HHS Public Access

Author manuscript

IEEE J Biomed Health Inform. Author manuscript; available in PMC 2019 September 08.

Published in final edited form as:

IEEE J Biomed Health Inform. 2019 September ; 23(5): 2164–2173. doi:10.1109/JBHI.2018.2885465.

Using Lexical Chains to Identify Text Difficulty: A Corpus Statistics and Classification Study

Partha Mukherjee,
Engineering Division, The Pennsylvania State University, Great Valley campus, Malvern, PA, USA (pom5109@psu.edu).

Gondy Leroy [Senior Member, IEEE],
Department at Eller College of Management, University of Arizona, Tucson, USA (gondyleroy@email.arizona.edu).

David Kauchak [Member, IEEE]
Department of Computer Science, Pomona College, California, USA (david.kauchak@pomona.edu)

Abstract

Our goal is data-driven discovery of features for text simplification. In this work, we investigate three types of lexical chains: exact, synonymous, and semantic. A lexical chain links semantically related words in a document. We examine their potential with 1) a document-level corpus statistics study (914 texts) to estimate their overall capacity to differentiate between easy and difficult text and 2) a classification task (11,000 sentences) to determine usefulness of features at sentence-level for simplification. For the corpus statistics study we tested five document-level features for each chain type: total number of chains, average chain length, average chain span, number of crossing chains, and the number of chains longer than half the document length. We found significant differences between easy and difficult text for average chain length and the average number of cross chains. For the sentence classification study, we compared the lexical chain features to standard bag-of-words features on a range of classifiers: logistic regression, naïve Bayes, decision trees, linear and RBF kernel SVM, and random forest. The lexical chain features performed significantly better than the bag-of-words baseline across all classifiers with the best classifier achieving an accuracy of ~90% (compared to 78% for bag-of-words). Overall, we find several lexical chain features provide specific information useful for identifying difficult sentences of text, beyond what is available from standard lexical features.

Correspondence to: Partha Mukherjee.

[1] https://www.nlm.nih.gov/research/umls/

[2] https://gate.ac.uk/

[3] https://en.wikipedia.org/

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

[5] https://cran.r-project.org/bin/windows/base/old/3.3.2/

Keywords

Health informatics; text difficulty; readability; classification; natural language processing; text simplification; logistic regression; decision trees; naïve Bayes; SVM; random forest

I. INTRODUCTION

Health literacy is an essential component of healthcare, and has been a national goal of the US since 1970 [1]. Providing information in an understandable but not oversimplified manner is essential to support patients and information consumers in their decision process.

Text-based information on preventive care, treatment, and recovery [2–5] is considered one of the most efficient approaches to improve patient health literacy [6]. Text written using clear and understandable language helps patients remember medical information [7], motivates them to read and understand text [8], and affects patient perception of medical staff [9].

Many different text features affect text difficulty [10], and many different text simplification approaches have been suggested to increase reader comprehension [11, 12]. Text complexity can be defined as a "three-part" model consisting of 1) a qualitative dimension, 2) a quantitative dimension, and 3) reader and task considerations [13]. The qualitative dimension includes the meaning, structure, language conventionality, and clarity experienced by human readers. The quantitative dimension refers to features typically measured computationally, such as: word length, word frequency, sentence length, number of syllables, and text cohesion. Reader and task considerations focus on the inherent complexity of the text specific to a group of readers (e.g., students, patients, etc.). We focus on the quantitative dimension of text complexity and examine the role of document-level features for identifying text difficulty.

Many features have been evaluated to measure text difficulty, ranging from features that focus on individual words or sentences to entire documents. For example, technical jargon and terms with low usage frequency are important features where replacement with more common synonyms is effective in increasing comprehension [14, 15]. Similarly, at the term level different types of negations are indicative of difficulty [16] and can be used to simplify text. Grammatical features such as sentence structure, for example measured by frequency of the parse structure [17], also affect difficulty [15, 18]. At the document level, cohesion has been shown to affect readability [19] and is important to describe how information is conveyed across multiple sentences [20, 21]. Semantic relatedness features in the form of semantic similarity between words between text segments have been used to classify text difficulty [22, 23]. These different types features have then been used in a range of different classifiers [24–27] and with a different feature processing, e.g. Bloehdorn et al. [28] propose smoothing kernels for text difficulty classification by implicitly encoding a super concept expansion and achieve satisfactory results under poor training data and data sparseness. The use of information technology tools has also been suggested to assist the understanding of information from clinical texts [29, 30].

In this paper, we evaluate how lexical chain features can be used to distinguish between easy and difficult text and how they can be used to identify sentences that are difficult. The type of chain that flags a sentence as difficult also provides information for the simplification process. We define a lexical chain as a sequence of semantically related lexical items, independent of the grammar structure of the text. They capture lexical cohesion structure [31] by highlighting the repetition of related concepts throughout the text. Different degrees of freedom can be employed to define semantic relatedness. We examine three notions of semantic relatedness, which define three types of lexical chains. The following text snippets show an example for each (exact, synonymous, and semantic lexical chains are highlighed in bold, underlined, and bold underlined, respectively).

Exact: {syndrome, syndrome, syndrome, syndrome}

Synonymous: {abnormalities, anomalies}

Semantic: {gene, nucleotide}

> "Aarskog-Scott **syndrome** is a rare disease inherited as autosomal dominant or x-linked and characterized by short stature, facial abnormalities, skeletal and genital anomalies. The Aarskog-Scott **syndrome** is also known as the Aarskog **syndrome** and faciogenital dysplasia. Aaarskog-Scott **syndrome** is due to mutation in the fgd1 **gene**. Fgd1 encodes a guanine **nucleotide** exchange factor(gef) that specifically activates cdc42, a member of rho (rashomology) family of p21 gtpases."

To utilize the lexical chains, we created five new features that allow us to quantify the lexical chains and differentiate between easy and difficult medical text. We measure the usefulness of these features using two tasks, document profiling through corpus statistics and sentence classification, both of which have been previously used to understand the usefulness of text features for measuring text difficulty [32]. In this research, we have initiated the text simplification study with lexical chain features which is purely quantitative and will look into the qualitative aspects to measure the usefulness of lexical chain features involving human experts.

II. LITERATURE REVIEW

Lexical chains are sequences of semantically related words in a document [33]. They usually span across sentences and provide a thread for the ideas throughout a text.

A. Construction of Lexical Chains

Lexical chains are based on lexical cohesion [20], which is exhibited via cohesive relations. The relations are: 1) repetition of the same word in the same sense, 2) the use of synonyms/hypernyms/hyponyms for a word, and 3) semantic relationships between words that often co-occur. Existing algorithms for identifying lexical chains are based upon the inclusion of candidate words such as nouns and compound nouns [34]. Several algorithms have been proposed for computing lexical chains. Though Morris and Hirst [33] first suggested the algorithm to construct lexical chains and introduced the idea of different lexical chain features such as length, density, span, etc., the widely used ones are those put forward by

[35–38]. Hirst and St-Onge [35] classified cohesive relations into extra-strong (i.e., identity and synonymy), strong (hypernymy and hyponymy), and medium strong (hypernymy and hyponymy path). They used a greedy strategy to add words to the chains with which the words had the strongest relations. On the other hand, Barzilay and Eldahad [36] created a list of interpretations exclusive of each other and selected chains according to the best interpretation that had the most connections with the words. Silber and McCoy [37] proposed a two pass algorithm with identity, synonymy, hypernymy/hyponymy, and hypernymy/hyponymy tree relations to construct lexical chains. The first pass identified the noun instances within the text and assigned each sense of that noun instance using the relations, then the best interpretation of the noun was found in the second pass. Jarmasz and Szpakowicz [38] proposed an algorithm that constructed lexical chains by a set of words linked via thesaural relations. Most approaches used WordNet [39] to identify relations among words to construct the lexical chains; although [38] used Roget's thesaurus of English words and phrases [40]. In our research, we constructed lexical chains using 1) repetitions, 2) synonyms, and 3) semantic relations between nouns using the Unified Medical Language System (UMLS[1]) database.

B. Application of Lexical Chains in Text Processing

Lexical chains have multiple applications in text processing and analytics. For example, lexical chains were used in text segmentation to identify semantic boundaries in a text where a transition occurs from one topic to another, and they/these can be examined at varying level of granularity [41–46]. All of the lexical chains used WordNet to identify the relations among words to segment the text except Tatar et al. [45], who used Roget's thesaurus-based relations. Secondly, lexical chains can measure the quality of text coherence [47] by examining coherence characteristics such as text unity [20], variety [48], elaboration and detailing [49], and organization [50]. This application used Lin's [51] thesaurus-based relations to construct the lexical chains. Thirdly, lexical chains have been used for text summarization in multilingual platforms [36, 52–55]. Among the lexical chain-based text summarization approaches, Fuentes and Rodriguez [53] used Spanish WordNet [56] for Spanish text summarization, while Chen et al. [54] used lexical chains for summarization of Chinese texts. Finally, lexical chains have been used to improve the performance of question answering systems [57–59]. These applications tend to use WordNet [39] to identify the relations between words. To our knowledge, lexical chains have not been used to classify text difficulty levels.

C. Use of Lexical Chains in the Medical / Disability Domain

Lexical chains have not been frequently used in the medical domain. Reeve [60] used the UMLS to identify concept chains for summarizing text on illnesses. They found that creating chains using concepts based on semantic types could be successfully applied for medical text summarizations using both abstracts and full texts. Feng et al. [61] used lexical chain features in a tool to automatically rate the readability for users with mental disabilities. The authors built features using term repetition-based lexical chains as proposed by Galley and McKeown [43, 62]. They created a corpus from Encyclopedia Britannica with easy and difficult articles and found that the number of lexical chains and average chain span are significantly higher in easy texts than in difficult texts.

III. COHESION-DRIVEN DIFFICULTY METRIC

For our approach, we focus on noun-based lexical chains since nouns tend to be the major content-bearing items in text and have good support in external resources. We created 1) exact lexical chains containing only identical nouns, 2) synonymous lexical chains containing nouns that are synonyms, and 3) semantic lexical chains containing nouns that belong to the same semantic tree.

A. Resources Used

We used GATE (General Architecture for Text Engineering[2]) to process our texts [63]. GATE is a Java suite of tools used for many different types of text analytics tasks. We used the built-in tokenizer and sentence splitter to pre-process the texts and the Stanford parser [64] to identify nouns.

To identify synonymous and semantic lexical chains, we used the UMLS, which contains three types of knowledge sources: 1) Metathesaurus, 2) Semantic Network, and 3) SPECIALIST lexicon. The Metathesaurus forms the base of the UMLS with over 1,000,000 biomedical concepts and 5,000,000 concept names. The biomedical concepts originate from 100 controlled vocabularies and classification systems. The Semantic Network is a catalog of semantic types and relationships with 127 semantic types and 54 relationships altogether. The SPECIALIST lexicon contains information about common English vocabulary, biomedical terms, and terms found in the UMLS Metathesaurus. The synonyms are identified using the Metathesaurus, while semantic relationships between words are identified with both the Metathesaurus and Semantic Network.

B. Exact Lexical Chain

Exact lexical chains are formed with repeated nouns (see Fig. 1). For each noun in a text, we check for repetitions allowing for plurality variation, e.g., “syndrome” and “syndromes” would be an exact match and would occur in the same chain. A lexical chain must contain at least two occurrences of a noun, i.e., a single occurrence of a word does not count as a lexical chain.

C. Synonymous Lexical Chain

Synonymous chains include nouns and their synonyms (see Fig. 2). Nouns are considered synonymous if they are not lexically identical and share the same Concept Unique Identifier (CUI) in the UMLS.

Example: “*In addition to the ultrasound or afp scanning, it is also necessary for children with this* ***disease*** *to be checked for other birth defects because genetic* ***disorders*** *are usually associated with some of the abdominal wall defects***.”**

Synonymous chain: disease → disorders. Both have CUI = **C0012634,** retrieved from the MRCONSO table in the UMLS.

D. Semantic Lexical Chain

Semantic chains are computed using nouns that are semantically related (see Fig. 3). For each noun we retrieve its semantic family tree using the UMLS. Two nouns are semantically related if they are not lexically identical, not synonyms, and they reside in the same semantic family tree.

Relationships in the UMLS Metathesaurus contain concepts and the nouns assigned to these concepts. For each noun, we retrieve the noun and its CUIs from the UMLS. Using the CUIs we retrieve the hierarchical context of the nouns from the MRHIER table in the Metathesaurus. The hierarchical path is represented as a list of Atomic Unique Identifiers (AUIs). If the AUI of any noun in the list exists in the hierarchical path of the noun in consideration, they are considered as semantically related nouns.

Example: ***"Cancer*** *is a group of* ***diseases*** *involving abnormal cell growth with the potential to invade or spread to other parts of the body."*

Semantic chain: cancer → diseases. The CUIs retrieved from the UMLS MRCONSO table for cancer and diseases are C0006826 and C0012634. Then, the retrieved hierarchical paths for cancer and diseases from the MRHIER table using the corresponding CUIs are A0398472.A1883168 and A0398472, respectively. The hierarchical paths of these two nouns share the same AUI (A0398472), thus they are semantically related.

E. Lexical Chain Features

To use lexical chains algorithmically, features are created that capture characteristics of the chains. We computed five different features that reflect the diversity of entities/concepts a reader must keep in mind when reading a document:

- Number of chains: the number of chains in the text.
- Average chain length: The length of a lexical chain is the number of nouns in the chain. For example, the exact chain in example snippet above would have length 4, i.e., 4 occurrences of the word "syndrome." We average the length of all chains in the text.
- Average chain span: The "span" of a chain is the number of words in the text between the first and the last noun in the chain, excluding the first and last noun. We average the span of all chains in the text.
- Number of cross chains: Two chains cross if the first/earlier chain partially or fully contains the second one.
- Number of half document length chains: Number of lexical chains whose span is greater than half of the text length.

IV. CORPUS STATISTICS STUDY

Our goal is to validate the importance of lexical chains to distinguish between easy and difficult text. Our first study examines document-level distributions of lexical chains to demonstrate how lexical chains differ between easy and difficult texts.

A. Method

We analyzed a medical text corpus from English Wikipedia and Simple English Wikipedia, representing difficult and easy texts, respectively. Table I provides summary statistics of the corpus.

- English Wikipedia[3]: We downloaded all 625 articles in English Wikipedia under the “List of Diseases” [65]. We extracted the texts from the corresponding Wikipedia page using crawler scripts written in Java.
- Simple English Wikipedia[4]: Simple English Wikipedia is written in basic English with easier words and grammatical structures than normal English Wikipedia. Many texts in Simple English Wikipedia were directly generated from the original counterpart, though often with some information omissions. The articles are meant to be accessible to a wider audience than the normal Wikipedia. We carried out a similar process on Simple English Wikipedia [66] and extracted the text from 289 medical texts listed under the “List of diseases” page [66].

We computed all five features per document and then normalized them by the number of words in the document to account for variation in document length, e.g., longer documents will on average have more lexical chains. We use Welch’s t-test [67] to measure statistical significance between feature values of easy and difficulty text, since the features have unequal variances. We use Bonferroni correction [68] to compensate for repeated testing.

B. Results

We compared the five features for each of the three different chain types between easy and difficult texts and found significant differences between many of the features.

1) Number of chains—Overall, there were more chains in the difficult texts than the simple texts (see Fig. 4). Note this is after normalizing for length, since the difficult texts are longer on average. This difference is significant for semantic chains ($p = 0.001$), but not for exact ($p = 0.049$) or synonymous chains ($p = 0.305$). The average number of exact chains is higher for synonymous and semantic chains in both difficult and simple texts.

Semantic chains are the least frequent of the three chain types, particularly in easy texts. Difficult text contains a higher proportion of different topics (reflected by number of semantic chains), while simple text has a lower distribution of topics.

2) Average chain length—There are longer chains in easy text than in difficult text (see Fig. 5). The average length of the lexical chains of all three types are more than three times longer in easy text than in difficult ($p = 0.000$). The average length of all three chain types are nearly the same in difficult texts. Although there are more topics in difficult text, the descriptions are shorter. Topics in simple text contain more repetitions, synonyms, and semantically related words (reflected in the length of the chains).

3) Average chain span—Overall, easy texts have a higher average span for synonymous and semantic chains than difficult text, but no differences exist for exact chains (see Fig. 6). The average span for synonymous chains is more than double the spans of the

other chains in easy text, while for difficult text the synonymous chain span is more than six times the span of semantic chains. For exact chains, average span doesn't show much difference between the two types of corpora ($p = 0.228$). For semantic chains, the average chain span in Simple Wikipedia is four times than in difficult texts ($p = 0.000$), but for synonymous chains it is not significant ($p = 0.013$). The result shows easy text contains synonymous and semantic chains that cover a higher proportion of text than the difficult text.

As simple text contains a lower percentage of different topics, the topic description length is longer. Therefore, semantic and synonymous chains in simple text cover more sentences compared to those in difficult text.

4) Average number of cross chains—Overall, the average number of cross chains is higher in the difficult text than the easy (see Fig. 7). In easy text, the number of synonymous cross chains are nearly double that for exact chains and four times that for semantic chains. For semantic chains, the average cross chains in difficult text is three times the cross chains in easy text ($p = 0.000$). For exact chains ($p = 0.003$) and synonymous chains ($p = 0.002$), the differences between easy and difficult texts are also significant: difficult text more intersections between the chains compared to easy text. Difficult text contains more topics with relatively shorter descriptions as reflected by the high number of lexical chains with shorter length. The higher percentage of intersections between chains in difficult text signifies a higher proportion of nesting of topics compared to simple text.

5) Average count of half document length chains—The average half document length of synonymous and semantic chains is higher in easier texts (see Fig. 8). The average half document length for synonymous chains in easy texts is more than six times that for exact and semantic chains. For exact chains, average half document length is the same for easy and difficult texts ($p = 0.814$).

The average half document length in for semantic chains is four times that for easy text than for difficult text ($p = 0.000$), but there is no difference for synonymous chains ($p = 0.203$). The result implies that easy text contains a higher proportion of semantic chains that run through more than half of the text length.

As easy text contains fewer different topics with relatively longer lengths, the proportion of synonymous and semantic lexical chains with longer length and span that traverse more than half of the text length is also longer compared to those in difficult text.

C. Corpus Statistics Study Conclusion

We are interested in the differences between easy and difficult texts, i.e., the differences in features of different types of lexical chains present in Simple English Wikipedia versus normal English Wikipedia.

To understand how different lexical chain features correspond to text difficulty, we compared their occurrence in simple texts versus difficult texts. Fig. 9 shows the log of the ratio of the feature values in the difficult text versus in the easy text. Positive log-ratios indicate features that have larger values in difficult text and negative log-rations indicate higher values in the

simple text. We found that all three lexical chain types are shorter in difficult text (chain length is greater in simple text), but with a higher proportion of intersections (i.e., average number of cross chains). Chains are more frequent in difficult text, but pass through a smaller proportion of the text as easy text has a higher average chain span and average half document length chains.

V. SENTENCE LEVEL CLASSIFICATION STUDY

We then evaluated whether the features can be used at a more fine-grained level to identify individual sentences as easy or difficult. We first evaluate the features for each type of chain separately and then for all combined. As a baseline, we compare the lexical chain features to a standard bag-of-words set of features [69, 70].

A. Datasets and Classifiers

We created a balanced set of 11,000 sentences to classify with 5,500 sentences from English Wikipedia (difficult) and 5,500 sentences from Simple English Wikipedia (easy). We make the strong assumption that all sentences from English Wikipedia are difficult sentences and all are easy from Simple English Wikipedia. This does not hold for every sentence and we therefore would not expect a classifier on this task to ever achieve perfect performance.

To generate this dataset, we started with our document-level texts on diseases described above. We removed all documents that were 5KB or less to avoid short documents, leaving 210 texts in Simple English Wikipedia and 435 in normal English Wikipedia from the original set in the corpus. We then selected 146 texts randomly from the set of 435 normal Wikipedia texts and randomly select 5,500 sentences each from this set and the 210 Simple English texts to get our final balanced dataset of 11,000 sentences.

Our goal is to understand the usefulness of the lexical chain features for measuring text difficult. To get a broad sample of the usefulness for classification, we used six different common classifiers to compare the accuracy of identifying easy and difficult sentences: logistic regression [71], decision trees [72], naïve Bayes [73], SVM with linear and RBF kernels [74], and random forests [75]. We used the R 3.3.2 libraries[5] to run the classifiers on the data. Since we are interested in the usefulness of the features, we focus on classifier accuracy for understanding effectiveness. The classifiers have different time and memory requirements, which are important for practical implications, but, given the focus of this paper, we leave that discussion for future implementation analysis.

B. Computing Sentence Level Features

We derived the sentence-level features from those features computed at the document level.

1) Bag-of-words features—For comparison, we calculated a standard bag-of-words features for each sentence. To determine the bag-of-words features we combined the normal and Simple English Wikipedia texts. After removing stop-words there were a total of 9,354 words. The bag-of-words features per sentence is a sparse vector of length 9,354 with a 1 or 0 depending on the presence of a word within the sentence.

2) Lexical chain features—We compute the average feature values at the sentence level for the three types of chains. The lexical chain features for the sentences are weighted features computed at the document level. The weight is computed as the frequency ratios of the nouns appearing in the sentence and the text, shown in formula 1a. The weighted average feature values at the sentence level are computed following formula 1b:

$$w_i = \frac{f_i}{\sum_{\forall i \in D} f_i} \quad (1a)$$

$$x_s^{i,j} = \frac{1}{n} \sum_{i=1}^{n} w_i . x_D^j \quad (1b)$$

$x_s^{i,j}$ is the value of feature j for noun i in sentence s. n is the total number of nouns appearing in sentence s. x_D^j is the value for feature j in document D. w_i and f_i are the weight and document-level frequency of noun i, respectively. This results in set of five normalized, real-valued features for each of the lexical chain types for each sentence. The lexical chain features are presented to the classifier as a feature vector per sentence. The feature vector length for each type of lexical chain is 5 while in a combined scenario the vector length is 15 (i.e. for three chains) per sentence. The average feature statistics of the three lexical chain types at the sentence level are shown in Table II.

The lexical chain features at the sentence level follow the same trends observed at the document level. Average number of chains and average cross chain features are higher in the difficult sentences while average chain length is higher in easy sentences for all three chain types. Average chain span and average half document length are higher in easy sentences for synonymous and semantic chains.

C. Results

Table II shows the accuracies for the classifiers' different chain types individually, combined and for the bag-of-words features. We performed 10-fold cross validation (10 rounds of training/testing) and averaged across the 10 folds.

1) Results for bag-of-words only—For all classifiers and all types of lexical features the bag-of-words features perform worse than the lexical chain features with the exception of exact chains with random forests (0.777 vs. 0.613): the lexical chain features are capturing information not captured by a standard set of bag-of-words features. All differences are significant ($p = 0.000$) based on a paired t-test over the 10-folds between the bag-of-words features and the features of all three chains combined for each classifier.

2) Results by chain type—Table II shows the accuracies by chain type. Overall, there is no single chain type that performs best across all classifiers. Three of the classifiers (logistic regression, decision tress, and naïve Bayes) perform the best with exact chains, both

SVM variants perform best with synonymous chains, and random forests perform best with semantic chains. The SVM performs better with the RBF kernel than with the linear kernel for all chain types. For all classifiers and all chain types, the accuracy is better using lexical chain features than bag-of-words features, except for random forests with exact chains, with improvements ranging from 1% absolute to as much as 17% absolute.

3) Results combining all three chain types—The sixth column of Table II shows the result when all 15 input features from the three types of chains are used together. The classification accuracy increases considerably in detecting text difficulty when all three types of chains are used together. We find that random forest (0.898) performs the best in this setting and naïve Bayes the worst (0.777). Using all three types of chains increases accuracy by ~12% absolute from the bag-of-words with the best classifier.

4) Results for bag-of-words with all three chains—As a final comparison, we combined all 15 lexical chain features with the 9,354 bag-of-words features. For all classifiers except naïve Bayes, this combined set of features performs better than only using the set of all three lexical chain features: the lexical chain features provide a strong set of features for classifying text difficulty, but the bag-of-words features do provide some additional complementary information. For example, for the random forest classifier, all chains combined outperforms bag-of-words along by 12.1% absolute, but adding the two together achieves an additional 1.7% improvement.

5) Follow-up analysis—We examined the coefficients of the chain features for logistic regression since the model is easy to interpret and provided good performance (second only to random forests). Table III shows the coefficients of the predictors. Since all features are positive, the sign of the coefficient indicates whether the feature contributed towards simple (positive) or difficult (negative) and the magnitude indicates the importance of the feature in the model. Average chain length, average chain span, and average chains with greater than half document have a negative correlation with sentence difficulty. On the other hand, number of chains and average cross chains have a positive correlation with sentence difficulty.

From the corpus statistics study, we observe that number of chains and average cross chains are higher in difficult text (i.e. positively correlated with text difficulty) while average chain length, average chain span, and average chains with half document length are in higher proportion in the simple texts (negatively correlated with text difficulty). The sentence level analysis shown in Table III support the observations seen in the corpus analysis study. The same trends are found for all feature coefficients of all the chain types for the logistic regression in the combined scenario.

VII. CONCLUSION AND DISCUSSION

Our overall goal is to simplify medical text in a semiautomatic manner. We aim to discover a variety of different text features that can be integrated into text simplification algorithms. In this work, we tested the usefulness of lexical chains to distinguish between easy and difficult text.

Both the corpus statistics study and the classification study showed that the lexical chain features can be used to distinguish between easy and difficult medical text. The easy texts and sentences contain longer chains, while difficult texts contain a higher volume of synonymous and semantic chains. This signifies that easy texts comprise relatively fewer, but longer, chains. Again, the easiness of text is positively correlated with number of semantically related words in the text, while chains in difficult text have shorter word cover. The high volume of shorter chains in difficult text implies higher volume of intersections between the chains (cross chains) in the text.

The classification result showed similar results regarding the usefulness of lexical chains for discriminating sentence difficulty. The second best classifier for sentence classification is analyzed as it is easy to interpret.

The correlations of the features for the exact, synonymous and semantic chains with the sentence level difficulty separately show the same tendency of the features at the text level observed in corpus statistics study. Information obtained from corpus statistics can be used to measure the sentence difficulty using lexical chain features derived at the sentence level. The classifier accuracy showed that the features of lexical chains could independently distinguish easy from difficult sentences in the text. The combination of all three types of chains increases the accuracy further. The coefficients learned by the logistic regression classifier further support this picture with similar trends. Finally, we find that the lexical chain features provide better information and complementary information to a standard bag-of-words classifier: the performance of the classifiers with the lexical chains features was significantly better than with the bag-of-words features, though combining all features did result in a small performance improvement.

Building from this work, there are a number of possible next steps to examine. Our goal for this work was to determine whether lexical chains provide information for determining text difficulty. We provided an initial feature set and analysis which is purely quantitative, but further research is required to examine other possible features and to provide a more in-depth classification study that involves checking the usefulness of lexical chain features to classify text using a gold standard data set created by more than one human experts (qualitative dimension). Using the lexical chains, we will develop simplification algorithms by exploiting the lexical chain features and will verify the simplification of the medical text done by the algorithms with the intervention of human experts before integrating them into a software system (our long term goal) to improve US health literacy along with other features we have found to affect text difficulty.

ACKNOWLEDGMENT

Research reported in this paper was supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM011975. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM011975. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Biography

Partha Mukherjee received his Bachelors in mechanical engineering from Jadavpur University, India in the year of 1995. He received his Master of Technology (M.Tech) diploma in Computer Science from Indian Statistical Institute, India in 2001. He earned his second Masters (M.S) in computer Science from University of Tulsa, USA in 2008. He completed his PhD in Information Sc. and Tech with minor in applied Statistics from Pennsylvania State University in 2016.

Currently he is an Assistant Professor in Engineering Division, The Pennsylvania State University, Great Valley campus, PA, USA. Previously he worked as Post-Doctoral research scholar in MIS department, Eller College of Management, The University of Arizona, and as Research Consultant in Indian Institute of Technology, (IITKGP) in designing and implementing combinational unit of ADSP 21020 microprocessor. He also served as an Assistant Professor in multiple institutes in India during 2002 to 2004 and 2008 to 2010.

He is a member of ACM, IEEE, AIS, AoIR and ASE. He has published papers in IEEE and ACM conferences. His research interests are in social computing, web analytics, data mining, and natural language processing.

Gondy Leroy earned a combined BS and MS (1996) in cognitive psychology from the Catholic University of Leuven (Belgium) and a MS (1999) and PhD (2003) from the University of Arizona's Management Information Systems (MIS) department.

She is a senior member of IEEE and Professor of MIS at the University of Arizona's MIS department. She serves on the editorial board of the Journal of Database Management, International Journal of Social and Organizational Dynamics in IT, Health Systems, Journal of Business Analytics and several conference program committees. She authored the book "Designing User Studies in Informatics" and conducts tutorials on evaluating artifacts in a design science context. Her research has been published in ACM Computing Surveys, JAMIA, JASIST, International Journal of Medical Informatics, and IEEE-TITB among others. She served as president for the AIS Special Interest Group on Health and as co-chair of the AIS Women's Network.

David Kauchak received his B.S. in computer science from the University of Utah (2000) and his M.S. (2002) and Ph.D. (2006) in computer science from the University of California San Diego.

He worked in industry from 2007–2009 as a research scientist and has been in academia since then. He is currently an Associate Professor in the computer science department at Pomona College. His research interests are in natural language processing with a recent focus on text simplification. He is a member of IEEE.

REFERENCES

[1]. Nutbeam D, “Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century”, Health promotion international, 15(3), pp. 259–267, 2000.

[2]. Mouradi O. Influence of text and participant characteristics on perceived and actual text difficulty; Proc.; 2013 46th Hawaii International Conference on System Sciences (HICSS); IEEE; 2013. 2464–2473.

[3]. Nielsen-Bohlman L et al., “Committee on Health Literacy, Health Literacy: A Prescription to End Confusion”, 2006, The National Academies Press Available at http://www.nap.edu/catalog/10883.html#toc,aufgesuchtam.

[4]. Ong LM et al., “Doctor-patient communication: a review of the literature”, Social science & medicine, 40(7), pp. 903–918, 1995.. [PubMed: 7792630]

[5]. Zolnierek KBH and DiMatteo MR, “Physician communication and patient adherence to treatment: a meta-analysis”, Medical care, 47(8), pp. 826, 2009. [PubMed: 19584762]

[6]. Egbert N and Nanna KM, “Health literacy: Challenges and strategies”, The Online Journal of Issues in Nursing, 14(3), 2009

[7]. Burgers C et al., “How (not) to inform patients about drug use: use and effects of negations in Dutch patient information leaflets.

[8]. Leroy G and Kauchak D, “The effect of word familiarity on actual and perceived text difficulty”, Journal of the American Medical Informatics Association, 21(e1), pp. e169–e172, 2013. [PubMed: 24100710]

[9]. Burgers C et al., “How the doc should (not) talk: When breaking bad news with negations influences patients’ immediate responses and medical adherence intentions”, Patient education and counseling, 89(2), pp. 267–273, 2012. [PubMed: 22938871]

[10]. Kauchak D. Text simplification tools: using machine learning to discover features that identify difficult text; Proc. 47th Hawaii International Conference on System Sciences (HICSS); IEEE; 2014. 2616–2625.

[11]. Carroll J. Practical simplification of English newspaper text to assist aphasic readers; Proc. AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology; 1998. 7–10.

[12]. Carroll J. Simplifying text for language-impaired readers; Proc. EACL; 1999. 269–270.

[13]. Carver L, “The Struggling reader and the Common Core”, The Apple Shouldn’t Fall Far from Common Core: Teaching Techniques to Include All Students, pp. 85, 2015.

[14]. Nagel K et al., “Using plain language skills to create an educational brochure about sperm banking for adolescent and young adult males with cancer”, Journal of Pediatric Oncology Nursing, 25(4), pp. 220–226, 2008. [PubMed: 18539907]

[15]. Leroy G et al., “The influence of text characteristics on perceived and actual difficulty of health information”, International journal of medical informatics, 79(6), pp. 438–449, 2010. [PubMed: 20202895]

[16]. Mukherjee Partha et al., “NegAIT: A New Parser for Medical Text Simplification Using Morphological, Sentential and Double Negation”, Journal of Biomedical Informatics, 69, pp. 55–62, 2017. [PubMed: 28342946]

[17]. Mukherjee P. The Role of Surface, Semantic and Grammatical Features on Simplification of Spanish Medical Texts: A User Study; Proc. AMIA Fall Symposium; Washington DC. 2017.

[18]. Heilman M. Combining lexical and grammatical features to improve readability measures for first and second language texts; Proc. Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics; 2007. 460–467.

[19]. Todirascu A. Are Cohesive Features Relevant for Text Readability Evaluation?; Proc. 26th International Conference on Computational Linguistics (COLING 2016); 2016. 987–997.

[20]. Halliday MAK and Hasan R, Cohesion in english, Routledge, 2014.

[21]. Graesser AC et al., “Coh-Metrix: Analysis of text on cohesion and language”, Behavior Research Methods, 36(2), pp. 193–202, 2004.

[22]. Nasir JA, et al. “A knowledge-based semantic kernel for text classification”, in International Symposium on String Processing and Information Retrieval 2011 Springe.

[23]. Siolas G and d’Alché-Buc F. “Support vector machines based on a semantic kernel for text categorization”, in Proc. IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN, 2000) 2000.

[24]. Altınel B, et al., “A corpus-based semantic kernel for text classification by using meaning values of terms”, Engineering Applications of Artificial Intelligence, 43, pp. 54–66, 2015.

[25]. Altlnel B, et al. “A novel higher-order semantic kernel for text classification”, in Proc. International Conference on Electronics, Computer and Computation (ICECCO 2013) 2013 IEEE.

[26]. Altinel B, et al. A semantic kernel for text classification based on iterative higher–order relations between words and documents. in International Conference on Artificial Intelligence and Soft Computing 2014 Springer.

[27]. Altinel B, et al. “A simple semantic kernel approach for SVM using higher-order paths”, in Proc. IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA, 2014), 2014, IEEE.

[28]. Bloehdorn S, et al. “Semantic kernels for text classification based on topological measures of feature similarity”, in Sixth International Conference on Data Mining, (ICDM ‘06), 2006 IEEE..

[29]. Johnson SB et al., “Data management in clinical research: Synthesizing stakeholder perspectives”, Journal of biomedical informatics, 60, pp. 286–293, 2016. [PubMed: 26925516]

[30]. Morid MA et al., “Classification of clinically useful sentences in clinical evidence resources”, Journal of biomedical informatics, 60, pp. 14–22, 2016. [PubMed: 26774763]

[31]. Xiong D. Lexical Chain Based Cohesion Modelsfor Document-Level Statistical Machine Translation; Proc. EMNLP; 2013. 1563–1573.

[32]. Nopales C and Dredze M, “ Learning simple Wikipedia: A cogitation in ascertaining abecedarian language”, in Proc NAACL HLT, Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids, ACL, 2010, pp. 42–50.

[33]. Morris J and Hirst G, “Lexical cohesion computed by thesaural relations as an indicator of the structure of text”, Computational linguistics, 17(1), pp. 21–48, 1991.

[34]. Jayarajan D. Lexical Chains as Document Features; Proc. IJCNLP; 2008. 111–117.

[35]. Hirst G and St-Onge D, “Lexical chains asrepresentations of context for the detection and correction of malapropisms”, WordNet: An electronic lexical database, 305, pp. 305–332, 1998.

[36]. Barzilay R and Elhadad M, “Using lexical chains for text summarization”, Advances in automatic text summarization, pp. 111–121, 1999.

[37]. Silber HG and McCoy KF, “Efficiently computed lexical chains as an intermediate representation for automatic text summarization”, Computational Linguistics, 28(4), pp. 487–496, 2002.

[38]. Jarmasz M and Szpakowicz S, “Not as easy as it seems: Automating the construction of lexical chains using roget’s thesaurus”. in Proc. Canadian Society for Computational Studies of Intelligence: Springer, 2003, pp. 544–549.

[39]. Miller GA, “WordNet: a lexical database for English”, Communications of the ACM, 38(11), pp. 39–41, 1995.

[40]. Jarmasz M and Szpakowicz S, “Roget’s thesaurus: A lexical resource to treasure”, arXiv preprint arXiv:1204.0258, 2012.

[41]. Manabu O and Takeo H, “Word sense disambiguation and text segmentation based on lexical cohesion”. in Proc. 15th conference on Computational linguistics-Volume 2: Association for Computational Linguistics, 1994, pp. 75–761.

[42]. Stokes N, “Spoken and written news story segmentation using lexical chains”. in Proc. HLT-NAACL student research workshop-Volume 3: Association for Computational Linguistics, 2003, pp. 49–54.

[43]. Galley M. Discourse segmentation of multi-party conversation; Proc. 41st Annual Meeting on Association for Computational Linguistics-Volume 1: Association for Computational Linguistics; 2003. 562–569.

[44]. Adarve F et al., “Topic Segmentation of Meetings Using Lexical Chains”, 2007.

[45]. Tatar D, et al., “Text Segmentation Using Roget-Based Weighted Lexical Chains”, Computing and Informatics, 32(2): p. 393–410, 2013.

[46]. Kauchak D and Chen F, “Feature-based segmentation of narrative documents”. in Proc. ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing: Association for Computational Linguistics, 2005, pp. 32–39.

[47]. Somasundaran S. Lexical Chaining for Measuring Discourse Coherence Quality in Test-taker Essays; Proc. COLING; 2014. 950–961.

[48]. Witte SP and Faigley L, “Coherence, cohesion, and writing quality”, College composition and communication, 32(2), pp. 189–204, 1981.

[49]. Hobbs JR, “Coherence and coreference”, Cognitive science, 3(1), pp. 67–90, 1979.

[50]. Perfetti CA and Lesgold AM, “Discourse comprehension and sources of individual differences”, 1977.

[51]. Lin D, “Automatic retrieval and clustering of similar words”. in Proc. 17th international conference on Computational linguistics-Volume 2: Association for Computational Linguistics, 1998, pp. 768–774s.

[52]. Boguraev BK and Neff MS, “Lexical cohesion, discourse segmentation and document summarization”. in Proc. Content-Based Multimedia Information Access- Volume 2: LE CENTRE DE HAUTES ETUDES INTERNATIONALES D’INFORMATIQUE DOCUMENTAIRE, 2000, pp. 962–979.

[53]. Fuentes M and Rodríguez H, “Using cohesive properties of text for automatic summarization”, JOTRI’02, 2002..

[54]. Chen Y et al., “Automatic text summarization based on lexical chains”, Advances in Natural Computation, pp. 418–418, 2005.

[55]. Brunn M. Text summarization using lexical chains; Proc. Document Understanding Conference; 2001.

[56]. Atserias J. Spanish WordNet 1.6: Porting the Spanish Wordnet Across Princeton Versions; Proc. LREC; 2004.

[57]. Moldovan D and Novischi A, “Lexical chains for question answering”. in Proc. 19th international conference on Computational linguistics- Volume 1: Association for Computational Linguistics, 2002, pp. 1–7.

[58]. Novischi A and Moldovan D, “Question answering with lexical chains propagating verb arguments”. in Proc. 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics: Association for Computational Linguistics, 2006, pp. 897–904.

[59]. Harabagiu SM. Employing Two Question Answering Systems in TREC 2005; Proc. TREC; 2005.

[60. Reeve L. BioChain: lexical chaining methods for biomedical text summarization; Proc. 2006 ACM symposium on Applied computing: ACM; 2006. 180–184.

[61]. Feng L. Cognitively motivated features for readability assessment; Proc. 12th Conference of the European Chapter of the Association for Computational Linguistics: Association for Computational Linguistics; 2009. 229–237.

[62]. Galley M and McKeown K, “Improving word sense disambiguation in lexical chaining”. in Proc. IJCAI, 2003, pp. 1486–1488.

[63]. Cunningham H et al., “Getting more out of biomedical documents with GATE’s full lifecycle open source text analytics”, PLoS computational biology, 9(2), pp. e1002854, 2013. [PubMed: 23408875]

[64]. De Marneffe M-C et al., “Generating typed dependency parses from phrase structure parses”. in Proc. Genoa Italy, 2006, pp. 449–454.

[65]. Wikipedia. [cited 2016 July 1]; Available: https://en.wikipedia.org/wiki/Lists_of_diseases.

[66]. Simple English Wikipedia. [cited 2016 July 1]; Available: https://simple.wikipedia.org/wiki/List_of_diseases.

[67]. Welch BL, “The significance of the difference between two means when the population variances are unequal”, Biometrika, 29(3/4), pp. 350–362, 1938.

[68]. Neyman J and Pearson ES, “On the use and interpretation of certain test criteria for purposes of statistical inference: Part I”, Biometrika, pp. 175–240, 1928.

[69]. McTear M, et al., “The conversational interface”, Springer, 6(94), pp. 102, 2016.

[70]. Zhang Y, et al., “Understanding bag-of-words model: a statistical framework”, International Journal of Machine Learning and ybernetics, 1(1–4), pp. 43–52, 2010.

[71]. Walker SH and Duncan DB, “Estimation of the probability of an event as a function of several independent variables”, Biometrika, 54(1–2), pp. 167–179, 1967. [PubMed: 6049533]

[72]. Quinlan JR, “Induction of decision trees”, Machine learning, 1(1), pp. 81–106, 1986.

[73]. Murphy KP, “Naive bayes classifiers”, University of British Columbia, 2006.

[74]. Cortes C and Vapnik V, “Support-vector networks”, Machine learning, 20(3), pp. 273–297, 1995.

[75]. Ho TK, “Random decision forests,” in Proc. Third International Conference on Document Analysis and Recognition (ICDAR 1995), 1995 IEEE.

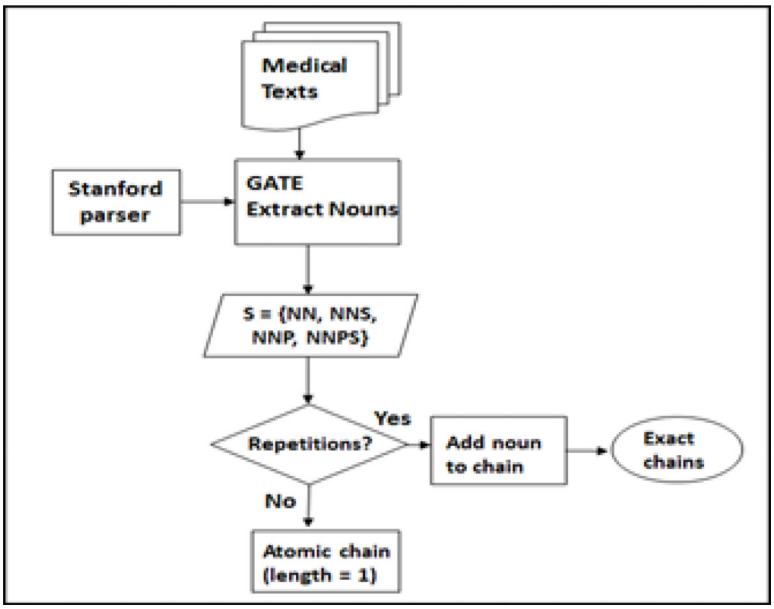

Fig. 1.
Computation of exact chains.

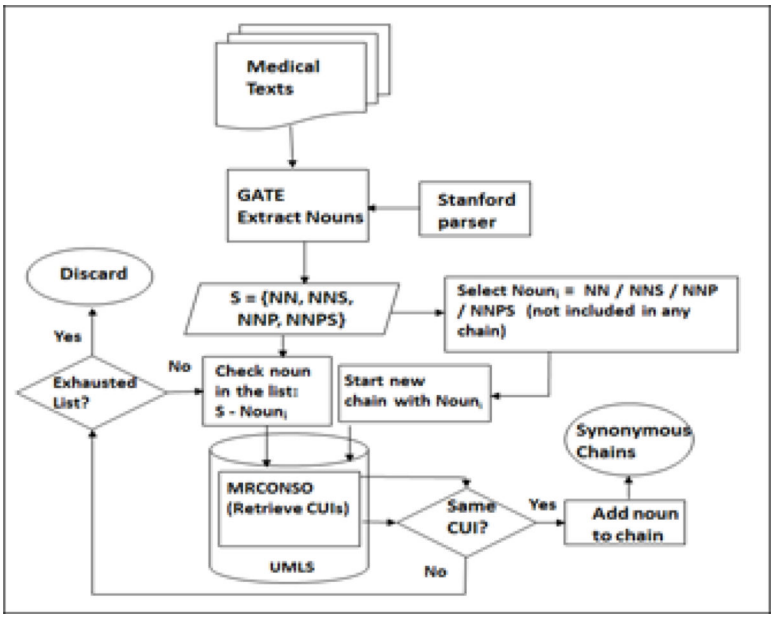

Fig. 2.
Computation of synonymous chains using the UMLS database.

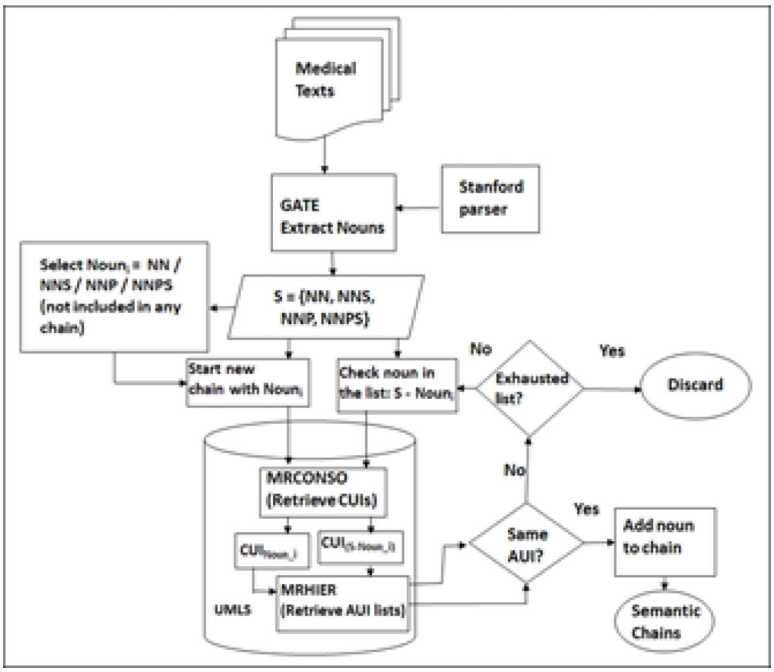

Fig. 3.
Computation of semantic chains using the UMLS database.

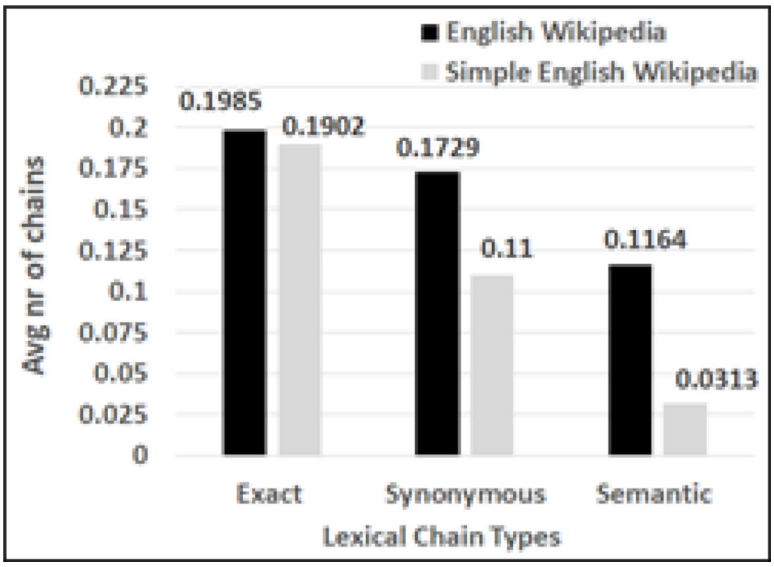

Fig. 4.
Number of lexical chains for medical texts from English and Simple English Wikipedia.

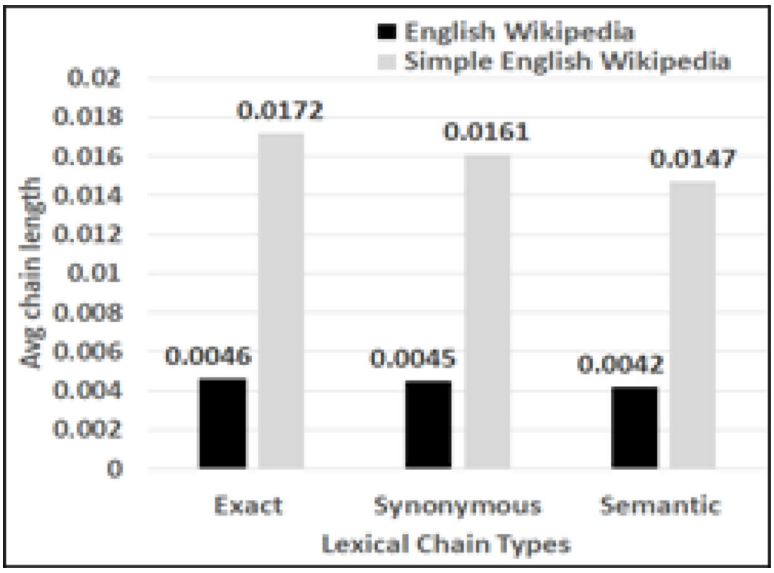

Fig. 5.
Average chain length in medical texts from English and Simple English Wikipedia

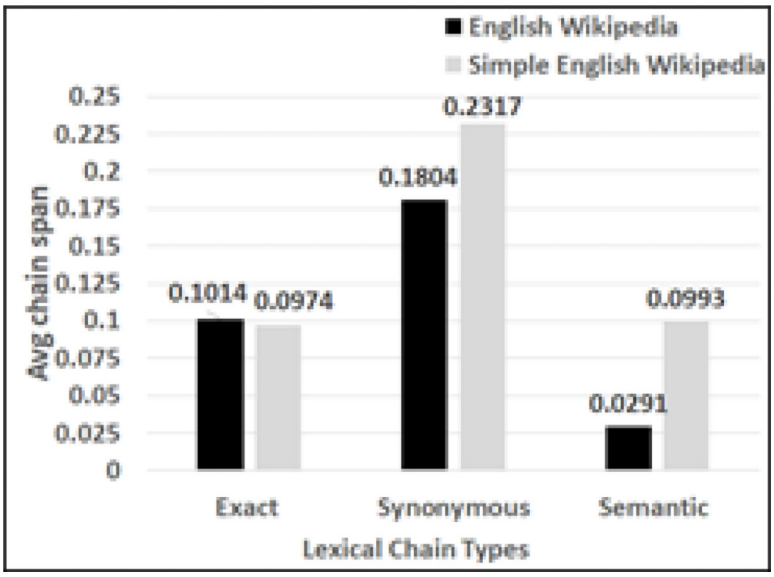

Fig. 6.

Average span of lexical chains for medical texts from English and Simple English Wikipedia.

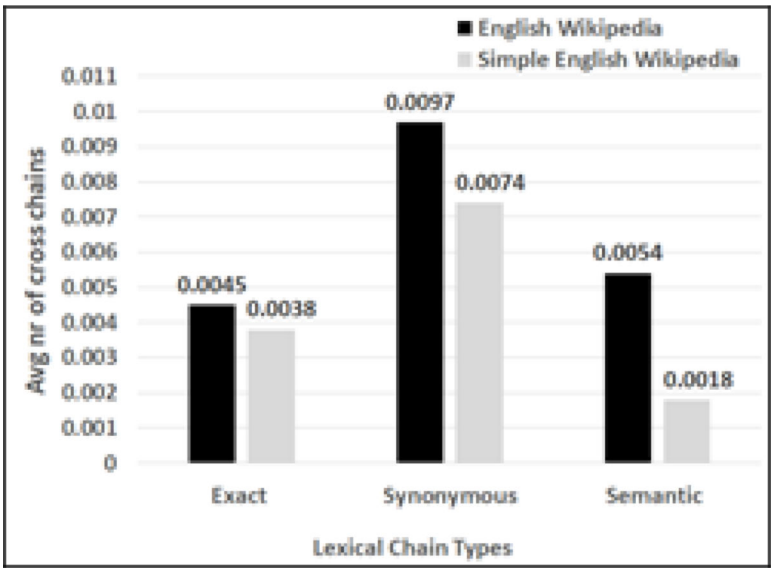

Fig. 7.
Average crossed chain for medical texts from English and Simple English Wikipedia.

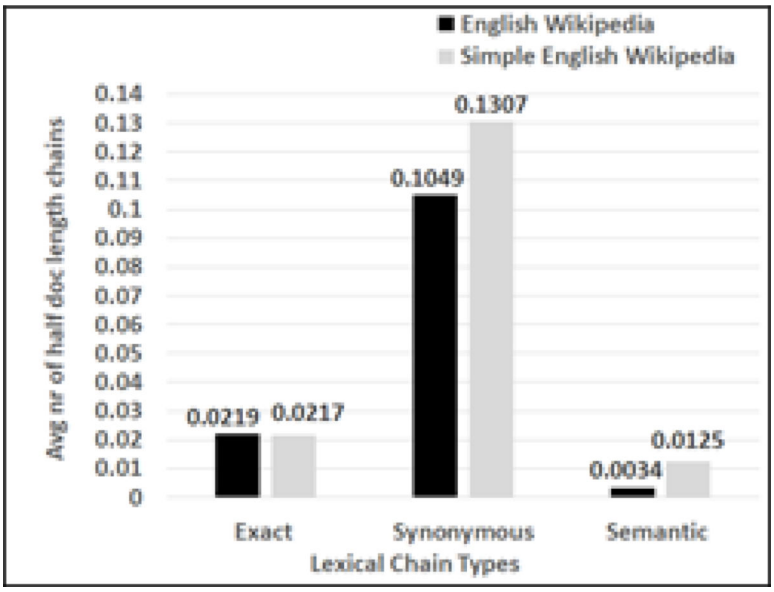

Fig. 8.

Average number of chains that are at least as long as half the document length for medical texts from English and Simple English Wikipedia.

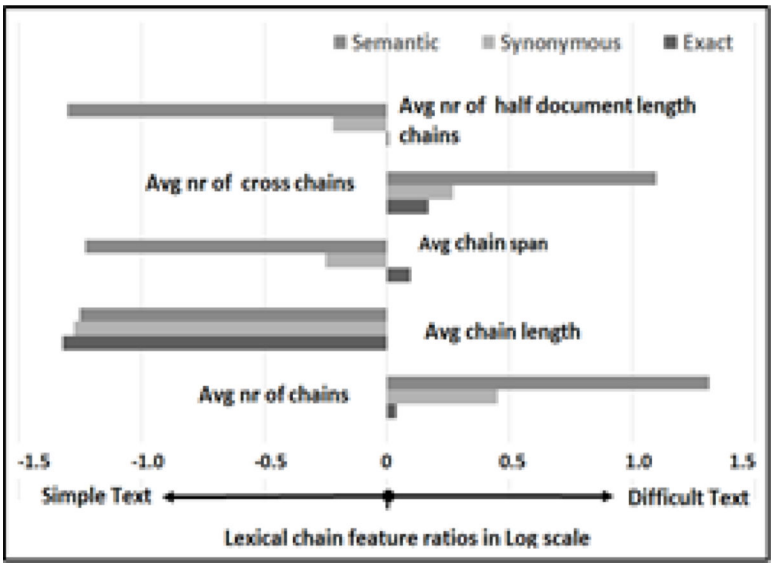

Fig. 9.

Lexical chain feature log-ratios between English and Simple English Wikipedia. Features to the left (negative) occur more in simple texts while those on the right (positive) more in difficult texts.

Author Manuscript

Author Manuscript

Author Manuscript

TABLE I

CORPORA FOR ANALYSIS

Corpus	# of texts	# of sentences	Average sentence length per text
ENGLISH WIKIPEDIA	625	60,108	96.17
SIMPLE ENGLISH WIKIPEDIA	289	6,591	22.81

TABLE II

LEXICAL CHAIN FEATURE STATISTICS AT THE SENTENCE LEVEL FOR DIFFICULT AND EASY TEXTS

Chain	Average number of chains		Average chain length		Average chain span		Average number of cross chain		Average number of chains with more than half document length	
	Diff	Easy	Diff	Easy	Diff	Easy	Diff	Easy	Diff	Easy
EXACT	0.0058	0.0052	0.0003	0.0013	0.0088	0.0067	0.0009	0.0004	0.0071	0.0069
SYNONYMOUS	0.0034	0.0021	0.0004	0.0009	0.0122	0.0226	0.0012	0.0005	0.0073	0.0085
SEMANTIC	0.0023	0.0010	0.0003	0.0007	0.0029	0.0045	0.0010	0.0007	0.0006	0.0009

IEEE J Biomed Health Inform. Author manuscript; available in PMC 2019 September 08.

TABLE III

ACCURACY OF CLASSIFIERS AVERAGED OVER 10-FOLD CROSS VALIDATION

	Bag of Words	Exact Chain	Synonymous Chain	Semantic Chain	All three Chains combined	Bag-of-words + All three Chains combined
LOGISTIC REGRESSION	0.590	0.765	0.738	0.683	**0.830**	0.895
DECISION TREE	0.644	0.772	0.741	0.730	**0.820**	0.905
NAÏVE BAYES	0.550	0.719	0.656	0.692	**0.770**	0.602
SVM (RBF)	0.723	0.756	0.764	0.739	**0.815**	0.786
SVM (LINEAR)	0.624	0.712	0.731	0.710	**0.775**	0.831
RANDOM FOREST	0.777	0.613	0.808	0.872	**0.898**	0.915

TABLE IV

LOGISTIC REGRESSION COEFFICIENTS FOR THREE TYPES OF CHAINS

Chain	Feature	Coefficient	Sig
EXACT	Number of chains	1.91	0.004
EXACT	Average span	−83.77	0.000
EXACT	Average length	−117.91	0.000
EXACT	Average cross-chains	2053.01	0.000
EXACT	Average half-doc-length	−0.63	0.000
SYNONYMOUS	Number of chains	45.96	0.000
SYNONYMOUS	Average span	−0.50	0.000
SYNONYMOUS	Average length	−3.84	0.025
SYNONYMOUS	Average cross-chains	13.97	0.000
SYNONYMOUS	Average half-doc-length	−0.82	0.000
SEMANTIC	Number of chains	6.52	0.000
SEMANTIC	Average span	−0.37	0.000
SEMANTIC	Average length	−1.59	0.000
SEMANTIC	Average cross-chains	13.83	0.000
SEMANTIC	Average half-doc-length	−2.79	0.000