Emmanuel Roche

Ph.D., U.S. Citizen

Education

Ph.D. in Computer Science from Universite Paris VII (1993)
MS in Mathematics from Ecole Polytechnique (1989)

Professional Activities

2014 – Present. CEO and Chief Scientist at Clover.AI. Clover.AI was founded to explore new types of NLP algorithm. Provides SOTA solutions for low-level NLP tools (tokenization, pos-tagging, entity extraction) as well as custom solutions.

2014 – Present. Chief Scientist and CEO of Gammakite, Inc. Gammakite is a language learning company focusing on "hard" languages (such as Chinese for English speakers). Uses NLP/AI to analyze text advances in pedagogy. Gammakite program have been used at many universities in the US such as MIT, Tufts University, Wellesley College.

2010-2014. Investor and Co-Founder and CEO, LavaRipples LLC . Chookka was a product of LavaRipples LLC, a simple yet powerful asynchronous and synchronous conversation tool that helps people communicate smarter.

2008 – 2010. Executive Director Teragram a Division of SAS Institute. Responsible for the R&D of Language Processing.

1997 – 2008. Executive Director and Co-Founder of Teragram Corporation, Cambridge MA. Responsible for Research and Development. Co-founder with Yves Schabes and 50% owner of Teragram Corporation. Teragram was the market leader in multilingual natural language processing until its acquisition by SAS Institute in 2008. We developed and licensed natural language tools in more than 30 languages with an emphasis on scalability and robustness to clients such as Google, Microsoft, Yahoo!, CNN, The New York Times, the World Bank, and Wolters Kluwer, among many others, and grew the company from two to fifty employees through internal growth only.

1997 – 2008. PicoSearch co-founder. PicoSearch was the first and market leader in advanced hosted search services for quickly, easily, and cost-effectively integrating search capabilities into any web site.

1992 – 1997. Research Scientist, Mitsubishi Electric Research Labs (MERL). Cambridge MA. Research in natural language processing and text analytics. Among other projects, worked on finite-state machine applied to text processing, bi-directional finite-state machines, high-speed pattern matching, and grammar checking technologies.

Coding

I love coding and never want to stop (probably >3M lines). I like to design fast and compact algorithms and follow-up with clean, robust, fast, and small implementations. My favorite languages are, in order of preference: C (I want to control memory allocation), Python (of course), Swift (nice design), C++, Java, Perl, … Also, many of my own languages since I like to write compilers.

Some Coding Achievements (in C):

- Designed and implemented first spelling correction for Search Engine for Altavista
- Designed and implemented the first "Did you mean…" feature on Google.
- Designed and implemented the first New York Times search engine.
- Co-Designed and Implemented the fastest POS tagger.
- Designed and implemented more that 200 novel algorithms related to NLP and Finite-State Computing

Scientific Research

I have investigated numerous formal aspects of rational functions, rational relations, finite-state transducers, bimachines, and subsequential functions on various semirings while applying those formal results to natural language processing. More recently, I have investigated how symbolic processing can be re-incorporated with recent language models (transformers,). I have been a program committee member of many international scientific conferences and the author of more than thirty international scientific publications.

Languages

English and French (native/bilingual), Chinese (fluent)

Selected Publications

Recent Technical Reports

Emmanuel Roche. *2023. Finite-State Representation of Geometric Transductions. DRAFT. (https://github.com/eroche/eroche.github.io/blob/main/geo_trans/geo_trans1/DRAFT_20230608_geo_trans1.pdf)*

Edited Book

Finite-State Language Processing, Roche and Schabes (editors). MIT Press. June 1997.

Book Chapters

Emmanuel Roche. *Finite state transducers: parsing free and frozen sentences.* In Andras Kornai, editors, Extended finite state models of language. Cambridge University Press, NY, November 1999.

Emmanuel Roche. *Parsing with Finite-State Transducers.* In Roche and Schabes, editors, Extended finite state models of language. MIT Press, June 1997

Emmanuel Roche and Yves Schabes. *Deterministic Part-of-Speech Tagging with Finite-State Transducers.* In Roche and Schabes, editors, Extended finite state models of language. MIT Press, June 1997

Others

Roche, Emmanuel, 1997. *Compact Factorization of Finite-State Transducers and Finite-State Automata*. Nordic Journal of Computer Science, Helsinki.

Roche, Emmanuel and Yves Schabes, 1995. *Deterministic Part-of-Speech tagging with Finite-State Transducers*. Computational Linguistics, 21(2):227-254.

Roche, Emmanuel, 1994. *Analyse syntaxique du Français avec transducteurs*. Linguisticae Investigationes XVIII:1. John Benjamins B. V., Amsterdam.

Roche, Emmanuel, 1993. *Une représentation par automates finis des textes et des propriét és transformationnelles des verbes*. Linguisticae Investigationes XVII:1. 189-222. John Benjamins B. V., Amsterdam.

Roche, Emmanuel, 1992. *Relations rationnelles et dictionnaires de mots composés*. Linguisticae Investigationes XV:1. John Benjamins B. V., Amsterdam.

Patents

11,243,651. 2022. E. Roche. *Guided operation of a language device based on constructed, time-dependent data structures.*

10,936,825. 2021. E. Roche. *Methods and apparatus to improve disambiguation and interpretation in automated text analysis using transducers applied on a structured language space*.

10,885,809. 2021. E. Roche. *Device for language teaching with time dependent data memory*

8,091,067. 2012. E. Roche and Y. Schabes. *Method and system for hosting a programming environment.*

7,873,657. 2011. E. Roche and Y. Schabes. *Method and system for processing, by an information retrieval system, user input modifying the information retrieval system.*

7,853,874. 2010. Y. Schabes and E. Roche. *Spelling and grammar checking system.*

7,809,724. 2010. E. Roche and Y. Schabes. *Method and system for responding to user-input based on semantic evaluations of user-provided expressions.*

7,805,444. 2010. E. Roche and Y. Schabes. *Method and system for responding to user-input based on semantic evaluations of user-provided resources.*

7,243,305. 2007. Y. Schabes and E. Roche. *Spelling and grammar checking system.*

7,120,627. 2006. Method for detecting and fulfilling an information need corresponding to simple queries

7,047,246. 2006. Search and index hosting system

6,859,800. 2005. System for fulfilling an information need

6,651,065. 2003. Search and index hosting system

6,535,842. 2003. Automatic bilingual translation memory system

6,424,983. 2002. Spelling and grammar checking system

6,336,116. 2002. Search and index hosting system

5,845,306. 1998. Context based system for accessing dictionary entries

5,799,269. 1998. System for correcting grammar based on parts of speech probability

5,610,812. 1997. Contextual tagger utilizing deterministic finite state transducer

5,537,317. 1996. System for correcting grammar based parts on speech probability

5,535,121. 1996. System for correcting auxiliary verb sequences

5,521,816. 1996. Word inflection correction system

5,485,372. 1996. System for underlying spelling recovery

5,477,448. 1995. System for correcting improper determiners