

# Michele Cafagna

MACHINE LEARNING RESEARCH SCIENTIST

Via Montello, 6 Pisa (PI), Italy

✉ m.cafagna26@gmail.com | 🌐 www.linkedin.com/in/michele-cafagna

## Education

### Università di Pisa

Pisa, Italy

MSC. IN COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE, 110/110

2017 - 2019

- Master Thesis titled Headline Generation and Analysis of Writing Styles in Journalism

### Università degli Studi di Bari

Bari, Italy

B.S. IN COMPUTER SCIENCE AND SOFTWARE PRODUCTION TECHNOLOGIES, FULL MARKS

2014 - 2017

- Bachelor Thesis titled Software system based on pattern recognition to features extraction of leukocyte cells

## Skills

### Professional

Machine Learning, Data Analysis, Computer Vision, Image Processing, Natural Language Processing, Natural Language Generation, Natural Language Understanding, Research

### Programming

Python, C/C++, JAVA, Matlab, LaTeX

### Software

Flask with Python, Docker, Git, Linux

### Languages

Italian, English

## Experience

### Aptus.AI

Pisa, Italy

MACHINE LEARNING RESEARCH SCIENTIST & SOFTWARE ENGINEER

Jul. 2019 - Present

- Research, development and deployment of Deep learning multimodal models
- Implement and design microservices architectures for AI-based applications
- Work closely to other team members to plan, design, and develop robust solutions, in an Agile workflow

### Research, Center For Language and Cognition (CLCG), University of Groningen

Groningen, The Netherlands

VISITING RESEARCHER, WORKED ON GENERATIVE LANGUAGE MODELS AND ANALYSIS OF WRITING STYLES IN

Apr. 2019 - Jul. 2019

JOURNALISM UNDER THE SUPERVISION OF PROF. MALVINA NISSIM

- Collected and analysed hundred of thousands of news scraped from Italian newspapers
- Experimented with various generative models for headline generation task, assessed by means of human evaluation
- Developed an automatic methodology based on classification, to evaluate the stylistic features learnt by a generative language model
- Devised a method to evaluate the styles of two corpora observing word shifts in the embedding space

## Publications

### Embeddings shifts as proxies for different word use in Italian newspapers

CLiC-it, Bari, Italy

MICHELE CAFAGNA, LORENZO DE MATTEI, MALVINA NISSIM

2019

- We study how words are used differently in two Italian newspapers at opposite ends of the political spectrum by training embeddings on one newspaper's corpus, updating the weights on the second one, and observing vector shifts. We run two types of analysis, one top-down, based on a preselection of frequent words in both newspapers, and one bottom-up, on the basis of a combination of the observed shifts and relative and absolute frequency. The analysis is specific to this data, but the method can serve as a blueprint for similar studies.

### Suitable doesn't mean attractive. Human-based evaluation of automatically generated headlines

CLiC-it, Bari, Italy

MICHELE CAFAGNA, LORENZO DE MATTEI, DAVIDE BACCIU AND MALVINA NISSIM

2019

- We train three different models to generate newspaper headlines from a portion of the corresponding article. The articles are obtained from two mainstream Italian newspapers. In order to assess the models' performance, we set up a human-based evaluation where 30 different native speakers expressed their judgment over a variety of aspects. The outcome shows that (i) pointer networks perform better than standard sequence to sequence models, creating mostly correct and appropriate titles; (ii) the suitability of a headline to its article for pointer networks is on par or better than the gold headline; (iii) gold headlines are still by far more inviting than generated headlines to read the whole article, highlighting the contrast between human creativity and content appropriateness.

## Invisible to People but not to Machines: Evaluation of Style-aware Headline Generation in Absence of Reliable Human Judgment (accepted)

LREC, Marseilles, France

LORENZO DE MATTEI, MICHELE CAFAGNA, FELICE DELL'ORLETTA AND MALVINA NISSIM

2020

- We automatically generate headlines that are expected to comply with the specific styles of two different Italian newspapers. Through a data alignment strategy and different training/testing settings, we aim at decoupling content from style and preserve the latter in generation. In order to evaluate the generated headlines' quality in terms of their specific newspaper-compliance, we devise a fine-grained evaluation strategy based on automatic classification. We observe that our models do indeed learn newspaper-specific style. Importantly, we also observe that humans aren't reliable judges for this task, since although familiar with the newspapers, they are not able to discern their specific styles even in the original human-written headlines. The utility of automatic evaluation goes therefore beyond saving the costs and hurdles of manual annotation, and deserves particular care in its design.

## Norm It! Lexical Normalization for Italian and Its Downstream Effects for Syntactic Analysis

LREC, Marseilles, France

ROB VAN DER GOOT, ALAN RAMPONI, TOMMASO CASELLI, MICHELE CAFAGNA, LORENZO DE MATTEI

2020

- Lexical normalization is the task of translating non-standard social media data to a standard form. Previous work has shown that this is beneficial for many downstream tasks in multiple languages. However, for Italian, there is no benchmark available for lexical normalization, despite the presence of many benchmarks for other tasks involving social media data. In this paper, we discuss the creation of a lexical normalization dataset for Italian. After two rounds of annotation, a Cohen's kappa score of 78.64 is obtained. During this process, we also analyze the inter-annotator agreement for this task, which is only rarely done on datasets for lexical normalization, and when it is reported, the analysis is usually very shallow. Furthermore, we utilize this dataset to train a lexical normalization model and show that it can be used to improve dependency parsing of social media data. All annotated data and the code to reproduce the results are available at:

## GePpeTto Carves Italian into a Language Model

arXiv

LORENZO DE MATTEI, MICHELE CAFAGNA, FELICE DELL'ORLETTA, MALVINA NISSIM, MARCO GUERINI

2020

- We develop GePpeTto, the first generative language model for Italian, built using the GPT-2 architecture. We provide a thorough analysis of GePpeTto's quality by means of both an automatic and a human-based evaluation. The automatic assessment consists in (i) calculating perplexity across different genres and (ii) a profiling analysis over GePpeTto's writing characteristics. We find that GePpeTto's production is a sort of *bonsai* version of human production, with shorter but yet complex sentences. Human evaluation is performed over a sentence completion task, where GePpeTto's output is judged as natural more often than not, and much closer to the original human texts than to a simpler language model which we take as baseline.

## Honors & Awards

### NATIONAL

2017 **Finalist**, e-Health Award. Research Competition or project on Digital Health, AiSDet Associazione Italiana di Sanità Digitale e Telemedicina

Bari, Italy

2019 **Best Paper Award Finalist with Special Mention**, Clic-it 2019 Sixth Italian Conference on Computational Linguistics

Bari, Italy

## Presentations

### Clic-it 2019 Sixth Italian Conference on Computational Linguistics

Bari, Italy

ORAL PRESENTER

Nov. 2019

- Embeddings shifts as proxies for different word use in Italian newspapers. Michele Cafagna, Lorenzo De Mattei, Malvina Nissim.