gaHealth: An English–Irish Bilingual Corpus of Health Data

Engaging Content
Engaging People

A World Leading SFI Research Centre

Science Foundation Ireland For what's next

Séamus Lankford
School of Computing,
Dublin City University, Dublin, Ireland

Haithem Afli
Department of Computer Science,
Munster Technological University, Cork, Ireland

Órla Ní Loinsigh
School of Computing,
Dublin City University, Dublin, Ireland

Andy Way
School of Computing,
Dublin City University, Dublin, Ireland

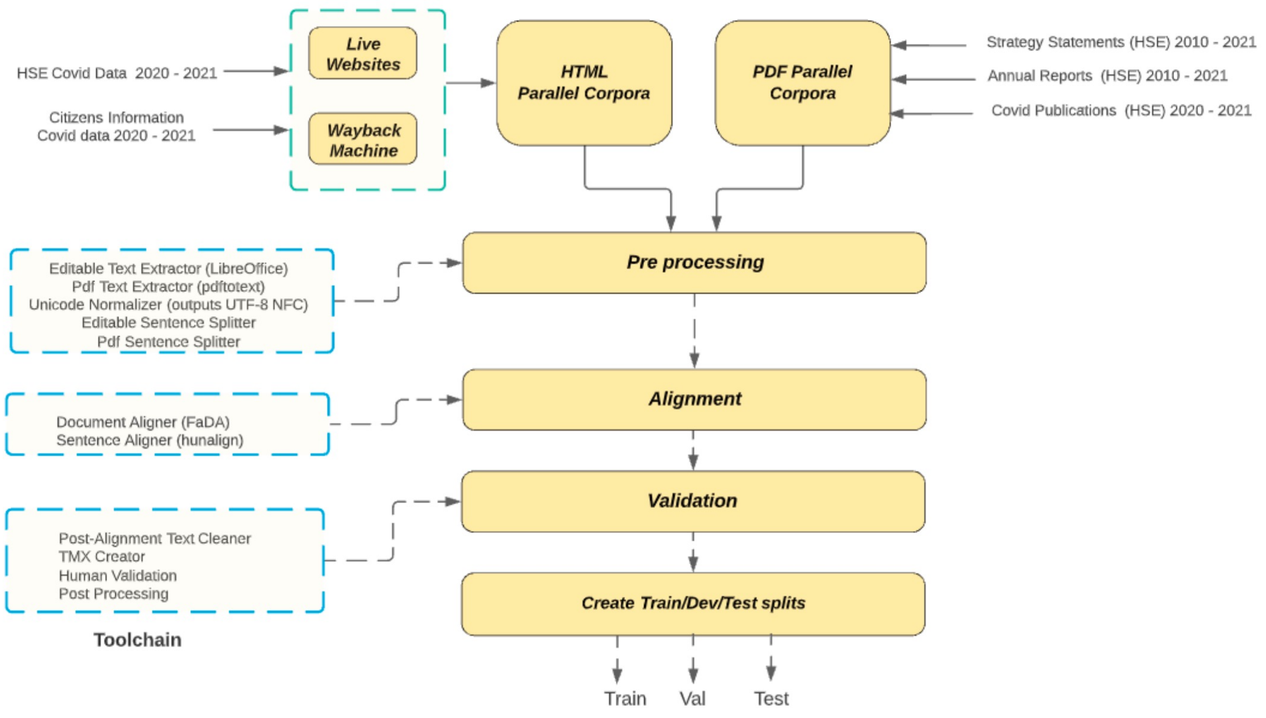
Introduction

A corpus, *gaHealth,* for the specific domain of health was developed for the low-resource English to Irish language pair. We define linguistic guidelines and outline the process used in developing the corpus. Our models, developed using the *gaHealth* corpus, demonstrated a maximum BLEU score improvement of 22.2 points (40%) when compared with top performing models from the LoResMT2021 Shared Task.

Proposed approach

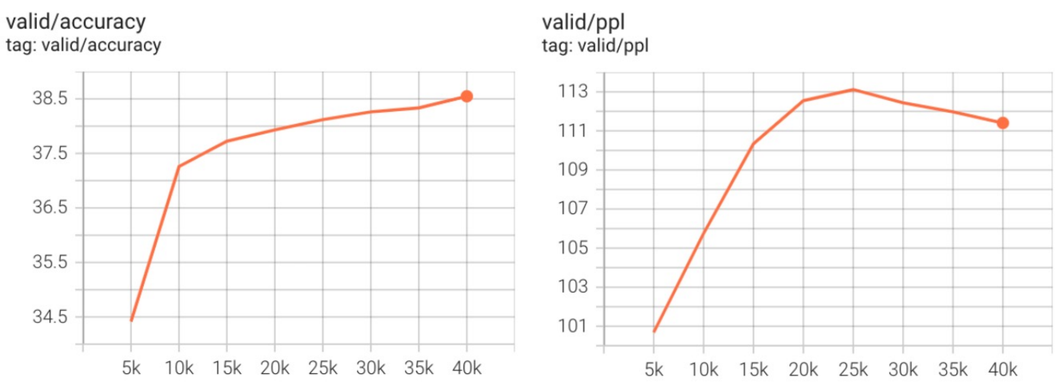
Documents	Source	Lines
Strategy Statement 2020	HSE	3k
Strategy Statement 2017	HSE	2.5k
Strategy Statement 2015	HSE	3k
Annual Report 2020	HSE	2k
Annual Report 2019	HSE	2k
Annual Report 2017	HSE	2k
Website (Covid)	Citizen's Advice	4k
Publications (Covid)	HSE	4k

Sources used in corpus development



Corpus development process

Empirical Evaluation: EN-GA MT model

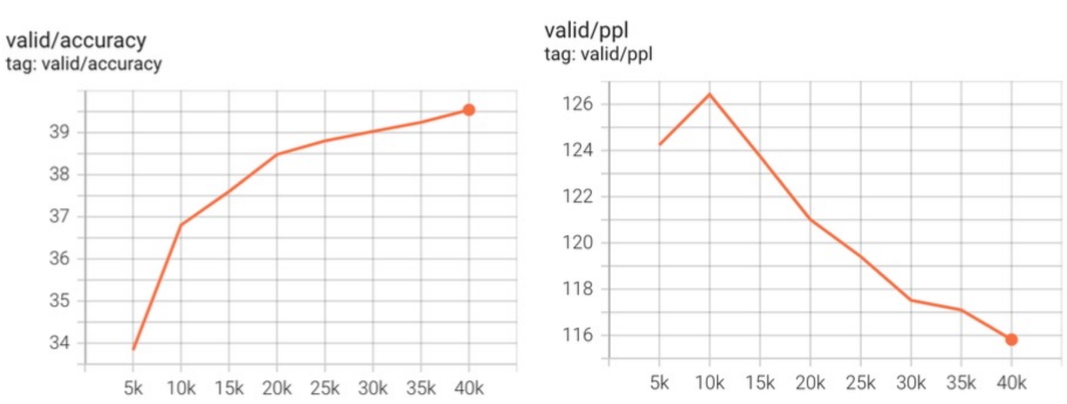


Training EN-GA model with gaHealth corpus

Team	System	BLEU ↑	TER ↓	ChrF3 ↑
UCF	en2ga-b	13.5	0.756	0.37
IIITT	en2ga-b	25.8	0.629	0.53
adapt	combined	32.8	0.590	0.57
gaHealth	en2ga	33.3	0.604	0.56
adapt	covid_extended	36.0	0.531	0.60
gaHealth	en2ga*	**37.6**	0.577	0.57

EN-GA *gaHealth* system compared with LoResMT 2021

Empirical Evaluation: GA-EN MT model



Training GA-EN model with gaHealth corpus

Team	System	BLEU ↑	TER ↓	ChrF3 ↑
UCF	ga2en-b	21.3	0.711	0.45
IIITT	ga2en-b	34.6	0.586	0.61
gaHealth	ga2en	**57.6**	0.385	0.71

GA-En *gaHealth* system compared with LoResMT 2021

Conclusion

- *gaHealth* is an ongoing translation project that built the first parallel corpus of health data for English to Irish translation
- We developed guidelines which help in the conversion process of raw source documents
- Transformer models, trained with gaHealth, achieved SOA MT performance in health domain: **EN-GA: BLEU score of 37.6 & GA-EN: BLEU score of 57.6**
- Open-source download: https://github.com/seamusl/gaHealth

This research was conducted with the financial support of Science Foundation Ireland at ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology at DCU [13/RC/2106_P2]. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Ollscoil Chathair Bhaile Átha Cliath
Dublin City University