Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230284

Clustering Similar Diagnosis Terms

Stefan SCHULZ[1], Akhila ABDULNAZAR and Markus KREUZTHALER
IMI, Medical University of Graz, Austria

Abstract. A large clinical diagnosis list is explored with the goal to cluster syntactic variants. A string similarity heuristic is compared with a deep learning-based approach. Levenshtein distance (LD) applied to common words only (not tolerating deviations in acronyms and tokens with numerals), together with pair-wise substring expansions raised F1 to 13% above baseline (plain LD), with a maximum F1 of 0.71. In contrast, the model-based approach trained on a German medical language model did not perform better than the baseline, not exceeding an F1 value of 0.42.

Keywords. Named Entity Normalization, Electronic Health Records

1. Introduction

Clinicians prefer telegram-style expressions over controlled terms from terminologies. Instead of "Malignant neoplasm of bronchus and lung, middle lobe, bronchus or lung" (ICD-10 C34.2) they write "Adenocarcinoma, middle lobe right", or "Adeno-Ca, R middle lobe", or even "Adneocarcinoma (sic!), right middle lobe". Clinical entity normalization (CEN) should assign the same code to term variants, being tolerant regarding typos, but strict regarding lexical differences ("Vitamin A" vs. "Vitamin B", or "Hepatectomy" vs. "Hepatotomy"). Increasingly, CEN combines neural approaches with dictionaries [1]. We processed about 20.5 million short (max. 50 chars) diagnosis descriptions annotated with ICD-10. A benchmark was created of 20 random entries. The baseline, Levenshtein Similarity (LS) is based on Levenshtein distance (LD) [2]. For strings S_1 and S_2 we define: $LS(S_1, S_2) = 1 - (2 * LD(S_1, S_2) / (Length(S_1) + Length(S_2)))$.

2. Methods

We introduced SLS (Selective Levenshtein Similarity), which ignores stop words and punctuation characters (except "."). SLS requires an exact, case sensitive match for all non-standard tokens (NST), i.e., tokens with non-alpha characters or any upper-case character beyond the first position. LS is applied to the standard tokens only. Exact string match between all NST of S_1 and S_2 is required, otherwise SLS is set to zero. Thus, *SLS* ("Type 1", "Type 2") equals zero as well as *SLS* ("EEG", "ECG"). We optionally consider variants with truncated tokens such as "chron." for "chronisch" (with or without period): for each token of a string pair S_1 and S_2 (after stop word and NST removal) token

[1] Corresponding Author: S. Schulz, Medical University, Auenbruggerplatz 2/V, 8036 Graz. E-mail: stefan.schulz@medunigraz.at.

S_{1i} is substituted by S_{2i} if the former is a left-sided substring of the latter, or S_{2i} is substituted by S_{1i} in the opposite case. We also compared the original word order with alphabetic word order (AWO). This algorithmic approach is then compared to a neural method. Top matches to the vector representations of strings from an embedding space are analysed, filled with ICD terms in their vector representations obtained by downstreaming a German medical language model leveraging SapBERT [3] on random pairs (max 50 of the same ICD code) of the list, enriched by official ICD-10 terms and synonyms. Training had been done for 50 epochs on an NVIDA GeForce GTX Titan X GPU. Similarity matching was based on a k-nearest neighbour approach using Faiss [4].

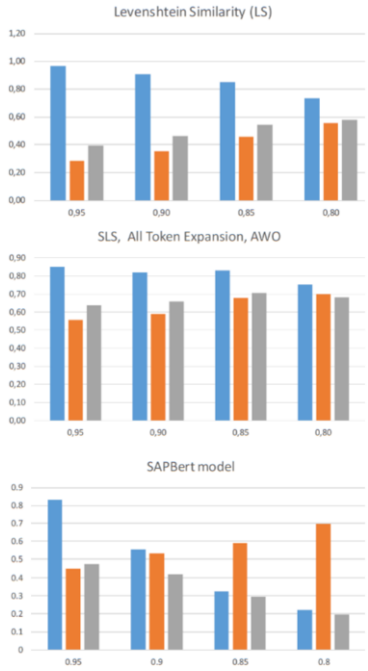

Figure 1. P, R, F_1 at several points

3. Results and Discussion

Regarding F_1, the neural model did not fare better than the LS baseline whereas the algorithmic approach yielded an F_1 13% above baseline (Fig. 1). We found that the precision drop of the model-based approach was mostly due to candidates with true variants plus additional modifiers, e.g. ("Duodenalstenose" vs. "St.p. Duodenalstenose", cosine 0.92) and small but significant variations in tokens with numerals ("Spinalkanalstenose L3,L4" vs. "Spinalkanalstenose L3-L5"). Variants with abbreviations ("rez. Erbrechen" vs. "rezidivierendes Erbrechen") had a lower cosine (here 0.85). To retrieve all variant candidates, the algorithmic approach took on average 4.5 min. compared to 23 sec. of the model-based one. The explaining power of these results is limited by small sample size and data heterogeneity. The importance to apply fuzzy string matching selectively as well as the potential of the resolution of truncation-based abbreviations is emphasised. In contrast, the SapBERT model lowercases all input and does not consider abbreviations. The coarse-grainedness of ICD-10 was the reason that the model did not learn many distinctions, even that "right" and "left" are not synonyms, because they occur in nearly the same distributional contexts. Future work should emphasise combinations of the two approaches, e.g., by using the algorithmic approach to optimise the sampling for SapBERT.

References

[1] Ferré, A, Deléger L, Bossy R et al. C-Norm: a neural approach to few-shot entity normalization. BMC Bioinformatics 2020 Suppl 23, 579
[2] Levenshtein VI et al. Binary codes capable of correcting deletions, insertions, and reversals. Soviet physics doklady. 1966; S. 707–710.
[3] Liu F, Shareghi E, Meng Z, Basaldella M, Collier N. Self-Alignment Pretraining for Biomedical Entity Representations. Proc. of NACL Human Language Technologies 2021 Jun, 4228–4238.
[4] Johnson J, Douze M, Jégou H. Billion-scale similarity search with GPUs. IEEE Transactions on Big Data. 2019;7(3):535–547.