Estimating and Comparing Entropies Across Written Natural Languages Using PPM Compression[1]

F. Behr, V. Fossum, M. Mitzenmacher, D. Xiao

We extend the previous work measuring the entropy of written English to include the following written natural languages: Arabic, Chinese, French, Japanese, Korean, Russian, and Spanish. We observe that translations of the same document have approximately the same size when compressed, even though they have widely varying uncompressed sizes. This provides further evidence of the popular linguistic postulate that different natural languages have the same descriptive ability. It also provides a possible tool to identify poor machine translations.

Our motivation is the following natural thought experiment. Linguistic theory has suggested that all natural languages are equally expressive. Under this assumption, and further assuming that PPM works well across many languages, we would expect that the same document translated into different languages would compress to approximately the same size. If this were not the case, we would have evidence either that different natural languages differ in expressiveness, that PPM is language-specific, or that the translations we are dealing with are poor. Each of these results would be interesting.

Our experiments follow those by Teahan and Cleary for English. We first clean the texts to a smaller alphabet. For English this alphabet consists of the 26 letters (without case) and the space character. We perform an analogous operation for each language. We then use efficient compression algorithms, in this case PPMD+, PPMZ, and BZIP2, to compress the given texts, and compare the resulting sizes.

We used two corpora to perform our experiments, the Bible and a set of United Nations treaties. We obtained human-generated translations of these texts in the languages mentioned above. For each corpus, the uncompressed size varies greatly across languages. For the Bible, the trend in compressed sizes is largely as we hypothesized: the better the compression algorithm, the closer the compressed sizes. The UN treaties do not match our hypothesis as closely. We speculate that this may be caused by the nature of the documents; legal jargon may be more concise in English. This result is an interesting starting point for future work.

We also performed similar experiments with machine translations. We found that machine translation often fails when it comes across an unknown word, in which case it outputs the word untranslated. This skews compression results, causing large variance in the compressed sizes. Based on our findings, we suggest that compression can be used as a tool to find poor translations.

The results of our experiments, while preliminary, support our hypothesis that translation preserves information content. We believe that our work opens the door for future research concerning the relationship between compression and translation.

---

[1] A full version of this paper is available as Harvard Computer Science Technical Report TR-12-02. Contact author: M. Mitzenmacher, michaelm@eecs.harvard.edu, Harvard University, Div. of Engineering and Applied Sciences. Supported in part by NSF grants CCR-9983832, CCR-0118701, CCR-0121154, and an Alfred P. Sloan Research Fellowship.