Multiword Frequency Analysis Based on the MEDLINE N-gram Set

Chris J. Lu, Ph.D.[1, 2], Destinee Tormey[1], Lynn McCreedy, Ph.D.[1] and Allen C. Browne [1]
[1]National Library of Medicine, Bethesda, MD [2]Medical Science & Computing, LLC, Rockville, MD

Abstract

Multiwords are vital to better precision and recall in NLP applications. The Lexical Systems Group (LSG) developed an effective approach to add multiwords to the SPECIALIST Lexicon from the MEDLINE n-gram set. This paper describes a frequency analysis on LexMultiwords (LMWs) and acronym expansions (e.g. blood pressure for BP) based on the word count (WC) in MEDLINE. Results show most LMWs locate in the low WC range with better precision and F1 score.

Introduction

LMWs are terms in Lexical records containing space(s). To be in the Lexicon, these terms must: 1) have a single part-of-speech (POS), 2) have inflections, and 3) be a special unit of lexical meaning by themselves. A set of filters and matchers based on empirical models has been developed to retrieve LMW candidates from the MEDLINE n-gram set[1]. This process generates high precision LMW candidates for efficient LMW building. An analysis of WC allows us to use the frequency filter effectively for better LMW acquisition.

Approach

First, all unique single words (464,781) and multiwords (431,432) in Lexicon.2015 are retrieved. Second, the acronym expansions are retrieved by applying the Acronym Expansion Pattern (AEP) matcher to the MEDLINE n-gram set. This AEP set includes 14,440 LMW candidates. They are tagged by LSG linguists and are added to the Lexicon if they are valid LMWs. The WC from the MEDLINE n-gram set is added to these three data sets to derive the frequency spectrum of WC class vs term number. Term number (TN) is the total terms in a WC class with a range of 100 incremental (Figure 1). The frequency spectrum of WC class vs local precision (valid tags/total tags), recall (valid tags/total valid tags), and F1 measurement (PRF) are derived for AEP (Figure 2). The recall is normalized between 0 and 1.

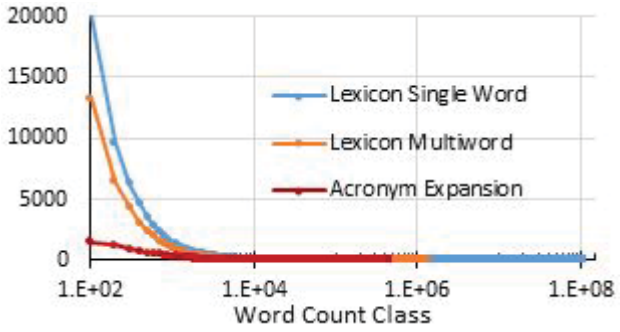

Figure 1. Frequency Spectrum - WC Class vs TN

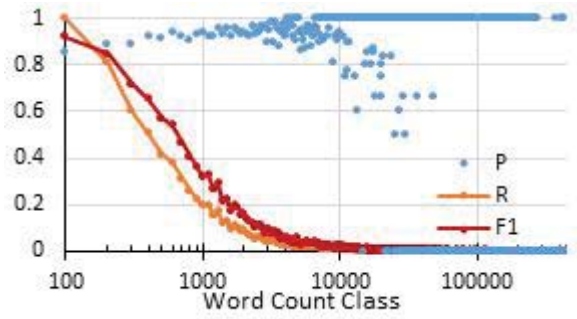

Figure 2. Frequency Spectrum – WC Class vs PRF

Conclusion

Figure 1 shows that most LMWs are located in the low WC range for both the Lexicon and the AEP. This result coincides with the distribution of single words in the Lexicon and "Alice in Wonderland"[2]. It seems both single words and multiwords share the common characteristic of distributing in the low frequency range. Figure 2 shows that low frequency n-grams in the AEP have higher normalized recall and F1 score, with precision above 0.8. Very few LMWs exist in the high WC range with a variation in precision of either 0 or 1. Accordingly, the frequency of LMW acquisition should be set on the lower WC range (100-10k) while the frequency of single word acquisition is set on the high WC range (because most unigrams are valid single words). This frequency strategy is applied with filters and matchers to generate LMW candidates from the MEDLINE n-gram set to enrich the coverage of LexMultiwords. Ultimately, this enhanced coverage provides better NLP results for projects that use the SPECIALIST Lexicon, and our WC results may guide MW acquisition efforts in others' datasets.

References

1. Lu CJ, Tormey D, McCreedy L, Browne AC. Generating the MEDLINE N-Gram Set, Proceedings of AMIA Annual Symposium, 2015 Nov. 14-18; San Francisco, CA; 2015. p. 1569.
2. Baayen RH. Word Frequency Distributions. 1st Edition. Springer Netherlands; 2001. p.10.