It's all Domain Adaptation: (Cross-lingual) Stance Detection and What We're Missing

Emily Allaway

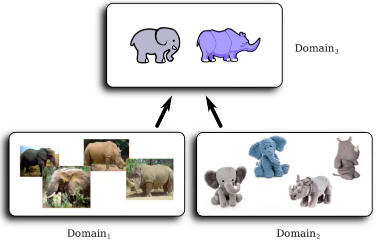
Candidacy Exam
Columbia University

Feb. 24 2021

Contact: eallaway@cs.columbia.edu

Introduction

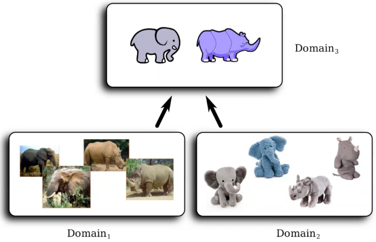
Tasks:



(Image credit: Baldwin (2021))

Introduction

Tasks:



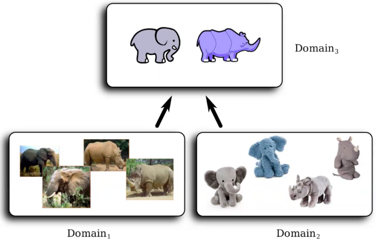
(Image credit: Baldwin (2021))

Topic: immigration **Stance:** against

Text: The jury's verdict will ensure that another violent criminal alien will be removed from our community for a very long period . . .

Introduction

Tasks:



(Image credit: Baldwin (2021))

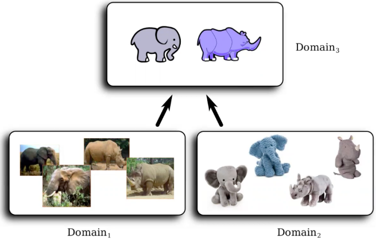
Topic: immigration **Stance:** against

Text: The jury's verdict will ensure that another violent criminal alien will be removed from our community for a very long period . . .

What is a **domain**?

Introduction

Tasks:



(Image credit: Baldwin (2021))

Topic: immigration **Stance:** against

Text: The jury's verdict will ensure that another violent criminal alien will be removed from our community for a very long period . . .

What is a **domain**?

Plank (2016)

- Variety: space of latent factors
- Domain: region in the variety

Introduction

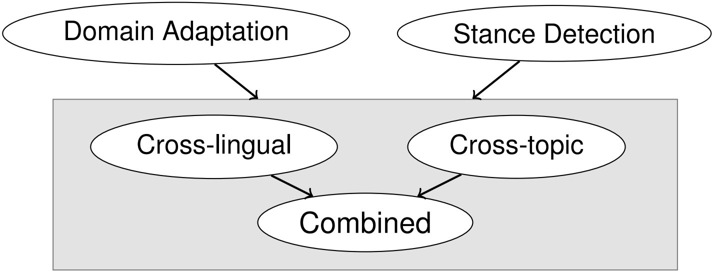
Domain Adaptation

Introduction

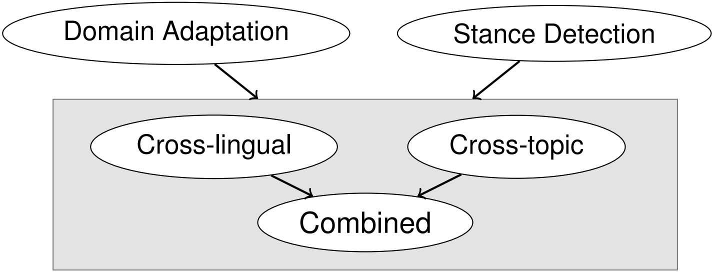
Domain Adaptation

Stance Detection

Introduction

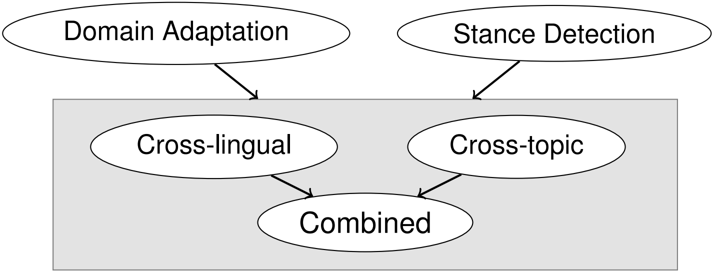


Introduction



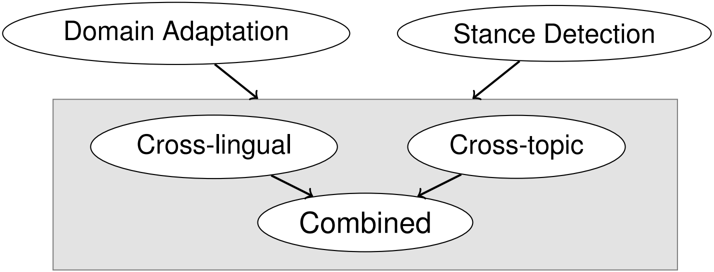
Why domain adaptation (DA)?

Introduction



Why domain adaptation (DA)?
$\Longrightarrow$ (arguably) *the* generalization task

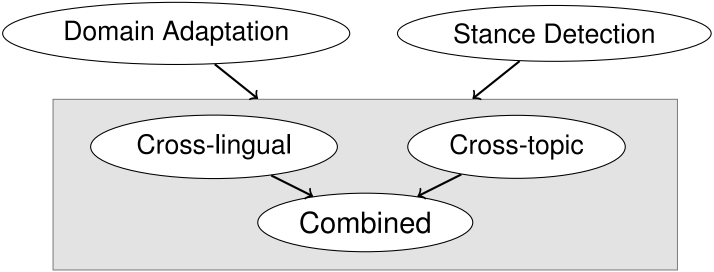
Introduction



Why domain adaptation (DA)?
$\Longrightarrow$ (arguably) *the* generalization task

Why stance detection?

Introduction



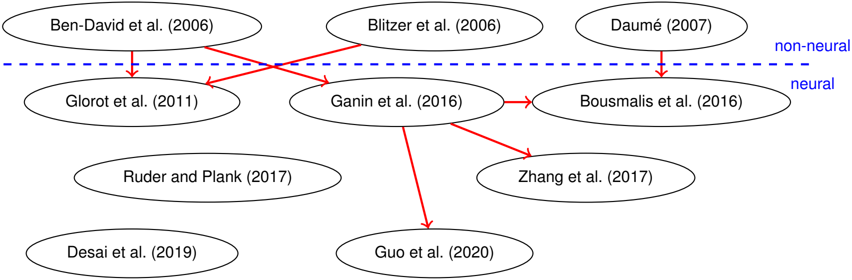
Why domain adaptation (DA)?
$\Longrightarrow$ (arguably) *the* generalization task

Why stance detection?
$\Longrightarrow$ great setting for exploring generalization along *many* latent dimensions

Outline

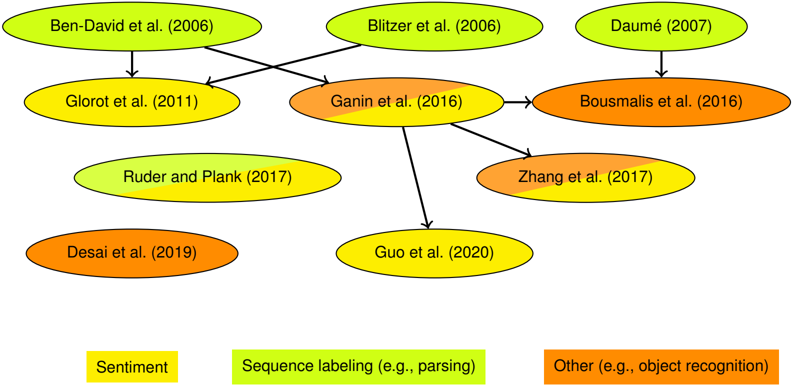
1. Domain Adaptation
2. Stance Detection
3. Domain Adaptation + Stance Detection

Domain adaptation: a sample



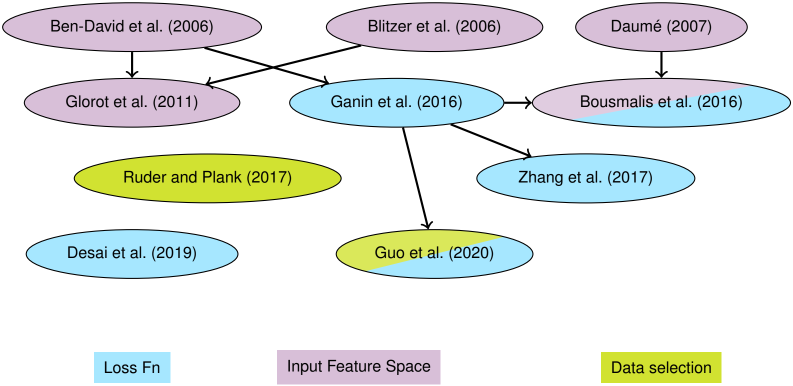
Domain adaptation: a sample

Tasks:



Domain adaptation: a sample

Adaptation method (modifying):



Early work on domain adaptation

- 2 types of features:
  1. shared *(domain-general)*
  2. private *(domain-specific)*

Early work on domain adaptation

- 2 types of features:
  1. shared *(domain-general)*
  2. private *(domain-specific)*
- Shared
  - Shared words, projected into a lower-dim space (Blitzer et al., 2006)
  - Extra copy of an instance in a particular location in the feature vector (Daumé, 2007)

$$\langle \text{shared}, \text{src}, \text{target} \rangle : \quad \langle \boldsymbol{x}, \boldsymbol{x}, 0 \rangle \quad \text{vs.} \quad \langle \boldsymbol{x}, 0, \boldsymbol{x} \rangle$$

Early work on domain adaptation

- 2 types of features:
  1. shared *(domain-general)*
  2. private *(domain-specific)*
- Shared
  - Shared words, projected into a lower-dim space (Blitzer et al., 2006)
  - Extra copy of an instance in a particular location in the feature vector (Daumé, 2007)

$$\langle \text{shared}, \text{src}, \text{target} \rangle : \quad \langle \boldsymbol{x}, \boldsymbol{x}, 0 \rangle \quad \text{vs.} \quad \langle \boldsymbol{x}, 0, \boldsymbol{x} \rangle$$

- Private
  - Copy the original features (for now)
  - Preserve domain-specific information (e.g., aspects of a specific product type)

Domain adaptation theory

What makes good **representations**?

Domain adaptation theory

What makes good **representations**?

Fix:
$\mathcal{H}$ and representation function $\mathcal{R}$ and apply $\mathcal{R}$ to an i.i.d sample of instances.

Domain adaptation theory

What makes good **representations**?

Fix:
$\mathcal{H}$ and representation function $\mathcal{R}$ and apply $\mathcal{R}$ to an i.i.d sample of instances.

Then with probability at least $1 - \delta$ for every $h \in \mathcal{H}$ (Ben-David et al., 2006)

$$\epsilon_T(h) \leq \hat{\epsilon}_S(h) + d_{\mathcal{H}}(\tilde{U}_S, \tilde{U}_T) + V$$

ϵ := error
$\tilde{U}$:= unlabeled data
$\mathcal{H}$:= Hypothesis class
V := other terms

Domain adaptation theory

What makes good **representations**?

Fix:
$\mathcal{H}$ and representation function $\mathcal{R}$ and apply $\mathcal{R}$ to an i.i.d sample of instances.

Then with probability at least $1 - \delta$ for every $h \in \mathcal{H}$ (Ben-David et al., 2006)

$$\epsilon_T(h) \leq \hat{\epsilon}_S(h) + d_{\mathcal{H}}(\tilde{U}_S, \tilde{U}_T) + V$$

- Target error depends on:

ϵ := error
$\tilde{U}$:= unlabeled data
$\mathcal{H}$:= Hypothesis class
V := other terms

Domain adaptation theory

What makes good **representations**?

Fix:
$\mathcal{H}$ and representation function $\mathcal{R}$ and apply $\mathcal{R}$ to an i.i.d sample of instances.

Then with probability at least $1-\delta$ for every $h \in \mathcal{H}$ (Ben-David et al., 2006)

$$\epsilon_T(h) \leq \hat{\epsilon}_S(h) + d_{\mathcal{H}}(\tilde{U}_S, \tilde{U}_T) + V$$

- Target error depends on:
  - src error $\hat{\epsilon}_S$

ϵ := error
$\tilde{U}$:= unlabeled data
$\mathcal{H}$:= Hypothesis class
V := other terms

Domain adaptation theory

What makes good **representations**?

Fix:
$\mathcal{H}$ and representation function $\mathcal{R}$ and apply $\mathcal{R}$ to an i.i.d sample of instances.

Then with probability at least $1 - \delta$ for every $h \in \mathcal{H}$ (Ben-David et al., 2006)

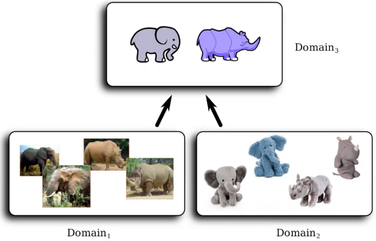
$$\epsilon_T(h) \leq \hat{\epsilon}_S(h) + d_{\mathcal{H}}(\tilde{U}_S, \tilde{U}_T) + V$$

- Target error depends on:
  - src error $\hat{\epsilon}_S$
  - domain distance $d_{\mathcal{H}}$

ϵ := error
$\tilde{U}$:= unlabeled data
$\mathcal{H}$:= Hypothesis class
V := other terms

Domain adaptation theory

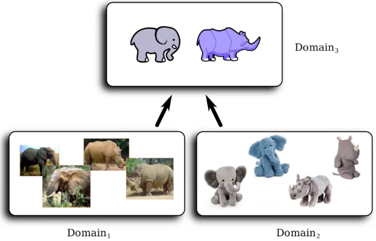
What is domain distance $d_{\mathcal{H}}$?



Domain adaptation theory

What is domain distance $d_{\mathcal{H}}$?

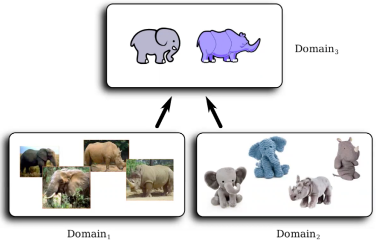
- $d_{\mathcal{H}} := \mathcal{A}$-distance on hypothesis space $\mathcal{H}$
  - Distance between distributions



Domain adaptation theory

What is domain distance $d_{\mathcal{H}}$?

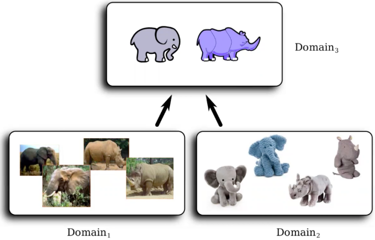
- $d_{\mathcal{H}} := \mathcal{A}$-distance on hypothesis space $\mathcal{H}$
  - Distance between distributions
- Minimize?



Domain adaptation theory

What is domain distance $d_{\mathcal{H}}$?

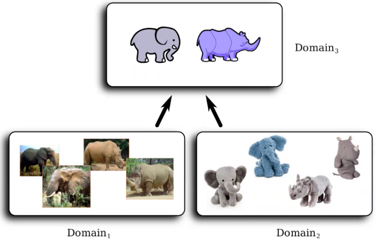
- $d_{\mathcal{H}} := \mathcal{A}$-distance on hypothesis space $\mathcal{H}$
  - Distance between distributions
- Minimize? $\longrightarrow$ find $h \in \mathcal{H}$ that has **max** error on domain discrimination



Domain adaptation theory

What is domain distance $d_{\mathcal{H}}$?

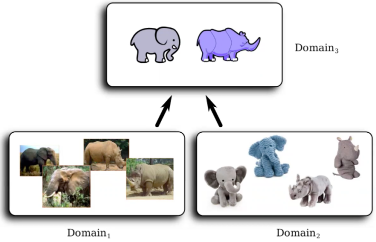
- $d_{\mathcal{H}} := \mathcal{A}$-distance on hypothesis space $\mathcal{H}$
  - Distance between distributions
- Minimize? $\longrightarrow$ find $h \in \mathcal{H}$ that has **max** error on domain discrimination
- Estimate?



Domain adaptation theory

What is domain distance $d_{\mathcal{H}}$?

- $d_{\mathcal{H}} := \mathcal{A}$-distance on hypothesis space $\mathcal{H}$
  - Distance between distributions
- Minimize? $\longrightarrow$ find $h \in \mathcal{H}$ that has **max** error on domain discrimination
- Estimate? $\longrightarrow$ Train domain discriminator



Directly minimize domain distance

Popular approach: add a distance loss term $\mathcal{L}_{dist}$

$$\min \mathcal{L} = \min(\mathcal{L}_{\text{task}} + \alpha \mathcal{L}_{\text{dist}})$$

Directly minimize domain distance

Popular approach: add a distance loss term $\mathcal{L}_{dist}$

$$\min \mathcal{L} = \min(\mathcal{L}_{\text{task}} + \alpha \mathcal{L}_{\text{dist}})$$

- Adversarial learning: (Ganin et al., 2016; Bousmalis et al., 2016; Zhang et al., 2017)
  - **Minimize** task error ($\hat{\epsilon}_S$): $\min \mathcal{L}_{task}$
  - **Minimize** domain distance ($d_{\mathcal{H}}$):
    - Maximize discriminator error: $\max \mathcal{L}_{dist}$ (so $\alpha < 0$)

Directly minimize domain distance

Popular approach: add a distance loss term $\mathcal{L}_{dist}$

$$\min \mathcal{L} = \min(\mathcal{L}_{\text{task}} + \alpha \mathcal{L}_{\text{dist}})$$

- Adversarial learning: (Ganin et al., 2016; Bousmalis et al., 2016; Zhang et al., 2017)
  - **Minimize** task error ($\hat{\epsilon}_S$): min $\mathcal{L}_{task}$
  - **Minimize** domain distance ($d_{\mathcal{H}}$):
    - Maximize discriminator error: $\max \mathcal{L}_{dist}$ (so $\alpha < 0$)
- Use other distance measures (Guo et al., 2020)
  - **Mimize** $\mathcal{L}$:
    - $\mathcal{L}_2$, cosine, Maximum Mean Discrepancy, Fisher Linear Discriminant

Minimizing distance: is it enough?

$$\min \mathcal{L} = \min(\mathcal{L}_{\text{task}} + \alpha \mathcal{L}_{\text{dist}})$$

Are distance-minimized shared features all we need?

Minimizing distance: is it enough?

$$\min \mathcal{L} = \min(\mathcal{L}_{\text{task}} + \alpha \mathcal{L}_{\text{dist}})$$

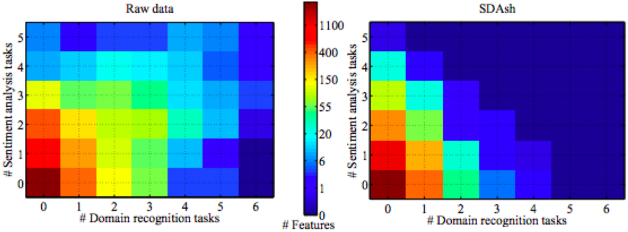
Are distance-minimized shared features all we need? Nope!

Minimizing distance: is it enough?

$$\min \mathcal{L} = \min(\mathcal{L}_{\text{task}} + \alpha \mathcal{L}_{\text{dist}})$$

Are distance-minimized shared features all we need? Nope!

- $d_{\mathcal{H}}$ can be misleading (Glorot et al., 2011)
  - Finding task-specific features may *increase* $d_{\mathcal{H}}$



Minimizing distance: is it enough?

$$\min \mathcal{L} = \min(\mathcal{L}_{\text{task}} + \alpha \mathcal{L}_{\text{dist}})$$

Are distance-minimized shared features all we need? Nope!

- $d_{\mathcal{H}}$ can be misleading (Glorot et al., 2011)
  - Finding task-specific features may *increase* $d_{\mathcal{H}}$
- Discriminator task ($\mathcal{L}_{\text{dist}}$) often easier to minimize than prediction
  - Prevent too much influence from $\mathcal{L}_{dist}$ (Zhang et al., 2017)

Minimizing distance: is it enough?

$$\min \mathcal{L} = \min(\mathcal{L}_{\text{task}} + \alpha \mathcal{L}_{\text{dist}})$$

Are distance-minimized shared features all we need? Nope!

- $d_{\mathcal{H}}$ can be misleading (Glorot et al., 2011)
  - Finding task-specific features may *increase* $d_{\mathcal{H}}$
- Discriminator task ($\mathcal{L}_{\text{dist}}$) often easier to minimize than prediction
  - Prevent too much influence from $\mathcal{L}_{dist}$ (Zhang et al., 2017)
- Domain-specific features are important for $\hat{\epsilon}_S$ (Blitzer et al., 2006; Daumé, 2007)
  - Encourage part of the feature space to embed domains orthogonally (Bousmalis et al., 2016)

Implicitly minimize domain distance

Can we handle $d_{\mathcal{H}}$ implicitly (just minimize $\hat{\epsilon}_S$)?

Implicitly minimize domain distance

Can we handle $d_{\mathcal{H}}$ implicitly (just minimize $\hat{\epsilon}_S$)?

- Smooth ***parameters*** to prevent minimizing $\hat{\epsilon}_S$ in a way that will unintentionally increase $d_{\mathcal{H}}$ (Desai et al., 2019)

Implicitly minimize domain distance

Can we handle $d_{\mathcal{H}}$ implicitly (just minimize $\hat{\epsilon}_S$)?

- Smooth ***parameters*** to prevent minimizing $\hat{\epsilon}_S$ in a way that will unintentionally increase $d_{\mathcal{H}}$ (Desai et al., 2019)
- Use distance measures as the ***features*** (Ruder and Plank, 2017)
  - Similarity (e.g., Jensen-Shannon divergence $\approx$ smoothed *KL*-divergence)
  - Diversity (e.g., Shannon entropy)
  - Focus on selecting training instances to minimize $\hat{\epsilon}_S$

Open questions

- Can we adapt by just choosing our training examples well?

Open questions

- Can we adapt by just choosing our training examples well?
  - *Maybe*, could aid *model/task/domain* transfer (Ruder and Plank, 2017)

Open questions

- Can we adapt by just choosing our training examples well?
    - *Maybe*, could aid *model/task/domain* transfer (Ruder and Plank, 2017)
- What if we have *multiple* source domains?

Open questions

- Can we adapt by just choosing our training examples well?
  - *Maybe*, could aid *model/task/domain* transfer (Ruder and Plank, 2017)
- What if we have *multiple* source domains?
  - Choose dynamically (Guo et al., 2020)
  - Or combine with data selection (Ruder and Plank, 2017)

Open questions

- Can we adapt by just choosing our training examples well?
  - *Maybe*, could aid *model/task/domain* transfer (Ruder and Plank, 2017)
- What if we have *multiple* source domains?
  - Choose dynamically (Guo et al., 2020)
  - Or combine with data selection (Ruder and Plank, 2017)
- What if a domain is temporal?

Open questions

- Can we adapt by just choosing our training examples well?
  - *Maybe*, could aid *model/task/domain* transfer (Ruder and Plank, 2017)
- What if we have *multiple* source domains?
  - Choose dynamically (Guo et al., 2020)
  - Or combine with data selection (Ruder and Plank, 2017)
- What if a domain is temporal?
  - Ensemble and curriculum learning (Desai et al., 2019)

Open questions

- Can we adapt by just choosing our training examples well?
  - *Maybe*, could aid *model/task/domain* transfer (Ruder and Plank, 2017)
- What if we have *multiple* source domains?
  - Choose dynamically (Guo et al., 2020)
  - Or combine with data selection (Ruder and Plank, 2017)
- What if a domain is temporal?
  - Ensemble and curriculum learning (Desai et al., 2019)
- How do we select hyperparameters?

Open questions

- Can we adapt by just choosing our training examples well?
    - *Maybe*, could aid *model/task/domain* transfer (Ruder and Plank, 2017)
- What if we have *multiple* source domains?
    - Choose dynamically (Guo et al., 2020)
    - Or combine with data selection (Ruder and Plank, 2017)
- What if a domain is temporal?
    - Ensemble and curriculum learning (Desai et al., 2019)
- How do we select hyperparameters?
    - Dev set for the target domain (Ganin et al., 2016; Bousmalis et al., 2016)

Open questions

- Can we adapt by just choosing our training examples well?
  - *Maybe*, could aid *model/task/domain* transfer (Ruder and Plank, 2017)
- What if we have *multiple* source domains?
  - Choose dynamically (Guo et al., 2020)
  - Or combine with data selection (Ruder and Plank, 2017)
- What if a domain is temporal?
  - Ensemble and curriculum learning (Desai et al., 2019)
- How do we select hyperparameters?
  - Dev set for the target domain (Ganin et al., 2016; Bousmalis et al., 2016)
  - is this fishy?

Summary

- Wide range of DA techniques that fall into 3 broad categories:
  - modify the training objective
  - modify the input features
  - modify how data is selected

Summary

- Wide range of DA techniques that fall into 3 broad categories:
  - modify the training objective
  - modify the input features
  - modify how data is selected
- Most techniques incorporate a notion of minimizing distance between domains

Summary

- Wide range of DA techniques that fall into 3 broad categories:
  - modify the training objective
  - modify the input features
  - modify how data is selected
- Most techniques incorporate a notion of minimizing distance between domains
- a number of challenging questions are still unanswered

Outline

1. Domain Adaptation
2. Stance Detection
3. Domain Adaptation + Stance Detection

Definition

public act by a social actor,

(Bois, 2007)

Definition

public act by a social actor,
achieved dialogically through overt communicative means,

(Bois, 2007)

Definition

public act by a social actor,
achieved dialogically through overt communicative means,
of simultaneously
    ***evaluating** objects [topics],*

(Bois, 2007)

Definition

public act by a social actor,
achieved dialogically through overt communicative means,
of simultaneously

> ***evaluating*** *objects [topics],*
> ***positioning*** *subjects, and*

(Bois, 2007)

Definition

public act by a social actor,
achieved dialogically through overt communicative means,
of simultaneously

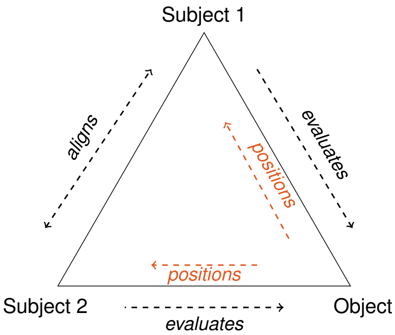
> ***evaluating*** *objects [topics],*
> ***positioning*** *subjects, and*
> ***aligning*** *with other subjects, with respect to any salient*

dimension of the sociocultural field

(Bois, 2007)

Stance: in theory

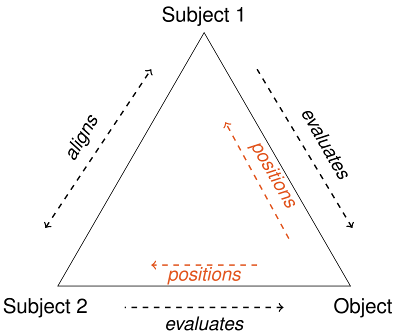
The stance triangle (Bois, 2007)



- a social interaction
- **Evaluate:**
  giving some value to Object
- **Position:**
  wrt to sociocultural value
- **Align:**
  wrt to other actors

Stance: in theory

The stance triangle (Bois, 2007)



- a social interaction
- **Evaluate:** giving some value to Object
- **Position:** wrt to sociocultural value
- **Align:** wrt to other actors

$\Longrightarrow$ **social** context is important

Stance: NLP definitions

Topic: *legalization of abortion*
Document: *The pregnant are more than walking incubators and have rights!*

Stance: NLP definitions

Topic: *legalization of abortion*
Document: *The pregnant are more than walking incubators and have rights!*

Input: document d and topic t
Output: label $y \in Y$

Stance: NLP definitions

Topic: *legalization of abortion*
Document: *The pregnant are more than walking incubators and have rights!*

Input: document d and topic t
Output: label $y \in Y$

- $Y^{(1)} = \{pro, con\}$ (Walker et al., 2012; Vamvas and Sennrich, 2020)
    - possibly with *neutral* (Mohammad et al., 2016)

Stance: NLP definitions

Topic: *legalization of abortion*
Document: *The pregnant are more than walking incubators and have rights!*

Input: document d and topic t
Output: label $y \in Y$

- $Y^{(1)} = \{pro, con\}$ (Walker et al., 2012; Vamvas and Sennrich, 2020)
  - possibly with *neutral* (Mohammad et al., 2016)
- $Y^{(2)} = \{for, against, observing\}$ (Ferreira and Vlachos, 2016)
  - possibly with *unrelated* (e.g., the FakeNewsChallenge in (Hanselowski et al., 2018))

Datasets (a sample): characteristics

- Genre:
  - $Y^{(1)}$: generally social media (e.g., forums, Twitter) (Mohammad et al., 2016; Walker et al., 2012; Vamvas and Sennrich, 2020)
  - $Y^{(2)}$: news/rumor articles (Ferreira and Vlachos, 2016; Mohtarami et al., 2019)

Datasets (a sample): characteristics

- Language:
  - Most English only (Walker et al., 2012; Mohammad et al., 2016; Ferreira and Vlachos, 2016)
  - Twitter only: many small (∼1 topic) datasets in other languages

    (e.g., Spanish, Catalan, Russian, French, Italian, Turkish, Arabic) [Links in appendix]
  - Limited non-Twitter non-English (Vamvas and Sennrich, 2020)

Datasets (a sample)

Range in # of topics and # of examples
Range in the amount of data available per topic (usually small)

	# topics	**# exs**	**langs**
Walker et al. (2012)	10	130*k*	en
Mohammad et al. (2016)	6	2*k*	en
Ferreira and Vlachos (2016)	300*	2.6*k*	en
Vamvas and Sennrich (2020)	194	67*k*	de, fr, it

(* topic is defined differently here)

Datasets (a sample)

Range in # of topics and # of examples
Range in the amount of data available per topic (usually small)

	# topics	# exs	langs
Walker et al. (2012)	10	130*k*	en
Mohammad et al. (2016)	6	2*k*	en
Ferreira and Vlachos (2016)	300*	2.6*k*	en
Vamvas and Sennrich (2020)	194	67*k*	de, fr, it

(* topic is defined differently here)

- 300 is misleading
- define a topic differently (a news headline), so they're very very specific
- basically an entailment task

Datasets (a sample)

Range in # of topics and # of examples
Range in the amount of data available per topic (usually small)

	# topics	# exs	langs
Walker et al. (2012)	10	130k	en
Mohammad et al. (2016)	6	2k	en
Ferreira and Vlachos (2016)	300*	2.6k	en
Vamvas and Sennrich (2020)	194	67k	de, fr, it

(* topic is defined differently here)

- From voting advice app
- Cross-lingual, will return to this later

Datasets (a sample)

Range in # of topics and # of examples
Range in the amount of data available per topic (usually small)

	# topics	**# exs**	**langs**
Walker et al. (2012)	10	130k	en
Mohammad et al. (2016)	6	2k	en
Ferreira and Vlachos (2016)	300*	2.6k	en
Vamvas and Sennrich (2020)	194	67k	de, fr, it

(* topic is defined differently here)

Datasets: topics

	Topics
Walker et al. (2012)	evolution, abortion, gun control, gay marriage, existence of god, healthcare, death penalty, climate change, communism vs. capitalism, marijuana legalization
Mohammad et al. (2016)	atheism, climate change is concern, feminist movement, Hillary Clinton, legalization of abortion, Donald Trump

Stance on forums (an *interesting* sample)

How much **social** context is actually used?

$\approx$ *none* *lots*

Stance on forums (an *interesting* sample)

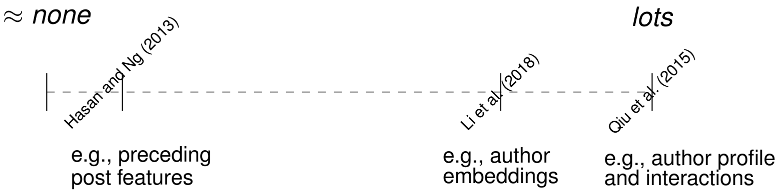
How much **social** context is actually used?

Stance on forums (an *interesting* sample)

How much **social** context is actually used?

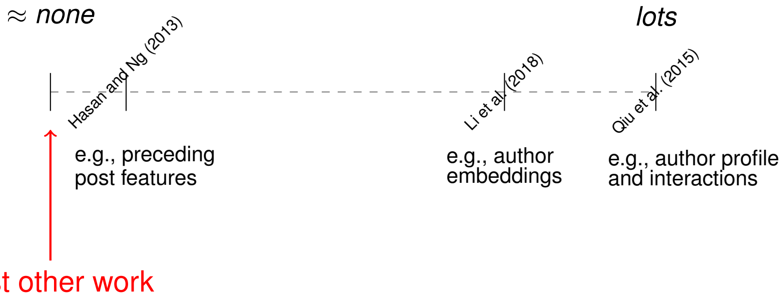
Stance on forums (an *interesting* sample)

How much **social** context is actually used?



Stance on forums (an *interesting* sample)

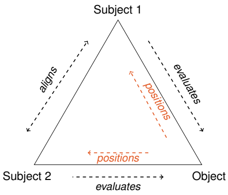
How much **social** context is actually used?



Aligning

Aligning in forums:

- User agreements/disagreements (Qiu et al., 2015; Li et al., 2018)
- Preceding post features (assume a reply) (Hasan and Ng, 2013)



Aligning

Aligning in forums:

- User agreements/disagreements (Qiu et al., 2015; Li et al., 2018)
- Preceding post features (assume a reply) (Hasan and Ng, 2013)

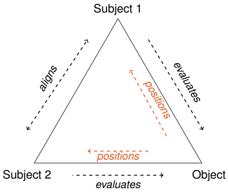
Difficulties:

- Information may be difficult to get, esp. in other domains (e.g., Twitter)
- Hard to model in NNs, instead use:
  - SVMs, NB, HMMs, etc. (Hasan and Ng, 2013)
  - graphical model (Qiu et al., 2015)
  - representation learning + ILP (Li et al., 2018)

Positioning

Positioning requires *subjectivity* (i.e., author identity)

- Author consistency constraint (Hasan and Ng, 2013; Li et al., 2018)
- Author embeddings (Li et al., 2018)
- Author attributes(Qiu et al., 2015; Li et al., 2018)
    - (e.g., gender, political party, religion)



Positioning

Positioning requires *subjectivity* (i.e., author identity)

- Author consistency constraint (Hasan and Ng, 2013; Li et al., 2018)
- Author embeddings (Li et al., 2018)
- Author attributes(Qiu et al., 2015; Li et al., 2018)
  - (e.g., gender, political party, religion)

Difficulties:

- Information may be hard to get
- Is using this ethical?
- Can a model with this information be misused?

Summary

- Stance is a *social* act and so requires social context
  - A lot of work *doesn't* use this context
  - Work that does use context makes other limited assumptions (e.g., the topics are known)

Summary

- Stance is a *social* act and so requires social context
  - A lot of work *doesn't* use this context
  - Work that does use context makes other limited assumptions (e.g., the topics are known)
- Most work assumes the following are fixed during training:
  - Topics
  - Language
  - Genre

Summary

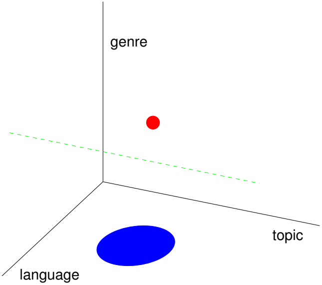
- Stance is a *social* act and so requires social context
  - A lot of work *doesn't* use this context
  - Work that does use context makes other limited assumptions (e.g., the topics are known)
- Most work assumes the following are fixed during training:
  - Topics
  - Language
  - Genre
- Need to **generalize** and go beyond these assumptions

Outline

1. Domain Adaptation
2. Stance Detection
3. Domain Adaptation + Stance Detection

How is stance domain adaptation?

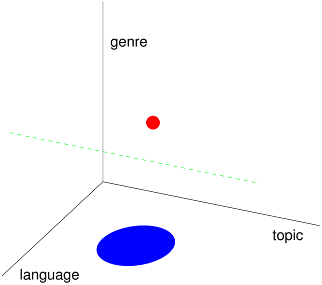
- **domain** := regions
  - red: language, genre, topic fixed
  - green: language and genre fixed (Mohammad et al., 2016)
  - blue: genre fixed (Vamvas and Sennrich, 2020)



at least 3 latent factors in the variety

How is stance domain adaptation?

- **domain** := regions
  - red: language, genre, topic fixed
  - green: language and genre fixed (Mohammad et al., 2016)
  - blue: genre fixed (Vamvas and Sennrich, 2020)
- **Goal:** move in any direction



at least 3 latent factors in the variety

Unsupervised stance detection

Infer stance and topic together (Gottipati et al., 2013)

- generative model that:
  - Identifies *latent* topics for a post
  - associates a post with a *side*
- topic distributions $\approx$ shared features (e.g., (Glorot et al., 2011))

Unsupervised stance detection

Infer stance and topic together (Gottipati et al., 2013)

- generative model that:
  - Identifies *latent* topics for a post
  - associates a post with a *side*
- topic distributions $\approx$ shared features (e.g., (Glorot et al., 2011))
- no social context used
- difficult to evaluate and interpret

Stance as domain adaptation

Cross-target stance detection (Augenstein et al., 2016; Xu et al., 2018; Zhang et al., 2020)

- Train on 1 topic t_i, test on 1 topic t_j where $i \neq j$

Stance as domain adaptation

Recall:

$$\epsilon_T(h) \leq \hat{\epsilon}_S(h) + d_{\mathcal{H}}(\tilde{U_S}, \tilde{U_T}) + \dots$$

Cross-target stance detection (Augenstein et al., 2016; Xu et al., 2018; Zhang et al., 2020)

- Train on 1 topic t_i, test on 1 topic t_j where $i \neq j$

Stance as domain adaptation

Recall:

$$\epsilon_T(h) \leq \hat{\epsilon}_S(h) + d_{\mathcal{H}}(\tilde{U}_S, \tilde{U}_T) + \ldots$$

$$\begin{array}{c} \text{feminism} \\ \Big\downarrow \textit{related} \\ \text{legalization} \\ \text{of abortion} \end{array}$$

Cross-target stance detection (Augenstein et al., 2016; Xu et al., 2018; Zhang et al., 2020)

- Train on 1 topic t_i, test on 1 topic t_j where $i \neq j$
- Assume t_i **related** to t_j
    - $\approx$ manually limit $d_{\mathcal{H}}$

Stance as domain adaptation

Recall:

$$\epsilon_T(h) \leq \hat{\epsilon}_S(h) + d_{\mathcal{H}}(\tilde{U}_S, \tilde{U}_T) + \ldots$$

$$\begin{array}{c} \text{feminism} \\ \Big\downarrow \textit{related} \\ \text{legalization} \\ \text{of abortion} \end{array}$$

Cross-target stance detection (Augenstein et al., 2016; Xu et al., 2018; Zhang et al., 2020)

- Train on 1 topic t_i, test on 1 topic t_j where $i \neq j$
- Assume t_i **related** to t_j
  - $\approx$ manually limit $d_{\mathcal{H}}$
- mostly on Twitter
  - other subjects in stance Δ implicit
  - social context not used

Using domain adaptation methods

Not really using techniques from the literature

- Xu et al. (2018): $\sim$ attempt to identify stance-specific features (as in Glorot et al. (2011))

Using domain adaptation methods

Not really using techniques from the literature

- Xu et al. (2018): $\sim$ attempt to identify stance-specific features (as in Glorot et al. (2011))
- learn domain (topic) shared features (e.g., Ganin et al. (2016)) using:
  - external knowledge (Zhang et al., 2020)
  - tuned embeddings (Augenstein et al., 2016)

Why are we not using domain adaptation methods?

Stance detection $\approx$ sentiment product reviews (classic DA task)

Why are we not using domain adaptation methods?

Stance detection ≈ sentiment product reviews (classic DA task)

⟹ **Why not use DA techniques??**

Why are we not using domain adaptation methods?

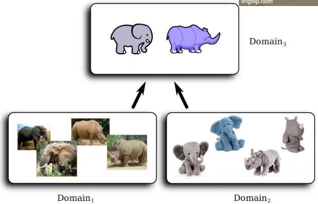
Stance detection $\approx$ sentiment product reviews (classic DA task)
$\Longrightarrow$ **Why not use DA techniques??**

- Short research memory? Limited research peripheral vision?

Why are we not using domain adaptation methods?

Stance detection $\approx$ sentiment product reviews (classic DA task)
$\Longrightarrow$ **Why not use DA techniques??**

- Short research memory? Limited research peripheral vision?
- More likely: difficult scenarios in stance detection
  - many-to-one (covered a bit in Guo et al. (2020))
  - many-to-many



Cross-lingual learning as domain adaptation

Language $\approx$ domain

Cross-lingual learning as domain adaptation

Language $\approx$ domain

Sample of existing datasets:

	Rasooli et al. (2017) Sentiment	Nooralahzadeh et al. (2020) XNLI	Pfeiffer et al. (2020) NER	XCOPA	XQuAD
# langs	16	15	16	12	11
# families	5	7	11	11	6

Cross-lingual learning as domain adaptation

Language $\approx$ domain

Sample of existing datasets:

	Rasooli et al. (2017) Sentiment	Nooralahzadeh et al. (2020) XNLI	Pfeiffer et al. (2020) NER	XCOPA	XQuAD
# langs	16	15	16	12	11
# families	5	7	11	11	6

common: Arabic, Chinese, German, Russian, Spanish, English

Cross-lingual embeddings

Cross-lingual embeddings $\approx$ shared features:

Cross-lingual embeddings

Cross-lingual embeddings $\approx$ shared features:

- From contextualized LM (Pfeiffer et al., 2020; Nooralahzadeh et al., 2020)
  - Large (unlabeled) monolingual corpora

Cross-lingual embeddings

Cross-lingual embeddings $\approx$ shared features:

- From contextualized LM (Pfeiffer et al., 2020; Nooralahzadeh et al., 2020)
  - Large (unlabeled) monolingual corpora
- Non-contextualized static (Rasooli et al., 2017)
  - parallel (or comparable) corpora
  - bilingual dictionaries

Using cross-lingual embeddings

Common approach (e.g., as in Glorot et al. (2011); Ganin et al. (2016))

- Treat embeddings as shared space
- Build classifiers directly on the shared features (Rasooli et al., 2017; Pfeiffer et al., 2020)

Using cross-lingual embeddings

Common approach (e.g., as in Glorot et al. (2011); Ganin et al. (2016))

- Treat embeddings as shared space
- Build classifiers directly on the shared features (Rasooli et al., 2017; Pfeiffer et al., 2020)
- Possibly also use shared features mapped to:
  - language-specific and task-specific features (Pfeiffer et al., 2020)
  - task-specific features (Nooralahzadeh et al., 2020)

Putting it all together

Cross-lingual stance detection as domain adaptation

Two types of corpora

Putting it all together

Cross-lingual stance detection as domain adaptation

Two types of corpora

1. *multiple* corpora with different languages and topics each

   Mohtarami et al. (2019)

Putting it all together

Cross-lingual stance detection as domain adaptation

Two types of corpora

1. *multiple* corpora with different languages and topics each

   Mohtarami et al. (2019)

2. *single* corpus with multiple languages and the same topics

   (Vamvas and Sennrich, 2020)

Putting it all together

Cross-lingual stance detection as domain adaptation

Two types of corpora

1. *multiple* corpora with different languages and topics each

   Mohtarami et al. (2019)

2. *single* corpus with multiple languages and the same topics

   (Vamvas and Sennrich, 2020)

Both require cross-lingual LM or embeddings

What do we see during training?

		New Topic(s)	
		Seen	Unseen
New Lang	Seen		
	Unseen		

What do we see during training?

		New Topic(s)	
		Seen	Unseen
New Lang	Seen	Mohtarami et al. (2019)	
	Unseen		

What do we see during training?

		New Topic(s)	
		Seen	Unseen
New Lang	Seen	Mohtarami et al. (2019)	Vamvas and Sennrich (2020)
	Unseen		

What do we see during training?

		New Topic(s) Seen	New Topic(s) Unseen
New Lang	Seen	Mohtarami et al. (2019)	Vamvas and Sennrich (2020)
	Unseen	Vamvas and Sennrich (2020)	

What do we see during training?

		New Topic(s) Seen	New Topic(s) Unseen
New Lang	Seen	Mohtarami et al. (2019)	Vamvas and Sennrich (2020)
	Unseen	Vamvas and Sennrich (2020)	GOAL

Why haven't we reached the goal?

What do we see during training?

		New Topic(s) Seen	New Topic(s) Unseen
New Lang	Seen	Mohtarami et al. (2019)	Vamvas and Sennrich (2020)
	Unseen	Vamvas and Sennrich (2020)	GOAL

Why haven't we reached the goal?

- Datasets (basically) don't exist
  - except Vamvas and Sennrich (2020)
  - combining genres another level of complexity

What do we see during training?

		New Topic(s)	
		Seen	Unseen
New Lang	Seen	Mohtarami et al. (2019)	Vamvas and Sennrich (2020)
	Unseen	Vamvas and Sennrich (2020)	GOAL

Why haven't we reached the goal?

- Datasets (basically) don't exist
  - except Vamvas and Sennrich (2020)
  - combining genres another level of complexity
- Topic adaptation alone is already hard

What do we see during training?

		New Topic(s) Seen	New Topic(s) Unseen
New Lang	Seen	Mohtarami et al. (2019)	Vamvas and Sennrich (2020)
	Unseen	Vamvas and Sennrich (2020)	GOAL

Why haven't we reached the goal?

- Datasets (basically) don't exist
  - except Vamvas and Sennrich (2020)
  - combining genres another level of complexity
- Topic adaptation alone is already hard
- How do we tune?

Summary

- Limited work on stance detection as DA
  - Tends to make simplifying assumptions
  - Doesn't really use DA techniques

Summary

- Limited work on stance detection as DA
  - Tends to make simplifying assumptions
  - Doesn't really use DA techniques
- Cross-lingual learning has a fair amount of work
  - Especially on embeddings/LMs

Summary

- Limited work on stance detection as DA
  - Tends to make simplifying assumptions
  - Doesn’t really use DA techniques
- Cross-lingual learning has a fair amount of work
  - Especially on embeddings/LMs
- Cross-lingual stance detection has very little work
  - Don’t really have the resources for this
  - Adding in cross-topic also very hard

Conclusion

Where we are:

Conclusion

Where we are:

- Domain adaptation is well studied but still leaves many questions unanswered
  - many-to-one, many-to-many, hyperparameter tuning

Conclusion

Where we are:

- Domain adaptation is well studied but still leaves many questions unanswered
  - many-to-one, many-to-many, hyperparameter tuning
- Stance detection is a clear domain adaptation task but isn't really treated as one
  - Models also tend to ignore important social context

Conclusion

Where we are:

- Domain adaptation is well studied but still leaves many questions unanswered
  - many-to-one, many-to-many, hyperparameter tuning
- Stance detection is a clear domain adaptation task but isn't really treated as one
  - Models also tend to ignore important social context
- Several tasks still minimally explored

Conclusion

Where we are:

- Domain adaptation is well studied but still leaves many questions unanswered
    - many-to-one, many-to-many, hyperparameter tuning
- Stance detection is a clear domain adaptation task but isn't really treated as one
    - Models also tend to ignore important social context
- Several tasks still minimally explored
    - Cross-lingual stance detection

Conclusion

Where we are:

- Domain adaptation is well studied but still leaves many questions unanswered
  - many-to-one, many-to-many, hyperparameter tuning
- Stance detection is a clear domain adaptation task but isn't really treated as one
  - Models also tend to ignore important social context
- Several tasks still minimally explored
  - Cross-lingual stance detection
  - Cross-topic stance detection

Conclusion

Where we should go:

Conclusion

Where we should go:

- Unexplored directions

Conclusion

Where we should go:

- Unexplored directions
  - Cross-lingual **and** cross-topic stance detection

Conclusion

Where we should go:

- Unexplored directions
  - Cross-lingual **and** cross-topic stance detection
  - Adding in cross-genre

Conclusion

Where we should go:

- Unexplored directions
  - Cross-lingual **and** cross-topic stance detection
  - Adding in cross-genre
- Expand language coverage in datasets

Conclusion

Where we should go:

- Unexplored directions
  - Cross-lingual **and** cross-topic stance detection
  - Adding in cross-genre
- Expand language coverage in datasets
- Fully utilize and expand on DA techniques

Conclusion

Where we should go:

- Unexplored directions
  - Cross-lingual **and** cross-topic stance detection
  - Adding in cross-genre
- Expand language coverage in datasets
- Fully utilize and expand on DA techniques

Thank you for listening!

References I

Augenstein, I., Rocktäschel, T., Vlachos, A., and Bontcheva, K. (2016). Stance Detection with Bidirectional Conditional Encoding. In EMNLP.

Baldwin, T. (2021). Nlp for user generated content. Lecture at Advanced Language Processing Winter School.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. C. (2006). Analysis of Representations for Domain Adaptation. In NIPS.

Blitzer, J., McDonald, R., and Pereira, F. C. (2006). Domain Adaptation with Structural Correspondence Learning. In EMNLP.

Bois, J. (2007). The stance triangle. Pragmatics and beyond. New series, 164:139–182.

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., and Erhan, D. (2016). Domain Separation Networks. In NIPS.

Daumé, H. (2007). Frustratingly Easy Domain Adaptation. In ACL.

Desai, S., Sinno, B., Rosenfeld, A., and Li, J. J. (2019). Adaptive Ensembling: Unsupervised Domain Adaptation for Political Document Analysis. In EMNLP/IJCNLP.

References II

Ferreira, W. and Vlachos, A. (2016). Emergent: a novel data-set for stance classification. In HLT-NAACL.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-Adversarial Training of Neural Networks. Journal of Machine Learning Research.

Glorot, X., Bordes, A., and Bengio, Y. (2011). Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. In ICML.

Gottipati, S., Qiu, M., Sim, Y., Jiang, J., and Smith, N. A. (2013). Learning Topics and Positions from Debatepedia. In EMNLP.

Guo, H., Pasunuru, R., and Bansal, M. (2020). Multi-Source Domain Adaptation for Text Classification via DistanceNet-Bandits. In AAAI.

Hanselowski, A., AvineshP.V., S., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C., and Gurevych, I. (2018). A retrospective analysis of the fake news challenge stance detection task. In COLING.

Hasan, K. and Ng, V. (2013). Stance Classification of Ideological Debates: Data, Models, Features, and Constraints. In IJCNLP.

References III

Li, C., Porco, A., and Goldwasser, D. (2018). Structured Representation Learning for Online Debate Stance Prediction. In COLING.

Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X.-D., and Cherry, C. (2016). SemEval-2016 Task 6: Detecting Stance in Tweets. In SemEval@NAACL-HLT.

Mohtarami, M., Glass, J. R., and Nakov, P. (2019). Contrastive Language Adaptation for Cross-Lingual Stance Detection. In SIGDAT.

Nooralahzadeh, F., Bekoulis, G., Bjerva, J., and Augenstein, I. (2020). Zero-Shot Cross-Lingual Transfer with Meta Learning. In EMNLP.

Pfeiffer, J., Vulić, I., Gurevych, I., and Ruder, S. (2020). MAD-X: An Adapter-based Framework for Multi-task Cross-lingual Transfer. In EMNLP.

Plank, B. (2016). What to do about non-standard (or non-canonical) language in NLP. In KONVENS.

Qiu, M., Sim, Y., Smith, N. A., and Jiang, J. (2015). Modeling User Arguments, Interactions, and Attributes for Stance Prediction in Online Debate Forums. In SDM.

Rasooli, M. S., Farra, N., Radeva, A., Yu, T., and McKeown, K. (2017). Cross-lingual sentiment transfer with limited resources. Machine Translation, 32:143–165.

References IV

Ruder, S. and Plank, B. (2017). Learning to select data for transfer learning with Bayesian Optimization. In EMNLP.

Vamvas, J. and Sennrich, R. (2020). X-stance: A Multilingual Multi-Target Dataset for Stance Detection. In KONVENS.

Walker, M., Tree, J. E., Anand, P., Abbott, R., and King, J. (2012). A Corpus for Research on Deliberation and Debate. In LREC.

Xu, C., Paris, C., Nepal, S., and Sparks, R. (2018). Cross-Target Stance Classification with Self-Attention Networks. In ACL.

Zhang, B., Yang, M., Li, X., Ye, Y., Xu, X., and Dai, K. (2020). Enhancing Cross-target Stance Detection with Transferable Semantic-Emotion Knowledge. In ACL.

Zhang, Y., Barzilay, R., and Jaakkola, T. (2017). Aspect-augmented Adversarial Networks for Domain Adaptation. Transactions of the Association for Computational Linguistics, 5:515–528.

Additional Datasets (non-English)

- Japanese: Murakami and Putra (2010)
- Chinese: Xu et al. (2016); Yuan et al. (2019)
- Spanish: Taulé et al. (2017)
- Catalan: Taulé et al. (2017)
- Arabic: Darwish et al. (2017); Baly et al. (2018)
- English-Hindi: Swami et al. (2018)
- Italian: Lai et al. (2018, 2020); Cignarella et al. (2020)
- French: Lai et al. (2020); Evrard et al. (2020)
- Czech: Hercig et al. (2017)
- Greek: Tsakalidis et al. (2018)
- Russian: Lozhnikov et al. (2018); Vychegzhanin and Kotelnikov (2019)

References I

Baly, R., Mohtarami, M., Glass, J. R., i Villodre, L. M., Moschitti, A., and Nakov, P. (2018). Integrating stance detection and fact checking in a unified corpus. In NAACL-HLT.

Cignarella, A. T., Lai, M., Bosco, C., Patti, V., and Rosso, P. (2020). Sardistance @ evalita2020: Overview of the task on stance detection in italian tweets. In EVALITA.

Darwish, K., Magdy, W., and Zanouda, T. (2017). Improved stance prediction in a user similarity feature space. Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017.

Evrard, M., Uro, R., Hervé, N., and Mazoyer, B. (2020). French tweet corpus for automatic stance detection. In LREC.

Hercig, T., Krejzl, P., Hourová, B., Steinberger, J., and Lenc, L. (2017). Detecting stance in czech news commentaries. In ITAT.

Lai, M., Cignarella, A. T., Farías, D. I. H., Bosco, C., Patti, V., and Rosso, P. (2020). Multilingual stance detection in social media political debates. Comput. Speech Lang., 63:101075.

Lai, M., Patti, V., Ruffo, G., and Rosso, P. (2018). Stance evolution and twitter interactions in an italian political debate. In NLDB.

References II

Lozhnikov, N., Derczynski, L., and Mazzara, M. (2018). Stance prediction for russian: Data and analysis. ArXiv, abs/1809.01574.

Murakami, A. and Putra, R. H. (2010). Support or oppose? classifying positions in online debates from reply activities and opinion expressions. In COLING.

Swami, S., Khandelwal, A., Singh, V., Akhtar, S., and Shrivastava, M. (2018). An english-hindi code-mixed corpus: Stance annotation and baseline system. ArXiv, abs/1805.11868.

Taulé, M., Martí, M., Pardo, F. M. R., Rosso, P., Bosco, C., and Patti, V. (2017). Overview of the task on stance and gender detection in tweets on catalan independence. In IberEval@SEPLN.

Tsakalidis, A., Aletras, N., Cristea, A., and Liakata, M. (2018). Nowcasting the stance of social media users in a sudden vote: The case of the greek referendum. Proceedings of the 27th ACM International Conference on Information and Knowledge Management.

Vychegzhanin, S. V. and Kotelnikov, E. V. (2019). Stance detection based on ensembles of classifiers. Programming and Computer Software, 45:228 – 240.

Xu, R., Zhou, Y., Wu, D., Gui, L., Du, J., and Xue, Y. (2016). Overview of nlpcc shared task 4: Stance detection in chinese microblogs. In NLPCC/ICCPOL.

References III

Yuan, J., Zhao, Y., Xu, J., and Qin, B. (2019). Exploring answer stance detection with recurrent conditional attention. In AAAI.