Proposal for an SD PRIM or an AI Project

Mining Textual Corrections from the Wikipedia Edit History

Antoine Amarilli <a3nm@a3nm.net>
DIG team, Télécom Paris

Wikipedia is a huge source of textual information which is available under a permissive license and can be easily downloaded using the Wikimedia dump service[1]. In addition to the current content of Wikipedia, this service provides a downloadable archive of the full history of all past versions of Wikipedia articles, from 2001 to the present day. This huge collection of edits contain precious information about how Wikipedia articles are corrected: for instance, it contains examples of typos being fixed across hundreds of thousands of pages, Wiki syntax being corrected, or article information being updated in multiple places in reaction to real-world events.

The goal of this project is to mine this information to learn interesting patterns of textual corrections.

Extracting the diffs. The first step of the project is to process the Wikimedia dumps to extract individual changes from the edit history. The edit history is provided as compressed files containing the successive versions of articles as XML documents. The first step will be to extract changes from these successive versions, using efficient implementations of an edit distance algorithm, either implemented by hand or using existing libraries, e.g., Diff Match Patch[2]. This poses substantial data management challenges because of the volume of data: for the English Wikipedia, there are over 900M edits, and 159 GB of highly compressed dumps which must be processed in streaming in a parallel fashion to achieve reasonable running times.

The resulting dataset (of individual changes to Wikipedia articles) would be a useful resource in itself for various applications, and can then be mined for edits in the next step.

If the running time for this first task is too high, we can either investigate variants (e.g., approximate edit distance or heuristics) or limit the study to smaller Wikipedias such as the French Wikipedia or Simple English Wikipedia.

Mining the diffs. Once we have computed a collection of diffs, the second step is to mine them for interesting patterns. One approach that we envision is to learn rewriting rules of the form "It's" → "It is", with contexts, e.g., "Their" → "There" when followed by "are". To mine such patterns efficiently, we propose to explore the dataset of mined edits using standard measures such as precision and recall, and an approach inspired by AMIE[3] to tractably explore and build interesting rules.

Extensions of this step are also possible, to investigate other ways to mine rules (e.g., different scoring functions), other notions of context information than fixed words (e.g., regular expressions), and other ways to generalize across contexts. It may also prove important to filter out vandalism,

[1] https://dumps.wikimedia.org/backup-index.html
[2] https://github.com/google/diff-match-patch
[3] https://suchanek.name/work/publications/vldbj2015.pdf

e.g., using some notion of trust on the editor who made a change, or filtering edits to only keep those that have not been reverted.

Applications. The dataset of edit rules can be applied for various tasks, but the most natural may be to search for places in the current state of Wikipedia where textual corrections can be applied. These corrections could then be applied automatically or semi-automatically with quick validation from human users, e.g., with a crowdsourced game similar to the Wikidata Distributed Game[4].

Related work. The Wikipedia edit history has already been leveraged for various uses: see for instance the related work review of Faruqui et al.[5] Most of these works do not process the Wikipedia edit history at scale, and focus on a subset of this data, or on non-optimal alignments computed by heuristics. They also usually focus on specific tasks (e.g., simplifications, entailment, etc.) that are less general than what we propose.

The most related projects is the JWPL library[6] and the edit distance computation that was performed by Cahill et al.[7], but this latter project is still rather old (so it does not provide up-to-date code or datasets), and it is again focused on a specific task.

The idea of mining the edit history of collaborative projects has also been investigated on structured information, e.g., on the Wikidata edit history[8].

Applicant requirement. The applicant should be interested in working with code to process large volumes of data. Prior experience with low-level code (e.g., C) and distributed processing (e.g., Hadoop, Spark) are not required but encouraged. Interest or involvement in Wikimedia projects is a plus.

Environment and supervision. The PRIM project will take place in the DIG team ("Data Intelligence and Graphs") of the INFRES department of Télécom Paris. It will be supervised by Antoine Amarilli[9], maître de conférences.

Applications should be sent by email to Antoine Amarilli.

[4] `https://tools.wmflabs.org/wikidata-game/distributed/`
[5] `https://arxiv.org/abs/1808.09422`, p8-9
[6] `https://github.com/dkpro/dkpro-jwpl`
[7] `http://aclweb.org/anthology//N/N13/N13-1055.pdf`
[8] https://thomas.pellissier-tanon.fr/papers/2019-WWW-corhist.pdf
[9] `https://a3nm.net/`