



Swansea University
Prifysgol Abertawe



Cronfa - Swansea University Open Access Repository

This is an author produced version of a paper published in:
International Journal of Population Data Science

Cronfa URL for this paper:

<http://cronfa.swan.ac.uk/Record/cronfa44398>

Conference contribution :

Fonferko-Shadrach, B., Lacey, A., Akbari, A., Thompson, S., Ford, D., Lyons, R., Rees, M. & Pickrell, O. (2018). *Using natural language processing to extract structured epilepsy data from unstructured clinic letters*. International Journal of Population Data Science, Banff, Canada: IPDLN 2018 conference.
<http://dx.doi.org/10.23889/ijpds.v3i4.699>

This item is brought to you by Swansea University. Any person downloading material is agreeing to abide by the terms of the repository licence. Copies of full text items may be used or reproduced in any format or medium, without prior permission for personal research or study, educational or non-commercial purposes only. The copyright for any work remains with the original author unless otherwise specified. The full-text must not be sold in any format or medium without the formal permission of the copyright holder.

Permission for multiple reproductions should be obtained from the original author.

Authors are personally responsible for adhering to copyright and publisher restrictions when uploading content to the repository.

<http://www.swansea.ac.uk/library/researchsupport/ris-support/>

Using natural language processing to extract structured epilepsy data from unstructured clinic letters

Fonferko-Shadrach, B¹, Lacey, A¹, Akbari, A², Thompson, S¹, Ford, D¹, Lyons, R³, Rees, M⁴, and Pickrell, O⁴

¹Swansea University

²Health Data Research UK - Wales and Northern Ireland, Swansea University Medical School

³Farr Institute, Swansea University Medical School

⁴Neurology and Molecular Neuroscience, Swansea University Medical School

Introduction

Electronic health records (EHR) are a powerful resource in enabling large-scale healthcare research. EHRs often lack detailed disease-specific information that is collected in free text within clinical settings. This challenge can be addressed by using Natural Language Processing (NLP) to derive and extract detailed clinical information from free text.

precision neurology research, in addition to potential applicability to other disease areas.

Objectives and Approach

Using a training sample of 40 letters, we used the General Architecture for Text Engineering (GATE) framework to build custom rule sets for nine categories of epilepsy information as well as clinic date and date of birth. We used a validation set of 200 clinic letters to compare the results of our algorithm to a separate manual review by a clinician, where we evaluated a “per item” and a “per letter” approach for each category.

Results

The “per letter” approach identified 1,939 items of information with overall precision, recall and F1-score of 92.7%, 77.7% and 85.6%. Precision and recall for epilepsy specific categories were: diagnosis (85.3%,92.4%), type (93.7%,83.2%), focal seizure (99.0%,68.3%), generalised seizure (92.5%,57.0%), seizure frequency (92.0%,52.3%), medication (96.1%,94.0%), CT (66.7%,47.1%), MRI (96.6%,51.4%) and EEG (95.8%,40.6%). By combining all items per category, per letter we were able to achieve higher precision, recall and F1-scores of 94.6%, 84.2% and 89.0% across all categories.

Conclusion/Implications

Our results demonstrate that NLP techniques can be used to accurately extract rich phenotypic details from clinic letters that is often missing from routinely-collected data. Capturing these new data types provides a platform for conducting novel

