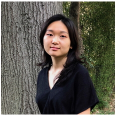
Multilingual Anchoring: Interactive Topic Modeling and Alignment across Languages

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

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

Authors



Michelle Yuan
UMD

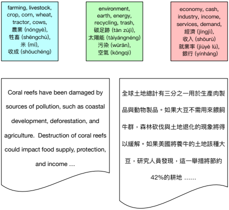
Benjamin Van Durme
JHU

Jordan Boyd-Gr
UMD

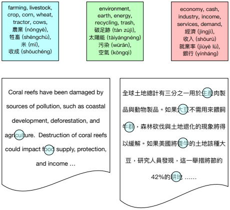
- Large text collections often require topic triage quickly in low-resource settings (e.g. natural disaster, political instability).

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

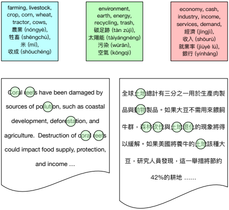
Modeling Multilingual Topics



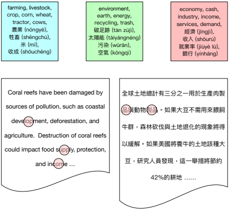
Modeling Multilingual Topics



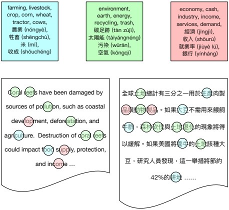
Modeling Multilingual Topics



Modeling Multilingual Topics



Modeling Multilingual Topics



Generative Approaches

- Polylingual Topic Model (Mimno et al., 2009)
- JointLDA (Jagarlamudi and Daumé, 2010)
- Polylingual Tree-based Topic model (Hu et al., 2014b)
- MCTA (Shi et al., 2016)

Generative Approaches

- Polylingual Topic Model (Mimno et al., 2009)
- JointLDA (Jagarlamudi and Daumé, 2010)
- Polylingual Tree-based Topic model (Hu et al., 2014b)
- MCTA (Shi et al., 2016)

These methods are slow, assume extensive knowledge about languages, and preclude human refinement.

Anchor words

Definition

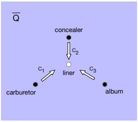
An **anchor word** is a word that appears with *high* probability in one topic but with *low* probability in all other topics.

From Co-occurrence to Topics

- Normally, we want to find $p(\text{word} \mid \text{topic})$ (Blei et al., 2003).
- Instead, what if we can easily find $p(\text{word} \mid \text{topic})$ through using anchor words and conditional word co-occurrence $p(\text{word } 2 \mid \text{word } 1)$?

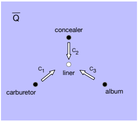
From Co-occurrence to Topics

$$\bar{Q}_{i,j} = p(w_2 = j \mid w_1 = i)$$



From Co-occurrence to Topics

$$\bar{Q}_{i,j} = p(w_2 = j \mid w_1 = i)$$



$$\begin{aligned}\bar{Q}_{\text{liner}} &\approx C_1\bar{Q}_{\text{carburetor}} + C_2\bar{Q}_{\text{concealer}} + C_3\bar{Q}_{\text{album}} \\ &= 0.4 * \begin{bmatrix}0.3\\ \cdots\\ 0.1\end{bmatrix} + 0.2 * \begin{bmatrix}0.1\\ \cdots\\ 0.2\end{bmatrix} + 0.4 * \begin{bmatrix}0.1\\ \cdots\\ 0.4\end{bmatrix}\end{aligned}$$

Anchoring

- If an anchor word appears in a document, then its corresponding topic is among the set of topics used to generate document (Arora et al., 2012).
- Anchoring algorithm uses word co-occurrence to find anchors and gradient-based inference to recover topic-word distribution (Arora et al., 2013).
- Runtime is **fast** because algorithm scales with number of unique word types, rather than number of documents or tokens.

Anchoring

1. Construct co-occurrence matrix from documents with vocabulary of size V:

$$\bar{Q}_{i,j} = p(w_2 = j \mid w_1 = i).$$

Anchoring

1. Construct co-occurrence matrix from documents with vocabulary of size V:

$$\bar{Q}_{i,j} = p(w_2 = j \mid w_1 = i).$$

2. Given anchor words $s_1, \ldots, s_K$, approximate co-occurrence distributions:

$$\bar{Q}_i \approx \sum_{k=1}^{K} C_{i,k} \bar{Q}_{s_k} \text{ subject to } \sum_{k=1}^{K} C_{i,k} = 1 \text{ and } C_{i,k}$$

Anchoring

1. Construct co-occurrence matrix from documents with vocabulary of size V:

$$\bar{Q}_{i,j} = p(w_2 = j \mid w_1 = i).$$

2. Given anchor words $s_1, \dots, s_K$, approximate co-occurrence distributions:

$$\bar{Q}_i \approx \sum_{k=1}^{K} C_{i,k} \bar{Q}_{s_k} \text{ subject to } \sum_{k=1}^{K} C_{i,k} = 1 \text{ and } C_{i,k}$$

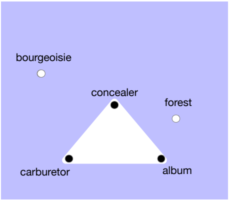
3. Find topic-word matrix:

$$\begin{aligned} A_{i,k} &= p(w = i \mid z = k) \propto p(z = k \mid w = i) p(w = i) \\ &= C_{i,k} \sum_{j=1}^{V} \bar{Q}_{i,j}. \end{aligned}$$

Finding Anchor Words

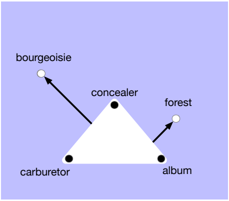
- So far, we assume that anchor words are given.
- How do we find anchor words from documents?

Finding Anchor Words



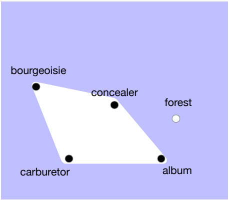
Anchor words are the vertices of the co-occurrence convex hull.

Finding Anchor Words



Anchor words are the vertices of the co-occurrence convex hull.

Finding Anchor Words



Anchor words are the vertices of the co-occurrence convex hull.

Issues with Topic Models

Topics
music concert singer voice chorus songs album
singer pop songs music album chorale jazz
cosmetics makeup eyeliner lipstick foundation primer eyesh

Issues with Topic Models

Topics
music concert singer voice chorus songs album
singer pop songs music album chorale jazz
cosmetics makeup eyeliner lipstick foundation primer eyesh

Duplicate topics.

Issues with Topic Models

Topics
music band art history literature books earth
bts taehyung idol kpop jin jungkook jimin

Issues with Topic Models

Topics
music band art history literature books earth
bts taehyung idol kpop jin jungkook jimin

Ambiguous topics.
Overly-specific topics.

Interactive Anchoring

- Incorporating interactivity in topic modeling has shown to improve quality of model (Hu et al., 2014a).
- Anchoring algorithm offers speed for interactive work, but single anchors are unintuitive to users.
- **Ankura** is an interactive topic modeling system that allows users to choose multiple anchors for each topic (Lund et al., 2017).
- After receiving human feedback, **Ankura** only takes a few seconds to update topic model.

Interactive Anchoring

- Incorporating interactivity in topic modeling has shown to improve quality of model (Hu et al., 2014a).
- Anchoring algorithm offers speed for interactive work, but single anchors are unintuitive to users.
- **Ankura** is an interactive topic modeling system that allows users to choose multiple anchors for each topic (Lund et al., 2017).
- After receiving human feedback, **Ankura** only takes a few seconds to update topic model.

These methods only work for monolingual document collections.

Linking Words

Definition

Language $\mathcal{L}$ is a set of word types w.

Linking Words

Definition

Language $\mathcal{L}$ is a set of word types w.

Definition

Bilingual dictionary $\mathcal{B}$ is a subset of the Cartesian product $\mathcal{L}^{(1)} \times \mathcal{L}^{(2)}$, where $\mathcal{L}^{(1)}, \mathcal{L}^{(2)}$ are two, different languages.

Linking Words

Definition

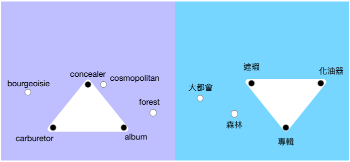
Language $\mathcal{L}$ is a set of word types w.

Definition

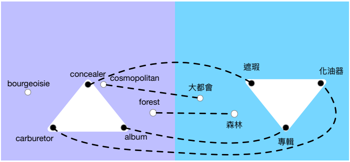
Bilingual dictionary $\mathcal{B}$ is a subset of the Cartesian product $\mathcal{L}^{(1)} \times \mathcal{L}^{(2)}$, where $\mathcal{L}^{(1)}, \mathcal{L}^{(2)}$ are two, different languages.

Idea: If dictionary $\mathcal{B}$ contains entry (w, v), create a link between w and v.

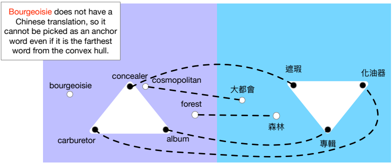
Finding Multilingual Anchors



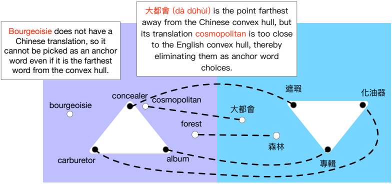
Finding Multilingual Anchors



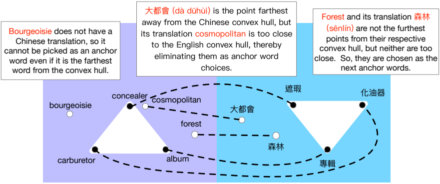
Finding Multilingual Anchors



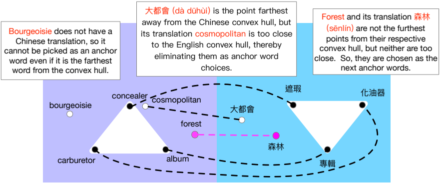
Finding Multilingual Anchors



Finding Multilingual Anchors



Finding Multilingual Anchors



Multilingual Anchoring

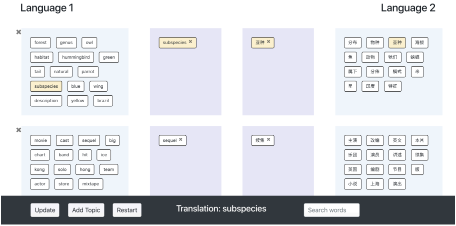
1. Given a dictionary, create links between words that are translations of each other.
2. Select an anchor word for each language such that the words are linked and span of anchor words is maximized.
3. Once anchor words are found, separately find topic-word distributions for each language.

- What if dictionary entries are scarce or inaccurate?
- What if topics aren't aligned properly across languages?

- What if dictionary entries are scarce or inaccurate?
- What if topics aren't aligned properly across languages?

Incorporate human-in-the-loop topic modeling tools.

MTAnchor



Language 1
Language 2
forest genus owl
habitat hummingbird green
tail natural parrot
subspecies blue wing
description yellow brazil
subspecies
亚种
分布 物种 亚种 海拔
鱼 动物 牠们
属下 分佈 模式 米
呈 印度 特征
movie cast sequel big
chart band hit ice
kong solo hong team
actor store mixtape
sequel
续集
主演 改编 英文 本片
乐团 演员 讲述 续集
英国 编剧 节目 版
小说 上海 演出
Update
Add Topic
Restart
Translation: subspecies
Search words

Experiments

Datasets:

1. Wikipedia articles (EN, ZH)
2. Amazon reviews (EN, ZH)
3. LORELEI documents (EN, SI)

Experiments

Metrics:

1. Classification accuracy
   - Intra-lingual: train topic model on documents in one language and test on other documents in the *same* languages
   - Cross-lingual: train topic model on documents in one language and test on other documents in a *different* language.
2. Topic coherence (Lau et al., 2014).
   - Intrinsic: use the trained documents as the reference corpus to measure local interpretability.
   - Extrinsic: use a large dataset (i.e. entire Wikipedia) as the reference corpus to measure global interpretability.

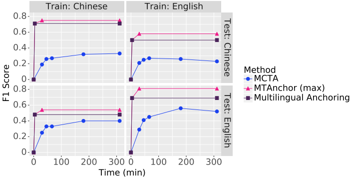
Comparing Models

		Classification accuracy			
Dataset	Method	EN-I	ZH-I SI-I	EN-C	ZH-C SI-C
Wikipedia	Multilingual anchoring	69.5%	71.2%	50.4%	47.8%
	MTAnchor (maximum)	**80.7%**	**75.3%**	**57.6%**	**54.5%**
	MTAnchor (median)	69.5%	71.4%	50.3%	47.2%
	MCTA	51.6%	33.4%	23.2%	39.8%
Amazon	Multilingual anchoring	**59.8%**	**61.1%**	**51.7%**	**53.2%**
	MCTA	49.5%	50.6%	50.3%	49.5%
LORELEI	Multilingual anchoring	**20.8%**	**32.7%**	**24.5%**	**24.7%**
	MCTA	13.0%	26.5%	4.1%	15.6%

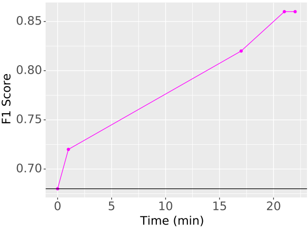
Comparing Models

		Topic coherence			
Dataset	Method	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

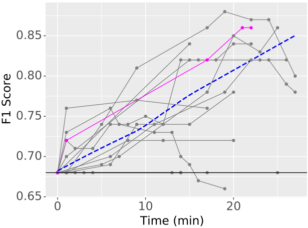
Multilingual Anchoring Is Much Faster



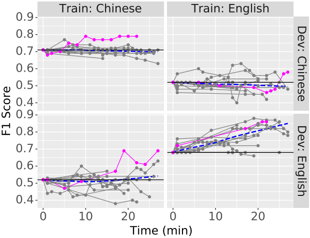
Improving Topics Through Interactivity



Improving Topics Through Interactivity



Improving Topics Through Interactivity



Comparing Topics

Dataset	Method	Topic
Wikipedia	MCTA	dog san movie mexican fighter novel california 主演 改編 本 小說 拍攝 角色 戰士
	Multilingual anchoring	**adventure** daughter bob kong hong robert movie 主演 改編 本片 飾演 **冒險** 講述 編劇
	MTAnchor	**kong hong movie** office martial box reception 主演 改編 飾演 本片 **演員 編劇** 講述
Amazon	MCTA	woman food eat person baby god chapter 來貨 頂頂 水 耳機 貨物 張傑 傑 同樣
	Multilingual anchoring	eat diet food recipe **healthy** lose weight **健康** 幫 吃 身體 全面 同事 中醫
LORELEI	MCTA	help need floodrelief please families needed victim
	Multilingual anchoring	aranayake warning landslide site missing nbro areas

Why Not Use Deep Learning?

- Neural networks are data-hungry and unsuitable for low-resource languages
- Deep learning models take long amounts of time to train
- Pathologies of neural models make interpretation difficult (Feng et al., 2018)

Summary

- Anchoring algorithm can be applied in multilingual settings.
- People can provide helpful linguistic or cultural knowledge to construct better multilingual topic models.

Future Work

- Apply human-in-the-loop algorithms to other tasks in NLP.
- Better understand the effect of human feedback on cross-lingual representation learning.

References I

Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. 2013. A practical algorithm for topic modeling with provable guarantees. In *ICML*.

Sanjeev Arora, Rong Ge, and Ankur Moitra. 2012. Learning topic models–going beyond SVD. In *Foundations of Computer Science (FOCS)*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *JMLR*.

Shi Feng, Eric Wallace, Alvin Grissom II, Pedro Rodriguez, Mohit Iyyer, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretation difficult. In *EMNLP*.

Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014a. Interactive topic modeling. *MLJ*.

Yuening Hu, Ke Zhai, Vlad Eidelman, and Jordan Boyd-Graber. 2014b. Polylingual tree-based topic models for translation domain adaptation. In *ACL*.

Jagadeesh Jagarlamudi and Hal Daumé. 2010. Extracting multilingual topics from unaligned comparable corpora. In *ECIR*.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *EACL*.

References II

Jeffrey Lund, Connor Cook, Kevin Seppi, and Jordan Boyd-Graber. 2017. Tandem anchoring: A multiword anchor approach for interactive topic modeling. In *ACL*.

David Mimno, Hanna M Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum. 2009. Polylingual topic models. In *EMNLP*.

Bei Shi, Wai Lam, Lidong Bing, and Yinqing Xu. 2016. Detecting common discussion topics across culture from news reader comments. In *ACL*.