3. Data

3.1 Data collection from Wikipedia dumps

Our dataset is based on French Wikipedia Dumps released on 1st of August, 2017 (frw, 2017). They are split among 89 xml files with an average size of 28 Go each and contain the complete edit history for each article available in the current French Wikipedia. We found no pattern concerning how these dumps were generated nor how the article edits are stored within them (alphabetically, topic-wise). Each file contains around three million edits. These datasets have the following advantages:

- Each edit is not hand-made by Mechanical Turk crowd-workers but by thousands of different Internet users that voluntarily contributed to the encyclopedia with the goal of making it more complete or accurate without predefined constraints. It enables one to have access to raw text data with a huge diversity of writing styles, topics and semantics.
- All articles are subjected to modifications or rectifications by any editor. This implies that erroneous editings are likely more diverse than if they were created artificially (ie. with a mere insertion or deletion of random words in a sentence). That approach proved to be successful for capturing diverse genuine grammatical errors (Grundkiewicz and Junczys-Dowmunt, 2014) and may be useful for semantic tasks.
- All entries are in French language that is also the language required within the DIT project. It also enables us to make an original corpus different from the most common NLP corpora that are only available in English.

While spanning each file, we compare two consecutive versions of the same article from which we extract several features:

- **id_file** : the id of the xml file (string),
- **id_modif**: the id of the edit (integer),
- **sample_id**: string containing id_file and id_modif (string),
- **parentid** and **id**: both ids of respectively previous and edited versions (integers),
- **timestamp**: the timestamp (string),
- **username**: the username of the editor (string, replaced by an IP address if the user is not registered),
- **registered**: boolean (True if registered user, False otherwise),
- **user_id**: user id (integer if registered user, empty otherwise),
- **comment**: comment left by the user if any, usually contains the title of the edited section within the article (string),
- **title**: the article title (string),
- **minor**: boolean (True if the user defined the edit as minor, False otherwise), refers to whether the modification is really unlikely to be reverted in the future, it usually applies to rectifications of typing or spelling mistakes,
- **modif**: output of the comparison of both text versions by library `difflib` (string). Each line/paragraph starting with "-" is a line removed in the previous version by the editor and each line starting with "+" is one that was added during the revision in the new version. Lines starting with "?" specify the positions where characters were added or removed. Unchanged paragraphs between versions are not returned by `difflib`.
- **modif_remove**: contains all lines from **modif** starting with "-" (string). It is considered as the text of past version that was subjected to editing.

- **modif_add**: contains all lines from **modif** starting with "+" (string). It is considered as the new version text provided by the editor.
- **format**: text format (string),
- **model**: text model (string),

From each pair of consecutive versions of the same article, we collect the features listed above. For each xml file, we extract around a million and a half pairs of versions. So far, we could preprocess 70 xml files and generate a total of 105 million version pairs divided into 70 tsv output files.

3.2 Filtering process

Once the version pairs (we also call them "edit samples" or "samples") are stored, they need to be classified depending on which type of modification was made. Initially, on the first tsv output files, we randomly sampled a few hundreds observations in order to define which edit samples should be set aside, what modifications classes could be defined and how they are distributed. We progressively set heuristics in order to exclude irrelevant pairs for semantic tasks:

- version pairs concerning articles targeted to Wikipedia editors and not to Wikipedia readers (titles start with "Discussion:", "Projet:", "Utilisateur:", "Wikipédia:", "Modèle:", "Catégorie:"),
- edits whose comments refer to a typing/grammatical/spelling mistake,
- edits defined as minor (boolean "minor" is True),
- edits made by bots (username usually contains a "bot" suffix such as "Sambot", "Herculesbot"), they generally add hyperlinks or standardize an entity denomination through articles but never affect the semantic content,
- pairs in which the new version contains a text inexistent in the previous one (feature modif_remove is an empty string),

- pairs in which the past version contains a text inexistent in the new one (feature modif_add is an empty string).

Once we exclude these observations, we are left with around 7% of all pairs. For instance, with a tsv file of 1.5 million pairs, we are left with over 100 thousands observations which is still a high number. A huge amount of version pairs irrelevant for semantic tasks still remains though.

Within each pair that was kept, each version contains a text that is either a simple sentence or a whole block of text.

One way of sampling could be to merge all version pairs in a single file and to draw samples from it, but it would be too memory-consuming. In order to guarantee a diversity of edit pairs, we sample at most two hundred samples over groups of five tsv output file each. Since the absolute number of version pairs is still high, we may ensure a random sampling does not draw a version pair twice from the same article.

3.3 Labelling process and definition of the annotation guide

We manually classified the edits samples at hand and had to define new classes or modify the existing ones as we discover new types of modifications. As several types of modifications may apply for the same version pair, we are in a multi-labelling approach instead of multi-classification.

For a set of 100 randomly sampled pair versions, we obtain an overall Cohen's kappa score of 0.32 between two annotators for all labels. The kappa scores are 0.17 and 0.11 for respectively **semant_simil** and **semant_diff**. After changing a few labels, updating the specifications of the annotation guide (see appendix, we also provide screenshots of several modification examples that were used for helping annotators in the labelling process) and annotating a new random set of 100 samples, we achieve an overall kappa score of 0.49 for all 14 labels. Scores for labels **semant_simil** and **semant_diff** rose to respectively 0.42 and 34.0. Although these last scores struggle to reach the 0.5 threshold, the inter-annotator agreement performance may be explained by the diversity of texts captured in the version pairs. Some samples contain mere sentences while others contain whole paragraphs. The existence of 14

different labels and the fact that several labels may apply for a single sample also explain the difficulty to increase inter-annotator agreement scores and to build a robust golden standard.

The annotation guide presents 14 labels for identifying modification(s) within each version pair. They are described in the following paragraphs. First, we have "superficial" edits that only relate to the form but not the semantic content of the article:

- **ortho_gram_typo** : typing spelling or grammar mistake,
- **wiki_formatting**: edit concerning Wikipedia formatting, hyperlinks, references (tags <ref>), titles or any change that does not change text bodies visible to the final reader,
- **vandal**: vandalism, deliberate erasure of whole paragraphs, graffitis, malevolent insertion of text completely irrelevant to the article topic (usually sexual comments),
- **reordering**: paragraphs change or switch positions, but the content of the paragraphs themselves is not modified,
- **revert_vandal**: a past version of the article is restored in order to remove vandalism,
- **revert_other**: a past version of the article is restored for any other reason than vandalism,
- **content_remove**: mere deletion of a few sentences or elements in the text body,
- **content_add**: mere addition of a few sentences or elements in the text body,

Next, we have labels that apply to the semantic modifications defined below:

- **nbr_wr**: a number (date, figure, amount) that was **W**rong in the past version is **R**ectified in the new one,
- **nbr_rw**: a number that was **R**ight in the past version is replaced with a **W**rong one,
- **nbr_x**: a number was changed from one version to the other but its veracity cannot be defined,
- **semant_simil**: the form, the style or the formulation of a sentence or an expression was changed but the general meaning, the semantic content is preserved,

- **semant_diff**: a sentence or an expression in the new version contradicts its past self in the previous version.

Another label named **other** was created for all other edits that correspond to none of the preceding labels. It usually applies to edits to non-textual content (images, tables, math formulas), edits made by bots and Wikipedia articles directed towards editors but not common readers, or to any modifications that were not excluded in the preliminary filtering process.

Along the labeling process and the creation of the dataset, we applied new filters in order to avoid irrelevant versions pairs and to maximize as much as possible the number of edit samples related to semantic modifications. For instance, in the filtering process, we excluded:

- edits whose comments refer to the restoration of a past version of an article (comments mention “Annulation des modifications”, “Révocation des modifications”),
- edits that only concern page formatting and not bodies of text (adding of hyperlinks leading to versions of the articles in other languages, changes to titles, infobox, or lists),
- edits that concern display of non-textual information in the article (images, tables).

We eventually gathered a total of 1900 labeled version pairs collected over 70 Wikipedia dump files. Table 3.1 shows labels distribution.

3.4 Filtering of semantic edits

We attempted to build a multi-labeling model in order to detect semantic edit samples only. The initial idea consists in training such a model on the already available hand-labeled data and use it on the rest of unlabeled data. The purpose of this semi-supervised approach is to increase the number of version pairs related to semantic edits.

In order to do so, we used a language model (it will be explained in further detail in the next section) trained on the latest version of French Wikipedia articles. Within each pair of texts, for each (past and new) version, we take the 300-dimensional vector of each word in the string and we sum them. From these two sums, we eventually compute a 300-dimensional

label	count	share of all observations (%)
ortho_gram_typo	506	26.63
wiki_formatting	696	36.63
vandal	114	6.00
reordering	91	4.79
revert_vandal	12	0.63
revert_other	6	0.32
content_remove	116	6.11
content_add	477	25.11
nbr_wr	39	2.05
nbr_rw	7	0.37
nbr_x	2	0.11
semant_simil	212	11.16
semant_diff	143	7.53
other	45	2.37

Table 3.1: Distribution of labels over the whole labeled dataset (1900 samples). Percentages do not sum up to 100 as several labels may apply to more than one observation.

vector difference between the new version and the past version. The same process is repeated for each edit sample.

The recent French Wikipedia articles used as training set for the language model covers a vocabulary of size $765,595$. For the words that were not found in this vocabulary (they amount to $65,607$), we retrieve 300-dimensional vectors from another pretrained Fasttext language model based on English Wikipedia and the skip-gram model from (Bojanowski et al., 2017). Words that did not exist in any of these models were skipped.

We could have used a pretrained Fasttext model for French language, but it appeared that the vocabulary was very noisy due to the fact that the tokenization was not performed properly (e.g. the same word with different punctuations is considered as distinct tokens like in *≪france* and *“france*).

We built two classification models, a logistic regression and a random forest (100 trees) and we performed a 5-fold cross-validation for each of the fourteen labels. Different types of features were successively used:

- bag-of-words: words are represented as one-hot vectors and each sample string is represented as a sparse vector with the size of the vocabulary and 1s for each word contained in the text fragment;
- TF-IDF: same as bag-of-words, but most frequent words are down-weighted.
- vector representations: we compute the difference between the respective mean vectors of both versions and use the dimensions of the resulting 300-dimensional real-valued vector as features.

In all cases, both models fail at detecting semantic modifications (metrics usually stick to 0) and have at most a mediocre performance for detecting content removal and adding, spelling/grammar mistakes and change in Wikipedia edition syntax (ie. hyperlinks). Classification performance results are shown in Table 3.2. Therefore, for our experiments, we only use the hand-labeled samples available for semantic tasks.

		ortho_gram_typo	wiki_formatting	vandal	reordering	revert_vandal	revert_other	content_remove	content_add	nbr_wr	nbr_rw	nbr_x	semant_simil	semant_diff	other
Logistic Regression	Precision	0.31 (0.03)	0.40 (0.05)	0.06 (0.01)	0.03 (0.01)	0.01 (0.01)	0.01 (0.01)	0.14 (0.01)	0.27 (0.04)	0.03 (0.01)	0.01 (0.01)	0.00 (0.00)	0.14 (0.04)	0.10 (0.02)	0.03 (0.02)
	Recall	0.36 (0.04)	0.48 (0.13)	0.49 (0.13)	0.20 (0.04)	0.40 (0.37)	0.50 (0.45)	0.74 (0.09)	0.45 (0.18)	0.40 (0.14)	0.33 (0.42)	0.00 (0.00)	0.39 (0.12)	0.35 (0.08)	0.34 (0.16)
	F1 score	0.33 (0.03)	0.43 (0.08)	0.10 (0.01)	0.06 (0.01)	0.02 (0.03)	0.02 (0.02)	0.23 (0.02)	0.33 (0.08)	0.05 (0.02)	0.01 (0.02)	0.00 (0.00)	0.20 (0.06)	0.15 (0.03)	0.06 (0.03)
Random Forest	Precision	0.60 (0.10)	0.76 (0.02)	0.00 (0.00)	0.30 (0.40)	0.00 (0.00)	0.00 (0.00)	0.62 (0.37)	0.62 (0.03)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Recall	0.16 (0.04)	0.59 (0.06)	0.00 (0.00)	0.03 (0.03)	0.00 (0.00)	0.00 (0.00)	0.15 (0.08)	0.33 (0.07)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	F1 score	0.25 (0.05)	0.66 (0.04)	0.00 (0.00)	0.05 (0.06)	0.00 (0.00)	0.00 (0.00)	0.23 (0.12)	0.43 (0.05)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)

Table 3.2: Metrics figures obtained for each label with either logistic regression or random forest. The first number is the mean of the metric and the one between parenthesis is the standard deviation across all folds.

3.5 Data preprocessing for DIT experiment

From all the 1900 version pairs that were labeled by annotators, 355 edit samples are related to semantic modifications and are actually used for our experiments. Among these edit samples, 143 are labeled as **semant_diff** and 212 as **semant_simil** (i.e. a 40/60% distribution). Within each version pair labeled as **semant_diff**, the past version is considered and labeled as incoherent (positive label 1) and the new version as coherent (negative label 0). We may consider that under the **semant_diff**, the new version is a rectification of past version (in our annotation guide we deliberately set aside everything that was related to vandalism). Within a pair labeled as semantically similar, both versions are individually considered as coherent. Therefore, we have a total of 710 different samples among which 567 are considered as coherent (around 80% of the whole dataset) and 143 as incoherent (around 20% of the whole dataset). We divide these samples into a development set (568 samples, around 80% of all semantic samples) and a test set (142 samples, around 20% of all semantic samples) that have overall the same coherent/incoherent samples distribution (respectively 80%/20% and 77%/23%).

In order to normalize the string contained in each sample, we use regex (regular expressions) for removing all characters related to Wikipedia editing: double squared brackets and vertical bars from hyperlinks (`[[ ]]`), series of apostrophes from text formatting(`'''''`), `http` links between squared brackets and references between angle brackets ($< >$, references are cited directly in the Wikipedia text source code). By doing so, we reduce the noise contained in our string samples and shape them in order to get phrases more suitable for our tasks. All letters are set to lower case and only textual content (oriented towards Wikipedia readers, unrelated to Wikipedia formatting) is kept. These samples have an average length of 600 characters with a standard deviation of around 640 (Table 3.3). In terms of number of tokens, the samples have a length of 120 tokens with a standard deviation around 128 (Table 3.4). These distribution are the same when comparing development and test sets or when comparing incoherent and coherent samples.

In order to make sure that there is no overlapping sentence between the Wikipedia training corpus (set of most recent versions of all articles in French used for training the skip-gram model) and our samples, we removed all entries in the corpus in which the title article matched

count	710
mean	594.6
std	636.1
min	22.0
25%	209.2
50%	401.0
75%	722.0
max	4795.0

Table 3.3: Statistics summary for character-wise length of string samples

count	710
mean	120.3
std	127.5
min	6.0
25%	43.0
50%	82.0
75%	148.7
max	969.0

Table 3.4: Statistics summary for token-wise length of string samples

that of a sample.