Automatically Detecting Corresponding Edit-Turn-Pairs in Wikipedia

Johannes Daxenberger[†] **and Iryna Gurevych**[†‡]

† Ubiquitous Knowledge Processing Lab
Department of Computer Science, Technische Universität Darmstadt

‡ Information Center for Education
German Institute for Educational Research and Educational Information

http://www.ukp.tu-darmstadt.de

Abstract

In this study, we analyze links between edits in Wikipedia articles and turns from their discussion page. Our motivation is to better understand implicit details about the writing process and knowledge flow in collaboratively created resources. Based on properties of the involved edit and turn, we have defined constraints for corresponding edit-turn-pairs. We manually annotated a corpus of 636 corresponding and non-corresponding edit-turn-pairs. Furthermore, we show how our data can be used to automatically identify corresponding edit-turn-pairs. With the help of supervised machine learning, we achieve an accuracy of .87 for this task.

1 Introduction

The process of user interaction in collaborative writing has been the topic of many studies in recent years (Erkens et al., 2005). Most of the resources used for collaborative writing do not explicitly allow their users to interact directly, so that the implicit effort of coordination behind the actual writing is not documented. Wikipedia, as one of the most prominent collaboratively created resources, offers its users a platform to coordinate their writing, the so called talk or discussion pages (Viégas et al., 2007). In addition to that, Wikipedia stores all edits made to any of its pages in a revision history, which makes the actual writing process explicit. We argue that linking these two resources helps to get a better picture of the collaborative writing process. To enable such interaction, we extract segments from discussion pages, called turns, and connect them to corresponding edits in the respective article. Consider the following snippet from the discussion page of the article "Boron" in the English Wikipedia. On February 16th of 2011, user `JCM83` added the **turn**:

> Shouldn't borax be wikilinked in the "etymology" paragraph?

Roughly five hours after that turn was issued on the discussion page, user `Sbharris` added a wikilink to the "History and etymology" section of the article by performing the following **edit**:

> '' borax '' → [[borax]]

This is what we define as a corresponding *edit-turn-pair*. More details follow in Section 2. To the best of our knowledge, this study is the first attempt to detect corresponding edit-turn-pairs in the English Wikipedia fully automatically.

Our motivation for this task is two-fold. First, an automatic detection of corresponding edit-turn-pairs in Wikipedia pages might help users of the encyclopedia to better understand the development of the article they are reading. Instead of having to read through all of the discussion page which can be an exhausting task for many of the larger articles in the English Wikipedia, users could focus on those discussions that actually had an impact on the article they are reading. Second, assuming that edits often introduce new knowledge to an article, it might be interesting to analyze how much of this knowledge was actually generated within the discourse on the discussion page.

The detection of correspondence between edits and turns is also relevant beyond Wikipedia. Many companies use Wikis to store internal information and documentation (Arazy et al., 2009). An alignment between edits in the company Wiki and issues discussed in email conversations, on mailing lists, or other forums, can be helpful to track the flow or generation of knowledge within the company. This information can be useful to improve communication and knowledge sharing.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 187–192,
Baltimore, Maryland, USA, June 23-25 2014.

In the limited scope of this paper, we will focus on two research questions. First, we want to understand the nature of correspondence between Wikipedia article edits and discussion page turns. Second, we want to know the distinctive properties of corresponding edit-turn-pairs and how to use these to automatically detect corresponding pairs.

2 Edit-Turn-Pairs

In this section, we will define the basic units of our task, namely edits and turns. Furthermore, we will explain the kind of correspondence between edits and turns we are interested in.

Edits To capture a fine-grained picture of changes to Wikipedia article pages, we rely on the notion of edits defined in our previous work (Daxenberger and Gurevych, 2012). Edits are coherent modifications based on a pair of adjacent revisions from Wikipedia article pages. To calculate edits, a line-based diff comparison between the old revision and the new revision is made, followed by several post-processing steps. Each pair of adjacent revisions found in the edit history of an article consists of one or more edits, which describe either inserted, deleted, changed or relocated text. Edits are associated with metadata from the revision they belong to, this includes the comment (if present), the user name and the time stamp.

Turns Turns are segments from Wikipedia discussion pages. To segment discussion pages into turns, we follow a procedure proposed by Ferschke et al. (2012). With the help of the Java Wikipedia Library (Zesch et al., 2008), we access discussion pages from a database. Discussion pages are then segmented into *topics* based upon the structure of the page. Individual turns are retrieved from topics by considering the revision history of the discussion page. This procedure successfully segmented 94 % of all turns in a corpus from the Simple English Wikipedia (Ferschke et al., 2012). Along with each turn, we store the name of its user, the time stamp, and the name of the topic to which the turn belongs.

Corresponding Edit-Turn-Pairs An edit-turn-pair is defined as a pair of an edit from a Wikipedia article's revision history and a turn from the discussion page bound to the same article. If an article has no discussion page, there are no edit-turn-pairs for this article.

A definition of correspondence is not straightforward in the context of edit-turn-pairs. Ferschke et al. (2012) suggest four types of explicit performatives in their annotation scheme for dialog acts of Wikipedia turns. Due to their performative nature, we assume that these dialog acts make the turn they belong to a good candidate for a corresponding edit-turn-pair. We therefore define an edit-turn-pair as corresponding, if: i) The turn is an *explicit suggestion, recommendation or request* and the edit performs this suggestion, recommendation or request, ii) the turn is an *explicit reference or pointer* and the edit adds or modifies this reference or pointer, iii) the turn is a *commitment to an action in the future* and the edit performs this action, and iv) the turn is a *report of a performed action* and the edit performs this action. We define all edit-turn-pairs which do not conform to the upper classification as non-corresponding.

3 Corpus

With the help of Amazon Mechanical Turk[1], we crowdsourced annotations on a corpus of edit-turn-pairs from 26 random English Wikipedia articles in various thematic categories. The search space for corresponding edit-turn-pairs is quite big, as any edit to an article may correspond to any turn from the article's discussion page. Assuming that most edit-turn-pairs are non-corresponding, we expect a heavy imbalance in the class distribution. It was important to find a reasonable amount of corresponding edit-turn-pairs before the actual annotation could take place, as we needed a certain amount of positive seeds to keep turkers from simply labeling pairs as non-corresponding all the time. In the following, we explain the step-by-step approach we chose to create a suitable corpus for the annotation study.

Filtering We applied various filters to avoid annotating trivial content. Based on an automatic classification using the model presented in our previous work (Daxenberger and Gurevych, 2013), we excluded edits classified as Vandalism, Revert or Other. Furthermore, we removed all edits which are part of a revision created by bots, based on the Wikimedia user group[2] scheme. To keep the class imbalance within reasonable margins, we limited the time span between edits and turns to 86,000

[1] www.mturk.com

[2] http://meta.wikimedia.org/wiki/User_classes

seconds (about 24 hours). The result is a set of 13,331 edit-turn-pairs, referred to as *ETP-all*.

Preliminary Annotation Study From ETP-all, a set of 262 edit-turn-pairs have been annotated as corresponding as part of a preliminary annotation study with one human annotator. This step is intended to make sure that we have a substantial number of corresponding pairs in the data for the final annotation study. However, we still expect a certain amount of non-corresponding edit-turn-pairs in this data, as the annotator judged the correspondence based on the entire revision and not the individual edit. We refer to this 262 edit-turn-pairs as *ETP-unconfirmed*.

Mechanical Turk Annotation Study Finally, for the Mechanical Turk annotation study, we selected 500 random edit-turn-pairs from ETP-all excluding ETP-unconfirmed. Among these, we expect to find mostly non-corresponding pairs. From ETP-unconfirmed, we selected 250 random edit-turn-pairs. The resulting 750 pairs have each been annotated by five turkers. The turkers were presented the turn text, the turn topic name, the edit in its context, and the edit comment (if present). The context of an edit is defined as one preceding and one following paragraph of the edited paragraph. Each edit-turn-pair could be labeled as "corresponding", "non-corresponding" or "can't tell". To select good turkers and to block spammers, we carried out a pilot study on a small portion of manually confirmed corresponding and non-corresponding pairs, and required turkers to pass a qualification test.

The average pairwise percentage agreement over all pairs is 0.66. This was calculated as $\frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{c=1}^{C} v_i^c}{C}$, where $N = 750$ is the overall number of annotated edit-turn-pairs, $C = \frac{R^2-R}{2}$ is the number of pairwise comparisons, $R = 5$ is the number of raters per edit-turn-pair, and $v_i^c = 1$ if a pair of raters c labeled edit-turn-pair i equally, and 0 otherwise. The moderate pairwise agreement reflects the complexity of this task for non-experts.

Gold Standard To rule out ambiguous cases, we created the Gold Standard corpus with the help of majority voting. We counted an edit-turn-pair as corresponding, if it was annotated as "corresponding" by least three out of five annotators, and likewise for non-corresponding pairs. Furthermore, we deleted 21 pairs for which the turn segmentation algorithm clearly failed (e.g. when the turn text was empty). This resulted in 128 corresponding and 508 non-corresponsing pairs, or 636 pairs in total. We refer to this dataset as *ETP-gold*. To assess the reliability of these annotations, one of the co-authors manually annotated a random subset of 100 edit-turn-pairs contained in ETP-gold as corresponding or non-corresponding. The inter-rater agreement between ETP-gold (majority votes over Mechanical Turk annotations) and our expert annotations on this subset is Cohen's $\kappa = .72$. We consider this agreement high enough to draw conclusions from the annotations (Artstein and Poesio, 2008).

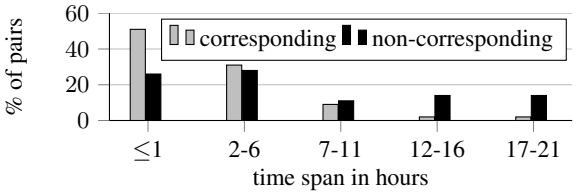

Figure 1: Percentage of (non-)corresponding edit-turn-pairs for various time intervals in ETP-gold.

Obviously, this is a fairly small dataset which does not cover a representative sample of articles from the English Wikpedia. However, given the high price for a new corresponding edit-turn-pair (due to the high class imbalance in random data), we consider it as a useful starting point for research on edit-turn-pairs in Wikipedia. We make ETP-gold freely available.[3]

As shown in Figure 1, more than 50% of all corresponding edit-turn-pairs in ETP-gold occur within a time span of less than one hour. In our 24 hours search space, the probability to find a corresponding edit-turn-pair drops steeply for time spans of more than 6 hours. We therefore expect to cover the vast majority of corresponding edit-turn-pairs within a search space of 24 hours.

4 Machine Learning with Edit-Turn-Pairs

We used DKPro TC (Daxenberger et al., 2014) to carry out the machine learning experiments on edit-turn-pairs. For each edit, we stored both the edited paragraph and its context from the old revision as well as the edited paragraph and context from the new revision. We used Apache

[3] http://www.ukp.tu-darmstadt.de/data/edit-turn-pairs

OpenNLP[4] for the segmentation of edit and turn text. Training and testing the classifier has been carried out with the help of the Weka Data Mining Software (Hall et al., 2009). We used the Sweble parser (Dohrn and Riehle, 2011) to remove Wiki markup.

4.1 Features

In the following, we list the features extracted from preprocessed edits and turns. The *edit text* is composed of any inserted, deleted or relocated text from both the old and the new revision. The *edit context* includes the edited paragraph and one preceding and one following paragraph. The *turn text* includes the entire text from the turn.

Similarity between turn and edit text We propose a number of features which are purely based on the textual similarity between the text of the turn, and the edited text and context. We used the cosine similarity, longest common subsequence, and word n-gram similarity measures. Cosine similarity was applied on binary weighted term vectors (L^2 norm). The word n-gram measure (Lyon et al., 2004) calculates a Jaccard similarity coefficient on trigrams. Similarity has been calculated between i) the plain edit text and the turn text, ii) the edit and turn text after any wiki markup has been removed, iii) the plain edit context and turn text, and iv) the edit context and turn text after any wiki markup has been removed.

Based on metadata of edit and turn Several of our features are based on metadata from both the edit and the turn. We recorded whether the name of the edit user and the turn user are equal, the absolute time difference between the turn and the edit, and whether the edit occurred before the turn. Cosine similarity, longest common subsequence, and word n-gram similarity were also applied to measure the similarity between the edit comment and the turn text as well as the similarity between the edit comment and the turn topic name.

Based on either edit or turn Some features are based on the edit or the turn alone and do not take into account the pair itself. We recorded whether the edit is an insertion, deletion, modification or relocation. Furthermore, we measured the length of the edit text and the length of the turn text. The 1,000 most frequent uni-, bi- and trigrams from the turn text are represented as binary features.

[4] http://opennlp.apache.org

	Baseline	**R. Forest**	**SVM**
Accuracy	.799 ±.031	**.866** ±.026†	.858 ±.027†
$F1_{mac.}$	NaN	**.789** ±.032	.763 ±.033
$Precision_{mac.}$	NaN	**.794** ±.031	.791 ±.032
$Recall_{mac.}$	.500 ±.039	**.785** ±.032†	.736 ±.034†
$F1_{non\text{-}corr.}$	.888 ±.025	**.917** ±.021	.914 ±.022
$F1_{corr.}$	NaN	**.661** ±.037	.602 ±.038

Table 1: Classification results from a 10-fold cross-validation experiment on ETP-gold with 95% confidence intervals. Non-overlapping intervals w.r.t. the majority baseline are marked by †.

4.2 Classification Experiments

We treat the automatic classification of edit-turn-pairs as a binary classification problem. Given the small size of ETP-gold, we did not assign a fixed train/test split to the data. For the same reason, we did not further divide the data into train/test and development data. Rather, hyperparameters were optimized using grid-search over multiple cross-validation experiments, aiming to maximize accuracy. To deal with the class imbalance problem, we applied cost-sensitive classification. In correspondence with the distribution of class sizes in the training data, the cost for false negatives was set to 4, and for false positives to 1. A reduction of the feature set as judged by a χ^2 ranker improved the results for both Random Forest as well as the SVM, so we limited our feature set to the 100 best features.

In a 10-fold cross-validation experiment, we tested a Random Forest classifier (Breiman, 2001) and an SVM (Platt, 1998) with polynomial kernel. Previous work (Ferschke et al., 2012; Bronner and Monz, 2012) has shown that these algorithms work well for edit and turn classification. As baseline, we defined a majority class classifier, which labels all edit-turn-pairs as non-corresponding.

4.3 Discussion and Error Analysis

The classification results for the above configuration are displayed in Table 1. Due to the high class imbalance in the data, the majority class baseline sets a challenging accuracy score of .80. Both classifiers performed significantly better than the baseline (non-overlapping confidence intervals, see Table 1). With an overall macro-averaged F1 of .79, Random Forest yielded the best results, both with respect to precision as well as recall. The low F1 on corresponding pairs is likely due to the small number of training examples.

To understand the mistakes of the classifier, we manually assessed error patterns within the model of the Random Forest classifier. Some of the false positives (i.e. non-corresponding pairs classified as corresponding) were caused by pairs where the revision (as judged by its comment or the edit context) is related to the turn text, however the specific edit in this pair is not. This might happen, when somebody corrects a spelling error in a paragraph that is heavily disputed on the discussion page. Among the false negatives, we found errors caused by a missing direct textual overlap between edit and turn text. In these cases, the correspondence was indicated only (if at all) by some relationship between turn text and edit comment.

5 Related Work

Besides the work by Ferschke et al. (2012) which is the basis for our turn segmentation, there are several studies dedicated to discourse structure in Wikipedia. Viégas et al. (2007) propose 11 dimensions to classify discussion page turns. The most frequent dimensions in their sample are requests for coordination and requests for information. Both of these may be part of a corresponding edit-turn-pair, according to our definition in Section 2. A subsequent study (Schneider et al., 2010) adds more dimensions, among these an explicit category for references to article edits. This dimension accounts for roughly 5 to 10% of all turns. Kittur and Kraut (2008) analyze correspondence between article quality and activity on the discussion page. Their study shows that both implicit coordination (on the article itself) and explicit coordination (on the discussion page of the article) play important roles for the improvement of article quality. In the present study, we have analyzed cases where explicit coordination lead to implicit coordination and vice versa.

Kaltenbrunner and Laniado (2012) analyze the development of discussion pages in Wikipedia with respect to time and compare dependences between edit peaks in the revision history of the article itself and the respective discussion page. They find that the development of a discussion page is often bound to the topic of the article, i.e. articles on time-specific topics such as events grow much faster than discussions about timeless, encyclopedic content. Furthermore, they observed that the edit peaks in articles and their discussion pages are mostly independent. This partially explains the high number of non-corresponding edit-turn-pairs and the consequent class imbalance.

While there are several studies which analyze the high-level relationship between discussion and edit activity in Wikipedia articles, very few have investigated the correspondence between edits and turns on the textual level. Among the latter, Ferron and Massa (2014) analyze 88 articles and their discussion pages related to traumatic events. In particular, they find a correlation between the article edits and their discussions around the anniversaries of the events.

6 Conclusion

The novelty of this paper is a computational analysis of the relationship between the edit history and the discussion of a Wikipedia article. As far as we are aware, this is the first study to automatically analyze this relationship involving the textual content of edits and turns. Based on the types of turn and edit in an edit-turn-pair, we have operationalized the notion of corresponding and non-corresponding edit-turn-pairs. The basic assumption is that in a corresponding pair, the turn contains an explicit performative and the edit corresponds to this performative. We have presented a machine learning system to automatically detect corresponding edit-turn-pairs. To test this system, we manually annotated a corpus of corresponding and non-corresponding edit-turn-pairs. Trained and tested on this data, our system shows a significant improvement over the baseline.

With regard to future work, an extension of the manually annotated corpus is the most important issue. Our classifier can be used to bootstrap the annotation of additional edit-turn-pairs.

Acknowledgments

The authors would like to give special thanks to Viswanathan Arunachalam and Dat Quoc Nguyen, who carried out initial experiments and the preliminary annotation study, and to Emily Jamison, who set up the Mechanical Turk task. This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806, and by the Hessian research excellence program "Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz" (LOEWE) as part of the research center "Digital Humanities". We thank the anonymous reviewers for their helpful suggestions.

References

Ofer Arazy, Ian Gellatly, Soobaek Jang, and Raymond Patterson. 2009. Wiki deployment in corporate settings. *IEEE Technology and Society Magazine*, 28(2):57–64.

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Leo Breiman. 2001. Random Forests. *Machine Learning*, 45(1):5–32.

Amit Bronner and Christof Monz. 2012. User Edits Classification Using Document Revision Histories. In *European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 356–366, Avignon, France.

Johannes Daxenberger and Iryna Gurevych. 2012. A Corpus-Based Study of Edit Categories in Featured and Non-Featured Wikipedia Articles. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 711–726, Mumbai, India.

Johannes Daxenberger and Iryna Gurevych. 2013. Automatically Classifying Edit Categories in Wikipedia Revisions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 578–589, Seattle, WA, USA.

Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych, and Torsten Zesch. 2014. DKPro TC: A Java-based Framework for Supervised Learning Experiments on Textual Data. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. System Demonstrations*, page (to appear), Baltimore, MD, USA.

Hannes Dohrn and Dirk Riehle. 2011. Design and implementation of the Sweble Wikitext parser. In *Proceedings of the International Symposium on Wikis and Open Collaboration (WikiSym '11)*, pages 72–81, Mountain View, CA, USA.

Gijsbert Erkens, Jos Jaspers, Maaike Prangsma, and Gellof Kanselaar. 2005. Coordination processes in computer supported collaborative writing. *Computers in Human Behavior*, 21(3):463–486.

Michela Ferron and Paolo Massa. 2014. Beyond the encyclopedia: Collective memories in Wikipedia. *Memory Studies*, 7(1):22–45.

Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar. 2012. Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 777–786, Avignon, France.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18.

Andreas Kaltenbrunner and David Laniado. 2012. There is No Deadline - Time Evolution of Wikipedia Discussions. In *Proceedings of the Annual International Symposium on Wikis and Open Collaboration*, Linz, Austria.

Aniket Kittur and Robert E. Kraut. 2008. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, pages 37–46, San Diego, CA, USA.

C. Lyon, R. Barrett, and J. Malcolm. 2004. A theoretical basis to the automated detection of copying between texts, and its practical implementation in the Ferret plagiarism and collusion detector. In *Plagiarism: Prevention, Practice and Policy Conference*, Newcastle, UK.

John C. Platt. 1998. Fast training of support vector machines using sequential minimal optimization. In Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods: Support Vector Learning*, pages 185–208. MIT Press.

Jodi Schneider, Alexandre Passant, and John G. Breslin. 2010. A Content Analysis: How Wikipedia Talk Pages Are Used. In *Proceedings of the 2nd International Conference of Web Science*, pages 1–7, Raleigh, NC, USA.

Fernanda B. Viégas, Martin Wattenberg, Jesse Kriss, and Frank Ham. 2007. Talk Before You Type: Coordination in Wikipedia. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, pages 78–78, Big Island, HI, USA.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco.