Academic Curriculum Vitae

Sandy Aoun

sandy.aoun@uni-graz.at - https://sandyaoun.github.io

Higher Education

Doctoral Student in Digital Humanities (Mar. 2024 - Present)

University of Graz, Faculty of Arts and Humanities, Graz, Austria

Research topic: Information extraction from late medieval charters

Doctoral Student in Computer Science (Dec. 2016 - Feb. 2017)

University of Rennes I, MATISSE Doctoral School, IRISA, Lannion, France

Research topic: Recording scripts optimization for the expressive reading (speech synthesis) of audiobooks

[Research] M.Sc. in Computer Science and Telecommunication (Oct. 2015 - Sept. 2016)

Lebanese University, Faculty of Sciences I, Hadath, Lebanon

University of Toulouse III, Faculty of Science and Engineering, IRIT, Toulouse, France

Double degree program, Track: IRDBM (Information Retrieval, Database, and Multimedia)

Master's thesis topic: Optimal constitution of the speech corpus for the speech synthesis of audiobooks

Completed with Honors - *Mention Assez Bien*, Rank: 1/10

First year of graduate studies in Computer Science (Oct. 2014 - June 2015)

Lebanese University, Faculty of Sciences II, Fanar, Lebanon

Completed with High Honors - *Mention Bien*, Rank: 3/19

B.Sc. in Computer Science (Oct. 2009 - June 2014)

Lebanese University, Faculty of Sciences II, Fanar, Lebanon

Graduated with Honors - *Mention Assez Bien*

Research Experience

Research Assistant (Aug. 2022 - Present)

Institution: University of Graz, Austrian Centre for Digital Humanities, Graz, Austria

Research topic: Information extraction from late medieval charters

Supervisor: Prof. Dr. Georg Vogeler

Disciplines: Natural Language Processing, Lexical and Relational Semantics, Machine Learning, Historical Linguistics

Research Assistant (Apr. 2022 - June 2022)

Institution: American University of Beirut, Faculty of Engineering and Architecture, Beirut, Lebanon

Research topic: Arabic graph-based named entity linking

Supervisor: Dr. Fadi A. Zaraket

Scientific disciplines: Natural Language Processing, Lexical and Relational Semantics, Knowledge Graphs

Research Assistant (Sept. 2019 - Nov. 2020)

Institution: American University of Beirut, Faculty of Arts and Sciences, Beirut, Lebanon

Research topic: Automatic speech recognition of Arabic speech using sequence-to-sequence models

Supervisor: Dr. Wassim El-Hajj

Scientific disciplines: Natural Language Processing, Machine Learning, Speech Processing

Research Internship of the Master's Program in Computer Science (Mar. 2016 - Sept. 2016)

Institution: University of Rennes I, ENSSAT, IRISA, Team EXPRESSION, Lannion, France

Research topic: Optimal constitution of the speech corpus for the speech synthesis of audiobooks

Supervisors: Dr. Jonathan Chevelu, Dr. Ali Choumane, and Dr. Damien Lolive

Scientific disciplines: Natural Language Processing, Combinatorial Optimization, Speech Processing

Engineering Experience

Natural Language Processing Engineer (June 2021 - Sept. 2021)

Institution: University of California at Berkeley, College of Letters & Science, United States

Topic: Construction of a biological database from 'Flora in the Eastern Mediterranean' reference books

Collaborator: Maryam Sedaghatpour (Ph.D. student in Integrative Biology)

Scientific disciplines: Natural Language Processing, Information Extraction, Data Cleansing

Final Project of the First Year of Graduate Studies in Computer Science (Mar. 2015 - May 2015)

Institution: Lebanese University, Faculty of Sciences II, Fanar, Lebanon

Topic: Conception and implementation of a JSON to XML compiler

Mentor: Prof. Dr. Kablan Barbar

Scientific disciplines: Formal Language Theory, Compiler Design, Compiler Construction

Teaching Experience

Computer Science Tutor (April 2021 - Sept. 2021)

I privately tutored programming and computer science-related concepts. Programming-related concepts: variables (built-in types and associated built-in functions); operators; basic programming syntax; basic manipulation of command line; control statements; looping/iterating; built-in functions; the "main" function; user-defined functions; external modules; specific modules. Computer science-related concepts: algorithm; pseudo-code; programming language theory; data structure; functional programming. Self-made appropriate programming examples and exercises are used to illustrate the concept(s) in question. [Programming language used: *Python*.]

Scholarly Output

Publications

International Conferences.

Sandy Aoun, Varvara Arzt, Daniel Luger, Georg Vogeler. "*Information Extraction from German Medieval Charters Abstracts*". 19th Annual International Conference of the Alliance of Digital Humanities Organizations (DH 2024), Washington, D.C., United States, August 6-10, 2024. (Accepted!)

Andreas Habring, Anguelos Nicolaou, Daniel Luger, Florian Atzenhofer-Baumgartner, Florian Lamminger, Franziska Decker, Sandy Aoun, Tamás Kovács, Georg Vogeler, Martin Holler. "*Probabilistic Modeling of Chronological Dates to Serve Machines and Scholars*". 18th Annual International Conference of the Alliance of Digital Humanities Organizations (DH 2023), Graz, Austria, July 10-14, 2023.

Tamás Kovács, Sandy Aoun, Georg Vogeler, Anguelos Nicolaou, Daniel Luger, Florian Atzenhofer-Baumgartner, Florian Lamminger, Franziska Decker. "*Few Shot Classification for Labeling of Medieval and Early Modern Charter Texts*". Poster. 18th Annual International Conference of the Alliance of Digital Humanities Organizations (DH 2023), Graz, Austria, July 10-14, 2023.

Daniel Luger, Anguelos Nicolaou, Franziska Decker, Florian Atzenhofer-Baumgartner, Florian Lamminger, Georg Vogeler, Sandy Aoun, Tamás Kovács. "*Digital contributions to a 300 years old methodology: Diplomatics & DH*". Poster. 18th Annual International Conference of the Alliance of Digital Humanities Organizations (DH 2023), Graz, Austria, July 10-14, 2023.

Georg Vogeler, Anguelos Nicolaou, Daniel Luger, Tamás Kovács, Florian Atzenhofer-Baumgartner, Sandy Aoun, Franziska Decker. "*Computational Methods in Studying Late Medieval Charters*". Poster. Third Conference on Computational Humanities Research (CHR 2022), Antwerp, Belgium, December 12-14, 2022.

Other Conferences.

Florian Atzenhofer-Baumgartner, Daniel Luger, Tamás Kovács, Johannes Laroche, Anguelos Nicolaou, Franziska Decker, Nicolas Renet, Sandy Aoun, Niklas Tscherne, Georg Vogeler. "*Formulaic Language in Diplomatics: Investigating Formulas as Charter Type Discriminators*". Conference on Formulaic Language in Historical Research and Data Extraction, Amsterdam, The Netherlands, February 7-9, 2024.

Georg Vogeler, Daniel Luger, Anguelos Nicolaou, Tamás Kovács, Florian Atzenhofer-Baumgartner, Florian Lamminger, Sandy Aoun, Franziska Decker. "*Building a virtual research environment to move from digital to distant Diplomatics (ERC project DiDip)*". Poster. 9. Tagung des Verbands Digital Humanities im deutschsprachigen Raum (DHd 2023), Belval, Luxembourg and Trier, Germany, March 13-17, 2023.

Other Scholarly Manuscripts

Research Proposal. Sandy Aoun. "*Automatic Speech Recognition of Arabic Speech Using Sequence-to-Sequence Models*". Submitted to the Grant Research Program which is jointly supported by the American University of Beirut (AUB) and the National Council for Scientific Research. 2019/2020 Academic Year. [I also prepared and performed a *30-minute Oral Presentation*.]

It is worth noting that such a proposal can only be submitted by faculty members (full-time AUB professors). The proposal is usually written by a professor(s) who competes against other AUB professors - who have already written and submitted their respective research proposals - for one of a few securable research fundings.

Master's Thesis. Sandy Aoun. "*Optimal Constitution of the Speech Corpus for the Speech Synthesis of Audiobooks*". Lebanese University and University of Toulouse III - M.Sc. Thesis in Computer Science. September 2016. Written in French. [Thesis defense: I also prepared and performed a *20-minute Oral Presentation*.]

Technical Report. Sandy Aoun. "*Conception and Implementation of a JSON to XML Compiler*". Lebanese University - Graduate Research Project in Computer Science. May 2015. Written in French. [Project defense: I also prepared and performed a *30-minute Oral Presentation*.]

Research Software

Constructing Bilad al-Sham Flora Database: I implemented software programs which transform unstructured factual text input into a valuable biological database. In essence, useful/specific information is extracted from encyclopedia-like PDF files covering flora in the Eastern Mediterranean. The extracted data is consecutively refined into a standardized database. 2021. [Programming language used: *Python*.]

Building End-to-End ASR Dataset: I carried out an experiment which addresses building datasets suitable for training end-to-end automatic speech recognition (ASR) systems of spoken Arabic dialects. Our proposed automatic dataset collection method consists of firstly crawling YouTube videos whose Arabic closed captions are provided by the channel owner (the most frequent words in Arabic tweets are used as search keywords); then secondly passing the videos and their associated captions through several filtering heuristics which ensure reaching a satisfactory outcome. I also developed a program which aims to assess the effectiveness of our approach by generating relevant statistics. 2020. [Technologies used: *Python, Bash, YouTube Search API, SoX*.]

Packaging MGB-2 Dataset: I implemented software programs which process the MGB-2 dataset [Ali+16] in order to ultimately convert it into a form readable by the pipeline of the high-performance speech recognition framework Wav2letter++ [Pra+19]. 2019. [Technologies used: *Python, SoX, Wav2letter++* [Pra+19], *Docker*.]

Optimal Constitution of TTS Speech Corpus: I carried out an experiment which aspires to optimize the process of constructing the speech corpus of unit selection text-to-speech (TTS) systems. In this context, I implemented a greedy algorithm (spitting strategy) to bring into view the trade-off between the amount of text to be recorded and the quality of obtained (synthesized) speech signals. The implementation is based on our formal theoretical analysis which essentially profits from concepts related to the following domains/sub-domains: Set Cover Problem; Approximation Algorithms; and Linear Algebra. 2016. [Technologies used: *Python*, *IRISA TTS System* [Ala+16], *ROOTS* [CLL14].]

Objective Evaluation of Speech Signals: I implemented a software which measures the objective distance between natural and synthesized speech signals. In our case, the objective distance corresponds to the normalized Dynamic Time Warping cost which is computed on the Euclidean Distance between the Mel-Generalized Cepstral sequences of the signals. 2016. [Technologies used: *Python, SPTK* (Speech Signal Processing Toolkit), *SoX* (Sound eXchange).]

JSON to XML Compiler: I implemented a compiler which translates a JSON-formatted document into an interchangeable XML-formatted document. The implementation is based on my theoretical analysis which amounted to firstly defining a formal grammar as well as formulating a lexical and syntactic analysis of the syntax of JSON, then subsequently devising a semantic analysis by coming up with a suitable attribute grammar. 2015. [Programming language used: *C++*.]

Professional Activities

Research Collaboration

From June 18, 2020 until September 24, 2020; **Dr. Ahmed Ali**, who is a principal engineer at the Arabic Language Technologies team of the **Qatar Computing Research Institute**, attended my virtual work meetings. He provided insightful feedback on the experiments I was conducting at the time.

Academic Service

Journal Reviewer.

ACM Transactions on Asian and Low-Resource Language Information Processing (2020)

Conference Reviewer.

4th Workshop on Open-Source Arabic Corpora and Processing Tools (of LREC 2020)

19th International Digital Humanities Conference (2024)

Community Service

I prepared and presented a **Talk** covering the "*Time, Clocks, and the Ordering of Events in a Distributed System*" paper [Lam78] to the **Papers We Love** community (Lebanon Chapter, May 2017).

Conference/Seminar Attendance

I attended the 2016 **Annual Seminar** of the **EXPRESSION Team** (University of South Brittany, June 2016).

I attended (and volunteered at) the **Digital Diplomatics Conference 2022** (University of Graz, September 2022).

I attended (and volunteered at) the **International Digital Humanities Conference 2023** (University of Graz, July 2023).

Professional Training

I completed two **Training Courses** while working at the American University of Beirut, namely the **Intersections: Supervisor Anti-Harassment & Title IX** (April 2020) and the **Diversity and Inclusion** (November 2020) training courses.

Skills

Natural Languages Proficiency

English: Full professional proficiency (IELTS Academic test: CEFR level C1)

Arabic: Native or bilingual proficiency

French: Professional working proficiency

Technical Skills

Programming languages: C, C++, Python, Java, Lisp, SQL

Specification and modeling languages: UML, Merise, Petri Nets, TLA+, LOTOS

Miscellany: HTML, CSS, LaTeX, Linux, LTL, Docker, HPC clusters, pandas, spaCy, scikit-learn

Analytical Skills

I possess superior **Researching**; **Critical Thinking**; **Logical Reasoning**; **Creative Thinking**; and **Communication** abilities.

Soft Skills

Dedication; **Perseverance**; **Cooperation**; **Intellectual Humility**; **Empathy**; **Cheerfulness**; and **Curiosity** are the traits which my former workmates used most when describing my personality.

Old-Time Work Experiences

Volunteer Experience

Arabic ↔ English Interpreter (February 2014)

Location: Syrian refugees' households, Dibbiyeh, Lebanon

Task: Conducting interviews (consecutive interpreting) with Syrian refugees about their living conditions

Miscellany

Children's Entertainer (December 2013)

Location: City Centre Beirut, Hazmiyeh, Lebanon

Tasks: Assisting kids in writing/coloring a letter to Santa Claus; gift wrapping; and storytelling

Product Promoter (July 2009 - Sept. 2009)

Location: Various supermarkets and shopping malls, Greater Beirut, Lebanon

Task: Marketing specific products to people browsing items in the supermarket or shopping mall

Background Actor (July 2006 - Sept. 2006 and July 2007 - Sept. 2007)

Location: Various studios and venues, Greater Beirut, Lebanon (such as StudioVision studios, Naccache, Lebanon)

Tasks: Playing the silent role assigned to me in a scene or following the orders given to the studio audience of a TV show

Leisure Time

Leisure Interests

History; **Philosophy**; **Finance**; and **Art** are the main extra interests that I am passionate to learn more about.

Leisure Activities

I enjoy spending my leisure time **Reading** non-fiction content; watching **Documentaries**; following **Courses** on e-learning platforms; checking out some **International News** coverage; **Debating** with a diverse group of people; and honing my **Cooking** skills.

Advocacy

Take Back Parliament (Lebanon) member: Take Back Parliament was an independent atypical political movement serving as an alternative to the ruling homogeneous populist sectarian Lebanese political parties. The purpose of the newly established Lebanese reform movement was to campaign for a progressive political ideology and agenda, thus to promote first and foremost secular; social justice; anti-corruption; nature conservation; sexual diversity; and feminist values. [I was actively involved in the work of the team *from December 2012 until March 2013*.]

Free Hugs Campaign (Lebanon) member: The Free Hugs Campaign was a newly established social movement in Lebanon. The flashmobs' sole goal is to present strangers with the possibility of getting a hug; i.e., the only intention is to make people feel better. [I used to take part in the free hugs flashmobs *throughout the 2012 and 2013 years*.]

References

References and recommendation letters are available upon request.

Bibliography

[Ala+16] Pierre Alain et al. "The IRISA text-to-speech system for the Blizzard Challenge 2016". In: *Blizzard Challenge 2016 workshop*. 2016.

[Ali+16] Ahmed Ali et al. "The MGB-2 challenge: Arabic multi-dialect broadcast media recognition". In: *IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 2016, pp. 279–284.

[CLL14] Jonathan Chevelu, Gwénolé Lecorvé, and Damien Lolive. "ROOTS: a toolkit for easy, fast and consistent processing of large sequential annotated data collections." In: *International Conference on Language Resources and Evaluation (LREC)*. 2014, pp. 619–626.

[Lam78] Leslie Lamport. "Time, Clocks, and the Ordering of Events in a Distributed System". In: *Commun. ACM* 21.7 (1978), pp. 558–565.

[Pra+19] Vineel Pratap et al. "Wav2letter++: A fast open-source speech recognition system". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 6460–6464.

Last revised: April 2024