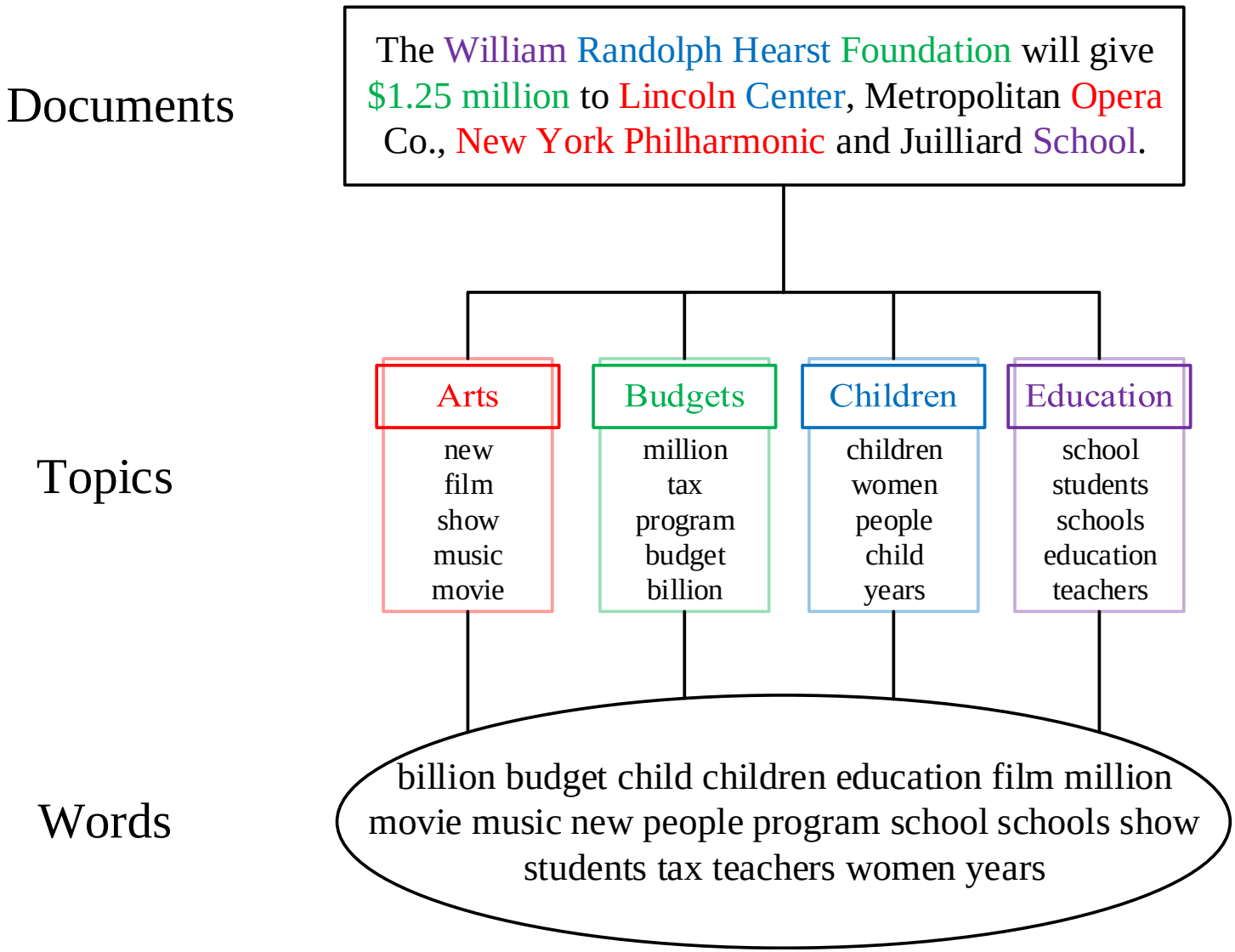
A Multilingual Topic Model for Learning Weighted Topic Links Across Corpora with Low Comparability

Weiwei Yang[1,2], Jordan Boyd-Graber[1,3], and Philip Resnik[1]

[1]University of Maryland, College Park
[2]Facebook, [3]Google AI Zürich

November 5, 2019

Topic Models



Topics Differ among Languages

Topics Differ among Languages

- In the same geographic area, e.g., London:

Topics Differ among Languages

- In the same geographic area, e.g., London:
  - Spanish: Chaos in Catalonia
  - Chinese: "Trade War" with the U.S.
  - English: Brexit deals

Topics Differ among Languages

- In the same geographic area, e.g., London:
  - Spanish: Chaos in Catalonia
  - Chinese: "Trade War" with the U.S.
  - English: Brexit deals
- On the same topic, e.g., Earthquake:

Topics Differ among Languages

- In the same geographic area, e.g., London:
  - Spanish: Chaos in Catalonia
  - Chinese: "Trade War" with the U.S.
  - English: Brexit deals
- On the same topic, e.g., Earthquake:
  - English: Earthquakes worldwide
  - Chinese: Earthquake in Sichuan Province in 2008

Topics Differ among Languages

- In the same geographic area, e.g., London:
  - Spanish: Chaos in Catalonia
  - Chinese: "Trade War" with the U.S.
  - English: Brexit deals
- On the same topic, e.g., Earthquake:
  - English: Earthquakes worldwide
  - Chinese: Earthquake in Sichuan Province in 2008
- When modeling topics multilingually, it is not a good idea to assume an aligned topic space.

Topics Differ among Languages

- In the same geographic area, e.g., London:
  - Spanish: Chaos in Catalonia
  - Chinese: "Trade War" with the U.S.
  - English: Brexit deals
- On the same topic, e.g., Earthquake:
  - English: Earthquakes worldwide
  - Chinese: Earthquake in Sichuan Province in 2008
- When modeling topics multilingually, it is not a good idea to assume an aligned topic space.
- Keep the topics of different languages separated and connect them by weighted topic links.

Weighted Topic Links

Weighted Topic Links

- Each language's topic distributions consist of the words in that language *only*.

sports, match, referee, tournament, champion	技术，信息，计算机，智能，系统
economics, dollars, million, invest, income	教育，大学，学校，学生，教师
politics, president, government, bill, vote	运动，比赛，裁判，锦标赛，冠军
technology, information, computers, smart, system	经济，美元，百万，投资，收入
education, universities, schools, students, teachers	政治，总统，政府，法案，投票

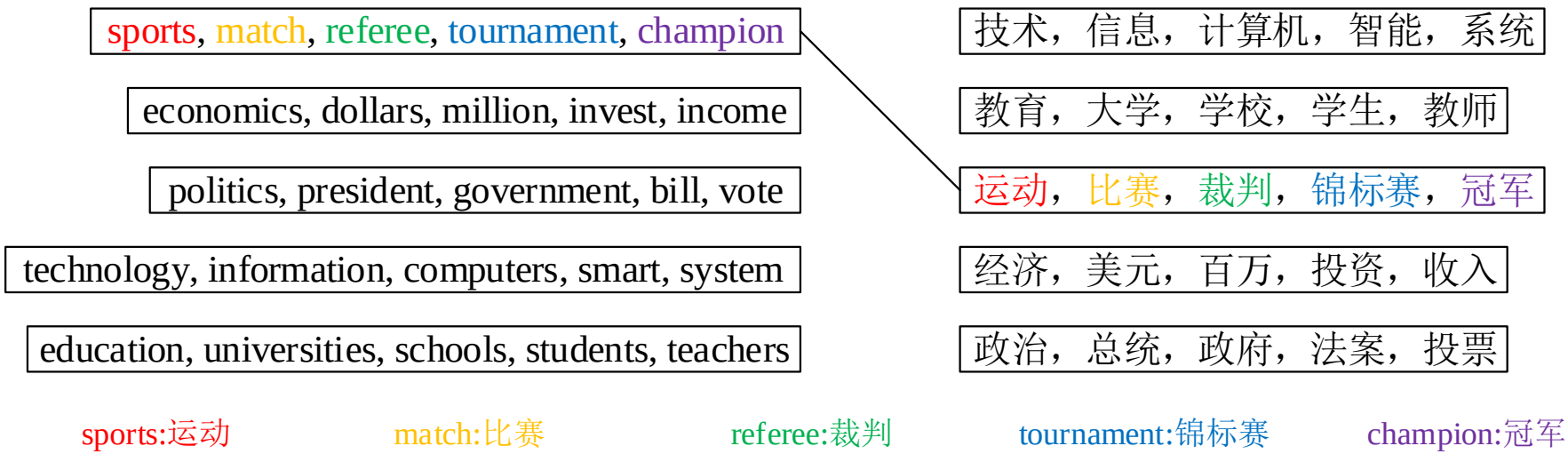
Weighted Topic Links

- Each language's topic distributions consist of the words in that language *only*.
- Weighted topic links connect topics *across* languages.

sports, match, referee, tournament, champion	技术，信息，计算机，智能，系统
economics, dollars, million, invest, income	教育，大学，学校，学生，教师
politics, president, government, bill, vote	运动，比赛，裁判，锦标赛，冠军
technology, information, computers, smart, system	经济，美元，百万，投资，收入
education, universities, schools, students, teachers	政治，总统，政府，法案，投票

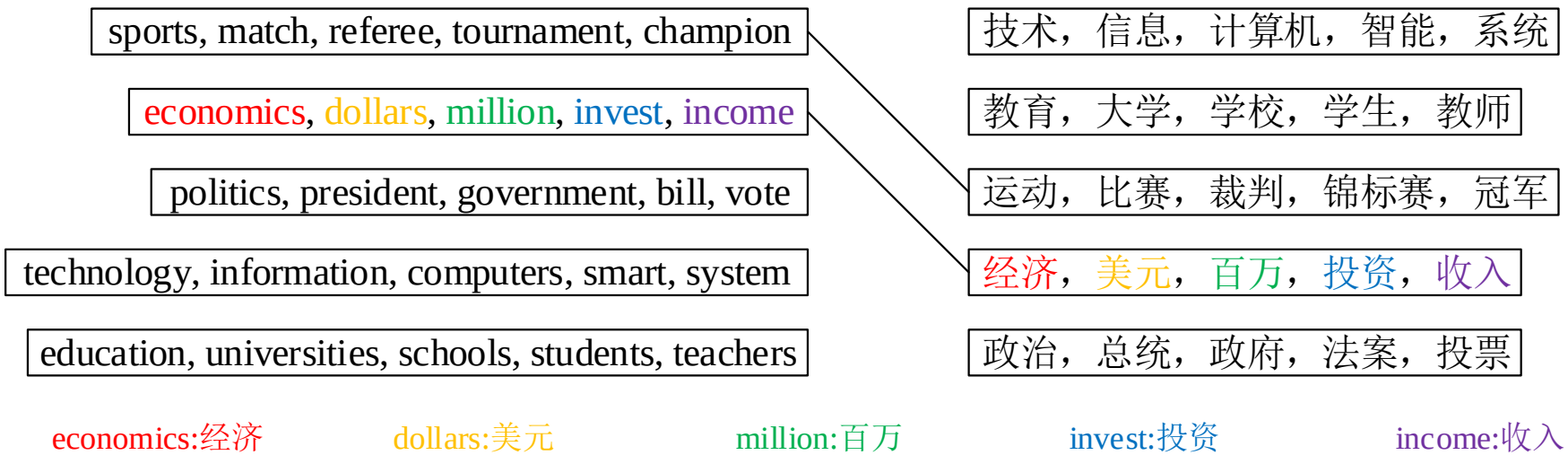
Weighted Topic Links

- Each language's topic distributions consist of the words in that language *only*.
- Weighted topic links connect topics *across* languages.
- Weighted topic links are learned based on word translations.



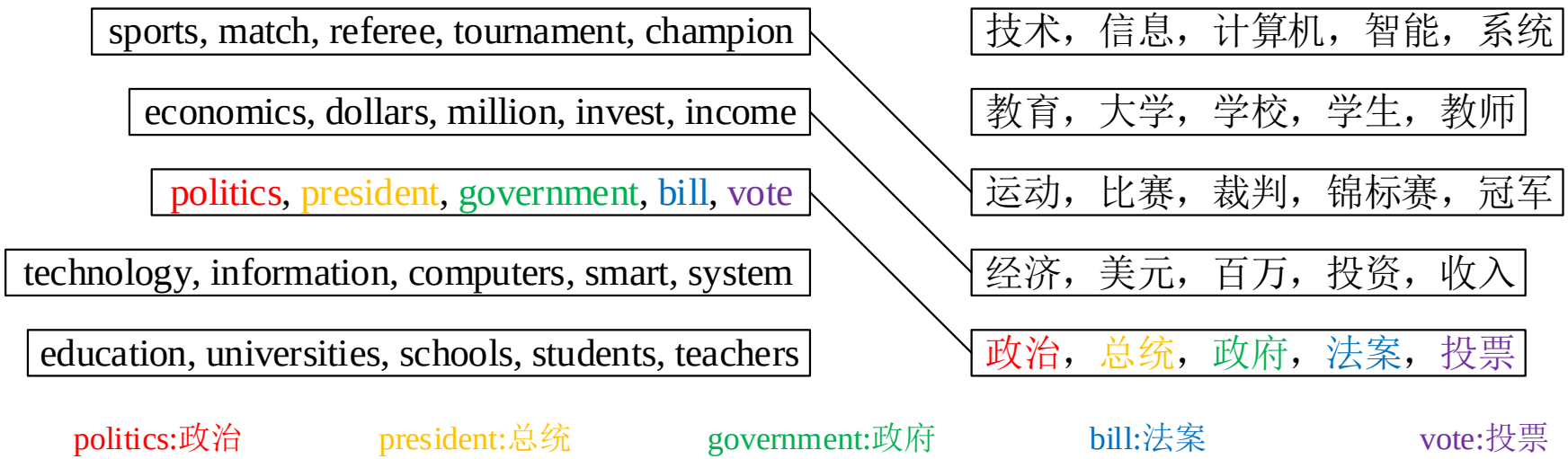
Weighted Topic Links

- Each language's topic distributions consist of the words in that language *only*.
- Weighted topic links connect topics *across* languages.
- Weighted topic links are learned based on word translations.



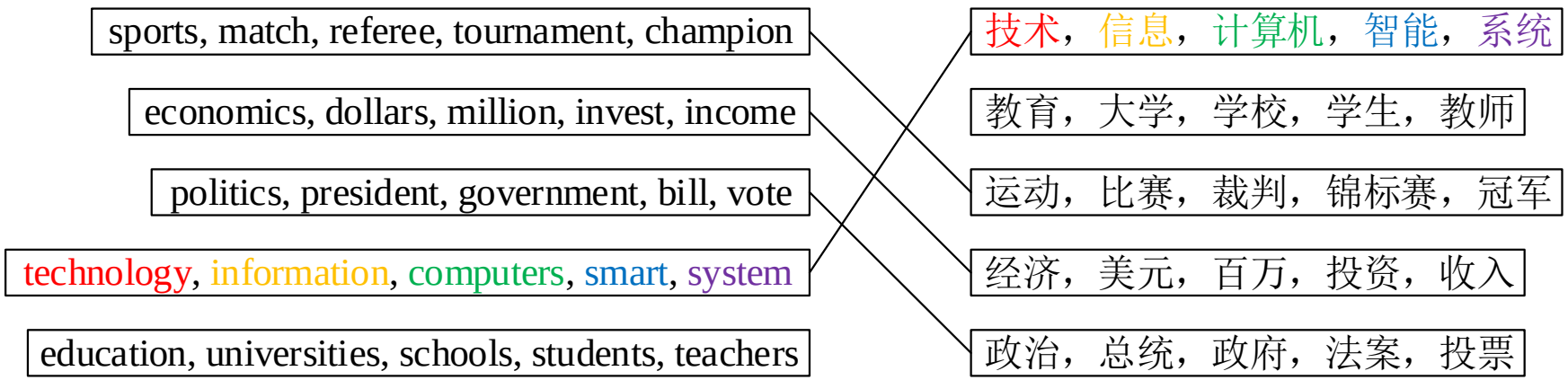
Weighted Topic Links

- Each language's topic distributions consist of the words in that language *only*.
- Weighted topic links connect topics *across* languages.
- Weighted topic links are learned based on word translations.



Weighted Topic Links

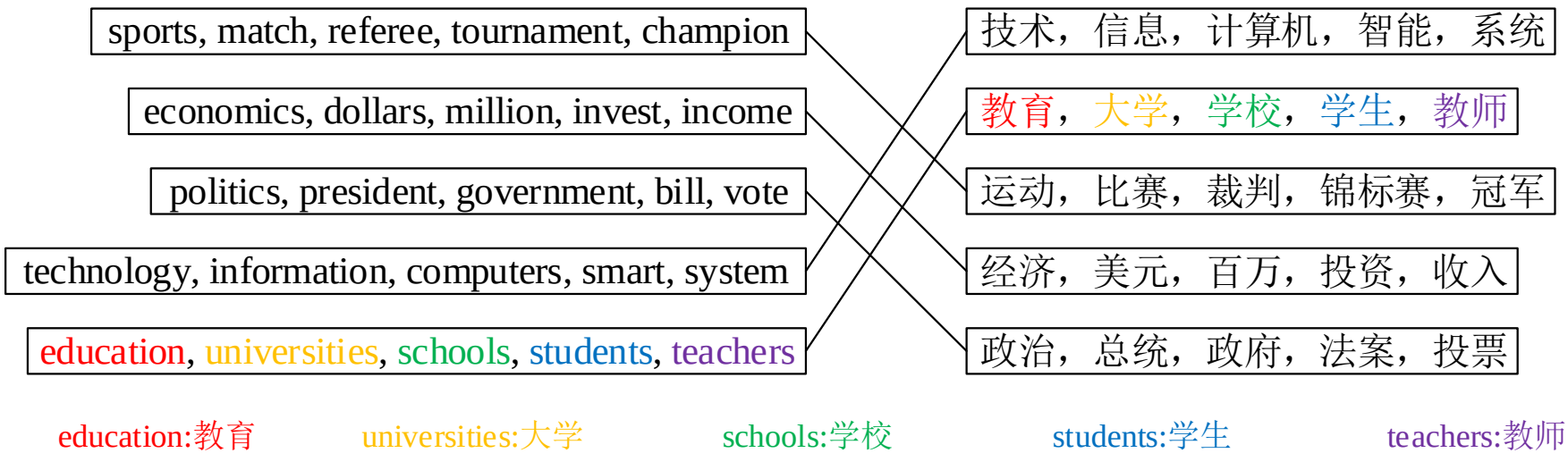
- Each language's topic distributions consist of the words in that language *only*.
- Weighted topic links connect topics *across* languages.
- Weighted topic links are learned based on word translations.



technology:技术 information:信息 computers:计算机 smart:智能 system:系统

Weighted Topic Links

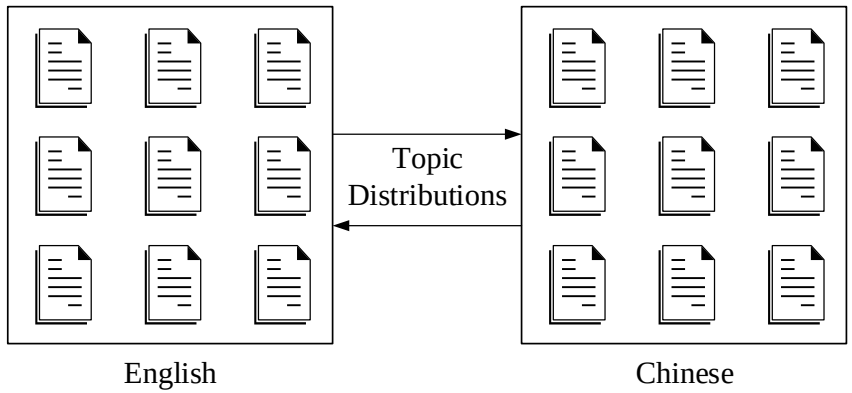
- Each language's topic distributions consist of the words in that language *only*.
- Weighted topic links connect topics *across* languages.
- Weighted topic links are learned based on word translations.



Why Weighted Topic Links?

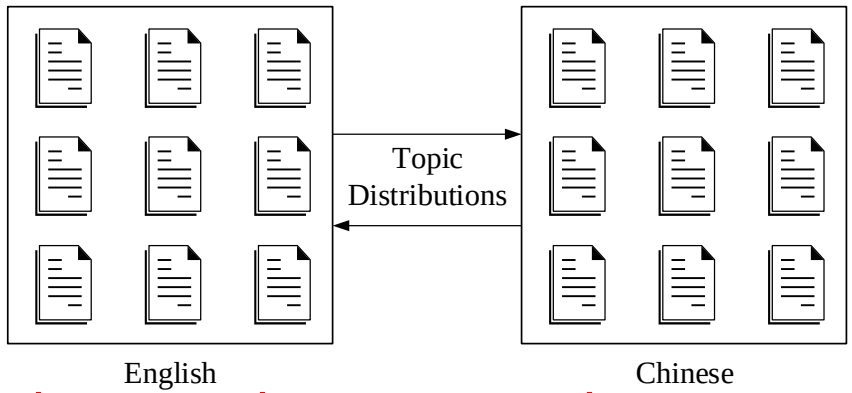
Why Weighted Topic Links?

- Transfer learned topic distributions from one language to another as prior knowledge

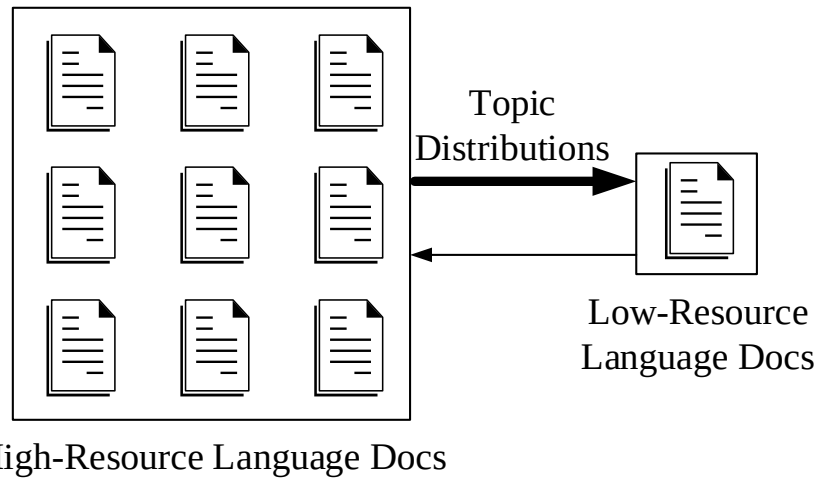


Why Weighted Topic Links?

- Transfer learned topic distributions from one language to another as prior knowledge

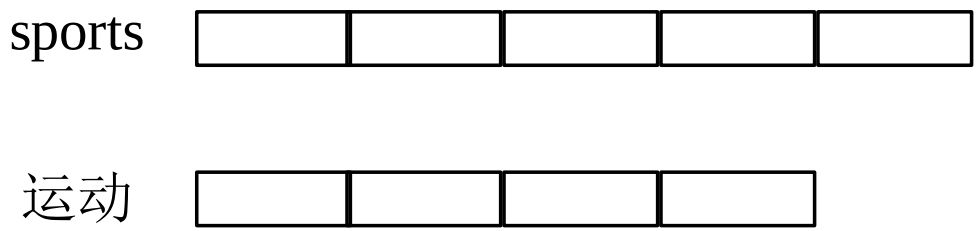


- Improve topic quality on a low-resource language with a high-resource one



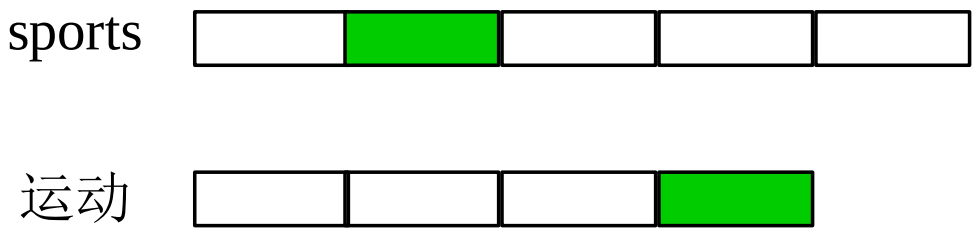
Learning Weighted Topic Links

- Weighted topic links are learned from translation pairs' topic distributions.



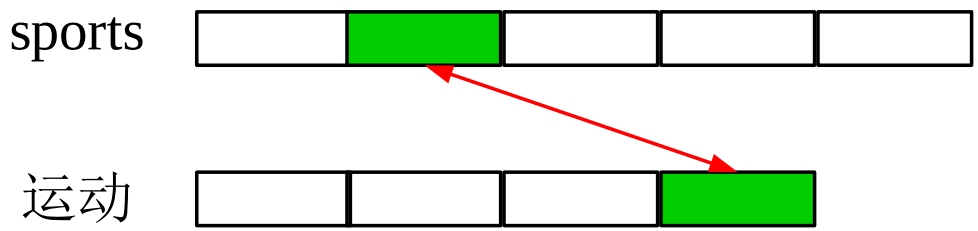
Learning Weighted Topic Links

- Weighted topic links are learned from translation pairs' topic distributions.
- For a pair of topics, if they receive high weights in the translation pair's topic distributions



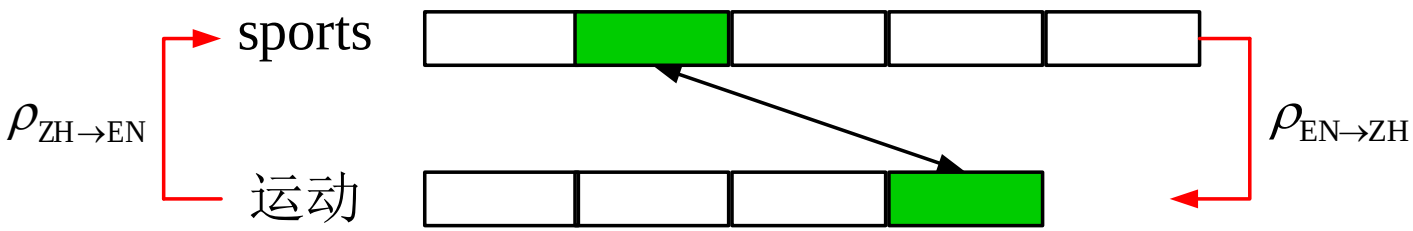
Learning Weighted Topic Links

- Weighted topic links are learned from translation pairs' topic distributions.
- For a pair of topics, if they receive high weights in the translation pair's topic distributions, they are likely to be corresponding topics.

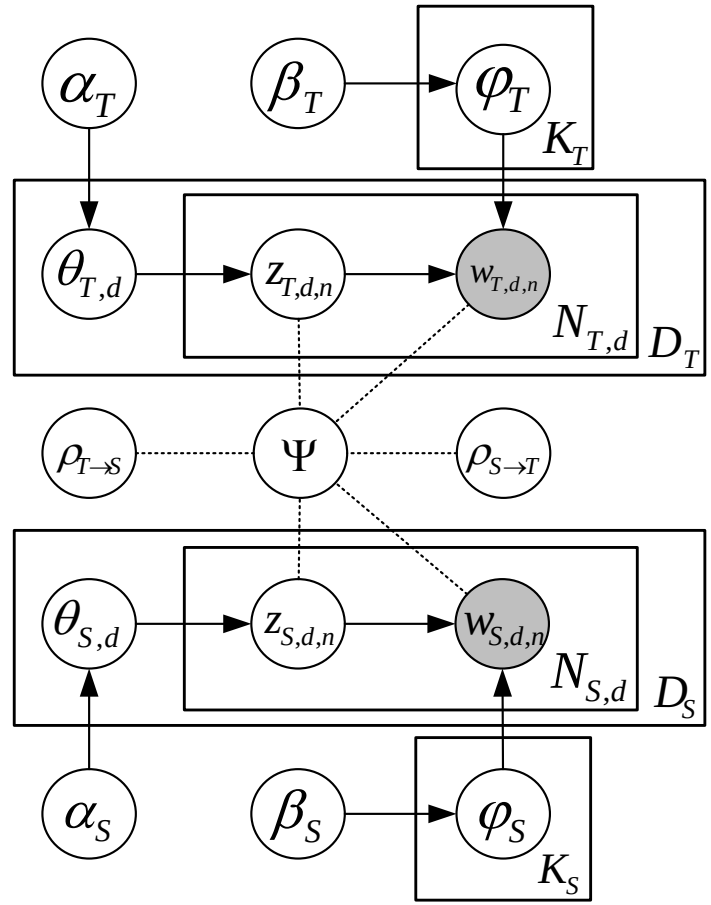


Learning Weighted Topic Links

- Weighted topic links are learned from translation pairs' topic distributions.
- For a pair of topics, if they receive high weights in the translation pair's topic distributions, they are likely to be corresponding topics.
- We use two matrices $\rho_{\text{EN}\to\text{ZH}}$ and $\rho_{\text{ZH}\to\text{EN}}$ to learn the topic relationships.

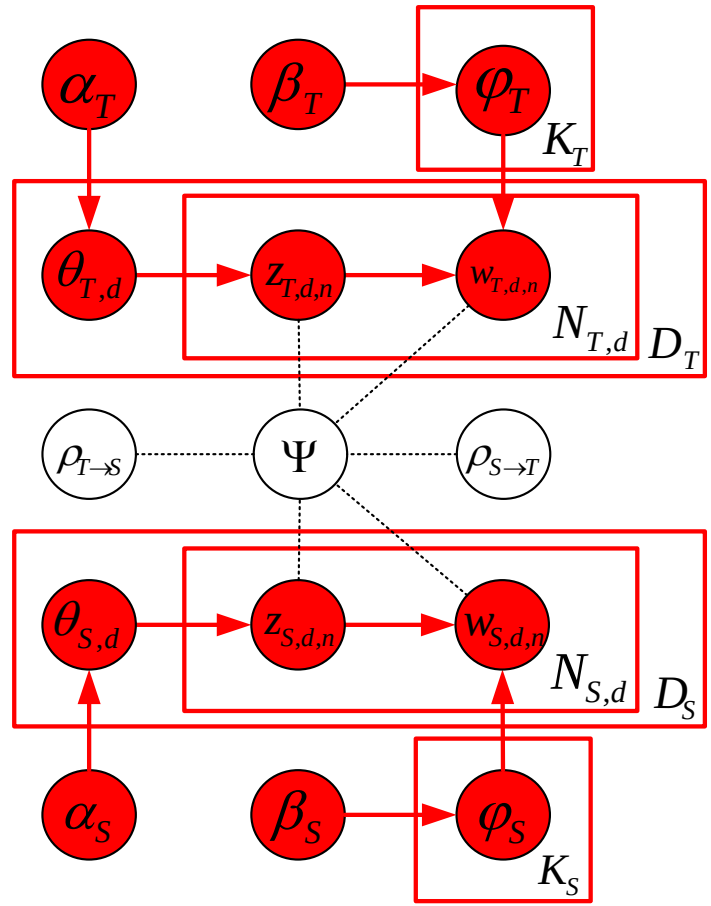


Multilingual Topic Model Details



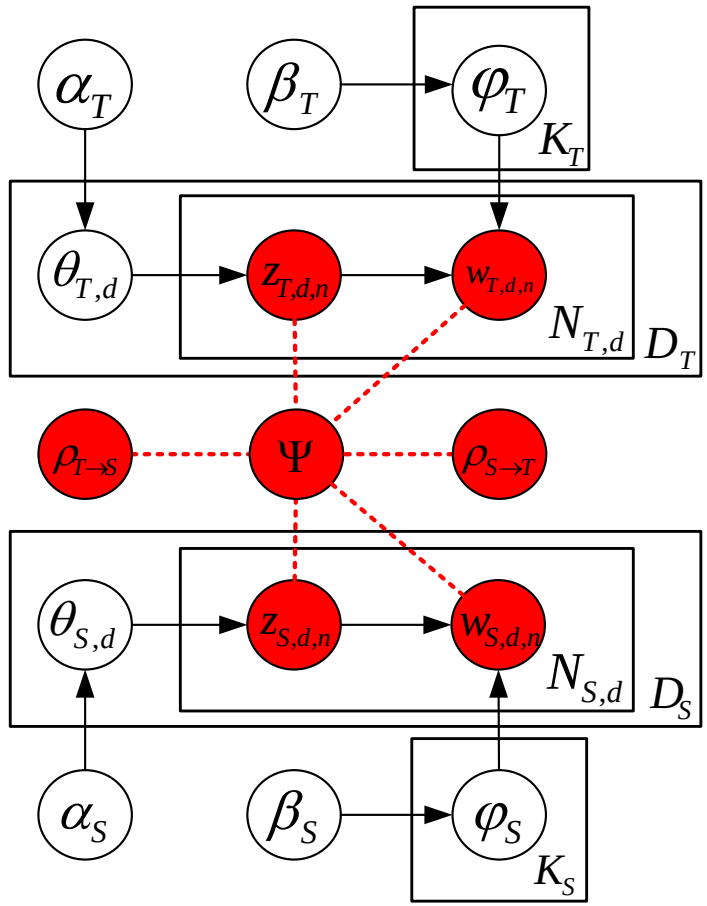
Multilingual Topic Model Details

- Two separate LDA generate the documents in languages S and T.



Multilingual Topic Model Details

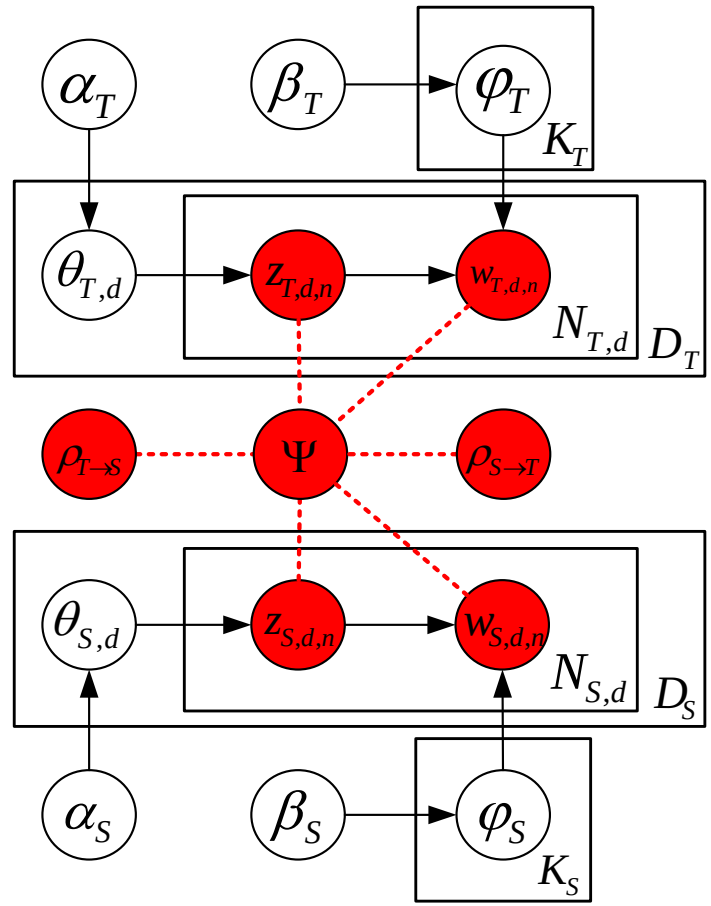
- Two separate LDA generate the documents in languages S and T.
- The posterior regularizer Ψ encodes translation information.



Multilingual Topic Model Details

- Two separate LDA generate the documents in languages S and T.
- The posterior regularizer Ψ encodes translation information.
  - We optimize Ψ to minimize the topic distribution distances of translation pairs after transformation.

$$\Psi = \left(\prod_{c=1}^{C} \left[\mathrm{Dis}\left(\boldsymbol{\Omega}_{S,c}, \boldsymbol{\rho}_{T \to S} \boldsymbol{\Omega}_{T,c} \right) \right]^{\eta_c} \right)^{-1} \times \left(\prod_{c=1}^{C} \left[\mathrm{Dis}\left(\boldsymbol{\rho}_{S \to T} \boldsymbol{\Omega}_{S,c}, \boldsymbol{\Omega}_{T,c} \right) \right]^{\eta_c} \right)^{-1} \quad (1)$$

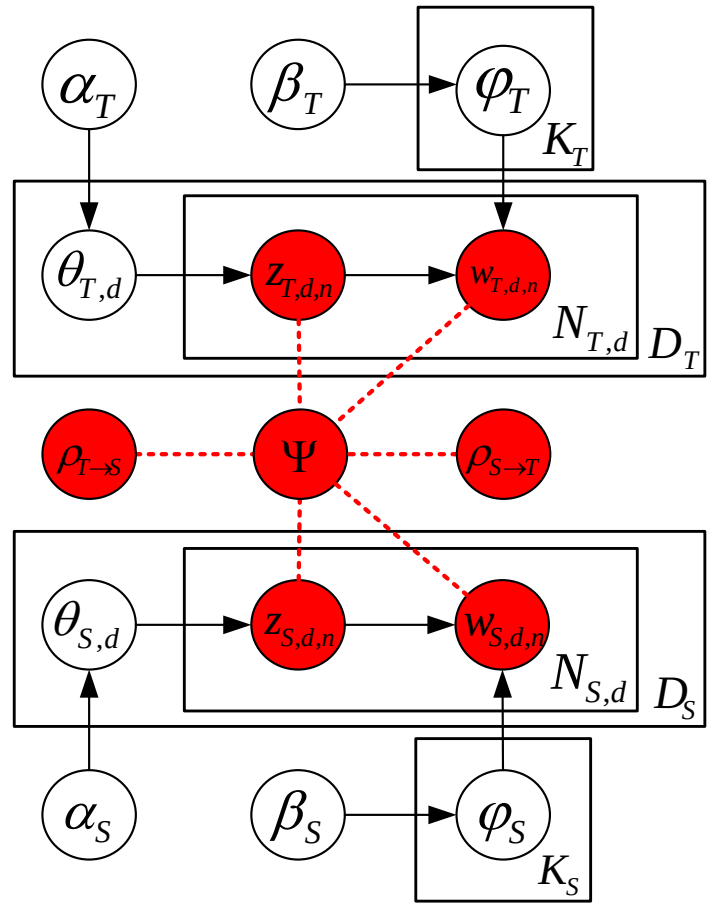


Multilingual Topic Model Details

- Two separate LDA generate the documents in languages S and T.
- The posterior regularizer Ψ encodes translation information.
  - We optimize Ψ to minimize the topic distribution distances of translation pairs after transformation.

$$\Psi = \left(\prod_{c=1}^{C} [\mathrm{Dis}\,(\boldsymbol{\Omega}_{S,c}, \boldsymbol{\rho}_{T\to S}\boldsymbol{\Omega}_{T,c})]^{\eta_c} \right)^{-1} \times \left(\prod_{c=1}^{C} [\mathrm{Dis}\,(\boldsymbol{\rho}_{S\to T}\boldsymbol{\Omega}_{S,c}, \boldsymbol{\Omega}_{T,c})]^{\eta_c} \right)^{-1} \quad (1)$$

  - The distance function can be Euclidean distance, KL divergence, etc.

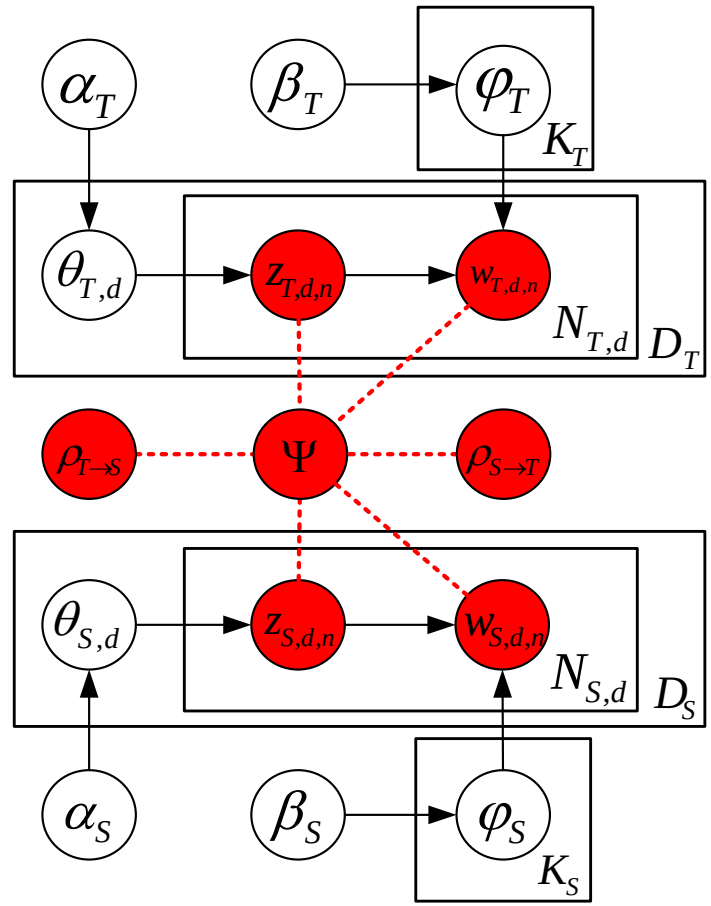


Multilingual Topic Model Details

- Two separate LDA generate the documents in languages S and T.
- The posterior regularizer Ψ encodes translation information.
  - We optimize Ψ to minimize the topic distribution distances of translation pairs after transformation.

$$\Psi = \left(\prod_{c=1}^{C} [\mathrm{Dis}\,(\boldsymbol{\Omega}_{S,c}, \boldsymbol{\rho}_{T\to S}\boldsymbol{\Omega}_{T,c})]^{\eta_c} \right)^{-1} \times \left(\prod_{c=1}^{C} [\mathrm{Dis}\,(\boldsymbol{\rho}_{S\to T}\boldsymbol{\Omega}_{S,c}, \boldsymbol{\Omega}_{T,c})]^{\eta_c} \right)^{-1} \quad (1)$$

  - The distance function can be Euclidean distance, KL divergence, etc.
  - Each translation pair can also be weighted.



Classification

- Datasets
  - Wikipedia (English/Chinese): Six-class classification of document categories
  - LORELEI (English/Sinhalese): Binary classification of need types

Classification

- Datasets
  - Wikipedia (English/Chinese): Six-class classification of document categories
  - LORELEI (English/Sinhalese): Binary classification of need types
- Baselines
  - LDA which runs monolingually (Blei et al., 2003)
  - Multilingual Cultural-common Topic Analysis (Shi et al., 2016, MCTA)
  - Multilingual Anchoring (Yuan et al., 2018, MTAnchor)
  - Tree LDA with tree priors of a word translation dictionary (Hu et al., 2014, tLDA)

Classification

- Datasets
  - Wikipedia (English/Chinese): Six-class classification of document categories
  - LORELEI (English/Sinhalese): Binary classification of need types
- Baselines
  - LDA which runs monolingually (Blei et al., 2003)
  - Multilingual Cultural-common Topic Analysis (Shi et al., 2016, MCTA)
  - Multilingual Anchoring (Yuan et al., 2018, MTAnchor)
  - Tree LDA with tree priors of a word translation dictionary (Hu et al., 2014, tLDA)
- All multilingual baselines assume aligned topic spaces.

Classification

- Datasets
  - Wikipedia (English/Chinese): Six-class classification of document categories
  - LORELEI (English/Sinhalese): Binary classification of need types
- Baselines
  - LDA which runs monolingually (Blei et al., 2003)
  - Multilingual Cultural-common Topic Analysis (Shi et al., 2016, MCTA)
  - Multilingual Anchoring (Yuan et al., 2018, MTAnchor)
  - Tree LDA with tree priors of a word translation dictionary (Hu et al., 2014, tLDA)
- All multilingual baselines assume aligned topic spaces.
- Translation pair weighting
  - Equal weights
  - TF-IDF weights

Intra-Lingual Classification Results

- Train and test classifiers on the same language.

Dataset	Method	EN	SI/ZH
LORELEI	MCTA	12.99	26.53
	MTAnchor	20.78	32.65
	LDA	27.78	24.01
	tLDA	12.77	18.18
	MTM	**42.86**	23.08
	MTM + TF-IDF	26.67	**38.10**
Wikipedia	MCTA	51.56	33.35
	MTAnchor	80.71	75.33
	LDA	92.08	83.37
	tLDA	91.58	83.33
	MTM	92.98	**86.48**
	MTM + TF-IDF	**94.07**	85.59

Intra-Lingual Classification Results

- Train and test classifiers on the same language.

Dataset	Method	EN	SI/ZH
LORELEI	MCTA	12.99	26.53
	MTAnchor	20.78	32.65
	LDA	27.78	24.01
	tLDA	12.77	18.18
	MTM	**42.86**	23.08
	MTM + TF-IDF	26.67	**38.10**
Wikipedia	MCTA	51.56	33.35
	MTAnchor	80.71	75.33
	LDA	92.08	83.37
	tLDA	91.58	83.33
	MTM	92.98	**86.48**
	MTM + TF-IDF	**94.07**	85.59

Cross-Lingual Classification Results

- Classifies on another language

Cross-Lingual Classification Results

- Classifies on another language
- We try two methods for our MTM to transform topic spaces.

Cross-Lingual Classification Results

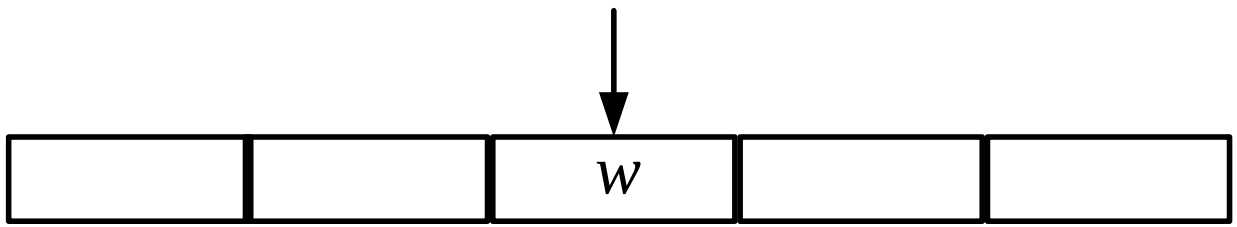
- Classifies on another language
- We try two methods for our MTM to transform topic spaces.
  - Multiply ρ's with topic distributions

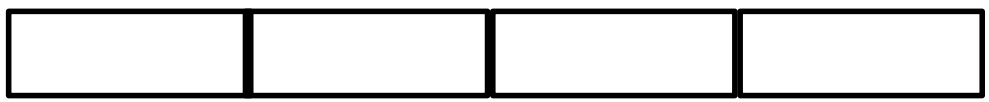
Cross-Lingual Classification Results

- Classifies on another language
- We try two methods for our MTM to transform topic spaces.
  - Multiply ρ's with topic distributions
  - Transfer a topic's weight to the topic in the other language with the highest topic link weight (TOP)

Cross-Lingual Classification Results

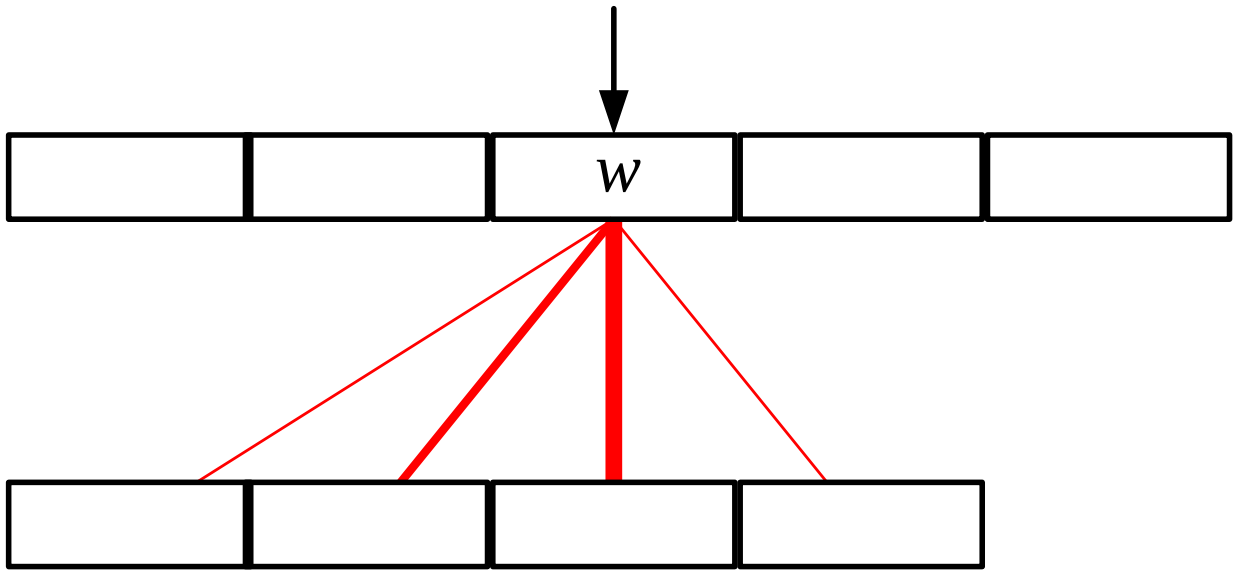
- Classifies on another language
- We try two methods for our MTM to transform topic spaces.
  - Multiply ρ's with topic distributions
  - Transfer a topic's weight to the topic in the other language with the highest topic link weight (TOP)





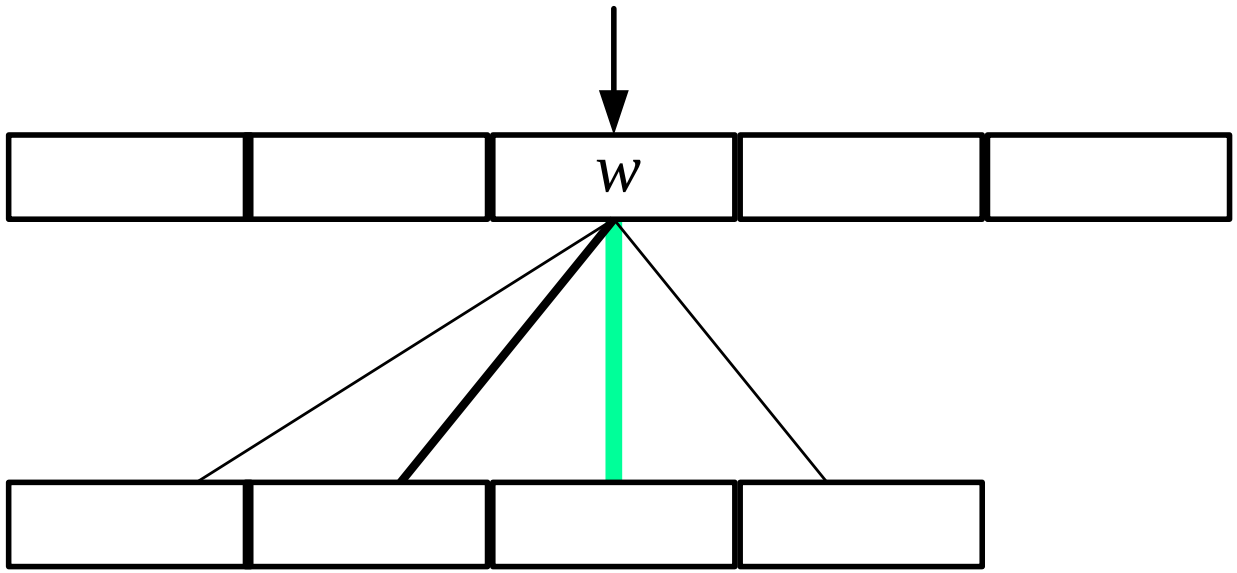
Cross-Lingual Classification Results

- Classifies on another language
- We try two methods for our MTM to transform topic spaces.
  - Multiply ρ's with topic distributions
  - Transfer a topic's weight to the topic in the other language with the highest topic link weight (TOP)



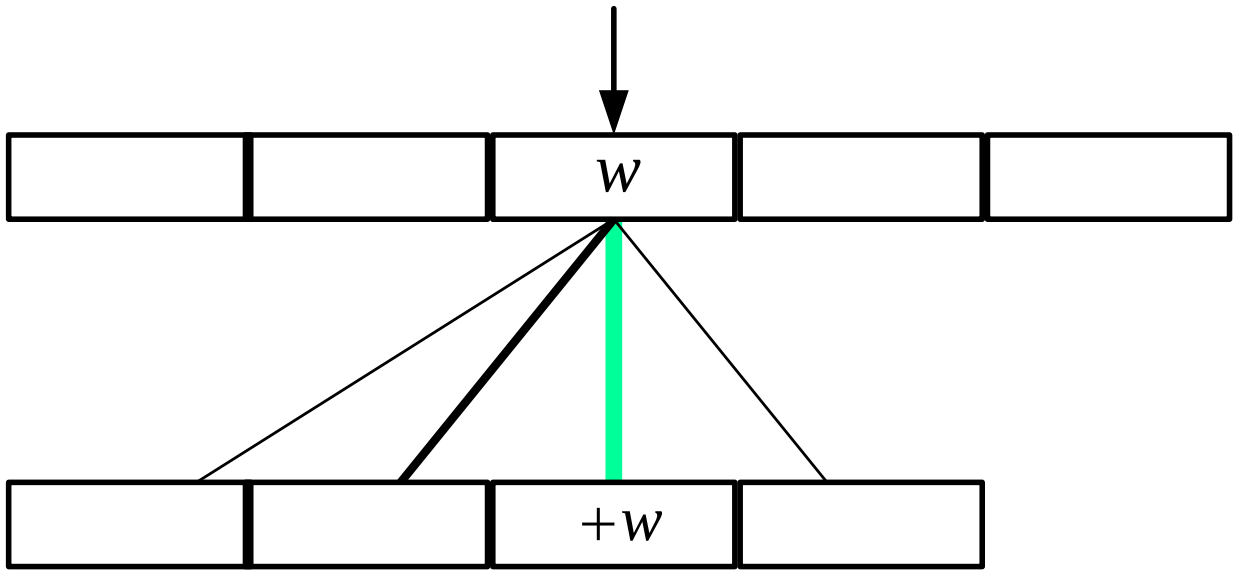
Cross-Lingual Classification Results

- Classifies on another language
- We try two methods for our MTM to transform topic spaces.
  - Multiply ρ's with topic distributions
  - Transfer a topic's weight to the topic in the other language with the highest topic link weight (TOP)



Cross-Lingual Classification Results

- Classifies on another language
- We try two methods for our MTM to transform topic spaces.
  - Multiply ρ's with topic distributions
  - Transfer a topic's weight to the topic in the other language with the highest topic link weight (TOP)



Cross-Lingual Classification Results

- Classifies on another language

Dataset	Method	EN	SI/ZH
LORELEI	MCTA	4.08	15.58
	MTAnchor	24.49	24.68
	LDA	22.86	21.05
	tLDA	16.01	15.09
	MTM	22.22	26.67
	MTM + TOP	**35.29**	**33.33**
	MTM + TF-IDF	14.46	15.09
	MTM + TF-IDF + TOP	14.46	11.43
Wikipedia	MCTA	23.24	39.79
	MTAnchor	57.62	54.54
	LDA	16.52	10.46
	tLDA	2.85	21.02
	MTM	74.69	64.48
	MTM + TOP	**78.13**	**83.08**
	MTM + TF-IDF	57.27	55.06
	MTM + TF-IDF + TOP	63.20	59.64

Cross-Lingual Classification Results

- Classifies on another language

Dataset	Method	EN	SI/ZH
LORELEI	MCTA	4.08	15.58
	MTAnchor	24.49	24.68
	LDA	22.86	21.05
	tLDA	16.01	15.09
	MTM	22.22	26.67
	MTM + TOP	**35.29**	**33.33**
	MTM + TF-IDF	14.46	15.09
	MTM + TF-IDF + TOP	14.46	11.43
Wikipedia	MCTA	23.24	39.79
	MTAnchor	57.62	54.54
	LDA	16.52	10.46
	tLDA	2.85	21.02
	MTM	74.69	64.48
	MTM + TOP	**78.13**	**83.08**
	MTM + TF-IDF	57.27	55.06
	MTM + TF-IDF + TOP	63.20	59.64

Cross-Lingual Classification Results

- Classifies on another language

Dataset	Method	EN	SI/ZH
LORELEI	MCTA	4.08	15.58
	MTAnchor	24.49	24.68
	LDA	22.86	21.05
	tLDA	16.01	15.09
	MTM	22.22	26.67
	MTM + TOP	**35.29**	**33.33**
	MTM + TF-IDF	14.46	15.09
	MTM + TF-IDF + TOP	14.46	11.43
Wikipedia	MCTA	23.24	39.79
	MTAnchor	57.62	54.54
	LDA	16.52	10.46
	tLDA	2.85	21.02
	MTM	74.69	64.48
	MTM + TOP	**78.13**	**83.08**
	MTM + TF-IDF	57.27	55.06
	MTM + TF-IDF + TOP	63.20	59.64

Selected Topics

Model	Lang.	Words
MCTA	ZH	主演 (starring), 改编 (adapt), 本 (this), 小说 (novel), 拍摄 (shoot), 角色 (role), 战士 (fighter)
	EN	dog, san, movie, mexican, fighter, novel, california
MTAnchor	ZH	主演 (starring), 改编 (adapt), 饰演 (act), 本片 (this movie), 演员 (actor), 编剧 (playwright), 讲述 (narrate)
	EN	kong, hong, movie, official, martial, box, reception
LDA	ZH	电影 (movie), 部 (movie quantifier), 美国 (USA), 上映 (release), 英语 (English), 剧情 (plot), 片 (movie)
	EN	film, star, direct, release, action, plot, character
tLDA	ZH	电影 (movie), 胶片 (film), 星 (star), 动作 (action), 释放 (release), 影片 (movie), 剧情 (plot)
	EN	film, star, direct, action, release, plot, write
MTM	ZH	电影 (movie), 部 (movie quantifier), 上映 (release), 动画 (animation), 故事 (story), 作品 (works), 英语 (English)
	EN	film, direct, star, release, action, plot, production
MTM + TF-IDF	ZH	电影 (movie), 部 (movie quantifier), 上映 (release), 美国 (USA), 英语 (English), 导演 (director), 片 (movie)
	EN	film, direct, star, action, release, plot, movie

Selected Topics

Model	Lang.	Words
MCTA	ZH	主演 (starring), 改编 (adapt), 本 (this), 小说 (novel), 拍摄 (shoot), 角色 (role), 战士 (fighter)
	EN	dog, san, movie, mexican, fighter, novel, california
MTAnchor	ZH	主演 (starring), 改编 (adapt), 饰演 (act), 本片 (this movie), 演员 (actor), 编剧 (playwright), 讲述 (narrate)
	EN	kong, hong, movie, official, martial, box, reception
LDA	ZH	电影 (movie), 部 (movie quantifier), 美国 (USA), 上映 (release), 英语 (English), 剧情 (plot), 片 (movie)
	EN	film, star, direct, release, action, plot, character
tLDA	ZH	电影 (movie), 胶片 (film), 星 (star), 动作 (action), 释放 (release), 影片 (movie), 剧情 (plot)
	EN	film, star, direct, action, release, plot, write
MTM	ZH	电影 (movie), 部 (movie quantifier), 上映 (release), 动画 (animation), 故事 (story), 作品 (works), 英语 (English)
	EN	film, direct, star, release, action, plot, production
MTM + TF-IDF	ZH	电影 (movie), 部 (movie quantifier), 上映 (release), 美国 (USA), 英语 (English), 导演 (director), 片 (movie)
	EN	film, direct, star, action, release, plot, movie

Selected Topics

Model	Lang.	Words
MCTA	ZH	主演 (starring), 改编 (adapt), 本 (this), 小说 (novel), 拍摄 (shoot), 角色 (role), 战士 (fighter)
	EN	dog, san, movie, mexican, fighter, novel, california
MTAnchor	ZH	主演 (starring), 改编 (adapt), 饰演 (act), 本片 (this movie), 演员 (actor), 编剧 (playwright), 讲述 (narrate)
	EN	kong, hong, movie, official, martial, box, reception
LDA	ZH	电影 (movie), 部 (movie quantifier), 美国 (USA), 上映 (release), 英语 (English), 剧情 (plot), 片 (movie)
	EN	film, star, direct, release, action, plot, character
tLDA	ZH	电影 (movie), 胶片 (film), 星 (star), 动作 (action), 释放 (release), 影片 (movie), 剧情 (plot)
	EN	film, star, direct, action, release, plot, write
MTM	ZH	电影 (movie), 部 (movie quantifier), 上映 (release), 动画 (animation), 故事 (story), 作品 (works), 英语 (English)
	EN	film, direct, star, release, action, plot, production
MTM + TF-IDF	ZH	电影 (movie), 部 (movie quantifier), 上映 (release), 美国 (USA), 英语 (English), 导演 (director), 片 (movie)
	EN	film, direct, star, action, release, plot, movie

Selected Topics

Model	Lang.	Words
MCTA	ZH	主演 (starring), 改编 (adapt), 本 (this), 小说 (novel), 拍摄 (shoot), 角色 (role), 战士 (fighter)
	EN	dog, san, movie, mexican, fighter, novel, california
MTAnchor	ZH	主演 (starring), 改编 (adapt), 饰演 (act), 本片 (this movie), 演员 (actor), 编剧 (playwright), 讲述 (narrate)
	EN	kong, hong, movie, official, martial, box, reception
LDA	ZH	电影 (movie), 部 (movie quantifier), 美国 (USA), 上映 (release), 英语 (English), 剧情 (plot), 片 (movie)
	EN	film, star, direct, release, action, plot, character
tLDA	ZH	电影 (movie), 胶片 (film), 星 (star), 动作 (action), 释放 (release), 影片 (movie), 剧情 (plot)
	EN	film, star, direct, action, release, plot, write
MTM	ZH	电影 (movie), 部 (movie quantifier), 上映 (release), 动画 (animation), 故事 (story), 作品 (works), 英语 (English)
	EN	film, direct, star, release, action, plot, production
MTM + TF-IDF	ZH	电影 (movie), 部 (movie quantifier), 上映 (release), 美国 (USA), 英语 (English), 导演 (director), 片 (movie)
	EN	film, direct, star, action, release, plot, movie

Selected Topic Links

Lang.	Weight	Words
ZH-0	–	学名 (scientific name), 它们 (they), 呈 (show), 白色 (white), 长 (long), 黑色 (black), 厘米 (centimeter)
EN-12	0.57	specie, bird, eagle, genus, white, owl, black
EN-19	0.13	breed, chicken, white, goose, bird, black, list
EN-10	–	album, release, record, music, song, single, feature
ZH-9	0.30	专辑 (album), 张 (album quantifier), 发行 (release), 音乐 (music), 首 (song quantifier), 唱片 (record), 歌手 (singer)
ZH-17	0.20	音乐 (music), 乐团 (musical group), 艺术 (art), 创作 (create), 奖 (prize), 演出 (perform), 担任 (serve)
ZH-14	–	主义 (-ism), 组织 (organization), 美国 (USA), 革命 (evolution), 运动 (campaign), 政府 (government), 人民 (people)
EN-16	0.32	sex, law, act, sexual, marriage, court, legal
EN-11	0.17	traffic, victim, government, trafficking, child, force, country

Selected Topic Links

- Most topics are linked based on mutual top words.

Lang.	Weight	Words
ZH-0	–	学名 (scientific name), 它们 (they), 呈 (show), 白色 (white), 长 (long), 黑色 (black), 厘米 (centimeter)
EN-12	0.57	specie, bird, eagle, genus, white, owl, black
EN-19	0.13	breed, chicken, white, goose, bird, black, list
EN-10	–	album, release, record, music, song, single, feature
ZH-9	0.30	专辑 (album), 张 (album quantifier), 发行 (release), 音乐 (music), 首 (song quantifier), 唱片 (record), 歌手 (singer)
ZH-17	0.20	音乐 (music), 乐团 (musical group), 艺术 (art), 创作 (create), 奖 (prize), 演出 (perform), 担任 (serve)
ZH-14	–	主义 (-ism), 组织 (organization), 美国 (USA), 革命 (evolution), 运动 (campaign), 政府 (government), 人民 (people)
EN-16	0.32	sex, law, act, sexual, marriage, court, legal
EN-11	0.17	traffic, victim, government, trafficking, child, force, country

Selected Topic Links

- Most topics are linked based on mutual top words.
- For some topics, our MTM can even learn the links beyond words.

Lang.	Weight	Words
ZH-0	–	学名 (scientific name), 它们 (they), 呈 (show), 白色 (white), 长 (long), 黑色 (black), 厘米 (centimeter)
EN-12	0.57	specie, bird, eagle, genus, white, owl, black
EN-19	0.13	breed, chicken, white, goose, bird, black, list
EN-10	–	album, release, record, music, song, single, feature
ZH-9	0.30	专辑 (album), 张 (album quantifier), 发行 (release), 音乐 (music), 首 (song quantifier), 唱片 (record), 歌手 (singer)
ZH-17	0.20	音乐 (music), 乐团 (musical group), 艺术 (art), 创作 (create), 奖 (prize), 演出 (perform), 担任 (serve)
ZH-14	–	主义 (-ism), 组织 (organization), 美国 (USA), 革命 (evolution), 运动 (campaign), 政府 (government), 人民 (people)
EN-16	0.32	sex, law, act, sexual, marriage, court, legal
EN-11	0.17	traffic, victim, government, trafficking, child, force, country

Topic Coherence on Low-Comparability Data

- Bilingual Wikipedia corpora
  - English
  - Arabic, Chinese, Spanish, Farsi, and Russian

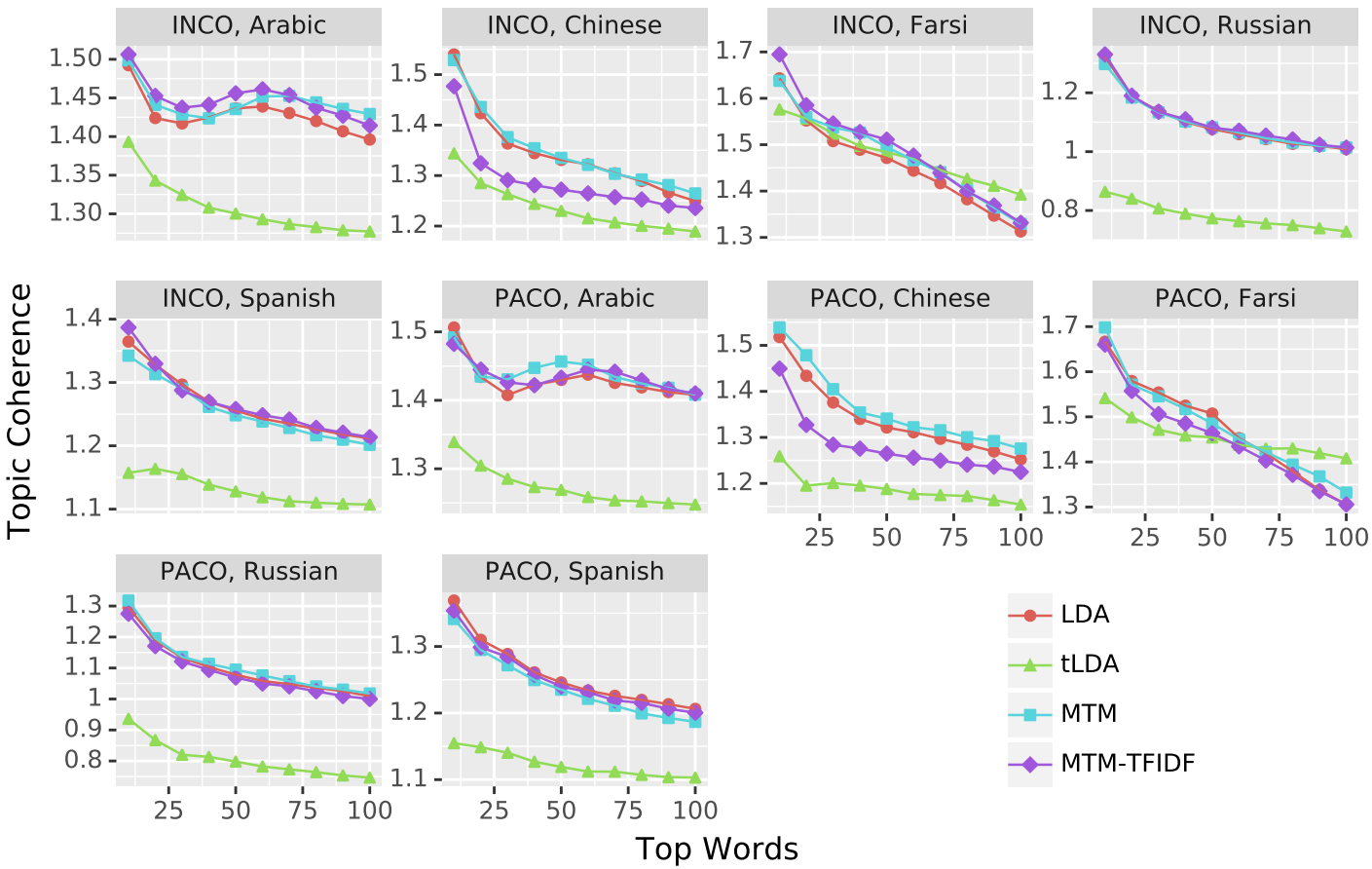
Topic Coherence on Low-Comparability Data

- Bilingual Wikipedia corpora
  - English
  - Arabic, Chinese, Spanish, Farsi, and Russian
- Each language pair has two corpora.
  - Partially comparable (PACO): 30% documents have direct translations in the other language.
  - Incomparable (INCO): No documents have direct translations.

Topic Coherence on Low-Comparability Data

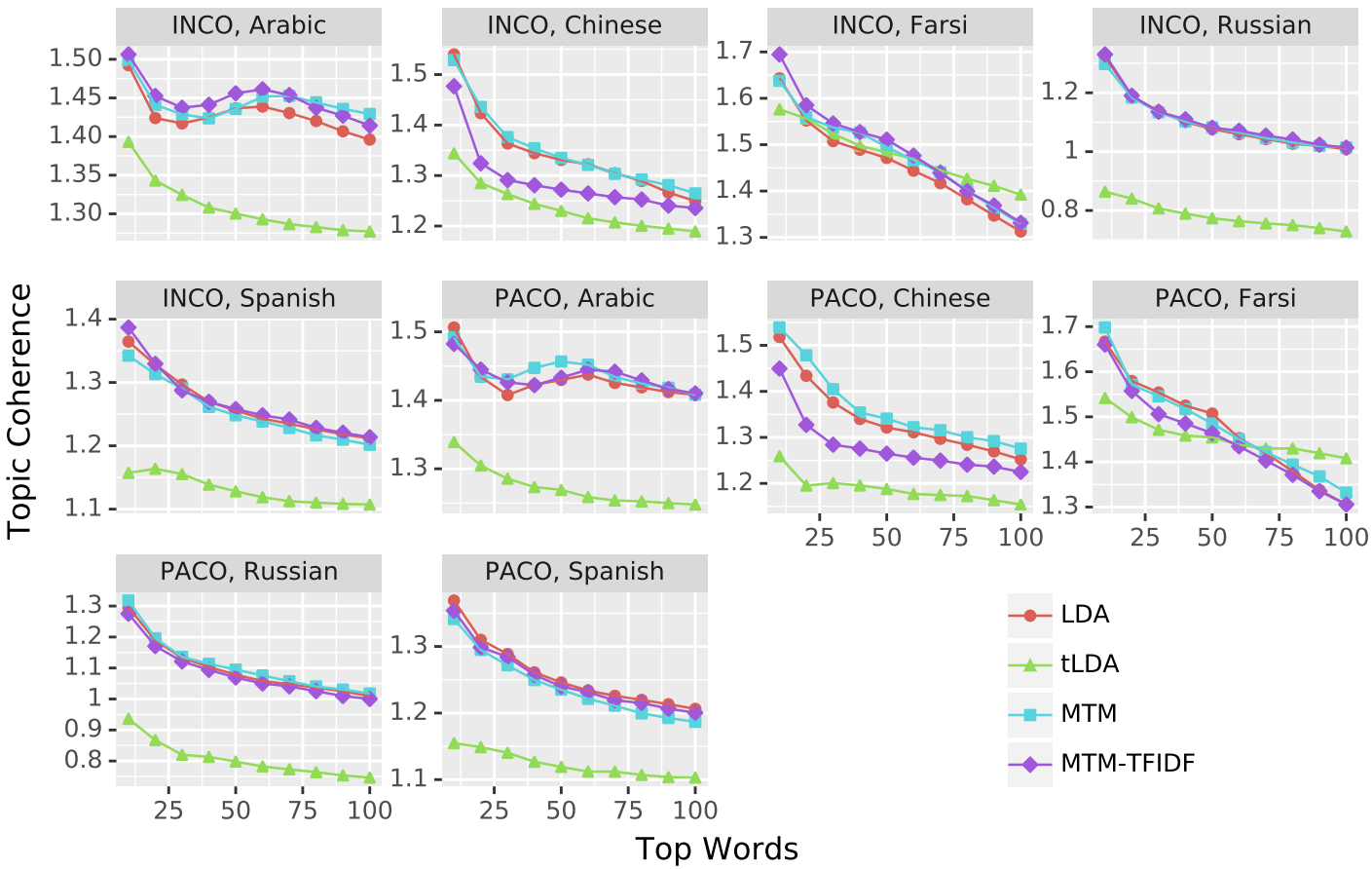
- Bilingual Wikipedia corpora
  - English
  - Arabic, Chinese, Spanish, Farsi, and Russian
- Each language pair has two corpora.
  - Partially comparable (PACO): 30% documents have direct translations in the other language.
  - Incomparable (INCO): No documents have direct translations.
- Baselines
  - Monolingual LDA
  - tLDA with tree priors of a word translation dictionary

Topic Coherence Results on Low-Comparability Data



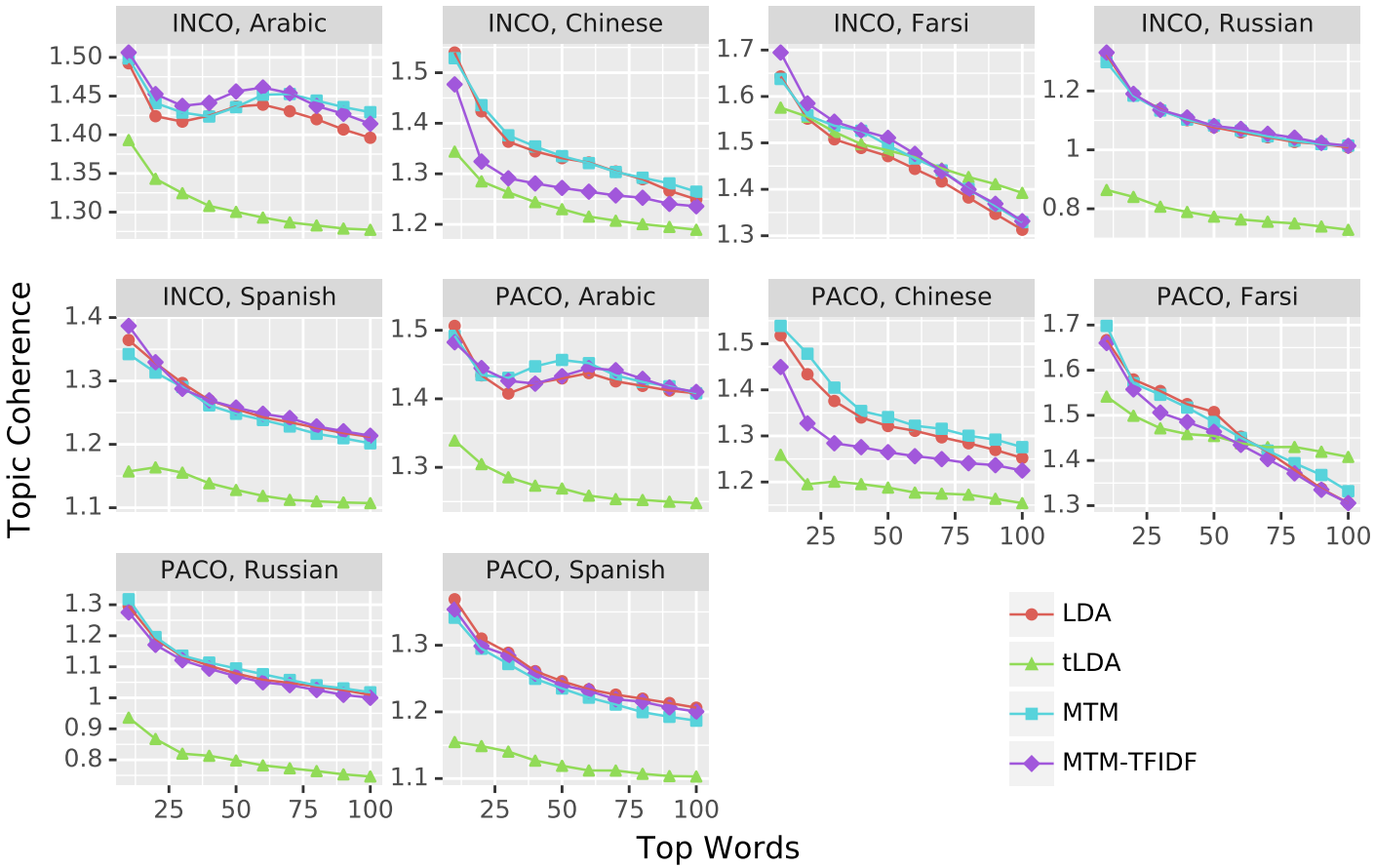
Topic Coherence Results on Low-Comparability Data

- MTM mostly performs as well as monolingual LDA.
  - Proves MTM's robustness on low-comparability data.



Topic Coherence Results on Low-Comparability Data

- MTM mostly performs as well as monolingual LDA.
  - Proves MTM's robustness on low-comparability data.
- tLDA sacrifices topic coherence for topic alignment.



Summary

- A multilingual topic model for learning weighted topic links
  - Does not force topic alignment and only connects topics when necessary
  - Improves classification performance both intra- and cross-lingually using the topic posteriors as features
  - Gives coherent topics and meaningful topic links
  - Robust when the data are less comparable or incomparable

Thanks

Collaborators

- Jordan Boyd-Graber (UMD/Google AI Zürich)
- Philip Resnik (UMD)

Funders

Raytheon
BBN Technologies

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *JMLR*, pages 993–1022.

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

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

Michelle Yuan, Benjamin Van Durme, and Jordan Boyd-Graber. 2018. Multilingual anchoring: Interactive topic modeling and alignment across languages. In *NIPS*.