François REMY

Software Engineer (NLP & W3C) – Assertive communicator – Passionate and Generous – Learning is part of my DNA

★★★★★ Curriculum Vitae ★★★★★

Personal information

Email	francois.remy@live.be	Phone	(+32) 472 21 12 41
Pseudo	FremyCompany	Blog	https://fremycompany.com/
Birthday	3rd of August 1991	Address	Maaltebruggestraat 234 B301 9000 Gent (Belgium)

Education + Certifications

2019-2024	Doctoral Student, Natural Language Processing, at IDLAB/UGent
2012-2014	Master in Software Engineering (Comp. Intelligence), at ULB – GD (A grade)
2012 Q3	Natural Language Processing, Stanford University (online at coursea.org)
2009-2012	Bachelor in Engineering (Polytechnic), at ULB – D,GD,D (A grade)
2009 Q4	Sixth place at Belgian Olympiads of Computer Science
2009 Q2	Fourth prize at Belgian Olympiads of Mathematics

Technical and soft skills

Designing, adapting, and training machine learning architectures for NLP (incl. LLM) and ASR

> Python, PyTorch, HuggingFace, Tensoflow, Tensorflow.js

> MongoDB, Excel, Azure Datalake U-SQL, MatLAB, R,

Designing software for the Web or modern operating systems

> HTML, CSS, ECMAScript, TypeScript, SVG, MathML

> Python, F#, VB.NET, C#, C++, Java, x86 assembly, others

Come up and validate hypotheses through experiments, and design data visualizations

Working with multiple people and achieve consensus through compromises

Proposing and discussing specifications based on use cases

Work Experiences

May 2024 : Multimodal NLP Engineer (6-months internship) @ NVIDIA, Gent, BE

2024-2024: Visiting Researcher on Low-Resource Language Modeling @ Leuven.AI (KU Leuven)

2015-2019 : Technical Program Manager (Web Platform) @ Microsoft, Seattle/Redmond, USA

2014-2015 : Web Application Developer @ Group S, Brussels, Belgium

July 2013 : Web Platform Summer Intern (12 weeks) @ Adobe Systems, San Francisco, USA

2010-2012 : Summer jobs at iRail (*PHP REST/JSON Backend, mobile app design*), People Web Solution (*SVG+JS diagrams, Exchange .NET API*), and M-Team (*InfoPath forms under SharePoint Server*)

Web standards specification editor for W3C Working Groups, Invited Expert at W3C since 2012, CSSWG since 2018

Native French   Good Dutch

Fluent English

Projects and highlights

Visiting Researcher @ Leuven.AI (KU Leuven)

Developing large language models for smaller languages is critical for delivering the latest NLP advances to the entire world. This collaboration aimed at developing model conversion techniques and efficient training strategies & evaluation frameworks for low-resource languages, to make this vision possible!

COLM 2024: A novel language adaptation technique called Trans-Tokenization, with stellar results.
Trans-Tokenization can successfully convert English LLMs to LLMs for endangered languages such as Tatar, by using a SMT-based token alignment technique. Combined with our novel Hydra LLMs, it also enables the creation of SOTA translation pipeline for such languages, without using any high-quality parallel data.

Doctoral Student @ IDLab's Text$_2$Knowledge Team (Ghent University)

The amount of data in a hospital is so huge that fully exploiting it is only possible with the help of computers. The ADAM project aims to develop an AI pipeline to transform unstructured clinical data building on top of the current state of the art techniques, by mapping unstructured data found in EHR text documents into structured representations conforming to the most recent international standards.

EMNLP 2022 Findings: A novel contrastive pre-training strategy for semantic concept representations.
BioLORD produces highly semantic concept representations that match more closely the hierarchical structure of ontologies than alternatives, thanks to its grounding of concept representations through concept definitions, as well as short concept descriptions automatically derived from a multi-relational knowledge graph consisting of biomedical ontologies.

MWE 2023 Best Paper Award: Detecting idiomatic multi-word expressions in clinical terminology.
This paper shines a light on the potential of definition-based semantic models for detecting idiomatic and semi-idiomatic multiword expressions (MWEs) in clinical terminology. Our study focuses on biomedical entities defined in the UMLS ontology and aims to help prioritize the translation efforts of these entities.

Other publications: Pharmacovigilance, ADE on Social Media, Low-Resource NLP, etc. (Semantic Scholar)
Collaborations with researchers from Stanford (USA), Udine (Italy), Leuven (Belgium), Aarhus (Denmark)

Technical Project Manager @ Microsoft's Web Platform Team

Ensure the work done by the developers of the EdgeHTML's Layout team is as effective as possible at enabling new experiences and improving the usability of the web platform for developers and users alike. Help shape the experiences we want to enable visually-impaired users to live browsing the web.

NLP 4 Good: Prototyping an NLP-based web experience for the blind and visually-impaired.
Built as a side project a complete and innovative screen reader experience for the web, including a fuzzy search-or-scan functionality trained using Wikipedia's articles and NLP datasets. Also initiated, supervised and successfully passed a real-word user study for the prototype.

Management: Engaging the web standards and open web communities.
Interacted with 5k Edge interop issues, incl 2k that have been closed in FY18 (700 fixed). Co-edited a 14k-words specification on the behavior of CSS Tables. Helped drive ~150 issues in the w3c css working group in FY17 alone, Attended w3c conferences in Lyon, San Francisco, Lisbon, Seattle, Berlin, Sydney.

Programming: Edit-and-continue prototype for Microsoft Edge codebase.
Showed that by hot-loading assembly code in the memory of a running process using windbg, it would be possible in some cases to cut the "edit-build-test" loop of minor changes to seconds, instead of minutes.

Programming: DOM Timeline – Time-travel debugging for web pages.
Built a complete undo-redo stack for arbitrary operations on a webpage's DOM as well as trace information useful to debug the webpage. Integrated the library into another team's debugging project (with new UI).

Other projects: Multi-Browser cloud-hosted web crawler. Code refactoring tools. Test creation platform.

Student @ Université Libre de Bruxelles (Software Engineering Program)

Thesis: Design of a graphical language for biological systems modeling

Thanks for your time!