Benchmarking the Simplificatin of Dutch Municipal Texts

Daniel Vlantis[1,2], Shuai Wang[1], Iva Gornishka[2]

[1]Vrije Universiteit Amsterdam, [2]City of Amsterdam

According to recent research, **16%** of the people between 16 and 65 in **Amsterdam** have **low literacy skills**. This **hinders societal participation** in tasks such as voting, paying taxes, reissuing documents, or applying for social benefits. Thus, as part of our Amsterdam for All project, we have set on a mission to research the ethical use of AI for measuring and improving the readability of municipal communication.

Related Work

Due to the lack of datasets and research in this domain, previous work on Dutch text simplification was limited to **lexical simplification** (Hobo et al., 2023), the **use of commercial tools or user studies for evaluation** (Harmsen and Van Raaij, 2023), or the **use of automatically-generated data without manual verification** (Van de Velde, 2023).

To alleviate the lack of sufficient training data, Evers (2021) proposed a **pivot-based approach** as an alternative. The aim is to take advantage of high-resource languages such as English by using them in an intermediate step. While Evers' work focuses on simplification in the **medical domain**, municipal communication covers **a large range of specialized domains** related not only to public health, but also taxes, regulations, social issues, etc.

Why not just GPT?

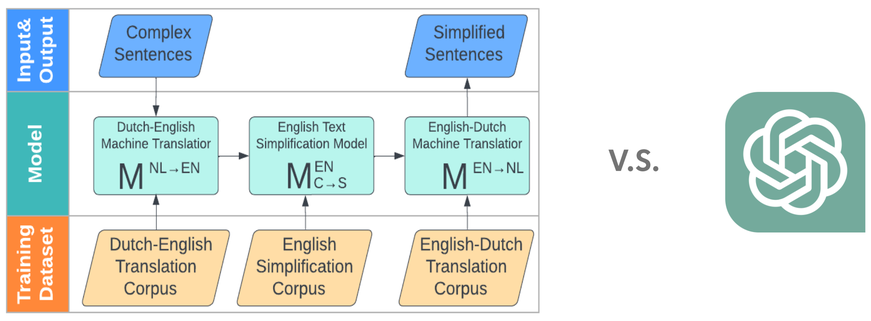
Applications of LLMs within the public sector bring a number of considerations related to **transparency**, **environmental impact** and **responsible development**, as well as the **societal biases** amplified by the model use at scale and the **lack of factuality** in generated content. We need to select methodologies that not only optimize performance, but also align with our own standards and values.

Research Questions

1. Can we improve the results of Evers' pivot approach by augmenting training data or using alternative translation corpora?
2. How does the pivot pipeline perform when transferred to the domain of Dutch municipal communication?
3. Can LLMs (such as GPT) outperform the pivot-based approach?

Methodology

We adopt the **original pipeline of Evers**, perform a number of **improvements** and compare the pivot pipeline to **GPT-3.5-Turbo** using a basic prompt template and minimal post-processing.



v.s.

1. First, we improve the results of the Dutch-to-English model, which translated well complex medical terminology but **failed to capture more general language**, by augmenting the NL-EN model's domain-specific training data (originally the EMEA dataset) with **data from the more general OpenSubtitles corpus**.

2. We test Evers' assumption that the **OpenSubtitles corpus is more suitable** for translating simple sentences in the EN-NL model by training on the **specialized EMEA corpus instead**.

3. We conduct experiments to evaluate the transferability of the model to a **new domain: Dutch municipal texts**.
   - **Europarl corpus** as a domain-specific corpus
   - data from the **Dutch Government Website** to extract a **municipal subset of OpenSubtitles**

Domain	Dataset	Purpose	#Pairs (x1000)
Encyclopedia	WikiSimple (Coster and Kauchak, 2011)	TS	283
Subtitles	OpenSubtitles (Lison and Tiedemann, 2016)	MT	1010
Medical	EMEA (Tiedemann, 2012)	MT	308
	WikiMed (Van et al., 2020)	Ref	3.39
	MedSubset (TF-IDF) (Lison Medical and Tiedemann, 2016)	MT	836
	MedSubset (BERT) (Lison and Tiedemann, 2016)	MT	379
	Medical Eval Set (Evers, 2021)	MT	0.1
Parliamentary	EuroParl (Koehn, 2005)	MT	1950
	Dutch Government Website (European Language Resource Coordination, 2015)	Ref	6.53
Municipal	MunSubset (TF-IDF) (Lison and Tiedemann, 2016)	MT	559
	MunSubset (BERT) (Lison and Tiedemann, 2016)	MT	531
	Municipal Eval Set	Eval	1.31

Dutch Municipal Evaluation Dataset

- **~50 documents manually edited by communication experts**
- provided by the City of Amsterdam
- diverse sources (e.g. **reports, citizen letters, newsletters**, etc.)
- variety of topics (**legal, medical, urban planning,** etc.).

- **create candidate complex-simple pairs** by aligning paragraphs and then sentences within them based on TF-IDF similarity
- **post-processing** to filer (near) duplicates, merge complex sentences which were simplified by splitting into 2 or more simple sentences and to preserve the simple version with lowest edit-distance

- This resulted in **1311 sentence pairs**
- The dataset was **manually anonimized** before publishing

Data & Paper

	Complex	Simple
NL	Op 17 maart 2021 **vindt** de verkiezing voor de Tweede Kamer **plaats**.	Op 17 maart 2021 **is** de verkiezing voor de Tweede Kamer.
EN	The elections for the Second Chamber **take place** on 17 March 2021.	The elections for the Second Chamber **are** on 17 March 2021.
NL	Een belangrijk onderdeel van de **circulaire ambities** is bouwen met hout.	Een belangrijk onderdeel van de **doelen voor blijvend hergebruik** is bouwen met hout.
EN	An important part of the **circular ambitions** is building with wood.	An important part of the **goals for sustainable reuse** is building with wood.
NL	Vervolgens is in 2007 een definitief ontwerp voor het park **gemaakt** en **bestuurlijk vastgesteld**.	Vervolgens **maakten we** in 2007 een definitief ontwerp voor het park dat **het bestuur vaststelde**.
EN	Then, in 2007, a final design for the park **was created** and **officially approved**.	Then, in 2007, **we created** a final design for the park, which **the officials approved**."

Results

- Combining the specialized EMEA corpus with a subset of OpenSubtitles yields best results
- The domain-specific corpus does not improve performance over the simpler one in the municipal experiments, likely due to differences in local and European terminology
- GPT outperforms the pivot-based approach

Numerical Information Preservation
- Example: **450** voting locations, **40%**
- Manually checked the 101 medical and 50 random municipal sentences
- GPT better at preserving them, especially in the municipal domain

Simplicity Evaluation
- Compared Flesch Reading Ease
- GPT achieves better scores in the medical domain
- closer scores in the municipal domain (65.20 vs 63.27)

System		SARI	BLEU	METEOR
Complex NL – EN	Simple EN – NL			
EMEA	OpenSubtitles	29.55	5.18	22.71
EMEA	MedSubset	30.59	6.82	27.79
EMEA	EMEA	32.52	11.44	35.27
EMEA	EMEA+MedSubset	32.69	11.45	35.49
EMEA+MedSubset	EMEA+MedSubset	34.14	15.41	40.89
GPT 3.5 Turbo		**40.26**	**21.23**	**47.49**

Table 1 Results in the **medical domain**

System		SARI	BLEU	METEOR
Complex NL – EN	Simple EN – NL			
Europarl	OpenSubtitles	24.64	7.72	29.34
Europarl	MunSubset	27.70	12.79	38.26
Europarl	Europarl	23.57	6.13	25.54
Europarl	Europarl+MunSubset	28.70	14.83	40.44
Europarl+MunSubset	Europarl+MunSubset	29.83	17.12	43.32
GPT 3.5 Turbo		**34.00**	**22.60**	**48.63**

Table 1 Results in the **municipal domain**

Conclusions & Future Work

- In the future, some **new metrics** could be designed that better reflect not only lexical similarity to reference sentences and fluency, but also simplicity, preservation of numbers and other factual information.

- The main bottleneck of Dutch municipal text simplification is the **lack of a suitable training dataset**. The availability of larger Dutch simplification datasets which were manually annotated, verified or thoroughly analyzed would allow the development of end-to-end models.

- Future work should also include experiments with **existing translation models** (e.g. M2M100, NLLB200)

- Finally, **experiments with domain experts** should validate the usefulness of the results.

City of Amsterdam

VU VRIJE UNIVERSITEIT AMSTERDAM

amsterdam intelligence