AlphaMWE-Arabic: Arabic Edition of Multilingual Parallel Corpora with Multiword Expression Annotations

Motivations

- Multiword Expressions (MWEs) have been a bottleneck for NLP due to their idiomaticity, ambiguity, and non-compositionality.
- Bilingual parallel corpora introducing MWE annotations are very scarce.
- Expanding AlphaMWE parallel corpus to Arabic: Dialectal Arabic (Tunisian and Egyptian) and Standard Arabic (MSA).

Strategies

- Machine Translation (MT), post-editing, and annotations for both standard Arabic.
- Mannual Translation from Scratch for dialectal Arabic Tunisian and Egyptian (MT was too poor to use).
- Analyse the MT errors when they meet MWEs-related content, both quantitatively using the human-in-the-loop metric HOPE and qualitatively.

Experimental Settings

- MT system comparisons using SysTran and GoogleMT.
- We added two new error types to accommodate our post-editing and evaluation tasks on English-to-Arabic MT output: **MWE Missed Chance (MMC)**: Indicate when the MT output on source MWEs is either wrong semantically or correct translation but without using the corresponding correct MWEs in the target (in the situation when there is indeed such MWE in target). **Skipped Word (SKP)**: Highlight when the MT system failed to translate a certain word that was important to the context.

Human Post-editing GoogleMT Outputs

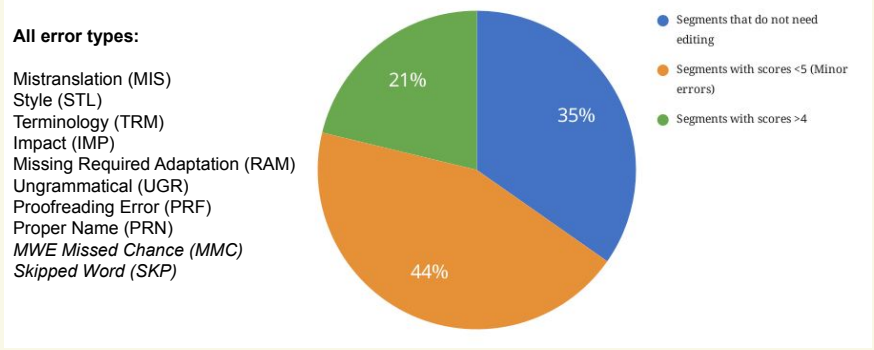

Figure: HOPE Metric evaluation on 150 Segments

MT System Selections

- when SysTran MT output makes mistakes, the errors are very severe, such as adding context out of the blue, while GoogleMT's output still makes some sense when it is wrong.
- SysTran has more correct translations on entities. To reduce the workload for the professional post-editing step; know more about how MT makes mistakes when translating MWEs and verbal idioms => We choose GoogleMT
- entity errors can be fixed more easily than out-of-the-blue errors;
- we can get more examples of how MT fails in translating MWE-related content => can be valuable for future research such as on guiding MT development.

Outcomes

- An Arabic corpus with MWEs annotated including three subsets:
- MSA corpus yielded 2,700 tokens
- Tunisian Arabic: 2,495 tokens translated
- Egyptian Arabic: 2,055 tokens translated MT Error Analysis with Qualitative and Quantitative Annotations

Statistics on each error type?

Error type	**MMC**	MIS	STL	TRM	**IMP**	UGR	PRF	SKP	All	PPS
Total Penalty Scores	76	68	69	39	114	37	46	6	455	
Ratio out of total segments	17%	15%	15%	9%	25%	8%	10%	1%		3.03

Table: Penalty Score Ratios of Each Error Type and Average Penalty Scores of Each Segment from 150 Segments using HOPE Metric

Najet Hadj Mohamed*, Malak Rassem*, Lifeng Han, Goran Nenadic

University of Tours, University of Sfax | University of Stuttgart | University of Manchester

Resources visit https://github.com/aaronlifenghan/AlphaMWE | Lifeng.Han@manchester.ac.uk