Comparing Top-Down and Bottom-Up Sorting of Enriched Micro-Texts by Humans and Machines

Markus Christen*[1], PhD; Thomas Niederberger[2], MS; Thomas Ott[2],PhD; Reto Aebersold[3]; Suleiman Aryobsei[4], MA; Reto Hofstetter[5], PhD

[1]University of Zurich, Switzerland; [2]Zurich University of Applied Sciences, Switzerland; [3]Atizo AG, Bern, Switzerland; [4]University of St. Gallen, Switzerland; [5]Università della Svizzera italiana, Lugano, Switzerland

*christen@ethik.uzh.ch

Large samples of micro-texts emerging from Facebook, Twitter, or crowdsourcing platforms have become an important source for research in psychology and related fields, but require machine text processing for classification. Due to low word-count and unstructured writing, micro-texts pose a challenge for automatized text processing. We present a semi-supervised learning system for micro text classification that includes optimal text enrichment, a bottom-up learning step to identify the best cluster(s), a supervised control step to define classifiers and an automatized top-down classification procedure. We used micro-texts emerging from the Swiss innovation platform "Atizo", where contributors submit ideas for solving a posted business problem. We tested how various preprocessing and enrichment-techniques (word-splitting, stemming, morphing, trilingual translation, synonyms; and combinations) improve the classification and we identified the best combination over several text types. Finally, we compared our results with human classification obtained by three large experiments ($n_1 = 875$; $n_2 = 901$, $n_3 = 895$) where the participants rated the pairwise similarity of micro texts. Sorting-quality has been measured against a predefined and validated benchmark classification using the Jaccard-coefficient. We find that our semi-supervised learning system achieves similar sorting quality compared to the predefined optimal classifier averaged over all enrichment techniques. Nevertheless, human bottom-up sorting was still stronger and achieved equal quality as the optimal classifier under best enrichment, although the subjects did not have a holistic perspective on the classification task. This outlines that simulating the human ability for context-dependent semantic similarity assessment is still the "holy grail" for automatized text classification.

Introduction

Modern communication platforms are sources of large samples of micro-texts that are in need of machine text processing for text classification and interpretation. On the Swiss innovation platform Atizo, contributors submit ideas or solutions to problems posed by companies. In such a crowdsourcing process, hundreds of people contribute a large amount of micro-text data that needs to be structured already during the process of idea generation in order to avoid repetitions and to optimize the solution space.

Due to low word-count and unstructured writing, microtexts pose a challenge for automatized text processing. Technically, the goal is to partition a growing set of microtexts into topical groups after vectorization of the texts and construction of a 'term frequency / inverse document frequency' (TF-IDF) matrix. In our research, we addressed the whole chain of issues related to data preprocessing and text enrichment, data classification, data visualization, as well as solution benchmarking and process improvement.

Methods

Benchmark: For our study, we used the data of three crowdsourcing contests on the Atizo innovation platform (P_1: 394 texts by 154 contributors; P_2: 314 texts by 129 contributors; P_3: 396 texts by 117 contributors). Each data set has been hand clustered to identify clear groups of ideas. Among those groups, 100 ideas per project have been selected such that for each project, the majority of the ideas belong to four clearly distinguishable groups of different sizes, whereas a minority of ideas belongs to neither group ("noise texts"). The mean numbers of words per bag was 58 for project 1, 56 for project 2 and 61 for project 3. 10 subjects per project have validated the benchmark clusterings with a coder reliability of 80%. The quality of a clustering C compared to the benchmark clustering C_o is calculated by the Jaccard coefficient defined as: $J(C, C_o) = a / (a + b + c)$, where a is the number of pairs of items that are both in C and C_o in same clusters, b is the number of pairs that are only in C in same clusters, and c is the number of pairs that are only in C_o in same clusters.

Text preprocessing and enrichment: We used various preprocessing and enrichment techniques: Baseline were the raw texts after elimination of standard stop words. We then used stemming (S), morphing (M), word splitting (H), synonym enrichment (Y), translation of the German texts into English and French (T) and combinations of those methods (see Fig. 2). We calculated, how each preprocessing and enrichment combination improved the Jaccard coefficient measure compared to the baseline for all three projects and both for bottom-up and top-down clustering.

Bottom-up Hebbian clustering: Our clustering method is based on a Hebbian network approach that is able to learn the principal components of the data in a continuous way based on the texts represented as TF-IDF matrix. See Niederberger et al. (2012) for the technical details (the paper is available on site upon request).

Top-down classification: Based on the benchmark clustering C_o for each project, we identified those words that had a high specificity for the single groups and created the optimial word bag for each group. By calculating the relative overlap of the word bag of the texts with each of the four word bags of the classifier (# of words that are contained in both sets divided by the size of the smaller set), each text is represented by a 4-dimensional vector with coordinates between 0 and 1. Classification has then been achieved using the clustering algorithm of Mathematica® (an adaptation of k-Means) for k=4.

Figures

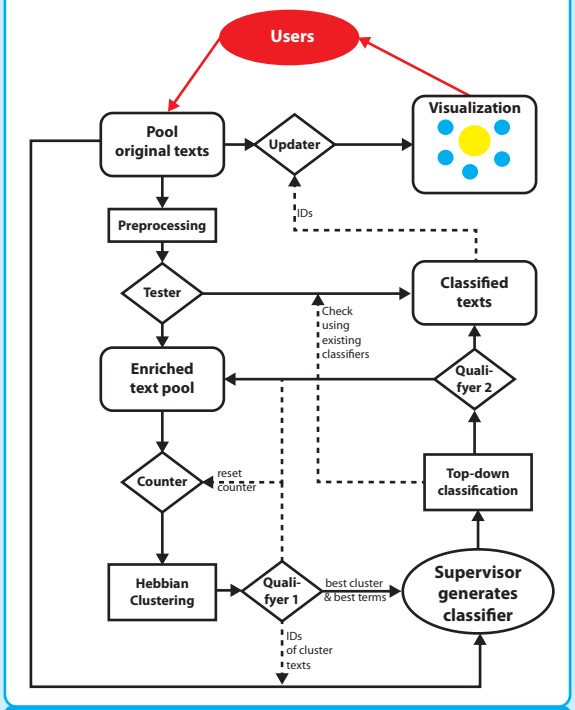

Figure 1: Text classification algorithm combining Hebbian bottom-up clustering and top-down classification. The human supervisor chooses keywords from a machine-generated selection. Qualifier 1 selects the cluster with the lowest inter-text distance, qualifier 2 selects the texts where the number of matches with the classifier word bag is above the mean. Identified clusters are visualized and allow new users to explore new parts of the idea space.

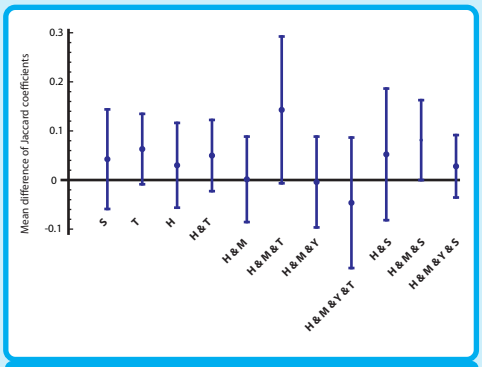

Figure 2: Mean quality improvements of text preprocessing and enrichment measured by the Jaccard coefficient. Abbreviations: see „methods“

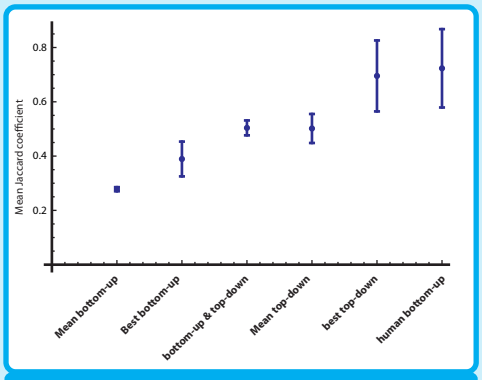

Figure 3: Mean quality of text classification methods applied to preprocessed and enriched texts measured by the Jaccard coefficient.

Methods (continued)

Combined bottom-up clustering and top-down classification: We implemented the procedure outlined in Figure 1 that combines bottom-up clustering for creating a proposal for a classifier and top-down classification. We used for all three project the preprocessing and enrichment procedure that lead to the most optimal results (see Results): H & M & T.

Human bottom-up classification: In the human bottom-up classification experiment, subjects rated the pairwise similarity of two ideas of a single project on a 7-point Likert scale, each subject rated 30 pairs. We aimed for five ratings per pair, so that in total almost 25'000 ratings per project were required. We used Amazon Mechanical Turk for recruiting participants, the study was cleared in accordance with the ethical review processes of the University of Zurich. After quality check, the data of 875 subjects in the first project (exclusion rate 14.8%), of 901 subjects in the second project (exclusion rate 18.8%), and of 895 subjects in the third project (exclusion rate 13.4%) have been included for the analysis. The data alllowed to calculate the normalized pairwise distance of all texts, i.e. each text was represented as a 100-d vector with coordinates between 0 and 1. Classification has been achieved using the clustering algorithm of Mathematica® for k=4.

Results

Optimal preprocessing and enrichment: We found that the effect of preprocessing and enrichment on classification is quite variable, i.e. depends on the type of texts (i.e. they are related to the project). It also shows that translation, which has been introduced as an alternative to synonyms for text enrichment, is a powerful enrichment technique, whereas synonyms generally increases the similarity of all texts and worsen cluster discrimination. In the mean, the most successful type is the combination of word splitting, morphing, and translation, by which up to 70% improvement was possible.

Comparing classification procedures: We find that the combination of bottom-up generated classifier with top-tow classification is equally good as the mean result of all top-down classifications over all preprocessing and enrichment procedures when using the optimal top-down classifier. As the result of text preprocessing and enrichment is strongly dependent on the type of texts (i.e. cannot be known a priori), we can conclude that the combination achieves an optimal result with rather small intervention by a human supervisor. The result of bottom-up clustering measured by the Jaccard index is only about half as good and even the mean of the best bottom-up results over all three projects is clearly worse than our semi-supervised system. A surprising result is that human bottom-up clustering is comparable to the mean of the best top-down classifications using an optimal classifier, although no subject had a holistic overview of all texts.

Conclusions

- Translations in combination with word splitting and morphing outperforms synonym-based text enrichment.
- A combination of bottom-up classifier identification and top-down classification with minimal human supervision achieves comparable results to top-down classification with a predefined optimal classifier.
- Human bottom-up clustering (mimicring the pairwise text-comparison of a distance function) still outperforms machine classification, probably due to the ability of context-dependent semantic similarity assessment.