gourmet

# Main Research Takeaways

User Event
18th November 2021

Felipe Sánchez-Martínez (University of Alicante)

GLOBAL UNDER-RESOURCED MEDIA TRANSLATION

EU H2020 GRANT AGREEMENT: 825299

EU H2020 GRANT AGREEMENT: 825299

# Content

(1) Low-resource neural machine translation

(2) GoURMET's contributions

(3) Surprise language challenge

EU H2020 GRANT AGREEMENT: 825299

# Low-resource neural machine translation

- ✔ Hot research topic in natural language processing
  - ✔ 18 workshops in the last 3 years
- ✔ Data resources
  - ✔ FLORES-101
  - ✔ Paracrawl

gourmet

EU H2020 GRANT AGREEMENT: 825299

# Low-resource neural machine translation

- ✔ Data augmentation: Use of synthetic data
  - ✔ Self-learning: Back and forward translation
  - ✔ Modification of exiting parallel sentences
- ✔ Pre-trained models
  - ✔ Huge multilingual models already trained and available for download

EU H2020 GRANT AGREEMENT: 825299

# GoURMET's contributions

- ✔ Lead of WMT shared tasks: Gujarati, Tamil, Hausa
- ✔ Research papers
  - ✔ + 40 research papers
  - ✔ Topics:
    - ✔ Data collection and augmentation, morphology, transfer learning, probabilistic modelling, etc.

EU H2020 GRANT AGREEMENT: 825299

# GoURMET's contributions

- ✔ Release of corpora crawled from the web
  - ✔ Bilingual and monolingual corpora
    - ✔ Swahili, Turkish, Amharic, Kyrgyz, Indian languages, etc.
- ✔ Software:
  - ✔ Cointribution to free/open-source tools
  - ✔ Release of code used in research papers
- ✔ Dokerised transaltion models

EU H2020 GRANT AGREEMENT: 825299

# GoURMET's contributions: some lessons learned

- ✔ Learning to generate the source (in addition to the target) improves models that learn from litle data, synthetic data or mix-domain data

EU H2020 GRANT AGREEMENT: 825299

# GoURMET's contributions: some lessons learned

- ✔ Learning to generate the source (in addition to the target) improves models that learn from litle data, synthetic data or mix-domain data

- ✔ Exploiting of  exisiting data on a multi-task learning framework by generating simple auxiliary tasks (eg. changing target word order, replace pairs of words, etc.) improves transaltion quality

EU H2020 GRANT AGREEMENT: 825299

# GoURMET's contributions: some lessons learned

- ✔ Learning to generate the source (in addition to the target) improves models that learn from litle data, synthetic data or mix-domain data

- ✔ Exploiting of  exisiting data on a multi-task learning framework by generating simple auxiliary tasks (eg. changing target word order, replace pairs of words, etc.) improves transaltion quality

- ✔ In extreme low-resource settings baseline systems are not good enough to generate synthetic data but a pivot language can be used to paraphrase English

gourmet

EU H2020 GRANT AGREEMENT: 825299

# Surprise language challenge

- ✔ Simulate sudden need of assimilation or dissemination of information in regions of the world with languages not included in the digital workflows
- ✔ NMT development period: February-March 2021

EU H2020 GRANT AGREEMENT: 825299

# Surprise language challenge


40-50 million speakers

gourmet

EU H2020 GRANT AGREEMENT: 825299

# Surprise language challenge: approach

- ✔ Crawling a manual list of 50 websites + 180 websites on the top-level domain
  - ✔ ~60k parallel sentences
- ✔ Existing corpora: ~340k parallel sentences

EU H2020 GRANT AGREEMENT: 825299

# Surprise language challenge: approach

- ✔ Crawling a manual list of 50 websites + 180 websites on the top-level domain
  - ✔ ~60k parallel sentences
- ✔ Existing corpora: ~340k parallel sentences

- ✔ Two NMT systems built:
  - ✔ From scracth: exploitation of monoligual corpora and parallel corpora in other languages
  - ✔ Fine-tuning of existing pre-trained models

gourmet

EU H2020 GRANT AGREEMENT: 825299

# Surprise language challenge: from scratch model

1. Train on easy related tasks
   - Monolingual gap-filling
   - German-English translation

EU H2020 GRANT AGREEMENT: 825299

# Surprise language challenge: from scratch model

1. Train on easy related tasks
   - Monolingual gap-filling
   - German-English translation
2. Fine-tuning on English-Pastho

EU H2020 GRANT AGREEMENT: 825299

# Surprise language challenge: from scratch model

1. Train on easy related tasks
   - Monolingual gap-filling
   - German-English translation
2. Fine-tuning on English-Pastho
3. Exploit monolingual corpora to generate synthetic data
   - Iterative process:
     - Translate from Pastho into English
     - Use output to train English→Pastho
     - Repeat the process in the other direction

gourmet

EU H2020 GRANT AGREEMENT: 825299

# Surprise language challenge: fine-tuning of pre-trained model

- Pre-trained model:
  - Multilingual system translating between English and 49 languages (both directions)
  - Released by Facebook on January 2021
  - Pastho included in the pre-trained model

- Start from already trained system
  - System gets better for English-Pastho and worse for the rest of language pairs

EU H2020 GRANT AGREEMENT: 825299

# Surprise language challenge: parameters

- From-scratch model
  - 62M parameters  - faster transaltion
- Fine-tuned pre-trained model
  - 610M parameters  - slower translation

EU H2020 GRANT AGREEMENT: 825299

# Surprise language challenge: human evaluation

| Score [0-100] | Commercial system | From-scratch model | Fine-tuned pre-trained model |
|---|---|---|---|
| English → Pastho | 68.5 | 67.5 | **92.3** |
| Pastho → English | 83.5 | 63.5 | **81.5** |

gourmet

# Thanks for your attention !!

Software, corpora, dockerised models, etc.: https://gourmet-project.eu/

GLOBAL UNDER-RESOURCED MEDIA TRANSLATION

EU H2020 GRANT AGREEMENT: 825299