A Recipe for Low-Resource NMT

Eryk Wdowiak
eryk@wdowiak.me
Arba Sicula

computing conference 2022

14-15 July 2022

Goal, Challenge and Solution

Our Goal

- to develop a neural machine translator for the Sicilian language

The Challenge

- low-resources – only 17,000 Sicilian-English sentence pairs

Our Solution

- model the language – incorporate theory into the modeling process

Low-Resource NMT

Previous literature suggests:

- Back-translation (Sennrich, Haddow and Birch, 2015)
- Multilingual trans. (Johnson et al., 2016) with "bridging" (Fan et al., 2020)
- Avoid overfitting by training:
  - a self-attentional Transformer model (Vaswani et al., 2017)
  - a smaller network with fewer layers (Sennrich and Zhang, 2019)
  - with small subword vocabularies (Sennrich, Haddow and Birch, 2016)
  - with high-dropout parameters (Srivastava et al., 2014)

Our innnovation – model the language

- incorporate theory into the dataset, tokenization and subword splitting
- train the model to learn like a human learns (ex.: by conjugating verbs)

Sicilian Translator

https://translate.napizia.com

What is the Sicilian Language?

- The Sicilian School of Poets at the imperial court of Frederick II:
  - created the first literary standard in Italy (13th century)
  - inspired Dante, the "father of the Italian language"
- Sicilian emerged as a literary language before Italian.
- The people of Sicily, Calabria and Puglia speak it everyday.
  - They speak Italian at work.
  - But at home – with family and friends – they speak Sicilian.
  - More precisely, their own dialect of the language.
- And Sicilian is a language spoken in Brooklyn, NY (where I live)
  - Mass migration – millions of people around the world speak Sicilian

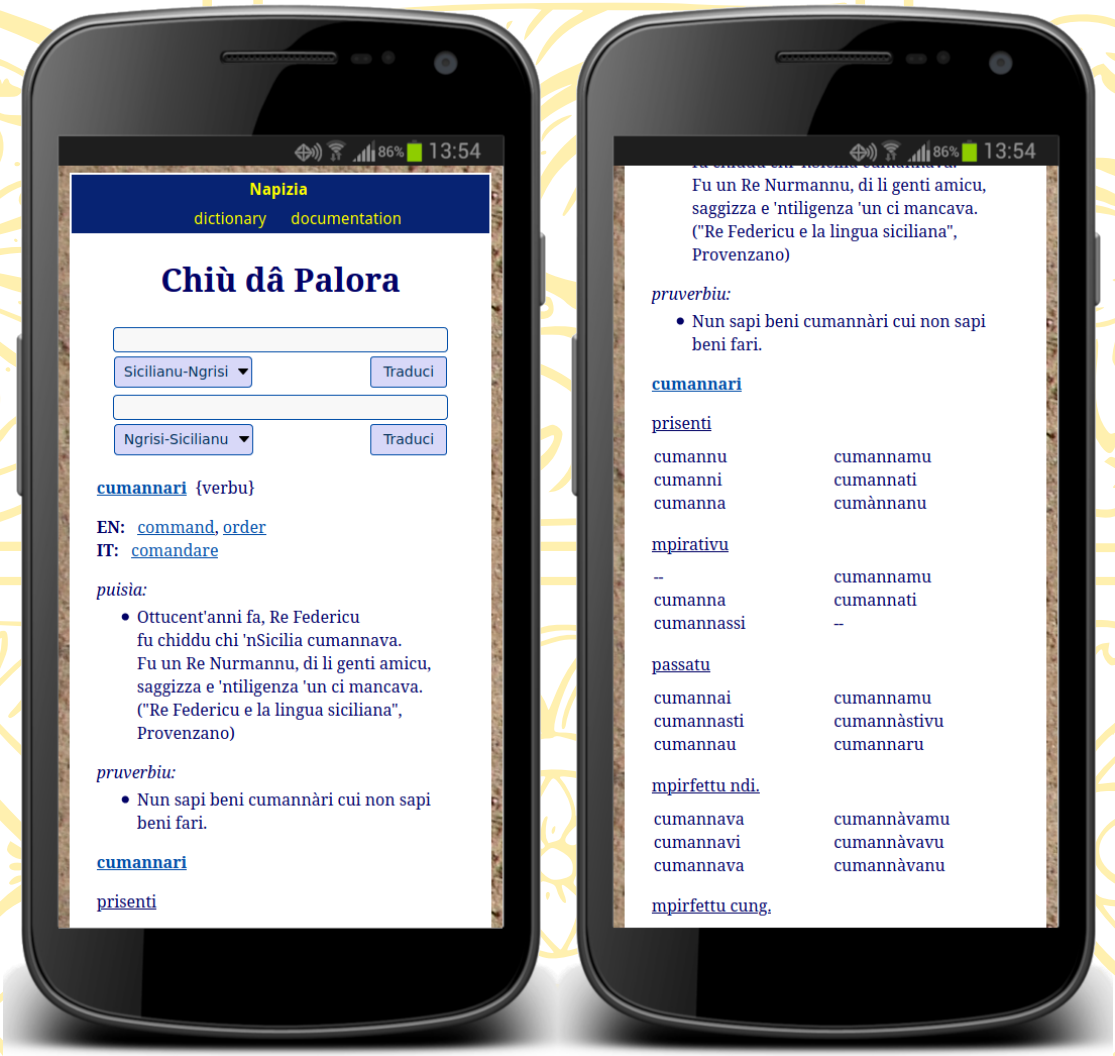
Arba Sicula

- In New York, since 1979, Arba Sicula has been:
  - organizing poetry recitals, concerts, cultural events and tours of Sicily
  - publishing books on Sicilian language, literature, history, cuisine, fiction, ...
  - translating Sicilian poetry and prose
  - publishing a bilingual journal, *Arba Sicula*
- So we assembled parallel text from:
  - the bilingual literary journal *Arba Sicula*
  - A. Dieli's translations of Sicilian poetry, proverbs and G. Pitrè's *Folk Tales*
  - examples from G. Cipolla's *Mparamu* and K. Bonner's *Introduction*

Standard Sicilian

- But we did **NOT** start with data collection
- We started by collecting the rules of Sicilian vocabulary and grammar.
  - Arthur Dieli's *Sicilian Vocabulary*
  - Kirk Bonner's *Introduction to Sicilian Grammar* (2001)
  - Gaetano Cipolla's *Mparamu lu sicilianu* (2013)
- And we created the *Chiù dâ Palora* (*More About the Word*) dictionary.
  - vocabulary annotated with grammar, proverbs, poetry, prose and examples
  - provides a reference for standardizing Sicilian language text
- There are many dialects of Sicilian. We present a common form.
  - G. Cipolla's *Learn Sicilian II* (2021) describes the diatopic variation
  - the local variations (usually) only appear in the spoken language
  - literary, written Sicilian is generally more homogeneous

More About the Word



https://www.napizia.com/cgi-bin/cchiu-da-palora.pl

Theory in the Dataset

- Data preparation
  - selected Sicilian language text that could be edited to Standard Sicilian
  - manually edited the text (both languages) for quality and standardization
- Standardization
  - a standard form like *beddu* replaces dialect forms like *bieddu* or *beddru*
  - enabling us to train our models on more examples of the standard form
- Textbook examples
  - from G. Cipolla's *Mparamu* and K. Bonner's *Introduction*
  - because "Language Models are Few-Shot Learners" (Brown et al., 2020)
- Our parallel corpus (so far):
  - 12,357 lines of bilingual text – 237,456 Sicilian words, 236,568 English words
  - 4,660 lines of trilingual textbook exercises – Sicilian, English and Italian
  - 121 hand-selected lines for validation (in all three languages)
  - the Italian-English subset of Farkas' *Books*

Theory in the Tokenization

- Standardization and Uncontraction
  - more replacement of dialect forms with standard forms
  - uncontraction of the spoken form into a literary form
    - *Hê parrari chî studenti.* → *Haiu a parrari cu li studenti.*
    - *I have to speak with the students.*
  - allows us to train our models on a single standard, literary form

- Lower-case and ASCII – further increases word frequencies
  - after uncontracting, we remove any remaining diacritics
  - and convert the text to lower case
  - creating an ASCII representation of the language

Theory in the Subword Splitting

Words are a sequence of subword units.

- Prefixes and Suffixes
  - *nchiudiri* = *n*+*chiudiri*
  - *enclose* = *en*+*close*
- Diminuitives and Augmentatives
  - cat: *jattu* = *jatt*+*u*
  - little cat: *jattareddu* = *jatt*+*ar*+*eddu*
- Verb conjugations
  - *parrari* → *parru*, *parri*, *parra*, *parramu*, *parrati*, *parranu*
- Biasing the subword splitting towards theoretical stems and desinences allows our models to learn a *theoretic* sequence of subword units.
- Easy to implement. Just append a vocabulary list to the dataset.

Evaluation Metrics

Theory improves translation quality

- At 13,839 lines of parallel text, biasing the subword distribution toward theoretical stems and desinences increased BLEU scores:
  - from 20.3 to 22.4 on English-to-Sicilian translation
  - from 21.4 to 24.1 on Sicilian-to-English translation

Putting it all together

- And at 17,017 lines plus back-translation and multilingual translation, our *Tradutturi Sicilianu* achieved BLEU scores of:
  - 35.0 on English-to-Sicilian, 36.8 on Sicilian-to-English
  - 36.5 on Italian-to-Sicilian, 30.9 on Sicilian-to-Italian
- albeit on a narrow domain

Conclusion

- Existing methods helped us train a neural machine translator
  - back-translation, multilingual translation and avoidance of overfitting
- *Modeling the language* helped us train a good one
  - we incorporated theory into dataset, tokenization and subword splitting
  - and our models learned to assemble a theoretic sequence of subword units
- If a model can learn like a human learns, then we should train our models to learn like a human learns.

Acknowledgements

- Thank you to Arba Sicula, Gaetano Cipolla and Arthur Dieli:
  - for developing the resources that made this project possible
  - for their support and encouragement
  - for sponsoring and developing Sicilian language and culture
- *Grazzi!*