Structural Alignment as the Basis to Improve Significant Change Detection in Versioned Sentences

Tan Ping Ping, Karin Verspoor and Timothy Miller
Department of Computing and Information Systems
University of Melbourne
Victoria 3010, Australia.
{pingt@student., karin.verspoor@, tmiller@}unimelb.edu.au

Abstract

Some revisions of documents can change the meaning of passages, while others merely re-phrase or improve style. In a multi-author workflow, assisting readers to assess whether a revision changes meaning or not can be useful in prioritising revision. One challenge in this is how to detect and represent the revision changes in a meaningful way to assist users in assessing the impact of revision changes. This paper explores a segmentation approach which utilises the syntactic context of revisions to support assessment of significant changes. We observe that length of normalised edit distance or Word Error Rate (WER) correlates better to the significance of the revision changes at sentence level compared to general sentence similarity approaches. We show that our proposed method, SAVeS, supports improved analysis of change significance through alignment of segments rather than words. SAVeS can be used as the basis for a computational approach to identify significant revision changes.

1 Introduction

Revision of documents is a common component of the writing process. In this work, we introduce an approach to analysing revisions that will support the identification of significant changes, such that attention can be focused on revisions that impact meaning.

We define a *versioned text* as a text document that has been revised and saved to another version, where the original version is directly available for comparison. An *edit* is defined as change that involves operations such as insertion, deletion or substitution of characters or words within a revised text. We define a *significant change* between versioned texts as a meaning altering change, which goes beyond string edit operations.

Faigley and Witte (1981) proposed a taxonomy to assist in evaluating the effect of revisions on meaning (Figure 1). They identify a range of revision types. On a general scale, they define *surface changes* as edits that improve readability without actually changing the meaning of the text, and *text-base changes* as edits that alter the original meaning of the text. These categories are subdivided. The subcategories for surface changes: *formal changes* includes copy editing operations such as correction in spelling, tense, format, etc., while *meaning preserving changes* includes rephrasing. For text-base changes, *microstructure changes* is meaning altering changes which do not affect the original summary of the text and *macrostructure changes* are major changes which alter the original summary of the text. Although they provided some examples, the definitions are insufficient for computational implementation.

Framed by this taxonomy, we consider significant change to be a macro-structure revision change while a minor meaning change is a microstructure revision. We adopt surface revision change to be no meaning change. Based on one original sentence, we provide examples of how we distinguish between meaning-preserving, microstructure and macro-structure revision changes in Table 1.

While some applications use tools like *diff* or come with 'track changes' capability that highlights changes, readers must manually assess the significance over a change, which can reduce efficiency when the number of revisions increases.

In this paper, we demonstrate empirically that general string similarity approaches have weak correlation to significance in revised sentences. We have conducted a preliminary study on a set of revised software use case specifications (UCS)

Ping Ping Tan, Karin Verspoor and Tim Miller. 2015. Structural Alignment as the Basis to Improve Significant Change Detection in Versioned Sentences . In *Proceedings of Australasian Language Technology Association Workshop*, pages 101−109.

Original Sentence	I paid a hundred dollars for the tickets to take my family to a movie.
Revision Type	**Example of Sentence Revisions**
Meaning preserving	I paid a hundred dollars to take my family to a movie.
Micro-structure	I paid a hundred dollars for the tickets, with popcorn and drinks, to bring my family to a movie.
Macro-structure	We decided to watch movie at home.

Table 1: Examples of sentence revision according to revision types

to provide insight into the identification of significant changes between versioned text documents, with particular focus on how impact of revision changes is assessed. The analysis highlights that an approach that considers the syntactic scope of revisions is required for meaning changes assessment.

We will present our proposed method, structural alignment of versioned sentences, SAVeS that addresses this requirement. We provide a performance comparison to three other word segmentation approaches. The broader aim of this research is to develop a computational approach to automatically identifying significant changes between versions of a text document.

2 Related Works

Research on revision concentrates on detecting edits and aligning sentences between versioned text documents. Considering sentences from the first and last draft of essays, Zhang and Litman (2014; 2015) proposed an automated approach to detect whether a sentence has been edited between these versions. Their proposed method starts with sentence alignment, and then identifies the sequence of edits (i.e., the edit operations of Add, Modify, Delete and Keep) between the two sentences. They further consider automated classification of the reason for a revision (i.e., claim, evidence, rebuttal, etc.), which they hypothesised can help writers to improve their writing. Classifying revisions based on the reasons of revision does not indicate the significance of revision changes. What we are attempting is to represent these revision changes in a meaningful way to assist in assessment of the significance. We concentrate on identification of significant revision changes, or revision changes that have higher impact of meaning change for the purpose of prioritising revision changes, especially in multi-author revision. Nevertheless, the work by Zhang and Litman (2014; 2015) provides insights to revisions from a different perspective.

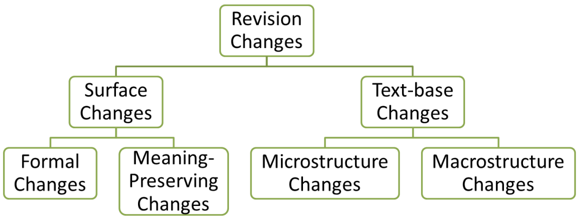

Figure 1: Taxonomy for revision analysis (Faigley and Witte, 1981)

Research has shown that predefined edit categories such as fluency edits (i.e. edits to improve on style and readability) and factual edits (i.e. edits that alter the meaning) in Wikipedia, where revision history data is abundant, can be classified using a supervised approach (Bronner and Monz, 2012; Daxenberger and Gurevych, 2013). The distinction of the edits can be linked to Faigley and Witte's (1981) taxonomy: fluency edits to surface changes and factual edits to text-base changes. Supervised classification would be difficult to apply to other types of revised documents, due to more limited training data in most domain-specific contexts. They too did not consider the significance of edits.

As our task is to align words between versioned sentences to assist in identification of significant changes between versioned texts, it is important to consider the semantics of sentences. Lee et. al. (2014) reviewed the limitations of information retrieval methods (i.e., the Boolean model, the vector space model and the statistical probability model) that calculate the similarity of natural language sentences, but did not consider the meaning of the sentences. Their proposal was to use link grammar to measure similarity based on grammatical structures, combined with the use of an ontology to measure the similarity of the meaning. Their method was shown to be effective for the problem of paraphrase. Paraphrase addresses detecting alternative ways of conveying the same information (Ibrahim et al., 2003) and we observe paraphrase problem as a subset to our task because sentence re-phrasing is part of revision. However, the paraphrase problem effectively try to normalize away differences, while versioned sentences analysis focuses more directly on evaluating the meaning impact of differences.

3 Dataset

The dataset that we study is a set of revised software requirements documents, the Orthopedic Workstation (OWS) Use Case Specifications (UCS) for Pre-Operative Planning for the Hip. We were provided with two versions, version 0.9 (original version, *O*) and version 1.0 (revised version, *R*). Version 1.0 has been implemented as software in a local hospital. Similar to most use case specification documents, the flow of software events, pre- and post-conditions as well as the list of glossary terms are available. The list of glossary terms contains 27 terms with 11 terms having more than one word.

A version that is created immediately following a previous version results in back-to-back versions; these tend to have high similarity to each other. Our dataset consists of back-to-back versions; previous works concentrate on the first and last drafts (Hashemi and Schunn, 2014) (Zhang and Litman, 2014). Therefore in this dataset, we observe more versioned sentences with minor edits that change the meaning substantially (Table 2). Such minor edits are more challenging to determine the significance, from a semantic perspective. These minor edits can be so specific that particular domain knowledge is required to comprehend the changes. We observe 23 pairs of versioned sentences, other than addition and deletion of sentences within this dataset.

4 Introspective Assessment of Revisions

In addition to the summary approach as defined by Faigley and Witte (1981), another approach to distinguish between macro- and micro-structure changes is to determine whether the concepts involved in a particular change affect the reading of other parts of the text. Their definitions are conceptual, for example, they use the notion of a 'gist' to distinguish micro- and macro-structure, but offer no concrete definition of this, such as whether the length of the summary is important, or how much reading of the other parts of the text is influences the summary. Thus, they are not directly suitable as a computational definition. Based on the example in Table 1, we argue that for most cases, micro- and macro-structure can be differentiated without reading the surrounding text, beyond the revised sentences. As our broader objective is to develop a computational method, we conduct our introspective assessment starting at the sentence level, where Zhang and Litman (2014) have demonstrated to work computationally.

Original Sentence, S_O	**Revised Sentence,** S_R
Store X-ray with Current Patient Information.	Store OWS X-ray as Annotated X-ray with Current Patient Record.
Calculate Offset of Non-Destroyed Hip.	Calculate Offset of Normal (Contra-lateral) Hip.
Select material for Insert.	Select material, internal diameter, and other attributes e.g. low profile, extended rim of Insert.

Table 2: Examples of Versioned Sentence Pairs

We observe that changes can be divided into the following three categories:

- **No change:** A pair of sentences which are identical between the versioned texts.
- **Local change:** A change (i.e. word or words added, deleted or modified) where the impact is confined to a pair of versioned sentences.
- **Global change:** A sentence (i.e. added or deleted) where the impact of change is beyond that sentence, for example, at the paragraph or document level.

We will show examples of local changes by considering the first sentence pair in Table 2. A *diff* identify the insertion of "OWS" and "as", "Annotated" and "X-ray", followed by substitution of "Information" to "Record". Based on these edits, readers can roughly estimate words that have changed but cannot assess how much of the meaning has changed. Readers will note that "X-ray" is changed to "OWS X-ray", "as Annotated X-ray" is added and "Patient Information" is substituted with "Patient Record". Readers can only deduce whether the change has any impact when they compare the two versions. “OWS” is the acronym of the system. Although both “OWS X-ray” and “Annotated X-ray” require auxiliary knowledge to identify and understand the changes, the assessment of the impact of the changes is confined within these two sentences or the text surrounding the edits but still within the two sentences. These are examples of *local changes*.

The edit operations observed correspond to the primitive edit operations identified by (Faigley and Witte, 1981; Zhang and Litman, 2014). In our data, there is a minimum of one edit per sentence pair and a maximum of three edits between the pairs. An edit itself can consist of one or multiple words. Substitution and deletion of words and sentences do occur, but a large number of the edits involve adding words to the later version. Most additions provide more clarification; 16 out of the local (i.e., word) additions contribute to either minor or major meaning change. Thus, local changes can be either significant or not.

Global changes have no matching or similar sentence between the two versions, unlike the other two changes. Most of the assessment of the impact of global changes is based on the preceding sentences, which can be either a revised sentence or an unchanged sentence. Even though we do not work on global changes in this paper, we provide an example differentiating local and global changes (Table 3).

Original, O	**Revised, R**
Label pathology on X-ray.	Label pathology on Annotated X-ray. Predefined Labels includes suggestions.
Local changes	‘X-ray’ to ‘Annotated X-ray’
Global Change	‘Predefined Labels includes suggestions.’

Table 3: Example of Local and Global Changes

Our introspection highlights three main things, which serve as motivation for this work:

- The need for local and global changes to be differentiated, before micro- and macro-structure differentiation.
- The way readers assess the impact of change depends upon both syntactic and semantic understanding of the changes.
- The words surrounding the edits are useful for assessment of impact of revision changes.

5 Structural Alignment of Versioned Sentences

Chomsky (2002) suggested that “structure of language has certain interesting implications for semantics study”. The idea of using sentence structure in natural language specification to describe program input data has been proposed by Lei (2013). Based on this notion, and the understanding of how local changes are assessed through our introspective study, we present a method to group words into segments. Specifically, we propose to use the sentence structure, corresponding to the syntactic context of the edited words, to assist in alignment of versioned sentences. Then we make use of these segments in assessing the impact of revision changes.

Our proposed Structural Alignment for Versioned Sentences (SAVeS) method starts by performing tokenization, where each word is treated as single token, for each of the sentences, producing T_{S_O} and T_{S_R}. Tokens that are the same between T_{S_O} and T_{S_R} are aligned, leaving the edited words from each sentence, E_{S_O} and E_{S_R}. In a separate process, each of the sentences serves as input to a syntactic parser, producing individual parse trees, PT_{S_O} and PT_{S_R}. SAVeS matches each of the edited words to the leaves of the parse trees, then extracts the head of the noun phrase for each edited word. The tokens in T_{S_O} and T_{S_R} are updated according to the grouped words (i.e. noun phrase of the edited words), producing T'_{S_O} and T'_{S_R}. Words that are not part of an edited phrase continue to be treated as individual tokens. Using S_O from the first example in Table 2, we provide a sample of how SAVeS captures the context of the edited word (in this case: ’information’) in Figure 2 and the full SAVeS algorithm appears in Table 4.

SAVeS uses general sentence structure, therefore, is applicable to different types of phrases. In this dataset, majority of phrases are noun phrases. As a preliminary, we work on noun phrasest.

Algorithm	Structural Alignment of Versioned Sentences
Input	Versioned Sentences: Original Sentence, S_O and Revised Sentence, S_R
Output	Word Error Rate, WER
	POS - Part Of Speech
	NP - Noun Phrase

```
For each sentence,
   $T_S$ = Tokenise each word in the sentence
End For
Align the words that are the same between $T_{S_O}$ and $T_{S_R}$,
   Extract the edited words for each of the sentence
For each of the sentence,
   $PT_S$ = Constituency-based parse tree
   For each of the edited word
      For each leaf
         If leaf value = edited word,
            While node POS not equal to NP,
               Get the POS of the parent of node
            End While
            Extract the NP
         End If
      End For
   End For
End For
For each of the extracted phrases
   $T'_S$ = Group the tokens based on the extracted phrase
End For
```

Table 4: Algorithm for Structural Alignment of Versioned Sentences

6 Experimental Setup

6.1 Measuring Revisions

The experiments measure revision changes at sentence and word segmentation level. String similarity is used to measure the surface similarity of two sentences, while semantic similarity measure whether two sentences have the same meaning. Therefore, we consider pairwise string and semantic similarity between sentences; pairs that are more different are considered to have more significant changes.

Given two strings, x and y, the edit distance between x and y is the minimum editing path to transform x to y, where edit path covers operations like substitution, insertion and deletion of word or character, taking into consideration of word order. Our work on revision sentences observes the transformation from the original sentence, S_O to the revised sentence, S_R. The length of the sentences can vary. Hence, we consider the length of normalised edit distance or Word Error Rate (WER) (Equation 1). WER is an automatic evaluation metric commonly used in machine translation to observe how far off the system output is from a target translation. In our case, it is used to automatically measures how different S_O and S_R is.

$$WER(S_O, S_R) = \frac{W(P)}{maximum\_length(S_O, S_R)} \tag{1}$$

Where:

P is minimum edit distance between S_O and S_R,

W(P) is the sum of the edit operations of P, where weight is added for edit operation involving word in the glossary for the weighted glossary experiment.

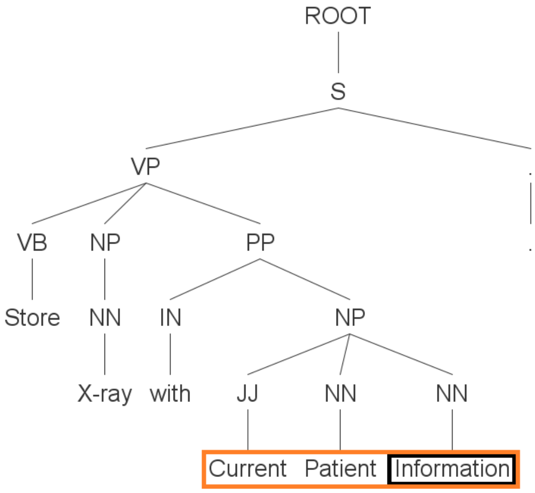

Figure 2: Example how SAVeS capture the context surrounding the edited word

6.2 Annotation

Before we can consider a suitable measurement for revision changes between versioned sentences, manual intuitive annotation is performed by an annotator, with review from one other. The versioned sentences are annotated based on significance of the changes, framed by Faigley and Witte's (1981) revision analysis taxonomy. We compared the original sentence, S_O, to the revised sentence, S_R, and for each sentence pair determined whether there is a meaning change. We first differentiate between surface and text-base revision changes. If the revision is a text-base change, we further distinguish between the micro- or macro-structure levels. The versioned sentences can have more than one local change; therefore, we annotate the sentence pair as *non significant, minor* and *significant* change based on the most significant change for that sentence pair.

Each of the measurements stated in Section 6.1 is plotted against this human annotation of significance, followed by the calculation of correlation coefficient, r values between the labels. If r value closer to 1, the measurement correlates better with the significance, while opposite correlation is observed for negative r value. When r value is closer to 0, weak correlation between the variables.

6.3 Similarities and Significant Revisions

The versioned sentence pairs serve as the input to the similarity approaches, and the output is the similarity values for each of the sentence pairs. For string similarity measurement, we used Jaro-Winkler proximity (Cohen et al., 2003). Automatic machine translation evaluation metrics, which normally integrate with linguistics knowledge, is used to measure how semantically similar between the translation output of a system to the parallel corpus without human judgement. This approach is also used for paraphrase evaluation (Madnani et al., 2012). For semantic similarity, we adopted one of the metrics, Tesla (Liu et al., 2010), which is linked to WordNet as our semantic similarity measurement between versioned sentences.

6.4 Word Segmentation impact on revision

For the task of word segmentation, we consider four scenarios. In each case, the alignment is computed using edit distance based on the relevant segmentation (considering insertions, deletions, and substitutions of *segments*). The word error rate (WER) or the length of normalised edit distance (Equation 1) is computed on the basis of this alignment.

- **Baseline:** We use the standard approach of treating a single word as a single token. In the alignment of S_O and S_R, matching tokens are aligned. We use this as the baseline approach.
- **Glossary:** In this approach, we consider changes in domain-specific terminology are more likely to impact the meaning of the sentence. Instead of just tokenizing on the individual terms as separate tokens, the terms that exist in the glossary terms are grouped together as a token, while the other words remained as single tokens.
- **Weighted Glossary:** Here, we consider that edited words in the versioned sentences that exist in the glossary list may have more importance. We added weights to these edited words in the edit distance calculation to emphasize their importance in aligning the glossary terms. In this scenario, similar to the second scenario, the glossary is used to guide tokenization, with addition that penalizes edits involving these glossary-based tokens more heavily. As there is no previous work on the optimal weight to use for aligning versioned sentences, we experimented with a weight value of $+2$.
- **SAVeS:** SAVeS is implemented based on the algorithm in Figure 4. The updated tokens are

Approach	r
String Similarity	-0.34
Semantic Similarity	-0.59
Tokenization approaches:	
Baseline	0.63
Glossary Terms	0.66
Weighted Glossary Terms	0.68
SAVeS	0.58

Table 5: Correlation coefficient (r) values between similarity measurement and significant changes, using various approaches to similarity assessment.

re-aligned based on the noun phrases. The Stanford parser (Klein and Manning, 2003) we used produced parse trees with minor errors in some sentences. To eliminate issues in the results related to the incorrect parsing, we manually corrected errors in the parse trees, thus assuming the existence of a 'perfect' parser.

7 Results and Discussion

Table 5 shows that semantic similarity has a stronger negative correlation to significant changes when compared to string similarity but the baseline approach of single word token alignment correlates better to significant changes. This result shows that semantic similarity could be used to filter out non-significant revised sentences before further evaluation of micro- and macro-structure assessment.

Using the weighted glossary term tokenization approach, the WER correlates best with the significance at sentence level, compared to the other tested approaches. A domain specific dataset clearly benefits from specific knowledge of terminology. However, we still do not understand the most appropriate weights to use. A more detailed study is required to fully determine the optimal weights for integrating the glossary to assist in producing an analysis of the impact of revision changes.

The human annotation of significance is based on the highest significance between the versioned sentence pair. Although for cases where there is more than one changes between the versioned sentence pairs, using WER evaluation cannot pinpoint which among the changes in that sentence pair is indeed significant.

Table 6 presents an analysis of the effect of different tokenization approaches and WER, based on the first example in Table 2, where the glossary terms are 'annotated x-ray‘ and 'patient information'. When we examine the changes after alignment more closely, the baseline approach outputs the edits between the two sentences without much indication of meaning changes. The glossary terms tokenization approach is able to treat 'annotated x-ray' as a single insertion and although 'patient record' appears as a segment but aligns to 'patient' it is not reflective of the meaning change, instead for this change, 'patient record' is substituted to 'patient information' should be a better representation to evaluate the meaning change.

Weighting glossary terms emphasizes the changes introduced by a shift in core terminology, the addition of 'annotated x-ray'. SAVeS identifies the main segments: 'annotated x-ray', which we can deduce as insertion of a noun phrase, 'x-ray' is substituted with 'ows x-ray', which we can be deduced is a type of X-ray and 'current patient information' is substituted with 'current patient record' which shows us, this is a possible meaning preserving change.

When we compare the relationship between these different tokenization approaches and the WER, we see that the weighted glossary term tokenization approach reflects a larger change between the sentences (i.e., WER = 0.78) compared to other tokenization approaches.

We examined the impact of the different tokenization approaches on the WER, according to the manually assigned significance category (Table 7). For the significance categories of *None* and *Minor*, the alignment using SAVeS measures less change (i.e. substitution, insertion and deletion) as compared to other tokenization approaches.

Consider the second example in Table 2. SAVeS extracted phrases that contain the edited words and aligned them, rather than individual words: the full phrase 'non-destroyed hip' is aligned by the phrase 'normal (contra-lateral) hip'. In this case, the WER for single word single token alignment (i.e., baseline) is 0.33 while SAVeS produces 0.25. SAVeS reflects that the scope of the edits is limited to one (syntactically bounded) portion of the sentence.

SAVeS highlights meaning changes by supplying the information that the full phrase 'non-

Tokenization Approach	Tokens	WER	Changes Detected
Baseline	S_O = {store, ows, x-ray, as, annotated, x-ray, with, current, patient, record} S_R = {store, x-ray, with, current, patient, information}	0.5	*insertion*: 'ows', 'as', 'annotated', 'x-ray' *substitution*: 'record' to 'information'
Glossary Terms	S_O = {store, ows, x-ray, as, annotated x-ray, with, current, patient record} S_R = {store, x-ray, with, current, patient, information}	0.56	*insertion*: 'ows', 'as', 'annotated x-ray' *substitution*: 'patient' to 'patient record' *deletion*: 'information'
Weighted Glossary Terms	S_O = {store, ows, x-ray, as, annotated x-ray, with, current, patient record} S_R = {store, x-ray, with, current, patient, information}	0.78	*insertion*: 'ows', 'as', 'annotated x-ray' (weight: +4) *substitution*: 'patient' to 'patient record' (weight: +4) *deletion*: 'information'
SAVeS	S_O = {store, ows x-ray, as, annotated x-ray, with, current patient record} S_R = {store, x-ray, with, current patient information}	0.67	*insertion*: 'as', 'annotated x-ray' *substitution*: 'x-ray' to 'ows x-ray', 'current patient information' to 'current patient record'

Table 6: An example of tokenization effect and WER.

destroyed hip' is substituted by 'normal (contralateral) hip'. Deduction of the impact can only be made if this substitution is analysed in more depth. Observe that the rightmost noun in the phrase (i.e., 'hip'; the syntactic and semantic head of the phrase) did not change; this too may have implications for the assessment of meaning. A few more other examples of the effect of SAVeS through analysis of the tokens alignment can be considered: 'surgeon authentication' is aligned to 'authentication' or 'labelled image' is aligned to 'labelled annotated x-ray' where other tokenization approaches cannot chunk and align these changes. The advantage of SAVeS over the glossary terms approach is that not all of the terms exist in the glossary list. Using the sentence syntactic structure, SAVeS is applicable to any sentence.

For the case of significant revision changes, the changes are small irrespective of the tokenization approach. This is due to the nature of our dataset; back-to-back versions. The small average WER across the category of significant changes shows that edits alone are insufficient to bring out the semantics of the changes.

Significance	SAVeS	BL	Gl	W-Gl
None	0.24	0.25	0.26	0.33
Minor	0.35	0.45	0.46	0.49
Significant	0.19	0.24	0.25	0.25

Table 7: The average WER by revision significance, based on each different tokenization approach (BL=baseline, Gl=glossary, W-Gl=weighted glossary)

.

We hypothesize that phrases will provide a better representation for meaning change analysis between versioned sentences than individual tokens, and further suggest that measuring edits at the phrasal level will lead to an improvement in our ability to computationally determine the significance of changes.

In a multi-author environment, the current tools only provide the edits of the revision but SAVeS indicates which of the noun phrases have changed. We hypothesise that this form of indicator is more useful to authors.

8 Conclusion

Our introspective assessment of revision changes in versioned use case specifications revealed that changes can be categorised into local and global changes, and that there exist versioned sentences which can be superficially similar and yet reflect substantial differences in meaning. In order to make direct comparison between changes for the purpose of assessment, we need to consider the context of a change. We empirically show that alignment of words between versioned sentences using word error rate correlates better to significance of a revision. In this paper, we have explored several approaches to aligning versioned sentences in this context. Our analysis of the alignment shows that incorporating structural information of the text affected by an edit is useful for taking into consideration the scope of an edit in its sentential context. We further demonstrate that similarity approaches are insufficient for our task.

We speculate that a phrasal representation of revisions will also be better for human readability of edits during manual assessment of the significance of changes, and plan to assess this in future work. This is a preliminary study and we plan to consider other kinds of versioned documents.

References

Amit Bronner and Christof Monz. 2012. User edits classification using document revision histories. In *EACL*, pages 356–366. Assoc. for Computational Linguistics.

Noam Chomsky. 2002. *Syntactic structures*. Walter de Gruyter.

WW Cohen, P Ravikumar, and S Fienberg. 2003. Secondstring: An open source java toolkit of approximate string-matching techniques. *Project web page: http://secondstring. sourceforge. net*.

Johannes Daxenberger and Iryna Gurevych. 2013. Automatically classifying edit categories in wikipedia revisions. In *EMNLP*, pages 578–589.

Lester Faigley and Stephen Witte. 1981. Analyzing revision. *College composition and communication*, pages 400–414.

Homa B Hashemi and Christian D Schunn. 2014. A tool for summarizing students changes across drafts. In *Intelligent Tutoring Systems*, pages 679–682. Springer.

Ali Ibrahim, Boris Katz, and Jimmy Lin. 2003. Extracting structural paraphrases from aligned monolingual corpora. In *Proceedings of the Second int'l workshop on Paraphrasing*, pages 57–64. Assoc. for Computational Linguistics.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *ACL*, pages 423–430. Assoc. for Computational Linguistics.

Ming Che Lee, Jia Wei Chang, and Tung Cheng Hsieh. 2014. A grammar-based semantic similarity algorithm for natural language sentences. *The Scientific World Journal*, 2014.

Tao Lei, Fan Long, Regina Barzilay, and Martin C Rinard. 2013. From natural language specifications to program input parsers. Assoc. for Computational Linguistics.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2010. Tesla: Translation evaluation of sentences with linear-programming-based analysis. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics*, WMT'10, pages 354–359, Stroudsburg, PA, USA. Assoc. for Computational Linguistics.

Nitin Madnani, Joel Tetreault, and Martin Chodorow. 2012. Re-examining machine translation metrics for paraphrase identification. In *NAACL/HLT*, pages 182–190. Assoc. for Computational Linguistics.

Fan Zhang and Diane Litman. 2014. Sentence-level rewriting detection. *ACL 2014*, page 149.

Fan Zhang and Diane Litman. 2015. Annotation and classification of argumentative writing revisions. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 133–143. Assoc. for Computational Linguistics.