# Multilingual Anchoring: Interactive Topic Modeling and Alignment Across Languages

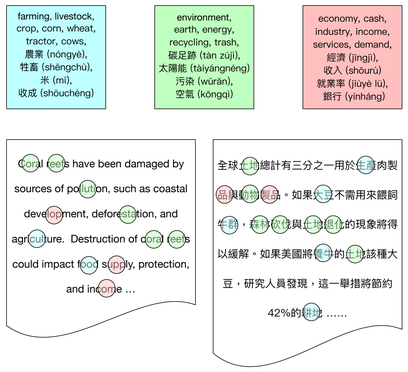
Michelle Yuan [1] Benjamin Van Durme [2] Jordan Boyd-Graber [1]

[1]University of Maryland [2]John Hopkins University

## Motivation

- Large text collections often require topic triage quickly in low-resource settings (e.g. natural disaster, political instability)
- Analysts need to examine multilingual documents but are scarce in one or more languages

## Modeling Multilingual Topics




## Anchor-based Topic Models

- An **anchor word** is a word that appears with high probability in one topic and low probability in all other topics
- Conditional co-occurrence matrix $\bar{Q}$ has entries such that $\bar{Q}_{i,j} = p(\text{word } 2 = j \mid \text{word } 1 = i)$
- Given anchor words $s_1, \ldots, s_K$, the algorithm approximates $\bar{Q}_i$ as the convex combination of $\bar{Q}_{s_1}, \ldots \bar{Q}_{s_K}$ and finds coefficients $C_{i,k}$ that estimate $p(\text{topic} = k \mid \text{word} = i)$

| $\bar{Q}$ matrix | carburetor | concealer | album | liner |
|---|---|---|---|---|
| carburetor | 0.80 | 0.05 | 0.05 | 0.10 |
| concealer | 0.13 | 0.60 | 0.07 | 0.20 |
| album | 0.05 | 0.05 | 0.70 | 0.20 |
| liner | 0.25 | 0.20 | 0.15 | 0.45 |

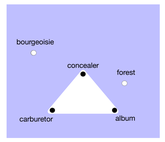
$$\bar{Q}_{\text{liner}} \approx C_1 \bar{Q}_{\text{carburetor}} + C_2 \bar{Q}_{\text{concealer}} + C_3 \bar{Q}_{\text{album}} \quad (1)$$

This decomposition (Eq. 1) resembles the topic distribution of ``liner'' over an automotive topic, a cosmetics topic, and a music topic.

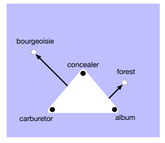
## Bridging Languages: How Do You Say Anchor in Chinese?

### Monolingual Anchoring

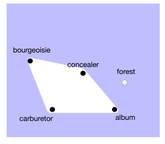
Rows in $\bar{Q}$ corresponding to the anchor words are the vertices of the convex hull formed by $\bar{Q}$.




To greedily find an anchor word, find a row in $\bar{Q}$ that is farthest away from the current span of anchor words.

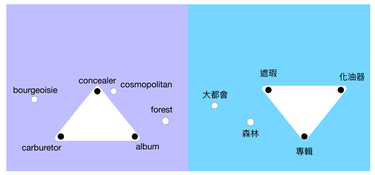



The goal is to maximize total span of anchor words so that each row in $\bar{Q}$ lies within this span and can be accurately approximated.

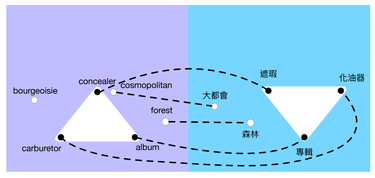



### Multilingual Anchoring

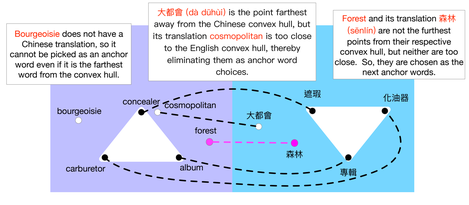
The challenge for multilingual topic modeling is to align topics cross-lingually even when words from different languages do not co-occur in the same documents.




Our algorithm first uses a dictionary to link words across languages.

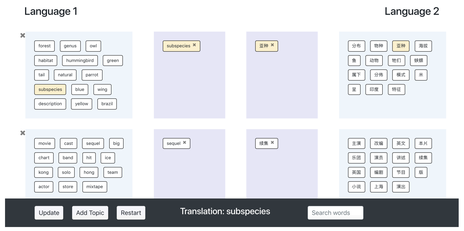



Then, it finds anchor words such that they are linked and can simultaneously expand the span of anchors words for both languages.

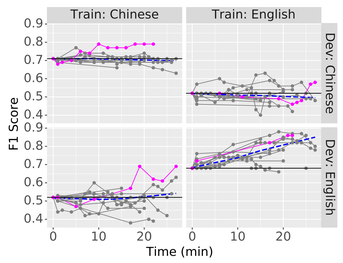


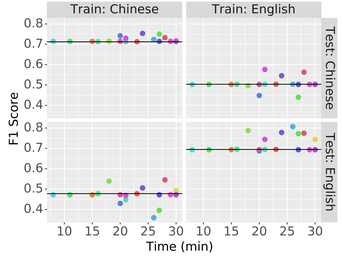

## MTAnchor: Interactive Topic Modeling

### User Interface




### User Study



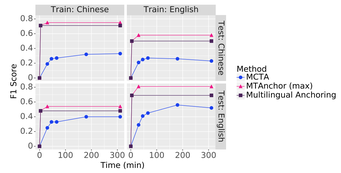




## More Information

- Code: `https://github.com/forest-snow/mtanchor_demo/`
- Author: `http://www.cs.umd.edu/~myuan/`
- This work was supported in part by the JHU Human Language Technology Center of Excellence (HLTCOE) and Raytheon BBN Technologies, by DARPA award HR0011-15-C-0113.

## Comparing Models

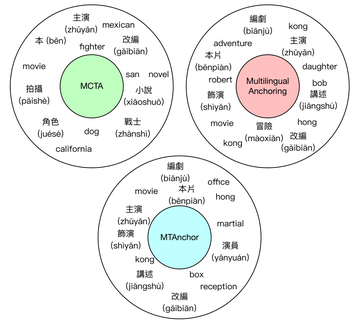
### Speed and Classification Accuracy




### Topic Coherence

| Dataset | Method | Topic coherence EN-I | ZH-I SI-I | EN-E | ZH-E SI-E |
|---|---|---|---|---|---|
| Wikipedia | Multilingual anchoring | 0.14 | 0.18 | 0.08 | 0.13 |
| | MTAnchor (maximum) | **0.20** | **0.20** | **0.10** | **0.15** |
| | MTAnchor (median) | 0.14 | 0.18 | 0.08 | 0.13 |
| | MCTA | 0.13 | 0.09 | 0.00 | 0.04 |
| Amazon | Multilingual anchoring | **0.07** | **0.06** | **0.03** | **0.05** |
| | MCTA | -0.03 | 0.02 | 0.02 | 0.01 |
| LORELEI | Multilingual anchoring | 0.08 | 0.00 | 0.03 | n/a |
| | MCTA | **0.13** | 0.00 | **0.04** | n/a |

### Sample Film Topic




## Conclusions

- Anchoring algorithm can be applied multilingually
- People can provide helpful linguistic and cultural knowledge to improve topic models