Into Language Technologies

Alberto Barrón-Cedeño
Alma Mater Studiorum - Università di Bologna
a.barron@unibo.it
@_albarron_

Session 1
19 December 2022

Into Language Technologies I

Alberto Barrón-Cedeño
Alma Mater Studiorum - Università di Bologna
a.barron@unibo.it
@_albarron_

Session 1
19 December 2022

What is this about?

- Language technologies in the job of a translator are a must
- [...] rely on machine translation to support their regular endeavours
- Technology [...] based on AI, NLP, ML
- a translator that is not only a user, but could eventually take part in the development of the technology

Rough programme

- An introduction to language technologies
- A path to jump from translation into language technologies
- Understanding IBM M1, the core of statistical machine translation
- A quick overview of the python programming language

Department of Interpreting and Translation

- Born in 2012 (merging the SITLeC Dept. and the *Scuola Superiore di Lingue Moderne per Interpreti e Traduttori*)
- Emphasis in applied research, theoretical, practical, and didactic aspects of translation and interpreting

Degrees

- Bachelor in Intercultural and Linguistic Mediation
- Masters in Interpreting
- Masters in Specialized Translation
- PhD in Translation, Interpreting, Interculturality

Why Language Technologies?

Overview

1. Why Language Technologies?
2. How CS is being *plugged* into DIT
3. Teaching initiatives
4. Three student projects

Language data and project specialist

- knowledge, skills and competences -

DISCIPLINARY
- Knowledge of specific languages
- Ability to conduct linguistic analysis
- Translation, interpreting, post-editing, localisation

(INTER)CULTURAL
- Awareness of specific cultural contexts and cultural differences
- Cultural agility

TRANSVERSAL
- Creative and innovative thinking
- **Problem-solving** skills
- Presentation skills
- Writing for different audiences

TECHNICAL
- Understanding and use of **language technologies** and resources
- Knowledge of a **programming language**
- Understanding of **computational linguistics / NLP**

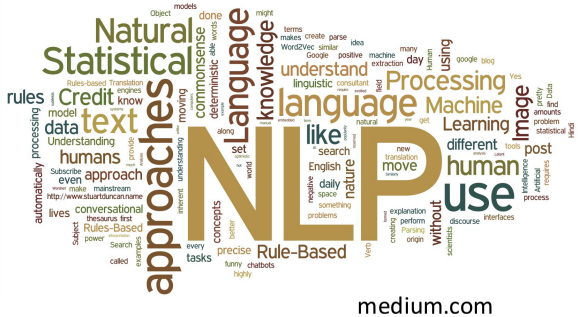
DATA-ORIENTED
- Ability to collect, manage, curate, analyse language data
- Knowledge of **statistics**
- Familiarity with data standards

RESEARCH-ORIENTED
- Knowledge of research design
- Analytical, logical and hypothetical thinking
- Accessing and **processing information** critically

ORGANISATIONAL
- Entrepreneurship
- Project management
- Quality control
- Planning
- Teamwork

Adapted from Freepik Storyset, https://storyset.com/data

Why Language Technologies?



medium.com

How CS has been

(re)*plugged* into DIT

Hiring of computing scientists

Spring 2019
Professor with NLP background

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation

Winter 2020
Research assistant with MT background

Microsoft Research

* but not only!

LT at DIT is (mostly) built on UPSKILLS

***UP**grading the*
***SKI**lls of*
***L**inguistics and*
***L**anguage*
***S**tudents*

An Erasmus+ strategic partnership September 2020 - August 2023

SKILLS

https://upskillsproject.eu

LT: Computational Thinking

1. Into computers and programming
2. Regular expressions
3. Basic programming (python)
4. Jupyter notebooks (programming *platform*)
5. Methods and classes (OOP)
6. Dealing with text

Teaching initiatives

Into Alma AI

DIT adhered to UniBO's Alma Mater Research Institute for Human Centered Artificial Intelligence

Foundations of AI	AI for health and well-being
AI and hard sciences	AI for law and governance
Humanistic AI	AI and education
AI for industry	AI and high performance computing

DIT Python course 2020, 2021, 2022*

Official objective 1. Giving a gentle introduction to programming in python to get students in the right position to go further on their own

Official objective 2. Paving the way to the NLP and other courses

Unofficial objective. Never again listen students say "that would be so awesome! But stop… it needs programming"

* 2020-2021: crash course, open to all; 2022: part of Profession-based research and PhD seminars

Initiatives with CS load

- Neural machine translation
- MT of academic websites
- Interaction with local companies
- Webinars and workshops on MT and related technologies
- Automatic identification of misconducts online (antenna)
- Automatic identification of MT suitability (MTsweet)
- Discussions on the curriculum in 5 years time

- Conveying that MT is not an enemy

DIT Python course 2020

Session 1. The basics

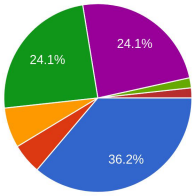
- What is a programming language?
- What is an algorithm
- ”Translating” from an algorithm into a program
- The characteristics of the python programming language
- Basic functions, variables, conditionals, loops

```
# my code
x = 0
while x < 50:
 for i in range(x):
   print('x', end="")
 print()
 x += 1
```


DIT Python course 2020

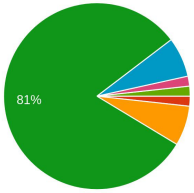
Pre-entry survey

A translator should know how to code



- Absolutely
- Yes, if (s)he wants to do research (not for industry)
- Yes, if (s)he wants to go for industry (n…
- Yes, if (s)he is targetting the software/…
- Maybe
- Nope. There is out-of-the-box softwar…
- Nope. Translation does not involve sof…
- knowing the theory when localizing so…
- Yes, for both industry and research pu…

Familiarity with Python

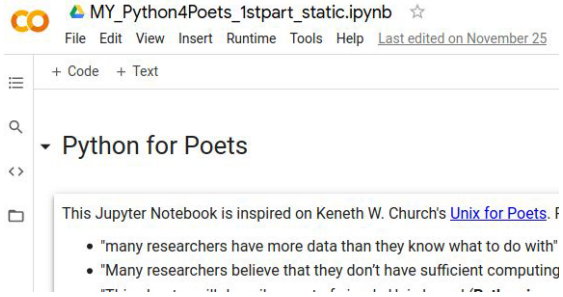


- I use it on a regular basis
- I can code a few routines, with a lot of effort
- I have passive knowledge (I can read it, but I cannot produce it)
- I've heard about it, but I don't know it…
- I hate it
- Isn't it a snake?
- last year I attended an online course o…
- So far I've only coded a few scripts for…

DIT Python course 2020

Session 2. Python 4 Poets (1/2)
(derived from K. Church’s Unix for poets)

- Opening text files
- Splitting into words (tokenisation)
- Obtaining vocabularies (types)
- Extracting *n*-grams



DIT Python course (PBR)

Course structure

Four+ sessions

1. Presentation of concepts with the support of slides
2. Live on-screen coding of task-specific routines
3. Take-home simple coding exercises

Coding *platform*

Jupyter notebooks on Google’s colab

http://colab.research.google.com

NLP course

Learning outcomes. [...] basic theoretical aspects of computational linguistics [...] acquire practical skills [all the way to] supervised models (machine learning)

Specialised translation: optative

Translation and technology: core

DIT Python course 2020

Session 3. Python 4 Poets (2/2)
(derived from K. Church's Unix for poets)

- Finding specific tokens/strings
- Finding palindromes
- String substitutions
- Functions
- Collocations

8. Mutual information to find collocations

From the Wikipedia articles on mutual information and collocations

In probability theory and information theory, the mutual information (MI) of two rar
between the two variables. More specifically, it quantifies the "amount of informat
obtained about one random variable through observing the other random variable.

Mutual information of words is often used as a **significance function for the comp**

A collocation is a series of words or terms that co-occur **more often than would be**

$$MI(x, y) = log_2 \frac{Pr(x,y)}{Pr(x)Pr(y)}$$

and, following Magerman and Marcus, in NLP it can be estimated as

$$MI(x, y) \approx log \frac{\frac{f(x,y)}{\sum_{(i,j)\in C} f(i,j)}}{\frac{f(x)}{\sum_{i\in C} f(i)} \frac{fy}{\sum_{i\in C} f(i)}}$$

where $\sum_{\cdot}$ is the sum over all instances of ·

```
[ ] from math import log

    bigrams = ngrams(tokens, 2)
    unigrams = ngrams(tokens, 1)
```


NLP course

Course structure

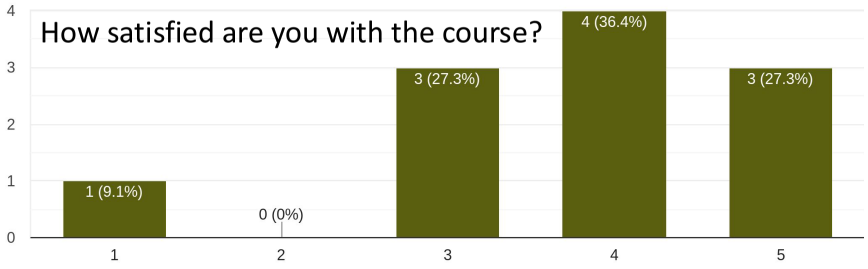
1. Presentation of concepts with the support of slides
2. Live on-screen coding with simple running routines + voluntary homework
3. Evaluation based on one final project (with potential to become a publication)

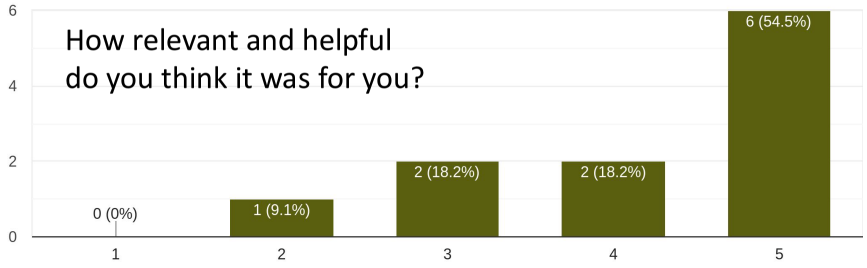
Coding *platforms*

jupyter notebooks + PyCharm + BASH

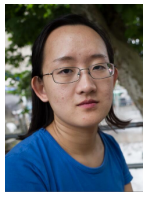
DIT Python course 2020

Closing perception





AriEmozione



Shibingfeng Zhang

Francesco Fernicola

Specialized Translation Masters

Objective. Identifying the emotion transmitted in 17th/18th-century Italian opera arias at the verse level

Developed in the context of UniBO's Centro per l'Interazione con le Industrie Culturali e Creative (https://site.unibo.it/cricc/it)

NLP course

Rough contents: coding, statistics, and machine learning applied to text

Building blocks…

1. Introduction to natural language processing
2. Tokens and the vector space model (algebra)
3. The Naïve Bayes classifier (statistics)
4. The training and evaluation process (machine learning)
5. Word vectors (algebra)
6. Latent semantic analysis (algebra)
7. Neural networks (machine learning)
8. Word embeddings (semantics)
9. Convolutional neural networks (machine learning)
10. Recurrent neural networks (machine learning)
11. Text generation

AriEmozione

AriEmozione 1.0 corpus

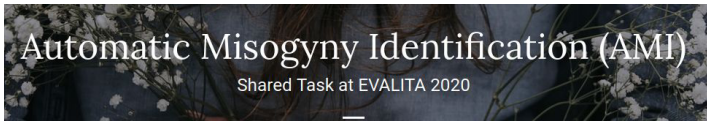
- 678 operas composed between 1655 and 1765
- All texts are written in Italian of the period and articulated in verses
- 2,473 verses manually annotated in six classes
  - Amore (Love)
  - Gioia (Joy)
  - Ammirazione (Admiration)
  - Rabbia (Anger)
  - Tristezza (Sadness)
  - Paura (Fear)

Three student projects

UniBO @ AMI 2020

Arianna Muti

Language, Society, and Communication Masters



Objective. Identifying whether a tweet is misogynous and, if it is,whether it is aggressive

AriEmozione

model representation	10-fold CV F_1	 Acc	test F_1	 Acc
***k*NN**				
char 3-grams	0.38	38.51	0.35	35.15
words	0.36	36.08	0.35	34.73
LDA char	0.30	29.97	0.31	30.54
SVM–RBF				
char 3-grams	0.44	43.70	0.43	43.00
words	0.42	42.00	0.44	44.00
LDA char	0.28	28.00	0.30	30.00
Log reg				
char 3-grams	0.44	**45.57**	0.42	43.10
words	0.41	43.20	0.41	43.10
LDA char	0.28	30.63	0.29	30.96
2-layers NN				
char 3-grams	0.42	43.61	**0.47**	**46.86**
words	0.42	42.91	0.43	43.10
LDA char	0.27	29.56	0.27	31.80
3-layers NN				
char 3-grams	**0.49**	41.86	0.40	41.84
words	0.47	42.60	0.40	41.84
LDA char	0.26	31.41	0.30	31.80
FastText				
char 3-grams	0.43	45.00	0.41	42.37
pre-trained chars	0.43	47.00	0.41	41.00
words	0.42	42.56	0.39	44.07
pre-trained words	0.38	41.00	0.40	42.00

Results: far from striking

Main drawbacks: short amount of data in archaic language

Enhancements: Creation of word embedding space and predictions at the verse level with deep learning

UniBO @ AMI 2020

Two classification tasks

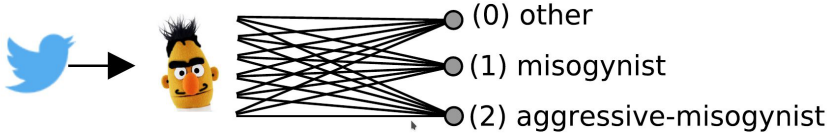
1. Is this tweet misogynous?

   YES/NO

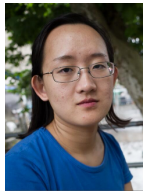
2. Is this misogynous tweet aggressive?

   YES/NO

Our solution: one multi-class network built on top of AlBERTo (an Italian version of BERT)



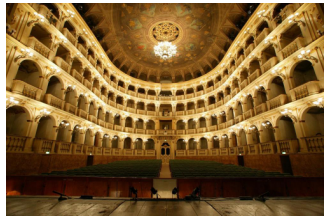
AriEmozione



Shibingfeng Zhang*

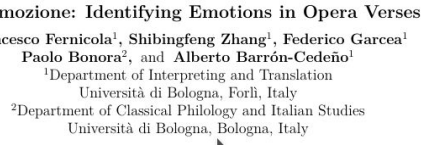
Francesco Fernicola+

Specialized Translation Masters



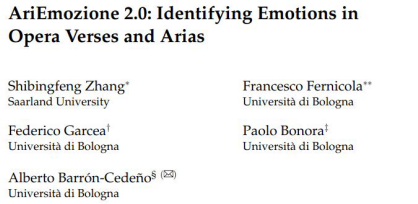
* NLP masters, Saarland
\+ PhD, UniBO

Papers published in CLIC-it 2020 and IJCOL



AriEmozione: Identifying Emotions in Opera Verses

Francesco Fernicola[1], Shibingfeng Zhang[1], Federico Garcea[1] Paolo Bonora[2], and Alberto Barrón-Cedeño[1]
[1]Department of Interpreting and Translation
Università di Bologna, Forlì, Italy
[2]Department of Classical Philology and Italian Studies
Università di Bologna, Bologna, Italy



AriEmozione 2.0: Identifying Emotions in Opera Verses and Arias

Shibingfeng Zhang*
Saarland University

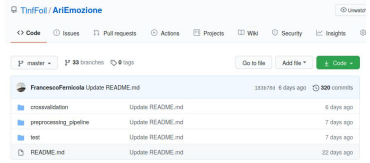
Francesco Fernicola**
Università di Bologna

Federico Garcea†
Università di Bologna

Paolo Bonora‡
Università di Bologna

Alberto Barrón-Cedeño§
Università di Bologna

Code available on github



Are fictional voices *different*?

Ettore Galletti

Specialized Translation Masters

Objective. Finding out if the authors of a specific play managed to create recognisable fixional voices

UniBO @ AMI 2020

team	run	constrained	score
UniBO^{a}	2	yes	**0.7438**
jigsaw	2	no	0.7406
jigsaw	1	no	0.7380
fabsam	1	yes	0.7343
YNU_OXZ	1	no	0.7314
fabsam	2	yes	0.7309
NoPlaceForHateSpeech	2	yes	0.7167
YNU_OXZ	2	no	0.7015
fabsam	3	yes	0.6948
NoPlaceForHateSpeech	1	yes	0.6934
AMI_the_winner	2	yes	0.6869
MDD	3	no	0.6844
PoliTeam	3	yes	0.6835
MDD	1	yes	0.6820
PoliTeam	1	yes	0.6810
MDD	2	no	0.6679
AMI_the_winner	1	yes	0.6653
PoliTeam	2	yes	0.6473
UniBO^{b}	1	yes	0.6343
AMI_the_winner	3	yes	0.6259
NoPlaceForHateSpeech	3	yes	0.4902

Misogyny: 0.8102

Aggressiveness: 0.6774

Results: top-performing model

Main drawbacks: still weak against aggressiveness

Further efforts: multimodal (memes) and multilingual models; implicit hate speech

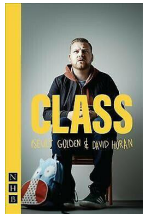
Are fictional voices *different*?

Identifying characters vs **Identifying groups of characters**

- McCafferty (teacher)
- Brian (father)
- Donna (mother)
- Jayden (son)
- Kaylie (Jayden's schoolmate)
- None (stage directions)

- Male
- Female
- None

- Adults
- Kids
- None



UniBO @ AMI 2020 Task A

Arianna Muti*

Language, Society, and Communication Masters

* PhD, UniBO

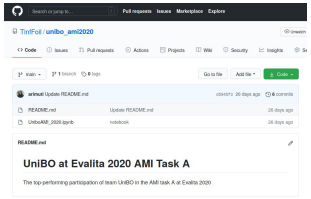
Papers at Evalita 2020, SemEval 2022, and more

UniBO@AMI: A Multi-Class Approach to Misogyny and Aggressiveness Identification on Twitter Posts Using AlBERTo

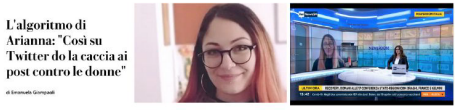
Arianna Muti
Department of Modern Languages, Literatures and Cultures - LILEC
Università di Bologna
Bologna, Italy
arianna.muti@studio.unibo.it

Alberto Barrón-Cedeño
DIT – Università di Bologna
Forlì, Italy
a.barron@unibo.it

Code available on github



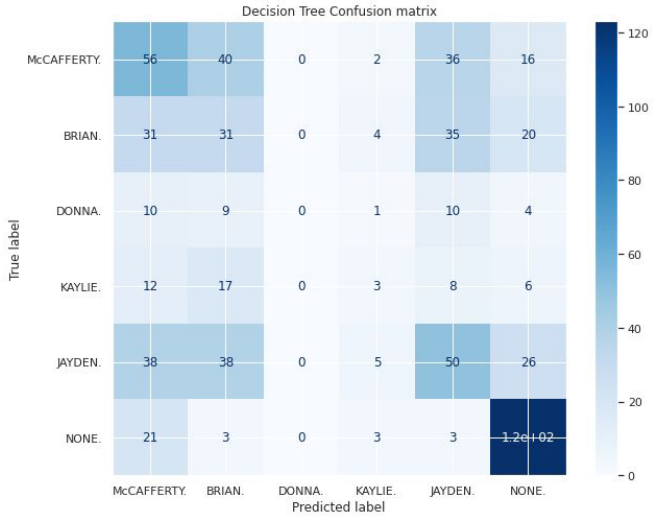
National media coverage



La Repubblica (Bologna); 31 March, 2021 RAI News 24; 8 April, 2021

Are fictional voices *different*?

Supervised approach. Build a multi-class character classifier and study whether it manages to label the interventions accurately

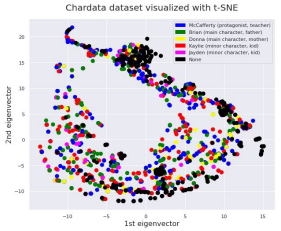


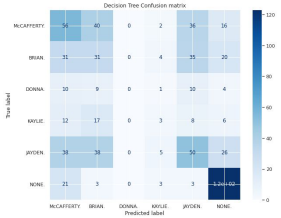
Are fictional voices *different*?

Core idea

1. Build a topic-independent representation of every character intervention
2. Observe if the representations of all interventions make the characters (clearly) differentiable

Are fictional voices *different*?





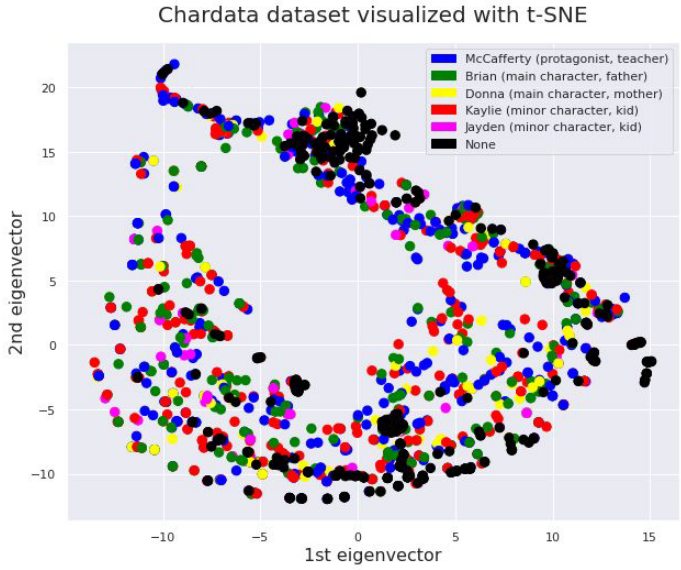
Results: non-conclusive

Main drawbacks: we observe some hints, but we need to study the problem further

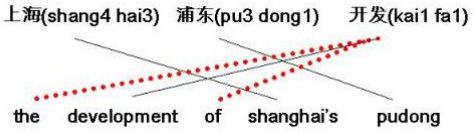
Italian translation: the characters are (a bit) more discriminative

Are fictional voices *different*?

Unsupervised approach. Cluster all the instances and analyse at what extent the clusters correlate with the characters



Next…



IBM M1

Python

Are fictional voices *different*?

Ettore Galletti*

Specialized
Translation
Masters

* Ufficio Rettorato, UniBO

Interested, questions?

GET IN TOUCH

a.barron@unibo.it

@_albarron_

Thanks!

Other (non-student-led) efforts

PR PPY

Automatic identification of propaganda in text

CheckThat!

Automatic prioritisation and verification of claims