MERGE: A new recursive approach towards multiword expression extraction and four small validation case studies

Stefan Th. Gries (University of California, Santa Barbara, USA) and
Alexander Wahl (Radboud University, The Netherlands)

In the first part of the paper, we introduce a new bottom-up approach to the identification/extraction of multi-word expressions (MWEs) in corpora. This approach is called MERGE (for Multi-word Expressions from the Recursive Grouping of Elements), and involves the successive merging of bigrams to form word sequences of various lengths. More specifically, the approach involves multiple iterations of looping over a corpus to count all tokens (words and n-grams from previous iterations), compute an association measure for all pairs of tokens (currently the log-likelihood measure but other association measures can also be used), and identifying the highest scoring pair of tokens. This pair then gets merged into a new single token, all frequency statistics in the corpus are updated to reflect the new token due to the merger, and the process reiterates for, say, a user-defined number of iterations. Some crucial advantages of this approach are that (i) MERGE is a thoroughly bottom-up procedure requiring few potentially subjective decisions other than the association measure which is computed (which in turn means it's flexible and includes more information than just cooccurrence frequency); (ii) MERGE does not just return n-grams of a certain pre-defined n, but can return all sorts of n-grams; and (iii) the method requires only a simple adjustment to deal with discontinuous sequences.

In the second part of the paper, we discuss four different validation studies to test/validate the performance of the proposed MERGE algorithm. First, we applied the algorithm to the combined Santa Barbara Corpus of Spoken American English and ICE-Canada to identify MWEs that, according to MERGE, are 'good' and 'bad' MWEs, which was operationalized on the basis of when during the iterations - early vs. late - a MWE was identified by MERGE; we let MERGE run for 20K iterations and allowed for 1-word gaps in the processing. We then had 20 native speakers of American English (students at UCSB) rate them on a Likert scale for how much the MWE stimuli constituted "a complete unit of vocabulary" to see whether the native speakers distinguished between 'good/early' and 'bad/late' MWEs with their ratings, thereby supporting the MWE-quality ratings implied by MERGE. We analyzed

the ratings with a linear mixed-effects model using the MWEs' lengths (as a polynomial to the 2nd degree) and MERGE's ranking *(early* vs. *late)* as fixed-effects independent variables and varying intercepts for all MWEs and varying slopes for length and rank for all MWEs and subjects. We found that MERGE's output indeed distinguishes significantly between 'good' and 'worse' MWEs ($t=-19.17$, $df=31$, $p<0.0001$) such that, as expected, 'good/early' MWEs receive better ratings than 'bad/late' ones, with ratings decreasing as MWEs become longer.

Second, we compared the output of MERGE when applied to the same corpus data as above to the output of Brook O'Donnell's (2011) Adjusted Frequency List, a conceptually similar approach, but one that does not take association strength between elements of MWEs/tokens into consideration. We took the 1000 first items returned by MERGE and the 1000 top items of the AFL and randomly sampled MWEs from them for another rating experiment. In that experiment, 20 (different) native speakers judged altogether 360 items on a Likert scale again. Again we analyzed the ratings with a linear mixed-effects model using MWE method *(MERGE* vs. *AFL)* and MWE length as independent variables (again as a polynomial to the 2nd degree) and varying intercepts for subjects as well as varying slopes for MWE method per participant. We found that, as hypothesized, when MWE length is controlled for, the MWEs returned by MERGE score significantly higher than those returned by the AFL ($t=2.128$, $df=23.4$, $p_{1\text{-tailed}}=0.022$).

Third, we applied both MERGE and AFL to the complete spoken component of the BNC to determine how well both methods can identify 388 expressions that the compilers of the BNC decided to tag as multi-word units (using the <mw></mw> tag). We took the top 10,000 items from either approach and used one-tailed binomial tests to compare the proportion of BNC multi-word units that either approach would identify, i.e. whether MERGE would perform better or worse than AFL on this task; given the previous results, we expected MERGE to find a higher percentage of multiword units than the AFL. Both approaches find mostly high-frequency rather than low-frequency MWEs but do not appear to have a preference for items with high degrees of dispersion. We did separate one-tailed binomial tests in both directions (against both baselines) and found that (i) MERGE found more BNC multi-word units (28.9%) than the AFL (24%) and (ii) that that difference is significant in either direction ($P_{\textbf{MERGE}\text{ vs. }\textbf{AFL}}=0.01522$, $P_{\textbf{AFL}\text{ vs. MERGE}} = 0.0178$).

Finally, we explored MERGE's performance using L1-acquisition corpus

data. We ran MERGE on both the adult and the child utterances of the Lara and Thomas corpora from the CHILDES collection. Specifically, we split adult and child data into an early (the first $^{2}/i$) and a late part (the last $^{1}/_{3}$) and compared the MERGE scores of the adult MWEs that the children used in the late partition to the MERGE scores of the adult MWEs that the children did not use later; more precisely, we computed the proportion of MWEs that the children learned within each of several dozens of bins as defined by log-likelihood scores. We then fit a linear model with the square roots of proportions of learned sequences as the dependent variable (we needed to take square roots to avoid violations of linear modeling assumptions) and we used log-likelihood bin as well as MWE lengths (as a polynomial to the second degree) and which child's data we were exploring as independent variables. We obtained a highly significant and highly explanatory model ($p<0.0001$, adj. $R^2=0.7801$) but, most importantly, we found a significant 3- way interaction of MWE length, MWE strength of association (as defined by log-likelihood bin), and child ($p=0.0076$): For both children (with slight differences between them), MWEs with higher MERGE scores are indeed those that children are more likely to learn even when length is controlled.

We conclude by integrating all results, discussing their implications, and suggesting future analyses.