Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History

Aurélien Max, Guillaume Wisniewski

LIMSI-CNRS and Université Paris-Sud 11
Orsay, France
{aurelien.max,guillaume.wisniewski}@limsi.fr

Abstract

Naturally-occurring instances of linguistic phenomena are important both for training and for evaluating automatic text processing. When available in large quantities, they also prove interesting material for linguistic studies. In this article, we present WiCoPaCo (Wikipedia Correction and Paraphrase Corpus), a new freely-available resource built by automatically mining Wikipedia's revision history. The WiCoPaCo corpus focuses on local modifications made by human revisors and include various types of corrections (such as spelling error or typographical corrections) and rewritings, which can be categorized broadly into meaning-preserving and meaning-altering revisions. We present an initial hand-built typology of these revisions, but the resource allows for any possible annotation scheme. We discuss the main motivations for building such a resource and describe the main technical details guiding its construction. We also present applications and data analysis on French and report initial results on spelling error correction and morphosyntactic rewriting. The WiCoPaCo corpus can be freely downloaded from http://wicopaco.limsi.fr.

1. Introduction

This paper describes the construction of a corpus of rewritings extracted from the revision history of Wikipedia, which includes spelling corrections, reformulations, and other local text transformations. Such rewritings are of interest for many NLP applications, including text correction and normalization, paraphrasing, summarization, etc. For many of these applications, only a few hand-crafted or artificial corpora of small size are available, which prevents researchers from using machine learning techniques requiring important amounts of training examples and questions the validity of evaluations that use them. For instance, the study reported in (Schroeder et al., 2009) shows the negative impact of using artificially produced sentential paraphrases in a multi-source Machine Translation experiment.

While the cost of the annotation effort has always been a burden to the creation of huge corpora of naturally-occurring rewritings, we believe that the growth of publicly editable wikis with high contribution rates allows us to easily collect large amounts of useful rewriting examples. Indeed, an important characteristics of Wikipedia (and other wikis) is the fact that users not only contribute new content but also improve the overall quality of the text collection (an encyclopedia in the case of Wikipedia), counteracting spam and making various types of corrections and improvements to the created texts.

The huge amounts of quality data in Wikipedia have triggered many works on automatic resource acquisition (e.g. acquisition of lexical-semantic knowledge (Zesch et al., 2008)). Closer to our work, (Nelken and Yamangil, 2008) exploit Wikipedia's revision history for acquiring instances of *eggcorns* (semantically plausible homophonic confusions) and their correction, as well as text spans and their compressed rewritings. These correspondences are found by applying a search for longest common subsequences (using the same algorithm as the `diff` command) between any two consecutive versions of articles and identifying substitutions in the results. In their work, a very simplifying assumption is made that such pairs correspond to instances of text compressions whenever the rewritten text is shorter than the original text.

There is, however, a much greater variety of natures of rewritings that are also of great interest for several NLP applications. In this work, we describe the construction of a new resource from Wikipedia's revision history. The raw resource contains all types of local rewritings found in the encyclopedia, with their context and various meta-data related to them. Independent annotation efforts can then assign labels to the data depending on a targeted application, yielding data suitable for supervised machine learning and for evaluation of NLP tools. Furthermore, the collected data constitutes a huge collection of naturally-occurring corrections and rephrasings of particular interest for writing studies.

2. Building the resource

In this section, we describe the main details behind the construction of our WiCoPaCo corpus[1] (Wikipedia Correction and Paraphrase Corpus). Our objective is to build a resource suitable for many types of studies on naturally-occurring local rewritings. However, we must set some practical bounds as not all rewritings will be of equal interest.

The construction of the corpus in done in two steps. In a first step, a set of local modifications is extracted by computing, from a Wikipedia dump stored in a local database, the differences among any two versions of all articles using the efficient longest common subsequence algorithm implemented by the `diff` standard command. All paragraphs containing at least one substitution are extracted, and their text is normalized (de-wikification, tokenization, etc.) As the aim is to extract local modifications, only rewriting implying at most 7 words are taken into account.

[1] WiCoPaCo can be freely downloaded from the website http://wicopaco.limsi.fr

This first step allows us to extract a very large number of local modifications. Note that we do not consider modifications that involve only additions or deletions of tokens, as this corpus is designed to support the study of text modifications where two text spans in context can be paired.

In a second step, we apply a set of hand-crafted filtering rules. In particular, we filter out modifications in which the ratio of common words in the original and the modified sentence (defined in a greedy sense) is under a given threshold and changes that concern only punctuation or case modifications. The first filter filters out modifications that may significantly change the meaning of the enclosing text unit. The second filter limits the size of the corpus, although these occurrences can be kept for studies requiring them, as for example studies on text ponctuation.

We record the full paragraph in which a local rewriting is found in order to allow application to exploit a larger context than that of the enclosing sentence.[2] Because contributors can make several edits when submitting a new revision, we record both the context of the original phrase and that of the modified phrase. We also record, as meta-data on the revision, all necessary identifiers from the Wikipedia database, including the identifier of the user who submitted the revision in which the text substitution was found and the number of revisions made by this user. We purposefully do not include revisions that were submitted by automatic bots, as we want to restrict the data to modifications that could be made by human contributors.[3] Anonymous and registered human users are distinguished, permitting further data mining to assess the reliability of a given local substitution based on its author's reputation (Adler et al., 2008).

An output XML file is finally produced with unique modification identifiers, which can be used in subsequent work to associate annotations to every instance.[4] Figure 1 shows an example of an entry of the WiCoPaCo corpus. Our initial work was carried out on the French version of the Wikipedia database. Figure 2 reports the main types of text substitutions that are found by manual inspection of the corpus, including many types which were not considered in (Nelken and Yamangil, 2008). We distinguished two main classes, namely that of modifications where the original and the modified text convey essentially the same meaning in context (according to human judgment), and that of modifications where meaning has been modified for various possible reasons.[5] Automatic classification of modifications, using the subclasses of Figure 2 or any application-oriented classes is part of our future work. For instance, the automatic detection of what is referred to as "subtle grammatical spamming" may be of great use for Wikipedia administrators to locate incorrect changes that pollute the textual database and might remain in the encyclopedia for long times before they are spotted and corrected.

3. Exploiting the Data

In this section, we illustrate some possible uses of the WiCoPaCo corpus by describing ongoing works on spelling error corrections and paraphrase generation selection.

3.1. Spelling error correction

The WiCoPaCo corpus can be used to easily build a corpus of spelling errors. Indeed, it can be assumed that most minor edits in documents (i.e. edits that only concern a few words) represent orthographic, grammatical or typographic corrections. As both the misspelled word and its correction are available[6], spelling errors can also be easily classified either as *non-word errors* that results in a non valid word (e.g. when "from" is spelled "rfom") or *real-word errors* in which a correctly spelled word is substituted for another word (e.g. "from" is spelled "form").

3.1.1. Building a Spelling Error Corpus

So far, we have considered editions that are limited to a single word, as most of the works on spell checking only deal with errors at that levela. We have also discarded all editions that involve either a punctuation sign, a digit, a word with more than one uppercase letter[7] or a number written in letters[8].

First, we used a spell checker[9] to detect whether the word involved in the edition (the *before word*) and the word that results from the edition (the *after word*) are erroneous or not with respect to the lexicon of the spell checker. This allows us to distinguish three kinds of editions:

- **non-word corrections**, when the before word is erroneous and the after word is correct;
- **real-word error corrections and reformulations**, when both the before and after word are correct;

[2] In addition to the shorter context, sentences are not good contextual units for this work as sentence segmentation is difficult to define, even on encyclopedic text. Taking paragraphs as units allows making use of unambiguous elements, such as blank lines or structural boundaries. In any case, all the necessary metadata that link to the original Wikipedia article and revision are kept.

[3] This, of course, does not mean that modifications programmed in bots are of no interest. Other reasons for not keeping them include the large quantities of modifications and their impact on modification frequency, the fact that bots can hardly take linguistic context into account and can introduce errors when their programmers did not anticipate cases where the bot's modifications should not apply.

[4] At the time of writing, the resource contains 408,816 modifications in context.

[5] Note that spam applied to our defined local modifications appears in this category.

[6] This is assuming that the last revision is the correct one. To make sure that no incorrect corrections are kept, one can keep modification pairs $A \rightarrow B$ which are significantly more frequent that the reverse modification pairs $B \rightarrow A$ (the latter may in fact never occur, in particular for most spelling error corrections). One may also consider exploiting metadata on the user responsible for the modification.

[7] Manual inspection of the corpus shows that words with more than one uppercased letter are mostly acronyms.

[8] Most of the editions that involve such numbers are semantic corrections, except when the number is *une* or *un* (*one*).

[9] We used the open source `hunspell` spell-checker (available from `http://hunspell.sourceforge.net`). In all our experiments we used the version 1.2.8 of `hunspell` with the version 3.4.1 of the French dictionary *Classique et réforme 90*.

```
<modif id="23" wp_page_id="7" wp_before_rev_id="4649540"
       wp_after_rev_id="4671967" wp_user_id="0"
       wp_user_num_modif="1096911" wp_comment="Définition">
  <before>On nomme <m num_words="1">Algebre</m> linéaire la branche
  des mathématiques qui se penche...</before>
  <after>On nomme <m num_words="1">Algèbre</m> linéaire la branche
  des mathématiques qui se penche...</after>
</modif>
```

Figure 1: Sample XML entry of WiCoPaCo

Same meaning	
Different spelling	
Encyclopedic normalizations	*[**Son 2ème disque –> Son deuxième disque**]*
Unknown words due to spelling	*c' est-à- dire la [**dernrière –> dernière**] année avant l' ère chrétienne*
Missing diacritics	*la jeune Natascha Kampusch ,[**agée –> âgée**] de 18 ans*
Homophonic confusions	*L' immense majorité de [**ses –> ces**] nobles vit dans des conditions*
Grammatical errors	*dans le but de [**sensibilisé –> sensibiliser**] sur les changements*
Different wording	
Syntactic rewriting	*Le tritium [**existe dans la nature . Il est produit –> se forme naturellement**] dans l' atmosphère*
Paraphrases	*'Gimme Gimme Gimme' et 'I Have A Dream' [**contribueront au gigantesque succès de –> viendront alimenter la gloire que connait**] Abba*
Translation	*Bertrand Russell , dans [**History of the Western Philosophy –> Histoire de la philosophie occidentale**]*
Different meaning	
Acceptable meaning changes	
Precision of meaning	*alors [**que l' ordinateur -> qu'un processeur de la famille x86**] reconnaîtra ce que l' instruction machine*
Simplification of meaning	*[**Le principal du collège M. Desdouets –> Un de ses professeurs**] dit de lui*
Change of point of view	*il présente sa démission le 18 janvier 2007 [**suite à un débordement télévisé inconvenant envers la candidate socialiste et –> après avoir lancé une plaisanterie sur**] François Hollande*
Questionable correction	*Des opérations de base sont disponibles dans [**tous les –> la plupart des**] jeux d' instructions*
Unquestionable correction	*textes [**de René Goscinny illustrés par –> et illustrations**] Albert Uderzo*
Spam	
Obvious agrammatical spamming	*Süleyman Ier s' [**empare de l' Arabie et fait entrer dans l' –> emp kikoo c moi ca va loll '**] empire ottoman Médine et La Mecque*
Obvious grammatical spamming	*pour promouvoir la justice , la solidarité et [**la paix –> l'apéro**] dans le monde*
Subtle grammatical spamming	*Inquiété par [**le gouvernement de Vichy –> la montée des prix du sucre**], Breton se réfugie en 1941 en Amérique*

Figure 2: Typology of the substitutions found in the data built from the French Wikipedia

- **proper noun or foreign word editions, spam insertion and wrong error corrections**, when the after word is erroneous (no matter what the before word is).

This simple step therefore allows us to identify non-word errors and to discard some uninteresting editions (especially proper noun and foreign words corrections). But we still have to distinguish real-word errors from reformulations and to remove some spam.

In a second step, we used the character edit distance between the before word and the after word to identify both spam editions and reformulations. Indeed, it is a well-known fact (Kukich, 1992) that most spelling errors are within a short edit distance of their correct form. Studying a sample of the corpus corroborates this result: it can be observed that in an edition with an edit distance strictly greater than 3, the word is usually completely re-written and the edition is therefore a paraphrase or a change of meaning. It also appears that an edition with an edit distance greater than 5 generally corresponds to various forms of spam introduction. That is why, for the non-word error corpus (resp. the real-word error corpus), we discarded all the editions that involve an edit-distance larger than 5 (resp. 3).[10]

By applying these two rules, we extracted 72,493 non-word errors and 74,100 real-word errors.

[10] In our experiments, we considered that all operations involved in an edition have a cost of 1.

3.1.2. French Spelling Error Patterns

The spelling error corpus we have gathered provides valuable information regarding spelling error patterns in French. We present here the results of our first analysis of that corpus.

Figure 3 shows the most frequent editions of non-word errors. Most of them involve a diacritic: in fact, 32.39% of non-word error corrections consist in only adding, changing or removing an accent. Apart from the correction of diacritic marks, most of the corrections are caused by the absence of a repeated consonant, which is consistent with many studies on spelling errors in French.

For real-word errors, forgetting diacritics and errors in plurals and feminine are causing most of the editions. It is also interesting to notice that if an edition is frequent, the opposite edition is also frequent (for instance, adding and removing a `s` are both frequent). Another general finding is that 46.96% of modifications occur at word endings, supporting our observation that many corrections involve plural or feminine marks.

3.1.3. Spell Checker Evaluation

Error corrections boils down to two subtasks: *i)* building the set of potential corrections of a given word (the *candidate set*) and *ii)* choosing the best correction among them. In most existing commercial or public spell-checkers, the decision of the correction to perform is left to the user rather than performed automatically. That is why we only only report here the evaluation of the quality of candidate sets, by counting the number of times the correct correction is in the candidate set. Three different candidate sets were considered:

1. The suggestion list of `hunspell` (denoted "hunspell" in the following). This list is built using a set of handcrafted rules that describe frequent error corrections and possible affixes.

2. A list of words built by applying the most frequent edition scripts to the word to correct (denoted "patterns"). This list is built by considering the most frequent error patterns in the corpus (see Table 3).

3. A list of spelling error corrections extracted from the corpus (denoted "pattern"). This list is built by gathering for each example in the train set the misspelled word and its correction.

To evaluate these approaches we randomly split our real-word error and non-word error corpora in a training set (80% of the examples) and a test set (20% of the examples). The training set is used to build the list of corrections; the test set is used to measure the different scores.

Table 1 presents the results for the three methods on the test set. Results clearly show that the combination of the three approaches to build the candidate set almost always produces a set that contains the correct spelling.

	non-word error		real-word error	
	#sugg.	corr.	#sugg.	corr.
hunspell	4.5	95.0%	8.6	65.1%
list	1.3	58.7%	8.3	75.7%
pattern	1.7	48.7%	2.3	53.2%
combi.	4.7	96.8%	14.9	92.6%

Table 1: Percentage of errors for which the correct spelling is in the candidate set and average number of suggestions

3.2. Paraphrase Generation Selection

Automatic paraphrasing of phrases is an active field of research, with applications in such diverse areas as text compression, information extraction or authoring aids. When new text is produced for human readers, it is particularly important to ensure that the resulting text is not only semantically equivalent to the original text but also grammatical, that is as if produced by a human. One way of ensuring local grammaticality is by reranking candidate paraphrases in context using a language model (Bannard and Callison-Burch, 2005) or a syntactic dependency conservation model that considers dependencies between the paraphrases and their context (Max, 2008). However, the first approach has been shown to select more semantically incorrect paraphrases, while the second approach takes a very conservative view on paraphrasing. (Callison-Burch, 2008) recently proposed to condition the probability of paraphrases on the syntactic context of an original phrase. But these probabilities have to be estimated indirectly via pivoting in other languages, as no large high quality representative corpora of phrasal paraphrases exist to derive valid rewriting syntactic patterns.

A stumbling block for research on paraphrasing and rewriting in general is the lack of available corpora for learning models and assessing their performance on naturally-occurring data. To our knowledge, to date no large resources of such rewritings have been made available. Most corpora are built with the aim to support specific research projects. For example, for their study on abstractive sentence compression (Cohn and Lapata, 2008), the authors artificially created their own corpus from 575 sentences. Other corpora, such as the Ziff-Davis corpus (Knight and Marcu, 2002), built from pairs of documents and abstracts, mostly focus on a single phenomenon (e.g. word deletion in 1067 sentence pairs).

As for spelling corrections, our corpus of modifications mined from Wikipedia's revision history can be used to increase by an order of magnitude the quantity of available data. Furthermore, the fact that these data are naturally-occurring pairs of rewritings is certainly the most important characteristic. Lastly, if data can be correctly classified, held-out data sets can be trivially extracted (and possibly validated by humans) to be used as evaluation sets.

This type of corpus was already used for the specific task of text compression (Nelken and Yamangil, 2008), where the authors report using a set of 380,000 pairs of full and compressed sentences extracted from a subset of the English Wikipedia. However, with no automatic classification of modifications, the authors made the simplifying assumption that all observed text compressions preserved meaning to build large lexicalized models for the task.

An interesting characteristic of our resource is that it allows associating valid *grammatical* rewritings, either at the

Non-Word Errors

e →é	6.7%	-l	1.9%
E →É	6.7%	+i	1.9%
oe →œ	4.6%	a →â	1.8%
+n	4.3%	-e	1.7%
+s	2.8%	-n	1.7%
+r	2.7%	+t	1.6%
é →è	2.7%	+m	1.6%
-s	2.5%	e →è	1.4%
+e	2.2%	+l	1.3%
é →e	2.1%	-r	1.3%

Real-Word Errors

+s	16.2%	-t	1.5%
+e	9.9%	e →a	1.4%
-s	8.8%	é →er	1.0%
A →À	5.6%	er →é	0.9%
-e	4.9%	u →ù	0.9%
i →î	2.7%	à →a	0.9%
a →à	2.2%	e →é	0.8%
+nt	1.9%	é →è	0.7%
+t	1.7%	s →t	0.7%
a →e	1.5%	û →u	0.7%

Figure 3: The 20 most frequent corrections. These corrections represent 65.0% of real-word and 53.5% of non-word errors

pos_1: ADJ		pos_1: ADJ ADJ		pos_1: DET ADJ NOM		pos_1: VER PRP DET NOM	
pos_2	$p(\text{pos}_2\|\text{pos}_1)$	pos_2	$p(\text{pos}_2\|\text{pos}_1)$	pos_2	$p(\text{pos}_2\|\text{pos}_1)$	pos_2	$p(\text{pos}_2\|\text{pos}_1)$
ADJ	0.4029	ADJ	0.2371	DET NOM	0.3081	VER	0.1666
NOM	0.1221	NOM	0.1666	DET ADJ NOM	0.0880	VER DET NOM	0.0555
VER	0.1116	NOM ADJ	0.0576	VER	0.0314	ADJ	0.0555
PRP NOM	0.0350	VER	0.0448	DET NOM ADJ	0.0314	VER VER	0.0476
NOM ADJ	0.0156	ADJ ADJ	0.0384	PRO	0.0251	DET NOM	0.0396
ADV ADJ	0.0126	ADJ PUN	0.0256	PRP NOM	0.0251	VER PRP NOM	0.0317

Figure 4: Distribution of part-of-speech sequences for rewritings for French phrases

lexical, morpho-syntactic, or syntactic level. We make the hypothesis that for local paraphrase generation most relevant data from our resource can be exploited, independently of the fact that some rewritings introduce some meaning change or not.[11]

For example, models of morpho-syntactic rewriting patterns can be built from a subset of the resource to build grammatical models based on valid morpho-syntactic patterns. An illustration is given on Figure 4, which provides some examples of the distribution of part-of-speech sequences for rewritings derived from our resource.[12] When several candidate paraphrases are produced automatically by some generative technique, information such as a sequence of two adjectives (ADJ ADJ) has a 0.2371 probability of being rewritten as a single adjective (ADJ) can be effectively exploited to assess the grammaticality of such rewritings. If the observed distribution of rewritings may adequately reflect natural rewriting patterns, further studies could also reveal some bias due to the genre of texts being rewritten or the types of contributors.

4. Conclusions and future work

In this article, we have introduced the freely available WiCoPaCo corpus of rewritings in context automatically extracted from Wikipedia's revision history. It is, to our knowledge, one of the largest corpus of naturally-occurring rewritings, which can be exploited, as was shown with our measures on spelling errors and morpho-syntactic rewriting patterns, on many levels.

Our future work will be in two areas. First, we want to carry out a more detailed analysis of the types of rewriting found in the resource in order to identify classes that may be of use for specific studies or applications. Automatic classification will be the next step, as the ability to classify rewritings will be needed for training algorithms or applications such as linguistic-aware text authoring or spam reporting. This type of work exploits latent data at very little cost (some CPU time and disk space) than can be very useful for linguistic studies and NLP applications. We therefore also plan to reproduce our methodology on other languages[13] and other wikis such as the WikiNews for news articles.

Acknowledgements

This work was funded by a LIMSI AI grant and the ANR Trace project. The authors wish to thank Julien Boulet and Martine Hurault-Plantet for their contribution to this work.

5. References

B.T. Adler, K. Chatterjee, L. de Alfaro, M. Faella, I. Pye, and V. Raman. 2008. Assigning Trust To Wikipedia Content. In *Proceedings of WikiSym 2008*, Porto, Portugal.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of ACL*, Ann Arbor, USA.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of EMNLP*, Hawai, USA.

[11] This is assuming that other models for meaning conservation are also used when assessing automatic paraphrases.

[12] To build those patterns, the `TreeTagger` analyzer (available from `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`) was used.

[13] Wikipedia is available in as many as 269 languages as of March, 2010, and our techniques can be applied on all languages with mostly clear word segmentation.

Trevor Cohn and Mirella Lapata. 2008. Sentence compression beyond word deletion. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Manchester, UK.

Kevin Knight and Daniel Marcu. 2002. Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107.

Karen Kukich. 1992. Techniques for automatically correcting words in text. *ACM Comput. Surv.*, 24(4):377–439.

Aurélien Max. 2008. Local rephrasing suggestions for supporting the work of writers. In *Proceedings of GoTAL*, Gothenburg, Sweden.

Rani Nelken and Elif Yamangil. 2008. Mining Wikipedia's Article Revision History for Training Computational Linguistics Algorithms. In *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, Chicago, USA.

Josh Schroeder, Trevor Cohn, and Philipp Koehn. 2009. Word lattices for multi-source translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 719–727, Athens, Greece, March. Association for Computational Linguistics.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morroco.