Lexical normalizzation of user-generated medical forum data

Anne Dirkson, Suzan Verberne, Gerard van Oortmerssen, and Wessel Kraaij

Universiteit Leiden

Data Science Research Programme

Introduction

Social media text is noisy and this is aggrevated in the medical domain [1]. It is plagued by:

- Typos
- Misspellings
- Domain-specific abbreviations

Lexical normalization of social media text has been addressed by Sarker [2], but does not deal with:

- Domain-specific abbreviations
- Medical OOV terms that should not be corrected

Data

	# Tokens	# Posts
(1) Gastro Intestinal Stromal Tumor forum	1,225,741	36,722
(2) Sub-reddit on cancer	4,520,074	274,532

Methods

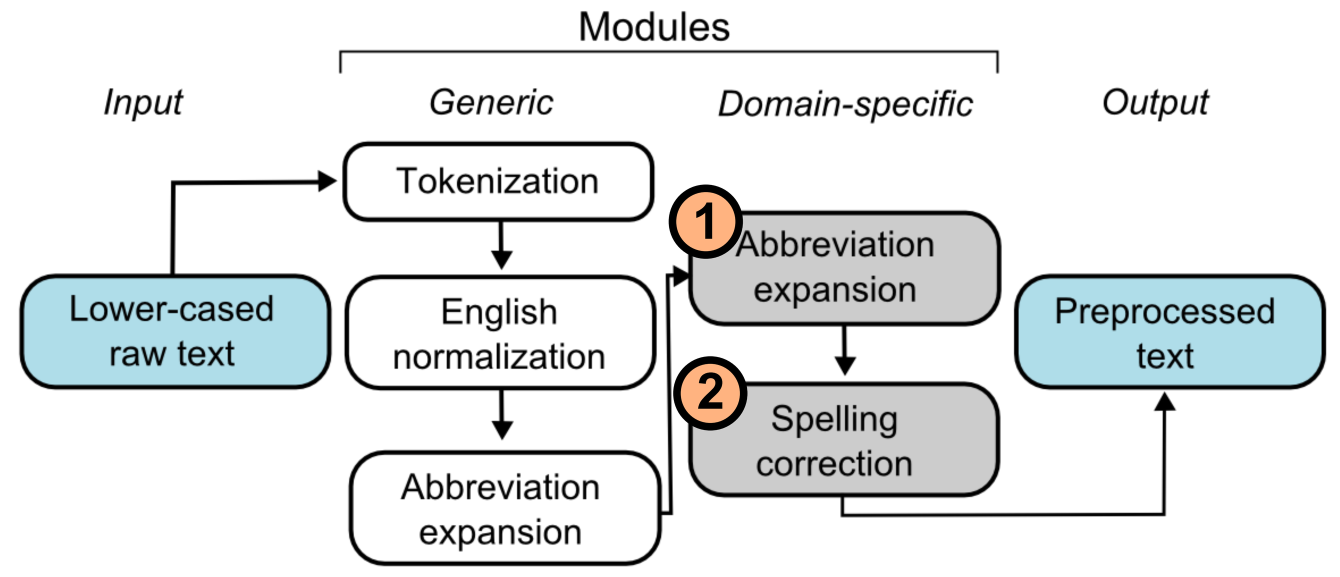

Figure 1: Sequential unsupervised preprocessing pipeline.

1 36 abbreviations found in 500 posts form the lexicon

2

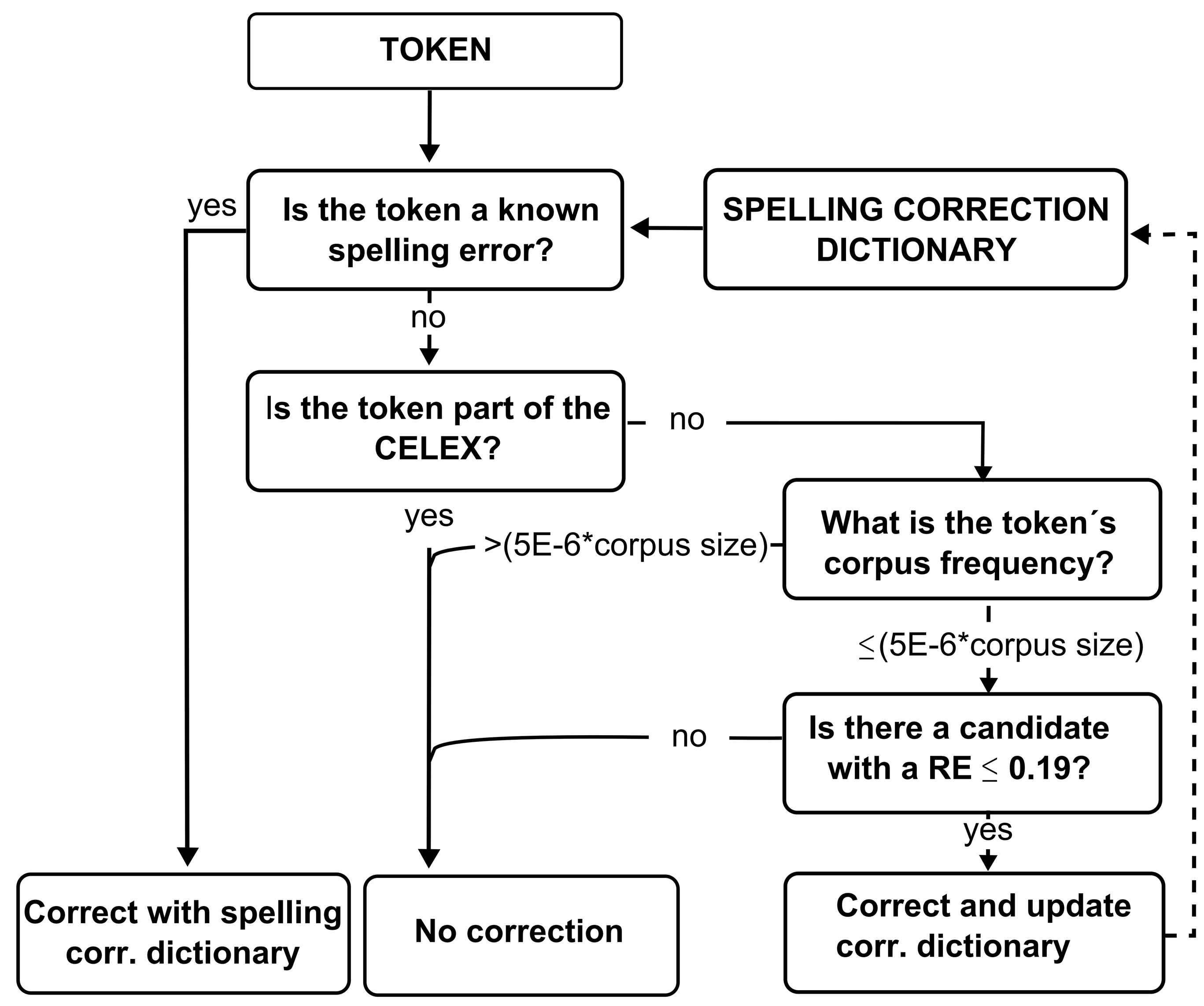

Figure 2: Decision process for spelling detection. RE: Relative Edit Distance. Correction candidates from CELEX [3] and corpus tokens > freq. threshold.

a.r.dirkson@liacs.leidenuniv.nl

SIDN fonds

Results

Our normalization pipeline:

- **is generalisable across cancer-related forums**
- **mainly targets medical concepts**

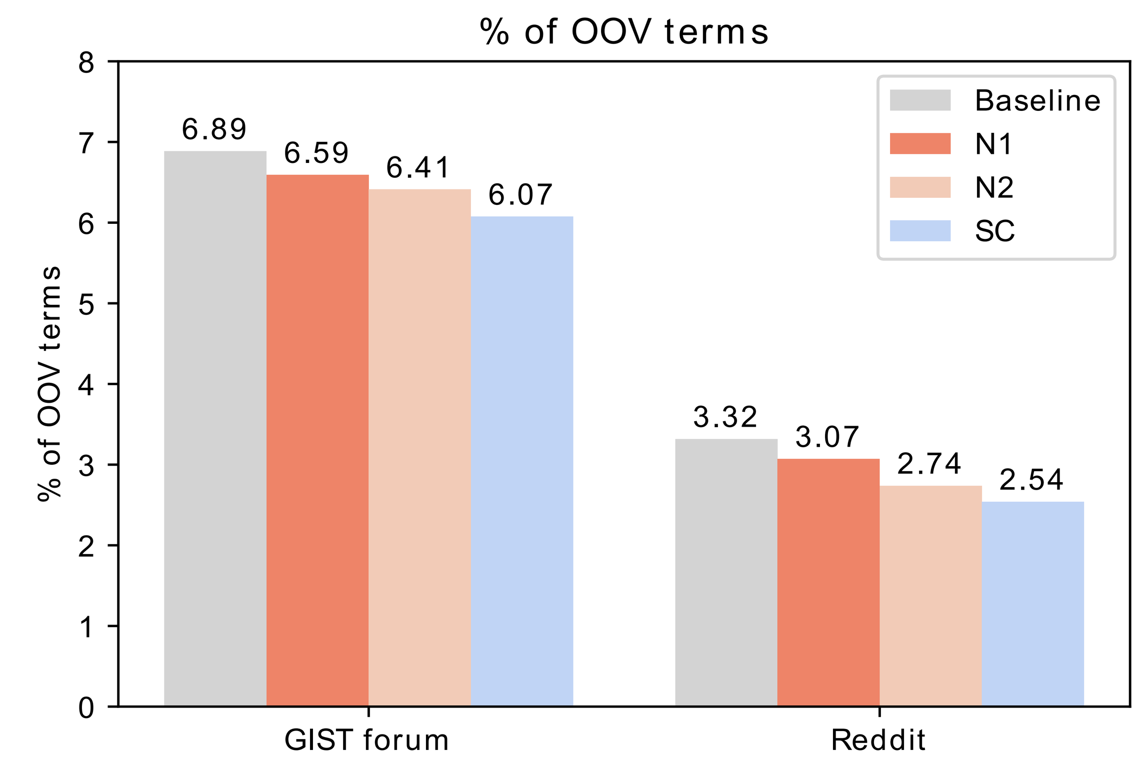

Figure 3: Number of OOV-terms with sequential modules. N1: Generic abbreviation expansion [2]. N2: Domain-specific abbreviation expansion. SC: Spelling correction.

GIST forum	gleevec	oncologist	diagnosed
Reddit	metastasized	treatment	diagnosed

Table 1: Most frequent spelling mistake corrections

- **Spelling correction for generic social media (S1) does not suffice**

	NAE	NAE+P	RE	RE+P	S1	S2
Accuracy	59.6%	59.6%	**66.0%**	**66.0%**	23.4%	19.1%

Table 2: Spelling correction algorithm comparison. NAE: normalized absolute edit distance. +P: with first-letter penalty. RE: relative edit distance. S1: Sarker´s algoritm [2]. S2: S1 without the language model.

- **Our method targets infrequent mistakes**
  → no false positives

	Recall	Precision	F_1	$F_{0.5}$	AUC
Decision process	0.38	1.0	0.55	0.75	0.69

Table 3: Detection of spelling mistakes in the test set.

Conclusion

Our pipeline can improve the quality of the text data from medical forum posts. Future work will explore its impact on text mining tasks.

References

1. G. Gonzalez-Hernandez, A. Sarker, K. O´Connor, and G. Savova. 2017. Capturing the Patient´s Perspective; A Review of Advances in Natural Language Processing of Health-related Text. Yearbook of Medical Informatics (2017, 214-217.
2. A. Sarker, 2017. A customizable pipeline for social media text normalization. Social Network Analysis and Mining, 7, 45(2017).
3. G. Burnage, R.H. Baayen, R. Piepenbrock, and H . Van Rijn. 1990. CELEX: A Guide for Users. (1990).