should be transformed to increase readability even under human-imposed constraints.

3 Methodology

3.1 Discriminative and Generative Models

Previous work on readability assessment has focused on the development of discriminative models. Those discriminative models are functions ϕ that map instantiations ℓ of a set of linguistic features $\mathscr{L}$ to a readability score $r \in \mathbb{R}$, $\phi : \mathscr{L} \rightarrow \mathbb{R}$. In this work, examples of linguistic features are "proportion of verbs to words", or "maximum number of active lexical chains" in a given text, and their instantiations are their actual values for that text. If we normalize the readability measure

[1] We will use the term *variables* interchangeably with *linguistic features*.

so that it assigns 1 to the whole space of possible feature instantiations, we can regard the readability score as a probabilistic measure, and without loss of generalization, reformulate the problem as:

$$\hat{r} = \underset{r}{\operatorname{argmax}} \Pr(r \mid \ell), \tag{2}$$

where we have to find the readability score r with maximum probability, given the instantiation ℓ of the set of linguistic features $\mathscr{L}$.

In this approach, the probability on all possible reading score assignments is well defined, but there is no attempt to model the probability of the instantiations ℓ of linguistic features $\mathscr{L}$. As it has been reported in related work, most explanatory effects on readability do not add up across all linguistic features. This suggests that linguistic features interact with each other and have mixed effects on readability prediction. There have been ablation and correlation studies to bring light on those feature interactions (Kate et al., 2010), but they were limited to a few feature combinations and no attempts were done to study causal relationships or other conditional independencies.

To attain diagnosis capabilities, we propose the use of Bayesian causal networks as an example of generative models $\Pr(r, \ell)$, where the readability score and the linguistic features are modeled *together* using a joint probability distribution. There are, however, some challenges associated to this model that are described below.

3.2 Independence Assumptions

To preserve generality, we will regard joint probability distributions as tables, where every row defines the probability of a discrete value assignment to all linguistic features and the readability score. The number of parameters to be estimated in the model is proportional to the number of possible assignments, which is exponential with the number of linguistic features. A simple approach is to consider the readability to be dependent on all features, but all features independent from each other. The joint probability distribution can be consequently defined as:

$$\begin{aligned}\Pr(r, \mathscr{L}) &= \Pr(r, l_1, \ldots, l_m) \\ &\approx p(r \mid l_1, \ldots, l_m) \cdot p(l_1) \cdots p(l_m),\end{aligned} \tag{3}$$

where $p(l_i)$ are the priors for every linguistic feature l_i, and the conditional probability distribution $p(r \mid l_1, \ldots, l_m)$ models the non-linear relationship between linguistic features and the readability score. The graphical representation of this model can be found in Figure 1a, where the gray circles are observed linguistic features, and edges encode probabilistic influence (or dependency). Due to the simplicity of this network, the number of dependencies in $p(r \mid l_1, \ldots, l_m)$ is large, which requires to estimate millions of parameters if there are more than twenty linguistic features.

In order to reduce the number of parameters without reducing the number of linguistic features, we will introduce language constructs in the form of hidden variables and set structural dependencies between the linguistic features and these language constructs. Guided by basic linguistic knowledge, we will detect sets of inter-dependent linguistic features and group them to a language construct consistent with the linguistic theory.

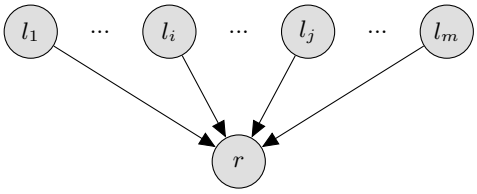
Examples of language constructs are the lexical difficulty $\mathbf{Lex}$, the syntactic difficulty $\mathbf{Syn}$, or the semantic difficulty $\mathbf{Sem}$. Those variables cannot be directly measured in a text (because they are not well defined), but are rather unknown functions of some other linguistic features, such as the length of a word (in characters or syllables), the amount of uppercase letters (e.g. in acronyms) or the presence of digits (e.g. protein names in biology). The graphical representation that introduces language constructs as hidden variables can be found in Figure 1b. Those hidden variables are typically introduced in joint probabilistic models as:

$$\begin{aligned}&\Pr(r, \mathscr{L}) = \Pr(r, l_1, \ldots, l_m) \\ &= \sum_{\mathbf{Lex}} \sum_{\mathbf{Syn}} \sum_{\mathbf{Sem}} \Pr(r, \mathbf{Lex}, \mathbf{Syn}, \mathbf{Sem}, l_1, \ldots, l_m)\end{aligned} \tag{4}$$

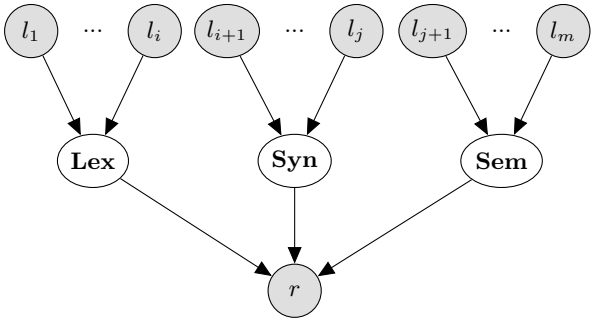
By inspecting Figure 1b, we can observe that readability score r is independent from observable linguistic features l_i given the language constructs $\mathbf{Lex}$, $\mathbf{Syn}$ and $\mathbf{Sem}$. Thus, we can rewrite Equation 4 to factorize over the graph in Figure 1b as:

$$\begin{aligned}&\Pr(r, l_1, \ldots, l_m) \\ &\quad\approx \sum_{\mathbf{Lex}} \sum_{\mathbf{Syn}} \sum_{\mathbf{Sem}} p(r \mid \mathbf{Lex}, \mathbf{Syn}, \mathbf{Sem}) \\ &\quad\quad \cdot p(\mathbf{Lex} \mid l_1, \ldots, l_i) \cdot p(\mathbf{Syn} \mid l_{i+1}, \ldots, l_j) \\ &\quad\quad \cdot p(\mathbf{Sem} \mid l_{j+1}, \ldots, l_m) \cdot p(l_1) \cdots p(l_m)\end{aligned} \tag{5}$$

where $l_1, \ldots, l_i$ are inter-dependent lexical features that somehow influence the lexical difficulty,



(a) Simple Bayesian network, where the readability score r depends on all linguistic features l_i. The number of parameters makes the estimation problem intractable.



(b) Structured Bayesian network that introduces language constructs (**Lex**, **Syn** and **Sem**) as hidden variables (white ellipses), with the purpose of reducing the dependencies of the readability score r from the rest of the linguistic features.

Figure 1: Graphical representations of causal networks. Arrows denote probabilistic influence.

$l_{i+1}, \ldots, l_j$ are syntactic features that influence syntactic difficulty, and the remaining are semantic features. Now the readability score r depends only on a small set of language constructs, which dramatically reduces the amount of parameters.

3.3 Estimating Parameter Values

Hidden variables and independency assumptions are often necessary to reduce the number of parameters that need to be estimated, specially when there are many variables or there is a limited amount of training data. In the factor graph of Figure 2, there is a conditional probability distribution (CPD) $\Pr(v \mid Pa_v)$ modeling the probability of every linguistic feature v given its parents Pa_v in the graph[2]. In this work, we make no assumptions on how a variable is related to its parents and we model this unknown relationship using non-parametric CPDs. The drawback is that we need to discretize the values of the linguistic variables and that the number of parameters[3] increases exponentially with the number of parent variables.

The estimation of the parameter values can be carried out using standard techniques that aim at optimizing the likelihood over the training data in presence of hidden variables. In this work, we used the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) for that purpose.

[2]If a variable v has no parents, then its CPD is $p(v)$.

[3]The term "non-parametric" might be misleading, since this type of CPDs have many parameters.

3.4 Querying the Model

Estimating the joint probability distribution $\Pr(r, \ell)$ has its advantages, since it gives us complete knowledge about the problem. In order to interpret the model, we can perform some insightful queries involving any variable.

Marginal Maximum a Posteriori is used to find the most probable value assignment to some linguistic features given some evidence. This query can mimic the functionality of discriminative models, where the objective is to find the most probable readability score $\hat{r}$ given the linguistic evidence ℓ^0, in presence of language constructs $\mathbf{L}$:

$$\mathrm{MAP}(\hat{r} \mid \ell^0) = \operatorname*{argmax}_r \sum_{\mathbf{L}} p(r, \mathbf{L} \mid \ell^0) \qquad (6)$$

where the conditional probability distribution $p(r, \mathbf{L} \mid \ell^0)$ can be found by using the Bayes rule:

$$p(r, \mathbf{L} \mid \ell^0) = \frac{p(r, \mathbf{L}, \ell^0)}{p(\ell^0)} \qquad (7)$$

Another application of marginal MAP queries is to gain linguistic insight about what characterizes unreadable texts. This insight could be obtained by querying the model in the opposite direction, i.e. $\mathrm{MAP}(\hat{\ell} \mid r^0)$, where we want to obtain the most plausible linguistic instantiation $\hat{\ell}$ given a certain readability level r^0. More complex queries can be similarly performed by conditioning the marginal MAP. For instance, the query $\mathrm{MAP}(\hat{\ell} \mid \mathbf{Lex}^{\mathrm{high}}, r^{\mathrm{good}})$ would result in the most plausible values of linguistic features that have a high lexical difficulty but a good readability.

Sensitivity analysis allows us to understand how sensitive a certain variable is to some observed linguistic features. In our study, we are interested in understanding what individual or combination of observable linguistic features influence most in the readability of a particular text. A common approach (Kjærulff and Madsen, 2007) is to compute the distance d between the joint probability distribution with different instantiations of the linguistic features under study.

4 Experiments

We first describe the data that we used to train our systems, and the data grounded on cognitive

effort that we used for validation. Then, we describe the full set of linguistic features and our baseline systems. Finally, we assess to what extent our Bayesian causal network is able to predict the specific parts of the documents that are difficult to read, and compare it to other systems.

4.1 Corpora

To estimate the parameters of the Bayesian causal network and our baseline systems, we opted to use texts from three corpora, namely Wikipedia Simple[4], Wikipedia English[5], and PubMed[6].

Wikipedia has been a valuable resource for the development of text transformation methods, such as summarization (Biadsy et al., 2008), or machine translation (Smith et al., 2010), among others. Wikipedia Simple is a relatively new version of the Wikipedia English, where articles are written in simple English[7]. Wikipedia English does not require any specific writing style other than clarity, precision and completeness. Finally, PubMed corpus is a large collection of academic biomedical articles, where readability is often sacrificed for precision and completeness. We assume that these three corpora have different expected readabilities (high, intermediate and low, respectively), and we use them as readability annotations at document level. Some linguistic features considered in our work are sensitive to text length (i.e. number of active lexical chains or average coreference distance). For this reason, we collected only abstracts from Wikipedia Simple that contain 10, 11 or 12 sentences, and randomly sampled from Wikipedia English and PubMed the same amount of long abstracts with the same text length distribution as Wikipedia Simple, totaling in $8,856$ abstracts.

Our hypothesis is that Bayesian causal networks are capable of recognizing specific parts of documents that make texts difficult to read. To test our hypothesis, we need documents with readability annotations at sub-document level. But such fine-grained annotations are difficult to obtain even for expert linguists because there are many linguistic variables involved in the annotation decisions.

In this work, we indirectly annotate the reading difficulty of every part of the text using an estimation of the expected cognitive effort required to understand that part of the text. There are several methods that have been proposed to measure moment-to-moment cognitive effort, such as functional magnetic resonance imaging (fMRI) to quantify activations of certain brain areas, or measurements in pupil size changes. However, those methods have difficulties in aligning cognitive effort spatially and temporally to segments in a text, and we opted to measure fixation time on individual words due to its relative simplicity. Thus, we work under the assumption that higher cognitive effort is reflected as longer fixation durations, since parts of the text that are difficult to read require longer cognitive processing time.

On the text side, we characterize a part of a text by a quantification of its linguistic features at word level. Let $f_{i,j}$ be the quantification of linguistic feature i at word w_j. As an example, linguistic feature "is noun", $f_{\text{noun},j} = 1$ if w_j is a noun. Non-binary linguistic features can be similarly quantified in the range $[0, 1]$ dividing their value by the maximum possible value. For features not defined at word level (e.g. "sentence length"), the feature quantification of words in the span are all equal to the quantification of the span.

In order to estimate fixation time T_i induced by every linguistic feature i, we accumulate fixation durations on words scaled by the quantification of every linguistic feature at those words, and normalize it by the total amount of fixation durations and total amount of feature quantification. Formally, let t_j be the total amount of fixation duration on word w_j. Then, fixation time T_i caused by linguistic feature i can be computed as:

$$T_i = \frac{\sum_j t_j \cdot f_{i,j}}{(\sum_j t_j) \cdot (\sum_j f_{i,j})} \tag{8}$$

We collected fixation durations on every word using the eye-tracker Tobii TX300, and used a text-gaze aligner (Martínez-Gómez et al., 2012) to correct the systematic errors introduced by the eye-tracker. There were 40 subjects participating in our study, and only the 20% of eye-tracking sessions with highest signal quality were selected for this study. Most subjects were non-native English speakers linked to academia, with varying language skills and background knowledge. They were asked to carefully read 2 documents on 3 topics (6 documents in total), about economics, nutrition and astronomy, and answer detailed questionnaires to assess their understanding. The average

[4] http://simple.wikipedia.org/

[5] http://en.wikipedia.org/

[6] http://www.ncbi.nlm.nih.gov/pubmed/

[7] Guidelines to write in simple English are proposed in Wikipedia Simple, but are not strictly enforced.

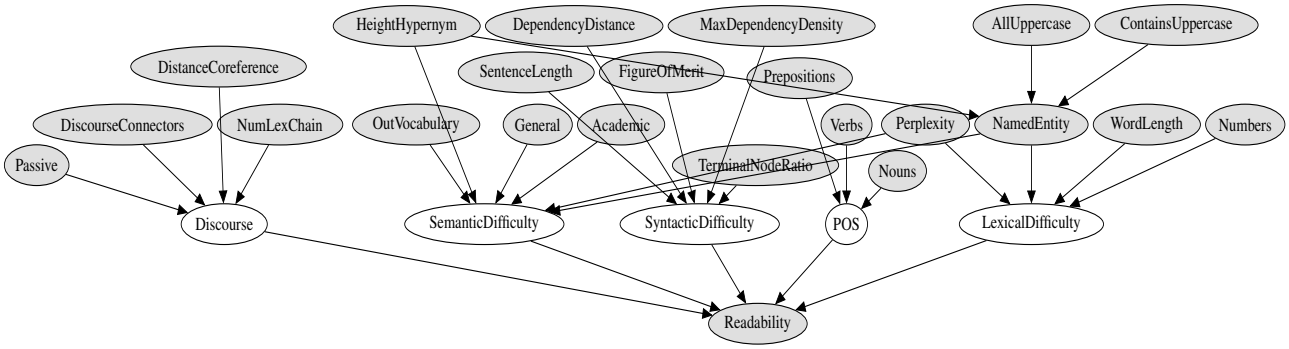


Figure 2: Graphical representation of our Bayesian causal network. Observable linguistic features are represented by white ellipses. Language constructs introduced as hidden variables are represented by gray ellipses. Directed edges indicate the direction of causality, and encode probabilistic influence.

duration of the reading and question-answering session was 1 hour, and every subject was compensated with the equivalent to 20 US dollars in cash at the end of the session. Documents contained 22.5 sentences and 469 words on average.

The objective of the Bayesian causal network will be to predict cognitive effort caused by each linguistic feature, and it will be compared to the results obtained by using discriminative methods.

4.2 Feature Set

Figure 2 shows the 22 linguistic features (gray ellipses) that were used in this work, the 5 language constructs that were introduced as hidden variables (white ellipses), and their probabilistic relationships (directed edges). The linguistic features that appear as ancestors of lexical difficulty and Part of Speech (POS) correspond to their average at token level (i.e. in the case of "Numbers", the percentage of tokens that are numbers).

Named entities were extracted using the NLTK toolkit (Bird et al., 2009), word lengths (in syllables) were computed by averaging the number of stresses in the CMU pronunciation dictionary (Weide, 1998). The perplexity was computed using Google 5-grams (Brants and Franz, 2006) with deleted interpolation tuned on a tokenized and non-lowercased separate subset of representative sentences from all three corpora. The percentage of prepositions, nouns and verbs was computed using the NLTK POS tagger.

Following the work in (Hudson, 1995) we considered the maximum dependency density and average distance between dependents as linguistic features that influence syntactic difficulty, computed using a dependency parser (Klein and Manning, 2003). Terminal node to non-terminal node ratio is another typical phrase-based measure of syntactic difficulty, and it was computed using an HPSG parser (Miyao and Tsujii, 2008). The figure of merit, as given by the same parser, is a function of the lexical probability rules that are triggered during the automatic parsing, and somehow represents the parsing surprise.

Height of hypernyms were computed as the average distance between token lemmas to the most abstract term in WordNet (Fellbaum, 2010) and measures how specific terms are. "General", "Academic" and "OutVocabulary" features denote the average number of words appearing in the General Word Service List (West and Jeffery, 1953), in the Academic Word List (Coxhead, 1998), or in none of them.

The average distance between mentions and their referents, and the maximum number of active lexical chains were computed using a coreference resolution system (Raghunathan et al., 2010) in a similar fashion to how the average dependency distance and maximum dependency density were computed to measure syntactic difficulty. Finally, the average number of passive clauses was computed using the output of the HPSG parser, and the percentage of tokens that are discourse connectors was measured checking the occurrence of every token in a hand-crafted list of 279 connectors.

4.3 Baseline

Using our `Bayesian` network, we computed the importance of each linguistic feature for every document as the sensitivity of the network conditioned on the observation of the rest of the variables. We compared our system to two baselines. The first baseline, `raw features`, measures the importance of linguistic features (across the cor-

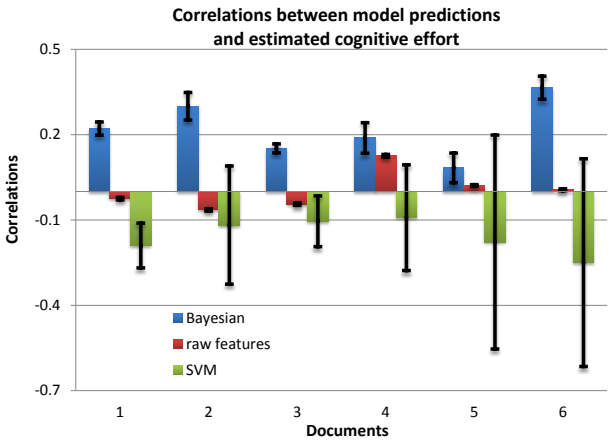


Figure 3: Correlations between predictions of feature impact on reading difficulty and the expected cognitive effort introduced by such features. Confidence intervals are computed at 95%.

pora) as the correlation between each linguistic feature and the readability score. As a second baseline, we chose SVM models due to their success in readability studies (François and Fairon, 2012). We measured their sensitivity to each linguistic feature, by observing the variations on the SVM response due to variations in each linguistic feature (Cortez, 2010), while holding the rest of linguistic variables to their average values. This baseline was built by training and tuning an SVM with a gaussian kernel in a cross-validation setup. Both baselines obtain quantifications of feature influence on readability independent of the document instantiation. For SVM sensitivity analysis, we also computed variability of SVM categorical response to changes in the linguistic feature under study while setting the rest of linguistic features to the instantiations on each document. However, there was a negligible variance in SVM response, and results are not reported for that experiment.

Confidence intervals for all systems were obtained by measuring prediction variability in 10 runs of random sampling with 90% of the data. All linguistic features were discretized in two intervals for the Bayesian network, except the readability score, which had three states (one for each class). This is an important loss of information for the Bayesian network, but it was necessary for computational reasons. SVM and raw features baselines, however, used continuous values.

4.4 Results

Figure 3 shows correlations between predictions of feature impact on reading difficulty and the expected cognitive effort introduced by such features. The $x-$axis corresponds to the identifier of each document for which we have estimated cognitive effort using the eye-tracker and the $y-$axis corresponds to correlations with the systems. Intervals for every prediction at a 95% confidence are displayed above and below each bar.

As it can be observed, raw features do not capture meaningfully cognitive effort and their correlations are close to zero, with a high confidence (narrow confidence intervals). The quantification on linguistic feature importance given by the SVM sensitivity analysis is slightly negative with large confidence intervals, which suggests that this type of analysis is not useful to predict reading difficulties in specific parts of the documents. The Bayesian causal network obtains mild, but consistent and positive correlations with the expected cognitive effort and its confidence intervals show strong significance.

Table 1 shows the most influential linguistic features on reading difficulty for documents 4 and 6. According to the cognitively-grounded reading difficulty, lexical perplexity (surprise), the occurrence of named entities, out of vocabulary words, passive clauses, academic words, nouns and abstraction (hypernyms) are the linguistic features that required longer fixation times in order to understand those documents. The Bayesian network ranked, on top 5, two and three of the most influential linguistic features for document 4 and 6.

5 Applications and Future Work

Bayesian causal networks for readability diagnosis have an immediate application to authoring systems, where the inference engine automatically detects text segments that make the text difficult to read. For that purpose, the average quantification of every linguistic feature has to be computed at document level. Then, causal reasoning (Bayesian sensitivity analysis) would be performed to find linguistic features with highest impact on reading difficulty for that specific document. Finally, instantiations of such linguistic features at segment level whose quantifications are above document average would be flagged for edition. Authors can then proceed to amend the text, or assert constraints. These constraints can take the form of "I want to increase readability without sacrificing the current lexical difficulty". Such constraints can be introduced using marginal MAPs as described

Document 4	Cognitive effort	Bayesian	SVM	raw features
feature 1	**Nouns**	General	Dependency density	General
feature 2	**Out Vocabulary**	**Out Vocabulary**	General	**Out Vocabulary**
feature 3	**Passive**	**Academic**	Contains uppercase	**Academic**
feature 4	**Academic**	Figure of Merit	Dependency distance	All uppercase
feature 5	**Height Hypernym**	Named entities	Verbs	Figure of Merit
Document 6	Cognitive effort	Bayesian	SVM	raw features
feature 1	**Perplexity**	**Named entities**	Dependency density	General
feature 2	**Named entities**	**Academic**	General	**Out Vocabulary**
feature 3	**Out Vocabulary**	General	Contains uppercase	**Academic**
feature 4	**Passive**	**Perplexity**	Dependency distance	All uppercase
feature 5	**Academic**	Figure of Merit	Verbs	Figure of Merit

Table 1: List of 5 most influential linguistic features for documents 4 and 6, sorted in descending order. The first column corresponds to the order given by cognitive effort. The rest of the columns correspond to predictions of systems. The Bayesian network finds 2 and 3 out of the 5 most influential features in documents 4 and 6. SVM and raw features provide constant estimations for all documents.

in Section 3.4. There are, however, features that cannot be tweaked individually and would require very complex user actions. Others are simply very difficult to handle by humans, as in the case of the terminal node to non-terminal node ratio.

In an automatic readability optimization setup, a set of transformation actions could be applied on a text, but discerning the most appropriate action can be challenging. Bayesian networks could be a solution to it, since they can infer the desirable configuration of linguistic values for a certain readability level in a given document, and what actions would lead to the largest readability gain.

The remaining challenges when working with non-parametric Bayesian networks are two. The first one is the necessary loss of information that occurs when discretizing features, and parametric models are possible solutions. Finding better network topologies is also an interesting challenge that brings linguistic insights into readability studies and increases the predictive power of the model. One approach is to refine the network using more thoughtful linguistic knowledge. Another possibility is to automatically estimate the optimal network topology driven by data, but causal properties could be difficult to preserve.

We used indirect measurements of cognitive effort that rely on the computation of a normalized fixation time on every linguistic feature. Fixation durations were recorded using a precise eye-tracker, but data collection is rarely exempt of systematic errors and new methods to estimate cognitive effort should account for this degraded calibration. Moreover, certain aspects of cognitive effort might not be reflected by fixation times, and other features of eye movements, such as regressions or changes in pupil diameter can be valuable.

Since estimations of feature impact on readability depends on each document, it was difficult to compare our findings to prior work. Future investigations in readability diagnosis would benefit from a combination of indirect measurements of cognitive effort and readability annotations by linguistic experts at sub-document level, that could be shared within the research community.

6 Conclusions

Discriminative models are built to predict readability and correlate well with human judgment. Those models are good readability predictors, but fail at explaining the causes of unreadability. With the intention of assisting humans to optimize readability or to fully automate it, we need methods able to infer the causes of readability.

We have presented the application of Bayesian causal networks to build generative readability models. To reduce the number of dependencies between linguistic features, we introduced language constructs as hidden variables and estimated the parameter values using the EM algorithm.

Using our proposed Bayesian causal network, we measured the impact of every linguistic feature in presence of all other variables, and compared the prediction accuracy to grounded cognitive effort. Our method showed significant and positive correlations with cognitive effort, suggesting that it is able to capture linguistic features that cause difficulties in reading for specific documents.

References

S. Aluísio, L. Specia, C. Gasperin, and C. Scarton. 2010. Readability assessment for text simplification. In *Proc. of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9. ACL.

F. Biadsy, J. Hirschberg, and E. Filatova. 2008. An unsupervised approach to biography production using Wikipedia. In *Proc. of ACL-08: HLT*, pages 807–815.

S. Bird, E. Klein, and E. Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

T. Brants and A. Franz. 2006. Web 1T 5-gram vers. 1.

J. Carroll, G. Minnen, D. Pearce, Y. Canning, S. Devlin, and J. Tait. 1999. Simplifying text for language-impaired readers. In *Proc. of EACL*, volume 99, pages 269–270.

P. Cortez. 2010. Data mining with neural networks and support vector machines using the r/rminer tool. In *Advances in Data Mining. Applications and Theoretical Aspects*, pages 572–583. Springer.

A. Coxhead. 1998. *An academic word list*, volume 18. School of Linguistics and Applied Language Studies, Victoria University of Wellington.

A. Davison and R.N. Kantor. 1982. On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly*, pages 187–209.

A.P. Dempster, N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society*.

S. Devlin, J. Tail, Y. Canning, J. Carroll, G. Minnen, and D. Pearce. 1999. The application of assistive technology in facilitating the comprehension of newspaper text by aphasic people. *Assistive Technology on the Threshold of the New Millennium*.

C. Fellbaum. 2010. WordNet. *Theory and Applications of Ontology: Computer Applications*, pages 231–243.

R. Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

T. François and C. Fairon. 2012. An AI readability formula for French as a foreign language. In *Proc. of the 2012 Joint Conference on EMNLP and CoNLL*, pages 466–477. ACL.

E. Fry. 1990. A readability formula for short passages. *Journal of Reading*, pages 594–597.

M. J. 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 *Proc. of NAACL HLT*, pages 460–467.

R.A. Hudson. 1995. Measuring syntactic difficulty. *Manuscript, University College, London*.

R.J. Kate, X. Luo, S. Patwardhan, M. Franz, R. Florian, R.J. Mooney, S. Roukos, and C. Welty. 2010. Learning to predict readability using diverse linguistic features. In *Proc. of the 23rd COLING*, pages 546–554. ACL.

J.P. Kincaid, R. Fishburne, R.L. Rogers, and B.S. Chissom. 1975. Derivation of new readability formulas for navy enlisted personnel.

U. B. Kjærulff and A. L. Madsen. 2007. *Bayesian Networks and Influence Diagrams*. Information science and statistics. Springer New York.

D. Klein and C. Manning. 2003. Accurate unlexicalized parsing. In *Proc. of the 41st ACL*, pages 423–430. ACL.

P. Martínez-Gómez, C. Chen, T. Hara, Y. Kano, and A. Aizawa. 2012. Image registration for text-gaze alignment. In *Proc. of the IUI'12*, pages 257–260.

G.H. Mc Laughlin. 1969. SMOG grading-a new readability formula. *Journal of reading*, pages 639–646.

Y. Miyao and J. Tsujii. 2008. Feature forest models for probabilistic HPSG parsing. *Computational Linguistics*, 34:35–80, March.

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

K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and C. Manning. 2010. A multi-pass sieve for coreference resolution. In *Proc. of the 2010 on EMNLP*, pages 492–501. ACL.

K. Rayner and S.A. Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 14(3):191–201.

L. Si and J. Callan. 2001. A statistical model for scientific readability. In *Proc. of the 10th CIKM*, pages 574–576. ACL.

A. Siddharthan. 2003. *Syntactic Simplification and Text Cohesion*. Ph.D. thesis, University of Cambridge.

J.R. Smith, C. Quirk, and K. Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *HLT: The 2010 NA-ACL*, pages 403–411. ACL.

R.L. Weide. 1998. The CMU pronunciation dictionary, release 0.6.

M. West and G.B. Jeffery. 1953. *A general service list of English words*. Longmans, Green London.

S. Williams and E. Reiter. 2005. Generating readable texts for readers with low basic skills. In *Proc. of the 10th European Workshop on Natural Language Generation*.