Automatic Text Simplification

Synthesis Lectures on Human Language Technologies

Grammatical Inference for Computational Linguistics
Jeffrey Heinz, Colin de la Higuera, and Menno van Zaanen
2015

Automatic Detection of Verbal Deception
Eileen Fitzpatrick, Joan Bachenko, and Tommaso Fornaciari
2015

Natural Language Processing for Social Media
Atefeh Farzindar and Diana Inkpen
2015

Semantic Similarity from Natural Language and Ontology Analysis
Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain
2015

Learning to Rank for Information Retrieval and Natural Language Processing, Second Edition
Hang Li
2014

Ontology-Based Interpretation of Natural Language
Philipp Cimiano, Christina Unger, and John McCrae
2014

Automated Grammatical Error Detection for Language Learners, Second Edition
Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault
2014

Web Corpus Construction
Roland Schäfer and Felix Bildhauer
2013

Recognizing Textual Entailment: Models and Applications
Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto
2013

Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax
Emily M. Bender
2013

Semi-Supervised Learning and Domain Adaptation in Natural Language Processing
Anders Søgaard
2013

Semantic Relations Between Nominals
Vivi Nastase, Preslav Nakov, Diarmuid Ó Séaghdha, and Stan Szpakowicz
2013

Computational Modeling of Narrative
Inderjeet Mani
2012

Natural Language Processing for Historical Texts
Michael Piotrowski
2012

Sentiment Analysis and Opinion Mining
Bing Liu
2012

Discourse Processing
Manfred Stede
2011

Bitext Alignment
Jörg Tiedemann
2011

Linguistic Structure Prediction
Noah A. Smith
2011

Learning to Rank for Information Retrieval and Natural Language Processing
Hang Li
2011

Computational Modeling of Human Language Acquisition
Afra Alishahi
2010

Introduction to Arabic Natural Language Processing
Nizar Y. Habash
2010

Cross-Language Information Retrieval
Jian-Yun Nie
2010

Automated Grammatical Error Detection for Language Learners
Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault
2010

Data-Intensive Text Processing with MapReduce
Jimmy Lin and Chris Dyer
2010

Semantic Role Labeling
Martha Palmer, Daniel Gildea, and Nianwen Xue
2010

Spoken Dialogue Systems
Kristiina Jokinen and Michael McTear
2009

Introduction to Chinese Natural Language Processing
Kam-Fai Wong, Wenjie Li, Ruifeng Xu, and Zheng-sheng Zhang
2009

Introduction to Linguistic Annotation and Text Analytics
Graham Wilcock
2009

Dependency Parsing
Sandra Kübler, Ryan McDonald, and Joakim Nivre
2009

Statistical Language Models for Information Retrieval
ChengXiang Zhai
2008

Automatic Text Simplification

Horacio Saggion

www.morganclaypool.com

ISBN: 9781627058681 paperback
ISBN: 9781627058698 ebook

DOI 10.2200/S00700ED1V01Y201602HLT032

A Publication in the Morgan & Claypool Publishers series
SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES

Lecture #32
Series Editor: Graeme Hirst, *University of Toronto*
Series ISSN
Print 1947-4040 Electronic 1947-4059

Automatic Text Simplification

Horacio Saggion
Department of Information and Communication Technologies
Universitat Pompeu Fabra

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES #32

MORGAN & CLAYPOOL PUBLISHERS

ABSTRACT

Thanks to the availability of texts on the Web in recent years, increased knowledge and information have been made available to broader audiences. However, the way in which a text is written—its vocabulary, its syntax—can be difficult to read and understand for many people, especially those with poor literacy, cognitive or linguistic impairment, or those with limited knowledge of the language of the text. Texts containing uncommon words or long and complicated sentences can be difficult to read and understand by people as well as difficult to analyze by machines. Automatic text simplification is the process of transforming a text into another text which, ideally conveying the same message, will be easier to read and understand by a broader audience. The process usually involves the replacement of difficult or unknown phrases with simpler equivalents and the transformation of long and syntactically complex sentences into shorter and less complex ones. Automatic text simplification, a research topic which started 20 years ago, now has taken on a central role in natural language processing research not only because of the interesting challenges it posesses but also because of its social implications. This book presents past and current research in text simplification, exploring key issues including automatic readability assessment, lexical simplification, and syntactic simplification. It also provides a detailed account of machine learning techniques currently used in simplification, describes full systems designed for specific languages and target audiences, and offers available resources for research and development together with text simplification evaluation techniques.

KEYWORDS

syntactic simplification, lexical simplification, readability measures, text simplification systems, text simplification evaluation, text simplification resources

To Sandra, Jonas, Noah, and Isabella

Contents

Acknowledgments

I am indebted to my fellow colleagues Stefan, Sanja, Biljana, Susana, Luz, Daniel, Simon, and Montserrat for sharing their knowledge and expertise with me.

Horacio Saggion
January 2017

CHAPTER 1

Introduction

Automatic text simplification is a research field in computational linguistics that studies methods and techniques to simplify textual content. Text simplification methods should facilitate or at least speed up the adaptation of available and future textual material, making accessible information *for all* a reality. Usually (but not necessarily), adapted texts would have information loss and a simplistic style, which is not necessarily a bad thing if the message of the text, which was in the beginning complicated, can in the end be understood by the target reader. Text simplification has also been suggested as a potential pre-processing step for making texts easier to handle by generic text processors such as parsers, or to be used in specific information access tasks such as information extraction. Simplifying for people is more challenging than the second use of simplification because the output of the automatic system could be perceived as inadequate in the presence of the least error.

The interest in automatic text simplification has grown in recent years and in spite of the many approaches and techniques proposed, automatic text simplification is, as of today, far from perfect. The growing interest in text simplification is evidenced by the number of languages which are targeted by researchers worldwide. Simplification systems and simplification studies exist at least for English [Carroll et al., 1998, Chandrasekar et al., 1996, Siddharthan, 2002], Brazilian Portuguese [Aluísio and Gasperin, 2010], Japanese [Inui et al., 2003], French [Seretan, 2012], Italian [Barlacchi and Tonelli, 2013, Dell'Orletta et al., 2011], Basque [Aranzabe et al., 2012], and Spanish [Saggion et al.].

1.1 TEXT SIMPLIFICATION TASKS

Although there are many text characteristics which can be modified in order to make a text more readable or understandable, including the way in which the text is presented, automatic text simplification has usually concentrated on two different tasks—lexical simplification and syntactic simplification—each addressing different sub-problems.

Lexical simplification will attempt to either modify the vocabulary of the text by choosing words which are thought to be more appropriate for the reader (i.e., transforming the sentence "The book was magnificent" into "The book was excellent") or to include appropriate definitions (e.g., transforming the sentence "The boy had tuberculosis." into "The boy had tuberculosis, a disease of the lungs."). Changing words in context is not an easy task because it is almost certain that the original meaning will be confused.

Syntactic simplification will try to identify syntactic phenomena in sentences which may hinder readability and comprehension in an effort to possibly transform the sentence into more readable or understandable equivalents. For example, relative or subordinate clauses or passive constructions, which may be very difficult to read by certain readers, could be transformed into simpler sentences or into active form. For example, the sentence "The festival was held in New Orleans, which was recovering from Hurricane Katrina" could be transformed without altering the original too much into "The festival was held in New Orleans. New Orleans was recovering from Hurricane Katrina."

As we shall later see, automatic text simplification is related to other natural language processing tasks such as text summarization and machine translation. The objective of text summarization is to reduce a text to its essential content which might be useful in simplification on occasions where the text to simplify has too many unnecessary details. The objective of machine translation is to translate a text into a semantic equivalent in another language. A number of recent automatic text simplification approaches cast text simplification as statistical machine translation; however, this approach to simplification is currently limited by the scarcity of parallel simplification data.

There is an important point to mention here: although lexical and syntactic simplification usually have been addressed separately, they are naturally related. If during syntactic simplification a particular syntactic structure is chosen to replace a complex construction, it also might be necessary to apply transformations at the lexical level to keep the text grammatical. Furthermore, with a text being a coherent and cohesive unit, any change at a local level (words or sentences) certainly will affect in one way or another textual properties (at the local and global level): for example replacing a masculine noun with a feminine synonym during lexical simplification will certainly require some languages to repair local elements such as determiners and adjectives, as well as pronouns or definite expressions in following or preceding sentences. Pragmatic aspects of the text, such as the way in which the original text has been created to communicate a message to specific audiences, are generally ignored by current systems.

As we shall see in this book, most approaches treat text simplification as a sequence of transformations at the word or sentence level, disregarding the global textual content (previous and following text units), thereby affecting important properties such as cohesion and coherence.

1.2 HOW ARE TEXTS SIMPLIFIED?

Various studies have investigated ways in which a given text is transformed into an easier-to-read version. In order to understand what text transformations would be needed and what transformations could be implemented automatically, Petersen and Ostendorf [2007] performed an analysis of a corpus of original and abridged CNN news articles in English (114 pairs), distributed by the Literacyworks organization,[1] aimed at adult learners (i.e., native speakers of English with poor reading skills). They first aligned the original and abridged versions of the news articles looking

[1] http://literacynet.org/

for the occurrence of an original-version sentence corresponding to a sentence in the abridged version. After having aligned the corpus, they observed that sentences from the original documents can be dropped (around 30%) or aligned to one (47% of same sentences) or more sentences (19%) in the abridged version (splits). The one-to-one alignments correspond to cases where the original sentence is kept practically untouched, cases where only part of the original sentence is kept, and cases of major re-writing operations. A small fraction of pairs of the original sentences were also aligned to a single abridged sentence, accounting for merges. Petersen and Ostendorf's study also tries to automatically identify sentences in the original document which should be split since those would be good candidates for simplification. Their approach consists of training a decision-tree learning algorithm (C4.5 [Quinlan, 1993]) to classify a sentence into split or non-split. They used various features including sentence length and several statistics on POS tags and syntactic constructions. Cross-validation evaluation experiments show that it is difficult to differentiate between the two classes; moreover, sentence length is the most informative feature, which explains much of the classification performance. Another interesting contribution is the study of dropped sentences, for which they train a classifier with some features borrowed from summarization research; however, the classifier is only slightly better than a majority baseline (i.e., not drop).

In a similar way, Bott and Saggion [2011b] and Drndarević and Saggion [2012a,b] identified a series of transformations that trained editors apply to produce simplified versions of documents. Their case in notably different from Petersen and Ostendorf [2007] given the characteristics of the language—Spanish—and target population of the simplified text version: people with cognitive disabilities. Bott and Saggion [2011b] analyzed a sample of sentence-aligned original and simplified documents to identify expected simplification operations such as sentence split, sentence deletion, and various types of change operations (syntactic, lexical, etc.). Moreover, additional operations such as insertion and reordering were also documented. Drndarević and Saggion [2012a,b] specifically concentrate on identifying lexical changes, in addition to synonym substitution, cases of numerical expression re-writing (e.g., rounding), named entity reformulation, and insertion of simple definitions. Like Petersen and Ostendorf [2007], Drndarević and Saggion train a Support Vector Machine (SVM) algorithm [Joachims, 1998] to identify sentences which could be deleted, improving over a robust baseline that always deletes the last sentence of the document.

1.3 THE NEED FOR TEXT SIMPLIFICATION

The creation of text simplification tools without considering a particular target population could be justifiable in that aspects of text complexity affect a large range of users with reading difficulties. For example, long and syntactically complex sentences are generally hard to process. Some particular sentence constructions, such as syntactic constructions which do not follow the canonical subject-verb-object (e.g., passive constructions), may be an obstacle for people with aphasia [Devlin and Unthank, 2006] or an autism spectrum disorder (ASD) [Yaneva et al., 2016b]. The

same is true for very difficult or specialized vocabulary and infrequent words which can also prove difficult to understand for people with aphasia [Carroll et al., 1998, Devlin and Unthank, 2006] and ASD [Norbury, 2005]. Moreover, there are also certain aspects of language that prove difficult to specific groups of readers. Language learners, for example, may have a good capacity to infer information, although they may have a very restricted lexicon and may not be able to understand certain grammatical constructions. Dyslexic readers, in turn, do not have a problem with language understanding *per se*, but with the understanding of the written representation of language. In addition, readers with dyslexia were found to read better when using more frequent and shorter words [Rello et al., 2013b]. Finally, people with intellectual disabilities may have problems processing and retaining large amounts of information [Fajardo et al., 2014, Feng et al., 2009].

In order to create adapted versions for specific populations, various initiatives exist which promote accessible texts. An early proposal is Basic English, a language of reduced vocabulary of just over 800 word forms and a restricted number of grammatical rules. It was conceived after World War II as a tool for international communication or a kind of interlingua [Ogden, 1937]. Other initiatives are Plain English (see "Language for Special Purposes" in Crystal [1987]), for English in the U.S. and U.K., and the Rational French, a French-controlled language to make technical documentation more accessible in the context of the aerospace industry [Barthe et al., 1999]. In Europe, there are associations dedicated to the adaptation of text materials (books, leaflets, laws, official documents, etc.) for people with disabilities or low literacy levels, examples of which are the Easy-to-Read Network in Scandinavian countries, the Asociación Lectura Fácil[2] in Spain, and the Centrum för Lättläst in Sweden.[3] These associations usually provide guidance or recommendation about how to prepare or adapt textual material. Some such recommendations are as follows:

- use simple and direct language;
- use one idea per sentence;
- avoid jargon and technical terms;
- avoid abbreviations;
- structure text in a clear and coherent way;
- use one word per concept;
- use personalization; and
- use active voice.

[2] http://www.lecturafacil.net/
[3] http://www.lattlast.se/

These recommendations, although intuitive, are sometimes difficult to operationalize (for both humans and machines) and sometimes even impossible to follow, especially in the case of adapting an existing piece of text.

1.4 EASY-TO-READ MATERIAL ON THE WEB

Although adapted texts have been produced for many years, nowadays there is a plethora of simplified material on the Web. The Swedish "easy-to-read" newspaper *8 Sidor*[4] is published by the Centrum för Lättläst to allow people access to "easy news." Other examples of similarly oriented online newspapers and magazines are the Norwegian *Klar Tale*,[5] the Belgian *l'Essentiel*[6] and *Wablie*,[7] the Danish *Radio Ligetil*,[8] the Italian *Due Parole*,[9] and the Finnish *Selo-Uutiset*.[10] For Spanish, the Noticias Fácil website[11] provides easy-to-read news for people with disabilities. The Literacyworks website[12] offers CNN news stories in original and abridged (or simplified) formats, which can be used as learning resources for adults with poor reading skills. At the European level, the Inclusion Europe website[13] provides good examples of how full text simplifications and simplified summaries in various European languages can provide improved access to relevant information. The Simple English Wikipedia[14] provides encyclopedic content which is more accessible than plain Wikipedia articles because of the use of simple language and simple grammatical structures. There are also initiatives which aim to give access to easy-to-read material in particular and web accessibility in general the status of a legal right.

The number of websites containing manually simplified material pointed out above clearly indicates a need for simplified texts. However, manual simplification of written documents is very expensive and manual methods will be not cost-effective, especially if we consider that news is constantly being produced and therefore simplification would, in turn, need to keep the same pace. Nevertheless, there is a growing need for methods and techniques to make texts more accessible. For example, people with learning disabilities who need simplified text constitute 5% of the population. However, according to data from the Easy-to-Read Network,[15] if we consider people who cannot read documents with heavy information load or documents from authorities or governmental sources, the percentage of people in need of simplification jumps to 25% of

[4] http://8sidor.lattlast.se
[5] http://www.klartale.no
[6] http://www.journal-essentiel.be/
[7] http://www.wablieft.be
[8] http://www.dr.dk/Nyheder/Ligetil/Presse/Artikler/om.htm
[9] http://www.dueparole.it
[10] http://papunet.net/selko
[11] http://www.noticiasfacil.es
[12] http://www.literacyworks.org/learningresources/
[13] http://www.inclusion-europe.org
[14] http://simple.wikipedia.org
[15] http://www.easytoread-network.org/

the population.[16] In addition, the need for simplified texts is becoming more important as the incidence of disability increases as the population ages.

1.5 STRUCTURE OF THE BOOK

Having briefly introduced what automatic text simplification is and the need for such technology, the rest of the book will cover a number of relevant research methods in the field which have been the object of scientific inquiry for more than 20 years. Needless to say, many relevant works will not be addressed here; however, we have tried to cover most of the techniques which have been used, or are being used, at the time of writing. In Chapter 2, we will provide an overview of the topic of readability assessment given its current relevance in many approaches to automatic text simplification. In Chapter 3, we will address techniques which have been proposed to address the problem of replacing words and phrases by simpler equivalents: the lexical simplification problem. In Chapter 4, we will cover techniques which can be used to simplify the syntactic structure of sentences and phases, with special emphasis on rule-based linguistically motivated approaches. Then in Chapter 5, machine learning techiques, optimization, and other statistical techniques to "learn" simplification systems will be described. Chapters 6 and 7 cover very related topics—in Chapter 6 we will present fully fledged text simplification systems which have as users specific target populations, while in Chapter 7, we will cover sub-systems or methods specifically based on targeted tasks or user characteristics. In Chapter 8, we will cover two important topics: the available datasets for experimentation in text simplification and the current text simplification evaluation techniques. Finally, in Chapter 9, we close with an overview of the field and critical view of the current state of the art.

[16]Bror Tronbacke, personal communication, December 2010.

CHAPTER 2

Readability and Text Simplification

A key question in text simplification research is the identification of the complexity of a given text so that a decision can be made on whether or not to simplify it. Identifying the complexity of a text or sentence can help assess whether the output produced by a text simplification system matches the reading ability of the target reader. It can also be used to compare different systems in terms of complexity or simplicity of the produced output. There are a number of very complete surveys on the relevant topic of text readability which can be understood as "what makes some texts easier to read than others" [Benjamin, 2012, Collins-Thompson, 2014, DuBay, 2004]. Text readability, which has been investigated for a long time in academic circles, is very close to the "to simplify or not to simplify" question in automatic text simplification. Text readability research has often attempted to devise mechanical methods to assess the reading difficulty of a text so that it can be objectively measured. Classical mechanical text readability formulas combine a number of proxies to obtain a numerical score indicative of the difficulty of a text. These scores could be used to place the texts in an appropriate grade level or used to sort text by difficulty.

2.1 INTRODUCTION

Collins-Thompson [2014]—citing [Dale and Chall, 1948b]—defines text readability as the sum of all elements in textual material that affect a reader's understanding, reading speed, and level of interest in the material. The ability to quantify the readability of a text has long been a topic of research, but current technology and the availability of massive amounts of text in electronic form has changed research in computational readability assessment, considerably. Today's algorithms take advantage of advances in natural language processing, cognition, education, psycholinguistics, and linguistics ("all elements in textual material") to model a text in such a way that a machine learning algorithm can be trained to compute readability scores for texts. Traditional readability measures were based on semantic familiarity of words and the syntactic complexity of sentences. Proxies to measure such elements are, for example, the number of syllables of words or the average number of words per sentence. Most traditional approaches used averages over the set of basic elements (words or sentences) in the text, disregarding order and therefore discourse phenomena. The obvious limitations of early approaches were always clear: words with many syllables are not necessarily complex (e.g., children are probably able to read or understand complex dinosaur names or names of Star Wars characters before more-common words are acquired) and short

sentences are not necessarily easy to understand (poetry verses for example). Also, traditional formulas were usually designed for texts that were well formatted (not web data) and relatively long. Most methods are usually dependent on the availability of graded corpora where documents are annotated with grade levels. The grades can be either categorical or ordinal, therefore giving rise to either classification or regression algorithmic approaches. When classification is applied, precision, recall, f-score, and accuracy can be used to measure classification performance and compare different approaches. When regression is applied, Root Mean Squared Error (RMSE) or a correlation coefficient can be used to evaluate the algorithmic performance. In the case of regression, assigning a grade of 4 to a 5th-grade text (1 point difference) is not as serious a mistake as it would be to assign a grade 7 to a 5th-grade text (2 points difference). Collins-Thompson [2014] presents an overview of groups of features which have been accounted for in the readability literature including:

- lexico-semantic (vocabulary) features: relative word frequencies, type/token ratio, probabilistic language model measures such as text probability, perplexity, etc., and word maturity measures;
- psycholinguistic features: word age-of-acquisition, word concreteness, polysemy, etc.;
- syntactic features (designed to model sentence processing time): sentence length, parse tree height, etc.;
- discourse features (designed to model text's cohesion and coherence): coreference chains, named entities, lexical tightness, etc.; and
- semantic and pragmatic features: use of idioms, cultural references, text type (opinion, satire, etc.), etc.

Collins-Thompson argues that in readability assessment it seems the model used—the features—is more important than the machine learning approach chosen. That is, a well-designed set of features can go a long way in readability assessment.

2.2 READABILITY FORMULAS

DuBay [2004] points out that over 200 readability formulas existed by the 1980s. Many of them have been empirically tested to assess their predictive power usually by correlating their outputs with grade levels associated with text sets.

Two of the most widely used readability formulas are the Flesch Reading Ease Score [Flesch, 1949] and the Flesch-Kincaid readability formula [Kincaid et al.]. The Flesch Reading Ease Score uses two text characteristics as proxies: the average sentence length *ASL* and the average number of syllables per word *ASW* which are combined in Formula (2.1):

$$\text{Score} = 206.835 - (1.015 * ASL) - (84.6 * ASW). \tag{2.1}$$

On a given text the score will produce a value between 1 and 100 where the higher the value the easier the text would be. Documents scoring 30 are very difficult to read while those scoring 70 should be easy to read.

The Flesch-Kincaid readability formula (2.2) simplifies the Flesch score to produce a "grade level" which is easily interpretable (i.e., a text with a grade level of eight according to the formula could be thought appropriate for an eighth grader).

$$\text{GL} = (0.4 * ASL) + (12 * ASW) - 15. \tag{2.2}$$

Additional formulas used include the FOG readability score [Gunning, 1952] and the SMOG readability score [McLaughlin, 1969]. They are computed using the following equations:

$$\text{FOG} = 0.4 * (ASL + HW) \tag{2.3}$$

$$\text{SMOG} = 3 + \sqrt{PSC}, \tag{2.4}$$

where *HW* is the percent of "hard" words in the document (a hard word is one with at least three syllables) and *PSC* is the polysyllable count—the number of words with 3 or more syllables in 30 sentences which shall be picked from the beginning, middle, and end of the document.

Work on readability assessment has also included the idea of using a vocabulary or word list which may contain words together with indications of age at which the particular words should be known [Dale and Chall, 1948a]. These lists are useful to verify whether a given text deviates from what should be known at a particular age or grade level, constituting a rudimentary form of readability language model.

Readability measures have begun to take center stage in assessing the output of text simplification systems; however, their direct applicability is not without controversy. First, a number of recent studies have considered classical readability formulas [Wubben et al., 2012, Zhu et al., 2010], applying them to sentences, while many studies on the design of readability formulas are based on considerable samples from the text to assess or need to consider long text pieces to yield good estimates; their applicability at the sentence level would need to be re-examined because empirical evidence is still needed to justify their use. Second, a number of studies suggest the use of readability formulas as a way to guide the simplification process (e.g., De Belder [2014], Woodsend and Lapata [2011]). However, the manipulation of texts to match a specific readability score may be problematic since chopping sentences or blindly replacing words could produce totally ungrammatical texts, thereby "cheating" the readability formulas (see, for example, Bruce et al. [1981], Davison et al. [1980]).

2.3 ADVANCED NATURAL LANGUAGE PROCESSING FOR READABILITY ASSESSMENT

Over the last decade, traditional readability assessment formulas have been criticized [Feng et al., 2009]. The advances brought forward in areas of natural language processing made possible a

whole new set of studies in the area of readability. Current natural language processing studies in the area of readability assessment rely on automatic parsing, availability of psycholinguistic information, and language modeling techniques [Manning et al., 2008] to develop more robust methods. Today it is possible to extract rich syntactic and semantic features from text in order to analyze and understand how they interact to make the text more or less readable.

2.3.1 LANGUAGE MODELS

Various works have considered corpus-based statistical methods for readability assessment. Si and Callan [2001] cast text readability assessment as a text classification or categorization problem where the classes could be grades or text difficulty levels. Instead of considering just surface linguistic features, they argue, quite naturally, that the content of the document is a key factor contributing to its readability. After observing that some surface features such as syllable count were not useful predictors of grade level in the dataset adopted (syllabi of elementary and middle school science courses of various readability levels from the Web), they combined a unigram language model with a sentence-length language model in the following approach:

$$P_c(g|d) = \lambda * P_a(g|d) + (1 - \lambda) * P_b(g|d), \tag{2.5}$$

where g is a grade level, d is the document, P_a is a unigram language model, P_b is a sentence-length distribution model, and λ is a coefficient adjusted to yield optimal performance. Note that probability parameters in P_a are words, that is the document should be seen as $d = w_1 \dots w_n$ with w_l the word at position l in the document, while in probability P_b the parameters are sentence lengths, so a document with k sentences should be thought as $d = l_1 \dots l_k$ with l_i the length of the i-th sentence. The P_a probability distribution is a unigram model computed in the usual way using Bayes's theorem as:

$$P_a(g|d) = \frac{P(d|g)P(g)}{P(d)}. \tag{2.6}$$

The probabilities are estimates obtained by counting events over a corpus. Where P_b is concerned, a normal distribution model with specific mean and standard deviation is proposed. The combined model of content and sentence length achieves an accuracy of 75% on a blind test set, while the Flesch-Kincaid readability score will just predict 21% of the grades correctly.

2.3.2 READABILITY AS CLASSIFICATION

Schwarm and Ostendorf [2005] see readability assessment as classification and propose the use of SVM algorithms for predicting the readability level of a text based on a set of textual features. In order to train a readability model, they rely on several sources: (i) documents collected from the *Weekly Reader*[1] educational newspaper with 2nd–5th grade levels; (ii) documents from the *Encyclopedia Britannica* dataset compiled by Barzilay and Elhadad [2003] containing original

[1] http://www.weeklyreader.com

encyclopedic articles (115) and their corresponding children's versions (115); and (iii) CNN news stories (111) from the LiteracyNet[2] organization available in original and abridged (or simplified) versions. They borrow the idea of Si and Callan [2001], thus devising features based on statistical language modeling. More concretely, given a corpus of documents with say grade k, they create a language model for that grade. Taking 3-gram sequences as units for modeling the text, the probability $p(w)$ of a word sequence $w = w_1 \dots w_n$ in the k-grade corpus is computed as:

$$p(w) = p(w_1)p(w_2|w_1) * \prod_{i=3}^{n} p(w_i|w_{i-1}, w_{i-2}). \tag{2.7}$$

where the 3-gram probabilities are estimated using 3-gram frequencies observed in the k-grade documents and smoothing techniques to account for unobserved events. Given the probabilities of a sequence w in the different models (one per grade), a likelihood ratio of sequence w is defined as:

$$LR(w, k) = \frac{p(w|k)p(k)}{\sum_{c \neq k} p(w|c)p(c)}, \tag{2.8}$$

where the prior $p(k)$ probabilities can be assumed to be uniform. The $LR(w, k)$ values already give some information on the likelihood of the text being of a certain complexity or grade. Additionally, the authors use perplexity as an indicator of the fit of a particular text to a given model where low perplexity for a text t and model m would indicate a better fit of t to m. Worth noting is the reduction of the features of the language models based on feature filtering by information gain (IG) values to 276 words (the most discriminative) and 56 part of speech tags (for words not selected by IG). SVMs are trained using the graded dataset (*Weekly Reader*), where each text is represented as a set of features including traditional readability assessment superficial features such as average sentence length, average number of syllables per word, and the Flesch-Kincaid index together with more-sophisticated features such as syntax-based features, vocabulary features, and language model features. Syntax-based features are extracted from parsed sentences [Charniak, 2000] and include average parse tree height, average number of noun phrases, average number of verb phrases, and average number of clauses (SBARs in the Penn Treebank tag set[3]). Vocabulary features account for out-of-vocabulary (OOV) word occurrences in the text. These are computed as percentages of words or word types not found in the most common 100, 200, and 500 words occurring in 2nd-grade texts. Concerning language model features, there are 12 perplexity values for 12 different language models computed using 12 different combinations of the paired datasets Britannica/CNN (adults vs. children) and three different n-grams: unigrams, bigrams, and trigrams (combining discriminative words and POS tags). The authors obtained better results in comparison to traditional readability formulas when their language model features are used in combination with vocabulary features, syntax-based features, and superficial indicators. Petersen and Ostendorf [2007] extend the previous work by considering additional non-graded

[2] http://literacynet.org/
[3] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

data from newspaper articles to represent higher grade levels (more useful for classification than for regression).

2.3.3 DISCOURSE, SEMANTICS, AND COHESION IN ASSESSING READABILITY

Feng et al. [2009] are specially interested in readability for individuals with mild-level intellectual disabilities (MID) (e.g., intelligence quotient (IQ) in the 55–70 range) and how to select appropriate reading material for this population. The authors note that people with MID are different from adults with low literacy in that the former have problems with working memory and with discourse representation, thereby complicating the processes of recalling information and inference as they read a text. The authors argue that appropriate readability assessment tools which take into account the specific issues of these users should therefore be designed. Their main research hypothesis being that the number of entity mentions in a text should be related to readability issues for people with MID, they design a series of features accounting for entity density. Where data for studying this specific population is concerned, they have created a small (20 documents in original and simplified versions) but rather unique ID dataset for testing their readability prediction model. The dataset is composed of news documents with aggregated readability scores based on the number of correct answers to multiple choice questions that 14 MID individuals had given after reading the texts. In order to train a model, they rely on the availability of paired and generic graded corpora. The paired dataset (not graded) is composed of original articles from *Encyclopedia Britannica* written for adults and their adapted versions for children and CNN news stories from the LiteracyNet organization available in original and abridged (or simplified) versions. The graded dataset is composed of articles for students in grades 2–5. Where the model's features are concerned, although many features studied were already available (or similar) in previous work, novel features take into account the number and the density of entity mentions (i.e., nouns and named entities), the number of lexical chains in the text, average lexical chain length, etc. These features are assessed on the paired datasets so as to identify their discriminative power, leaving all but two features outside the model. Three rich readability prediction models (corresponding to basic, cognitively motivated, and union of all features) are then trained on the graded dataset (80% of the dataset) using a linear regression algorithm (unlike the above approach). Evaluation is carried out on 20% of the dataset, showing considerable error reduction (difference between predicted and gold grade) of the models when compared with a baseline readability formula (the Flesch-Kincaid index [Kincaid et al.]). The final user-specific evaluation is conducted on the ID corpus where the model is evaluated by computing the correlation between system output and human readability scores associated with texts.

Feng et al. [2010] extended the previous work by incorporating additional features (e.g., language model features and out-of-vocabulary features from Schwarm and Ostendorf [2005] and entity coreference and coherence-based features based on those of Barzilay and Lapata [2008] and Pitler and Nenkova [2008]), assessing performance of each group of features, and comparing their

model to state-of-the-art competing approaches (i.e., mainly replicating the models of Schwarm and Ostendorf [2005]). Experimental results using SVMs and logistic regression classifiers show that although accuracy is still limited (around 74% with SVMs and selected features) important gains are obtained from the use of more sophisticated linguistically motivated features.

Heilman et al. [2007] are interested in the effect of **pedagogically motivated features** in the development of readability assessment tools, especially in the case of texts for second language (L2) learners. More specifically, they suggest that since L2 learners acquire lexicon and grammar of the target language from exposure to material specifically chosen for the acquisition process, both lexicon and grammar should play a role in assessing the reading difficulty of the L2 learning material. In terms of lexicon, a unigram language model is proposed for each grade level so as to assess the likelihood of a given text to a given grade (see Section 2.3.1 for a similar approach). Where syntactic information is concerned, two different sets of features are proposed: (i) a set of 22 grammatical constructions (e.g., passive voice, relative clause) identified in sentences after being parsed by the Stanford Parser [Klein and Manning, 2003], which produces syntactic constituent structures; and (ii) 12 grammatical features (e.g., sentence length, verb tenses, part of speech tags) which can be identified without the need of a syntactic parser. All feature values are numerical, indicating the number of times the particular feature occurred per word in the text (note that other works take averages on a per-sentence basis). Texts represented as vectors of features and values are used in a k-Nearest Neighbor (kNN) algorithm (see Mitchell [1997]) to predict the readability grade of unseen texts: a given text t is compared (using a similarity measure) to all available vectors and the k-closest texts retrieved, the grade level of t is then the most frequent grade among the k retrieved texts. While the lexical model above will produce, for each text and grade, a probability, the confidence of the kNN prediction can be computed as the proportion of the k texts with same class as text t. The probability of the language model together with the kNN confidence can be interpolated yielding a confidence score to obtain a joint grade prediction model. In order to evaluate different individual models and combinations, the authors use one dataset for L1 learners (a web corpus [Collins-Thompson and Callan, 2004]) and a second dataset for L2 learners (collected from several sources). Prediction performance is carried out using correlation and MSE, since the authors argue regression is a more appropriate way to see readability assessment. Overall, although the lexical model in isolation is superior to the two grammatical models (in both datasets), their combination shows significant advantages. Moreover, although the complex syntactic features have better predictive power than the simple syntactic features, their slight difference in performance may justify not using a parser.

Although these works are interesting because they consider a different user population, they still lack an analysis of the effect that different automatic tools have in readability assessment performance: since parsers, coreference resolution systems, and lexical chainers are imperfect, an important question to be asked is how changes in performance affect the model outcome.

Crossley et al. [2007] investigate three Coh-Metrix variables [Graesser et al., 2004] for assessing the readability of texts from the Bormuth corpus, a dataset where scores are given to texts

based on aggregated answers from informants using cloze tests. The number of words per sentence as an estimate of syntactic complexity, argument overlap—the number of sentences sharing an argument (noun, pronouns, noun phrases)—, and word frequencies from the CELEX database [Celex, 1993] were used in a multiple regression analysis. Correlation between the variables used and the text scores was very high.

Flor and Klebanov [2014] carried out one of the few studies (see Feng et al. [2009]) to assess lexical cohesion [Halliday and Hasan, 1976] for text readability assessment. Since cohesion is related to the way in which elements in the text are tied together to allow text understanding, a more cohesive text may well be perceived as more readable than a less cohesive text. Flor and Klebanov define lexical tightness, a metric based on a normalized form of pointwise mutual information by Church and Hanks [1990] (NPMI) that measures the strength of associations between words in a given document based on co-occurrence statistics compiled from a large corpus. The lexical tightness of a text is the average of NPMIs values of all content words in the text. It is shown that lexical tightness correlates well with grade levels: simple texts tend to be more lexically cohesive than difficult ones.

2.4 READABILITY ON THE WEB

There is increasing interest in assessing document readability in the context of web search engines and in particular for personalization of web search results: search results that, in addition to matching the user's query, are ranked according to their readability (e.g., from easier to more difficult). One approach is to display search results along with readability levels (Google Search offered in the past the possibility of filtering search results by reading level) so that users could select material based on its reading level assessment; however, this is limited in that the profile or expertise of the reader (i.e., search behavior) is not taken into consideration when presenting the results. Collins-Thompson et al. [2011] introduced a tripartite approach to personalization of search results by reading level (appropriate documents for the user's readability level should be ranked higher) which takes advantage of user profiles (to assess their readability level), document difficulty, and a re-ranking strategy so that documents more appropriate for the reader would move to the top of the search result list. They use a language-model readability assessment method which leverages word difficulty computed from a web corpus in which pages have been assigned grade levels by their authors [Collins-Thompson and Callan, 2004]. The method departs from traditional readability formulas in that it is based on a probabilistic estimation that models individual word complexity as a distribution across grade levels. Text readability is then based on distribution of those words occurring in the document. The authors argue that traditional formulas which consider morphological word complexity and sentence complexity (e.g., length) features and that sometimes require word-passages of certain sizes (i.e., at least 100 words) to yield an accurate readability estimate appear inappropriate in a web context where sentence boundaries are sometimes nonexistent and pages can have very little textual content (e.g., images and captions). To estimate the reading proficiency of users and also to train some of the model parameters and

evaluate their approach, they rely on the availability of proprietary data on user-interaction behaviors with a web search engine (containing queries, search results, and relevance assessment). With this dataset at hand, the authors can compute a distribution of the probability that a reader likes the readability level of a given web page from web pages that the user visited and read. A re-ranking algorithm, LambdaMART [Wu et al., 2010], is then used to improve the search results and bring results more appropriate to the user to the top of the search result list. The algorithm is trained using reading level for pages and snippets (i.e., search results summaries), user reading level, query characteristics (e.g., length), reading level interactions (e.g., snippet-page, query-page), and confidence values for many of the computed features. Re-ranking experiments across a variety of query-types indicate that search results improve at least one rank for all queries (i.e., the appropriate URL was ranked higher than with the default search engine ranking algorithm). Related to work on web documents readability is the question of how different ways in which web pages are parsed (i.e., extracting the text of the document and identifying sentence boundaries) influence the outcome of traditional readability measures. Palotti et al. [2015] study different tools for extracting and sentence-splitting textual content from pages and different traditional readability formulas. They found that web search results ranking varies considerably depending on different readability formulas and text processing methods used and also that some text processing methods would produce document rankings with marginal correlation when a given formula is used.

2.5 ARE CLASSIC READABILITY FORMULAS CORRELATED?

Given the proliferation of readability formulas, one may wonder how they differ and which one should be used for assessing the difficulty of a given text. Štajner et al. [2012] study the correlation of a number of classic readability formulas and linguistically motivated features using different corpora to identify which formula or linguistic characteristics may be used to select appropriate text for people with an autism-spectrum disorder.

The corpora included in the study were: 170 texts from Simple Wikipedia, 171 texts from a collection of news texts from the METER corpus, 91 texts from the health section of the British National Corpus, and 120 fiction texts from the FLOB corpus.[4] The readability formulas studied were the Flesch Reading Ease score, the Flesch-Kincaid grade level, the SMOG grading, and FOG index. According to the authors, the linguistically motivated features were designed to detect possible "linguistic obstacles" that a text may have to hinder readability. They include features of structural complexity such as the average number of major POS tags per sentence, average number of infinitive markers, coordinating and subordinating conjunctions, and prepositions. Features indicative of ambiguity include the average number of sentences per word, average number of pronouns and definite descriptions per sentence. The authors first computed over each

[4] http://www.helsinki.fi/varieng/CoRD/corpora/FLOB/