CODET: A Benchmark for Contrastive Dialectal Evaluation of Machine Translation

Md Mahfuz Ibn Alam, Sina Ahmadi, Antonios Anastasopoulos
{malam21, antonis}@gmu.edu sina.ahmadi@uzh.ch

Motivation

- The performance of NMT systems degrade when faced with even slight deviations in language.
- How to evaluate NMT systems on dialectal variations?

Contributions

1. Extract contrastive data from previous studies **Italian, Basque, and Swiss German.**
2. Re-purpose contrastive data from other sources in seven languages: **Arabic, Occitan, Tigrinya, Farsi, Malay-Indonesian, Swahili, and Greek**.
3. Create contrastive data in additional languages: **Bengali and Central Kurdish**.
4. Benchmark dialects of the target languages using SOTA MT models
5. Quantify the discrepancies across varieties.

Highlights

- Use CODET to evaluate the resilience of your models on MT for non-standard varieties
- Significant progress is needed before MT can effectively handle non-standard languages.

Repository

Code and Data:
https://github.com/mahfuzibnalam/dialect_mt

CODET Benchmark

There is currently no benchmark for the evaluation of MT of dialects and varieties. We create one by

- Utilizing Existing Datasets
- Scraping Syntactic Atlases
- Creating parallel corpora

Standard Italian Variant:		
Source:	*Hanno rubato il quadro*	
GTranslate:	They stole the painting	✓
Alassio Variant:		
Source:	*I han rubbau u quaddru*	
GTranslate:	I han rubbau u quaddru	✗

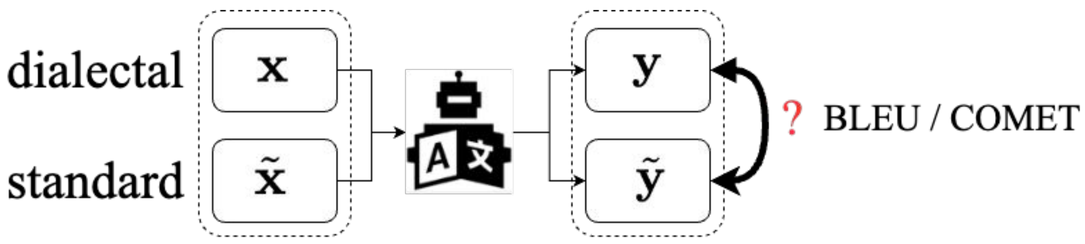
(Correct translation: "They stole the painting")

CODET encompasses 25,930 contrastive sentences of 891 distinct varieties spanning 12 diverse languages.

Evaluation

We evaluate CODET in the X → English direction using four different-sized NLLB-200 in two setups:

- With reference: compare with the reference standard
- Without references: consider non-standard sentences as **adversarial or non-native noisy** inputs

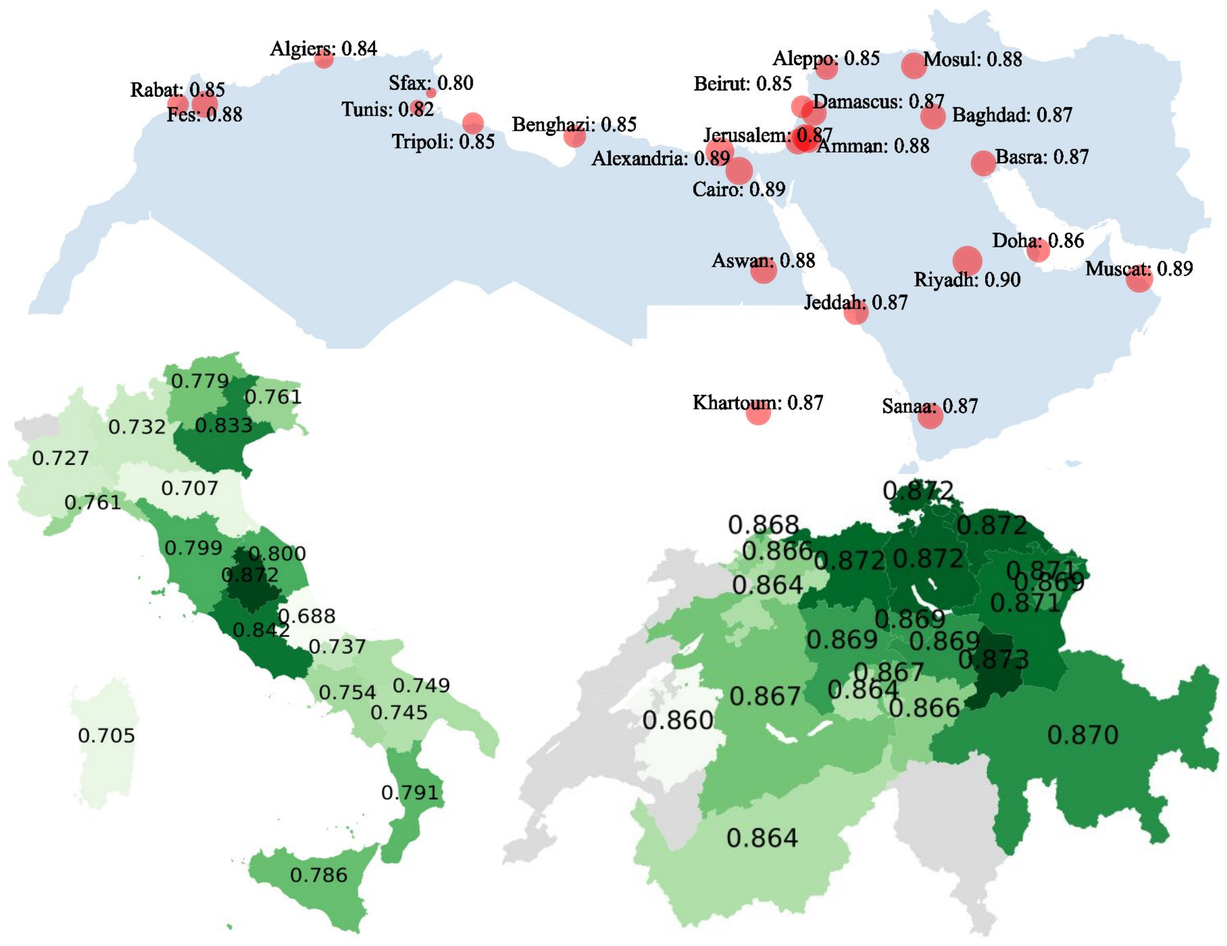


⇒ A robust MT system should produce the same output for dialectal inputs regardless of variations.

Languages/Varieties	# Sents	# Varieties
Italian Varieties	792	439
Swiss German Varieties	118	368
Basque Varieties	370	39
Arabic Vernaculars	12,000	25
Bengali Varieties	200	5
Central Kurdish Varieties	300	4
Farsi Varieties	3071	2
Malay-Indonesian	3071	2
Swahili	1919	2
Tigrinya Varieties	3071	2
Aranese	476	1
Central Occitan	379	1
Griko	163	1

Results

- MT systems excel at handling standard variants 😊
- As dialectal variations deviate further from the standard, the quality of translations decreases 😞
- Possible causes:
  - spelling variations
  - inadequate representation of morphosyntactic and lexical information of nonstandard dialects
  - terminologies and code-switching across borders



Future Work

- There are significant performance discrepancies across different varieties in MT
- More robust metrics to not penalize spelling variations
- **Still a lot of room for improvement**