An Empirical Analysis of Edit Importance between Document Versions

Tanya Goyal
tanyagoyal.93@gmail.com
Big Data Experience Lab
Adobe Research

Sachin Kelkar
sachinkel19@gmail.com
Indian Institute of Technology, Roorkee

Manas Agarwal
manasagarwal1993@gmail.com
Indian Institute of Technology, Roorkee

Jeenu Grover
groverjeenu@gmail.com
Indian Institute of Technology, Kharagpur

Abstract

In this paper, we present a novel approach to infer significance of various textual edits to documents. An author may make several edits to a document; each edit varies in its impact to the content of the document. While some edits are surface changes and introduce negligible change, other edits may change the content/tone of the document significantly. In this paper, we perform an analysis of the human perceptions of edit importance while reviewing documents from one version to the next. We identify linguistic features that influence edit importance and model it in a regression based setting. We show that the predicted importance by our approach is highly correlated with the human perceived importance, established by a Mechanical Turk study.

1 Introduction

In collaborative content authoring, multiple authors make changes to the same document, which results in the final version being significantly different from the base draft. Often there is a need to review the edits made to the original document, which can be a long and arduous task. Tools like Microsoft Word (mic) and Adobe Acrobat (ado) provide reviewers with a list of edits, in the form of insertions and deletions. While helpful, these tools do not differentiate between the different types of edits, or consider the varying impact of edits. For instance, change from numeric '18' to word 'eighteen' may be a minor change and less crucial for the author to review, as compared to an edit that alters the facts of the document. Thus, in our work, we focus on automatically inferring the impact/change introduced by edits, and predict the perceived importance of such edits by authors.

In this paper, we perform a linguistic analysis of how humans evaluate the significance of edits while reviewing documents. Our algorithm assigns scores to edits between two versions of a document, which indicate the significance of the specified edit as perceived by the reviewer. We demonstrate the efficacy of our approach by comparing our algorithm generated edit importance scores with the human perceived ground truth importance, established through a Mechanical Turk survey.

2 Related Work

There has been significant amount of work on defining the importance of a keyword or a sentence in the context of document summarization (Mihalcea and Tarau, 2004). Some prior work has also been done on inferring the type of edits between Wikipedia versions. Bronner et al. (Bronner and Monz, 2012) proposed a supervised approach to classify Wikipedia edits as factual or fluency. Daxenberger et al. (Daxenberger and Gurevych, 2013) propose an approach to classify these edits into a 21-category taxonomy. However, none of the the prior work studies the impact or significance of the edit to the content of the document. They do not take the context of the change into account, neither do they study how edits are perceived by reviewers and the significance associated to each edit type. To the best of our knowledge, there is no prior work that evaluates the importance of an edit between document versions as perceived by human reviewers, which is the novel contribution of our work here.

All authors have equal contribution in this paper.

This work was done as part of an internship at Adobe Research

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2780–2784
Copenhagen, Denmark, September 7–11, 2017.

3 Discussion on types of edits

Before trying to automatically infer the importance of individual edits, we first identified the broad categories of textual edits made by authors. Bronner et al. (Bronner and Monz, 2012) broadly classify text edits into two categories, namely, *Factual Edits* and *Fluency Edits*. Factual edits refer to those that modify, add or delete information in the document while fluency edits mainly deal with changes in writing styles or paraphrasing. To obtain finer granularity edit categories, we subclassified factual edits into *Information Modify, Information Delete* and *Information Insert*. To further classify fluency changes, we looked at linguistic literature (Honeck, 1971) and identified subcategories of paraphrase changes. Based on this, fluency edits were further classified into *Lexical Paraphrase* (change of textual elements by synonymous words/phrases/numbers) and *Transformational Paraphrase* (change in the structure of the sentence, e.g. active to passive voice).
For the purpose of this paper, we assume these edit categories to be exhaustive and consequently classify all changes as belonging to one of these.

4 Data and Annotation

Due to the unavailability of an appropriately annotated dataset, we performed an online survey on Amazon Mechanical Turk[1] to capture people's perception of edit importance. To achieve this, we used an available corpus of news articles[2]. We created newer versions for these articles by manually introducing multiple changes to each article. Fig 1 provides statistics for the types of edits (based on the discussion in the previous section) across this entire corpus of 52 article pairs. There are a total of 523 changed sentence pairs in the document corpus, and an average of 1.2 edits per sentence pair.

For annotation of edit importance, we asked Mechanical Turk workers to assign an importance score to each pair of changed sentences. We first provide each turker with the initial (original) version of a news article. After the turker finishes reading the article, he is presented with a list of sentences that were changed between the initial and the final version, along with the changed sentences. The worker classifies each of these

[1]https://www.mturk.com/mturk/
[2]http://literacynet.org/cnnsf/archives.html

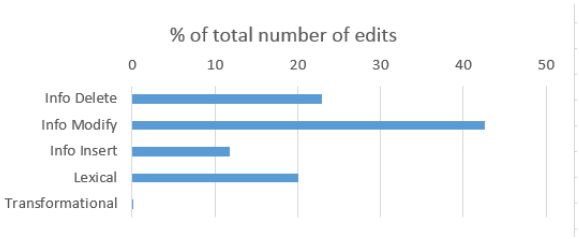

Figure 1: Number of edits of each type as a percentage of the total number of edits in the entire document corpus

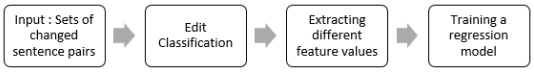

Figure 2: Steps in training a supervised model to scores sentences

sentence pairs as belonging to one of the following importance classes (a) Very Important, (b) Moderately Important, (c) Important, (d) Neutral, (e) Not necessary for review. To avoid introducing biases based on our own notions of edit importance, we provide only a brief description of the task and encourage annotators to follow their own intuitions of importance. Each sentence pair is annotated by 3 annotators and the final importance score is calculated as the mean of the three scores.

5 Solution Description

This section describes the methodology followed to obtain importance scores for text edits to documents. Fig 2 shows the overall workflow of the proposed approach.

The input to the algorithm is a corpus of documents D, which consists of pairs of documents (d, d') corresponding to the initial and final document versions respectively. We use the sentence alignment module proposed by Zhang et al. (Zhang and Litman, 2014) to obtain the mapped pairs of sentences (s, s'), where s represents a sentence in the first version d and s' is its modified variant in d'.

5.1 Classification of Edits

The first step of the algorithm is to determine the number and types of the various edits between each sentence pair (s, s'). In this module, we assign an edit type label, as discussed in Section 3, to all edits between a pair of sentences.

Our analysis of the document corpus revealed that most text edits to sentences do not significantly change the structure of the sentence. Thus, a simple heuristic based approach can be used to identify edit types. We use the Stanford Parser to extract the POS tag sequences of the two sentences, with words backed off to their named entities wherever possible. We identify the longest common subsequence between these to obtain a word to word mapping for the sentence pair. If the ratio of the LCS and the mean of the sentence lengths (original and the modified) is above a threshold, we assume that the sentence structure is preserved. A simple word comparison between the similarly tagged words reveals instances of Information-Modify and Lexical changes; additions and deletions are identified as Information-Insert and Information-Delete respectively. In case the structure of the sentence is not preserved, the above heuristic fails, and we tag the sentence pair as Transformational Paraphrase. For such sentence pairs, we employ the method outlined by Bronner et al. (Bronner and Monz, 2012) and train a supervised classifier to differentiate between factual and fluency edits, without bothering about the subtype.

Following the outlined heuristic, we were able to correctly classify 92% of all edits in the document corpus, without using the supervised classifier.

5.2 Feature Extraction

Next, we extract linguistic features for supervised modeling of edit importance. We hypothesized that the importance of edits would be affected by both the nature of the edits, characterized by the aforementioned categories, as well as the relevance of the sentence to the content of the document. Thus, we chose features that capture both these aspects and have divided them into two groups, namely, change-related features and relevance-related features.

5.2.1 Change-related features

These set of features account for the factual differences between sentence pairs caused due to the edits. The complete list of such features is as follows:

- One-hot feature for type of edits identified in the Edit Classification module. We conjectured that different types of changes will have different perceived importance. For example, factual changes may be more important for the author to review compared to paraphrasing changes.
- One-hot feature for the POS tags and Named Entities whose count changes between the initial and the final version of the sentences. We also include one-hot features for those tags whose corresponding word changes between the two versions. These aim to capture the importance associated with deletion, insertion or modification of specific POS tags and Named Entities.
- One-hot features for the following dependency tuples that change between the two versions, with lexical items backed off to POS tags: (gov,typ, dep), (gov, typ), (typ, dep), (gov, dep).
- Count for the number of edits between the two versions.
- Absolute difference in the Flesch Kinkaid readability scores of the two sentences. We hypothesized that human perception of degree of change may be correlated with the change in ease of readability of content.

5.2.2 Relevance-related features

These features aim to score the sentences where the edit occurred. We conjectured that edit importance must also depend on the relevance of the underlying sentence to the content of the document. For instance, in an article about the monarchy in the United Kingdom, an edit that occurs in a sentence discussing the Queen may potentially be more important than one that provides generic facts about the country. The features we consider are :

- **TextRank Score**: We use the TextRank algorithm (Mihalcea and Tarau, 2004) to extract keywords from the document along with the PageRank score attached to them. Each sentence is scored based on the cumulative

scores of all keywords that occur in it. Explicitly, the score of a particular sentence is calculated as:

$$Score(s) = \frac{\sum_{w \in W \cap S} KeywordScore(w)}{|S|} \quad (1)$$

where S is the set of words in the sentence and W is the set of keywords extracted by the TextRank algorithm.

- **Position of the sentence in the document**: The importance of sentence position has been studied in (Edmundson, 1969). We expect more important sentences to have a higher edit importance score attached to them.

We train a ridge regression model with the model parameters tuned using cross validation on the training data. We report the Spearman ρ correlation (Spearman, 1904) of the predicted edit importance scores with the human annotated scores on the test data.

6 Experiments and Results

In this section, we discuss the various experiments performed, and the results obtained. **Baselines**: To the best of our knowledge, our work is the first that attempts to infer importance/impact of text edits between document versions. Thus, we did not have established baselines to compare against. Instead we use the following features as baselines:

- **Sentence Order** - Sentences are ordered according to their position in the document, with the first sentence assigned most importance. This is also the order in which a reviewer would normally view edits.
- **Readability Score** - Sentence edit importance scores are calculated as being proportional to the change in their readability scores.
- **Text Rank** - We expect sentences with higher TextRank score to have higher edit importance attached to them.

Table 1 outlines the Spearman ρ correlation of our model and the above baselines with human judgments. We are able to achieve significant improvement over the baselines using the full set of features. An interesting observation was that sentence position correlates poorly with the human annotated importance scores. This indicates that the order/position of sentences has negligible effect on the perceived significance. Both readability score and TextRank have reasonable influence on edit importance, though neither of them is able to match the performance of the full set of features.

Approach	Spearman ρ
Sentence Position	0.067
Readability Score	0.306
Text Rank	0.208
Proposed Approach	0.979

Table 1: Spearman ρ of the predicted importance score with the human annotated importance scores.

Feature	Spearman ρ
Type of Change	0.47907
Readability Score	0.311989
Change in POS tags	0.978821
Change in NE	0.189176
Change in Dependency Tuples	0.96417
Sentence Position	0.06846
Text Rank	0.209058

Table 2: Performance of each feature group in isolation. Numbers reflect the performance (Spearman ρ) of the model when using only the specified feature group, relative to the performance when using all features.

Contribution of feature groups

In order to gain better insight into individual feature performance, we look more closely at the performance of each feature group in isolation. Table 2 shows the performance of the model when using only a specific feature group, relative to the performance when using all features. This provides us with a number of interesting insights. First, it is evident that change-related features contribute more to edit importance than relevance-related features.

According to our results, humans perceive change in number and types of POS tags to be the most significant indicator of edit importance. For further insight, we looked at the coefficient values of individual POS tags in the ridge regression model trained using only POS tags as features. Our investigations revealed that change in proper nouns, nouns, present participle verbs and modal are most highly correlated with edit importance. Contrary to our expectation, modification of named entities does not significantly influence edit importance. This may be due to the fact that named entity

changes occur in only a small subset of sentences, and hence cannot be good predictors of edit importance when used as a feature by themselves.

7 Conclusion and Future Work

In this paper we present a novel approach to infer the importance of text edit between two document versions. We present an empirical analysis of the relevance of various linguistic features for the task of scoring edit importance and model it using a regression model. AMT is used to collect human annotated data for edit importance, and a comparison against those establish the superiority of our proposed approach over several baselines.

References

Adobe acrobat. https://acrobat.adobe.com/in/en/acrobat/pdf-reader.html. Accessed: 2017-07-05.

Microsoft word. https://products.office.com/en-in/word. Accessed: 2017-07-05.

Amit Bronner and Christof Monz. 2012. User edits classification using document revision histories. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 356–366. Association for Computational Linguistics.

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

Harold P Edmundson. 1969. New methods in automatic extracting. *Journal of the ACM (JACM)*, 16(2):264–285.

Richard P Honeck. 1971. A study of paraphrases. *Journal of Verbal Learning and Verbal Behavior*, 10(4):367–381.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. Association for Computational Linguistics.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101.

Fan Zhang and Diane Litman. 2014. Sentence-level rewriting detection. *Grantee Submission*.