

NRC Publications Archive Archives des publications du CNRC

Automatic annotation of disposition counts in news articles

Rodier, Simon; Carter, Dave

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. / La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version acceptée du manuscrit ou la version de l'éditeur.

For the publisher's version, please access the DOI link below. / Pour consulter la version de l'éditeur, utilisez le lien DOI ci-dessous.

Publisher's version / Version de l'éditeur:

<https://doi.org/10.3233/SHTI250502>

Intelligent Health Systems: From Technology to Data and Knowledge: Proceedings of MIE 2025, Studies in Health Technology and Informatics, pp. 906-907, 2025-05-15

NRC Publications Archive Record / Notice des Archives des publications du CNRC :

<https://nrc-publications.canada.ca/eng/view/object/?id=402e0893-6d85-4fa3-8771-0b0e13360b9f>

<https://publications-cnrc.canada.ca/fra/voir/objet/?id=402e0893-6d85-4fa3-8771-0b0e13360b9f>

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at

<https://nrc-publications.canada.ca/eng/copyright>

READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site

<https://publications-cnrc.canada.ca/fra/droits>

LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Questions? Contact the NRC Publications Archive team at

PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

Vous avez des questions? Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

Automatic Annotation of Disposition Counts in News Articles

Simon RODIER^{a,1} and Dave CARTER^a

^aDigital Technologies, National Research Council Canada

ORCID ID: Simon Rodier <https://orcid.org/0000-0002-6797-629X>,

Dave Carter <https://orcid.org/0000-0003-3503-7615>

Abstract. News media aggregate and report disposition counts during crises: how many people are affected, suspected affected, have died, and have recovered or been recovered; and they tend to do so in a timely and trustworthy manner. We present and evaluate a method for identifying these counts in unstructured natural language text, supporting downstream tasks such as automatic creation of epidemic curves.

Keywords. Case counting; named entity recognition; natural language processing

1. Introduction

News media often describe the evolution of public health concerns by reporting on the number of people affected (cases, casualties, etc.). We present a system that identifies these *disposition counts*² with their certainty and location. Figure 1 presents an example.

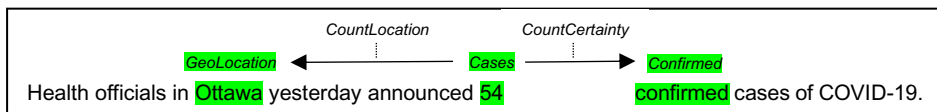


Figure 1. A sample sentence with a count (54), type (Cases), certainty (Confirmed), and location (Ottawa).

2. Methods and Results

The task of automatically annotating these counts can be modelled as one of finding relations between entities in text, where the entities are the counts, certainties and locations. We use Named Entity Recognition (NER; [1][2]) to tag the entities, and Relation Extraction (RE; [3][4]) to link counts to certainties and locations. We developed a dataset³ of 850 news articles from the Global Public Health Intelligence Network [5] platform. We preprocess the text using Stanford CoreNLP [7] to tag tokens, parts-of-

¹ Corresponding Author: Simon Rodier; E-mail: simon.rodier@cnrc-nrc.gc.ca.

² We define *disposition* as a medical state in which a person finds themselves, and a *disposition count* to be a reported number of people in this state at a given time.

³ Hand-annotated (with BRAT [6]) 50 highly-relevant COVID-19 articles per month from 2020/02 to 2021/06. We annotated all occurrences of disposition counts with their type, and the certainty and location words that relate back to a disposition count. Total of ~14 000 annotated entities.

speech, locations and numbers; and replace specific numbers and locations with generic tokens to improve generalizability. We train the NER and RE classifiers⁴ with the dataset.

We evaluate the system components using 5-fold cross-validation, and obtain an F_1 score⁵ of 0.751 for the NER task and 0.912 for RE. We evaluate the whole system on five metrics⁶ also with 5-fold cross-validation, measuring the system's correct annotation of: the count (e.g. 54; F_1 : 0.812); count and type (e.g. 54, cases; F_1 : 0.782); count, type and certainty (e.g. 54, cases, confirmed; F_1 : 0.750); count, type and location (e.g. 54, cases, Ottawa; F_1 : 0.541); and count, type, certainty and location (e.g. 54, cases, confirmed, Ottawa; F_1 : 0.517).

3. Discussion and Conclusions

We have presented a system for automatic disposition count annotation in news articles, using NER to identify entities and RE to link them together. Our NER score is lower than general-purpose NER taggers, such as huggingface's (F_1 : 0.913) [10] and CoreNLP's (F_1 : 0.923) [8]. However, those largely target noun entities (e.g. Person, Organization, Location), a task that is more linguistically constrained than this one. Even with the added complexity, our NER system provides useful results. Our RE score is in line with huggingface's (F_1 : 0.95) [11]. Overall tagging is strongest at identifying count, type and certainty, and weaker with location, which proved harder to automatically tag. Future work could further disambiguate counts by establishing the time covered by a count, and whether it is incremental or total.

References

- [1] Nadeau D, Sekine S. A survey of named entity recognition and classification. *Linguisticae Investigationes*. 2007;30(1):3-26.
- [2] Nasar Z, Jaffry SW, Malik MK. Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys (CSUR)*. 2021;54(1):1-39.
- [3] Angeli G, Premkumar MJJ, Manning CD. Leveraging linguistic structure for open domain information extraction. In: *Proceedings of the 53rd Annual Meeting of the ACL*; 2015. p. 344-54.
- [4] Nguyen TH, Grishman R. Combining neural networks and log-linear models to improve relation extraction. *arXiv preprint arXiv:151105926*. 2015.
- [5] Carter D, Stojanovic M, Hachey P, Fournier K, Rodier S, Wang Y, et al. Global public health surveillance using media reports: redesigning GPHIN. In: *Digital Personalized Health and Medicine*; 2020. p. 843
- [6] Stenetorp P, Pyysalo S, Topic G, Ohta T, Ananiadou S, Tsujii J. BRAT: a web-based tool for NLP-assisted text annotation. In: *Demonstrations at the 13th Conference of the EACL*; 2012. p. 102-7.
- [7] Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit. In: *ACL System Demonstrations*; 2014. p. 55-60.
- [8] Finkel JR, Grenager T, Manning CD. Incorporating non-local information into information extraction systems by gibbs sampling. In: *Proceedings of the 43rd annual meeting of the ACL*; 2005. p. 363-70.
- [9] Surdeanu M, McClosky D, Smith M, Gusev A, Manning CD. Customizing an information extraction system to a new domain. In: *ACL 2011 Workshop on Relational Models of Semantics*; 2011. p. 2-10.
- [10] Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*. 2018.
- [11] Orlando R, Huguet Cabot PL, Barba E, Navigli R. Retrieve, Read and LinK: Fast and Accurate Entity Linking and Relation Extraction on an Academic Budget. *ACL 2024. Bangkok, Thailand*; 2024.

⁴ NER classifier is a CRF classifier trained with CoreNLP [8], RE classifier also trained with CoreNLP [9]. Feature ablation experiments for both classifiers showed strong resilience.

⁵ We chose F_1 as a scoring metric as a balanced measure of true and false positives and negatives.

⁶ Partially-correct tags scored as 0.5 false positive and 0.5 false negative.