Disentangling the Thoughts: Latest News in Computational Argumentation

The European Conference on Data Analysis (ECDA)
Paderborn, Germany, 4th - 6th July, 2018

Why Do People Love to Argue?

When people are engaged in debate, they are effectively joined together in a search for truth.

It is a primary learning method in yeshivas. Even modern science is a sort of one long multi partner multi strand argument over time.

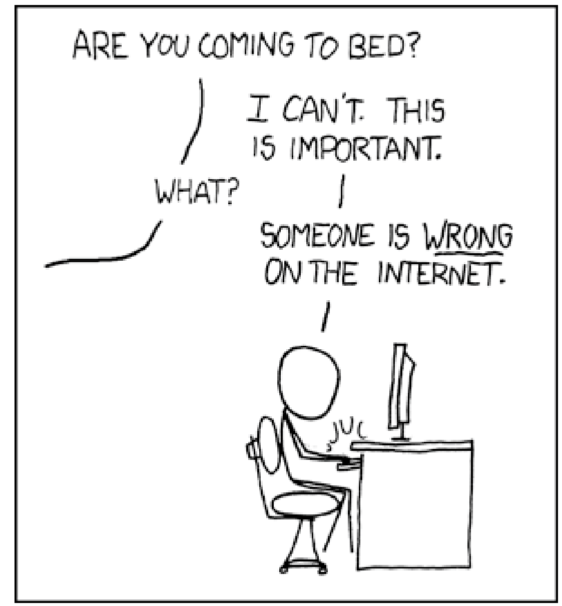
Arguing is an extremely effective way to gain knowledge, learn about another person, understand yourself, and just practice communicating.
Also, it is exciting.

For thousands of years in many unrelated cultures and traditions many serious thinkers have held that the best way to get closer to the truth, justice, insight or reality is through types of argument.

From Sentiment to Argumentation

https://xkcd.com/1098/

Argumentation:
Verbal, social, and rational activity aimed at **convincing** a reasonable critic of the acceptability of a **standpoint** by putting forward a constellation of one or more propositions to justify this standpoint (van Eeemer et al., 2014)



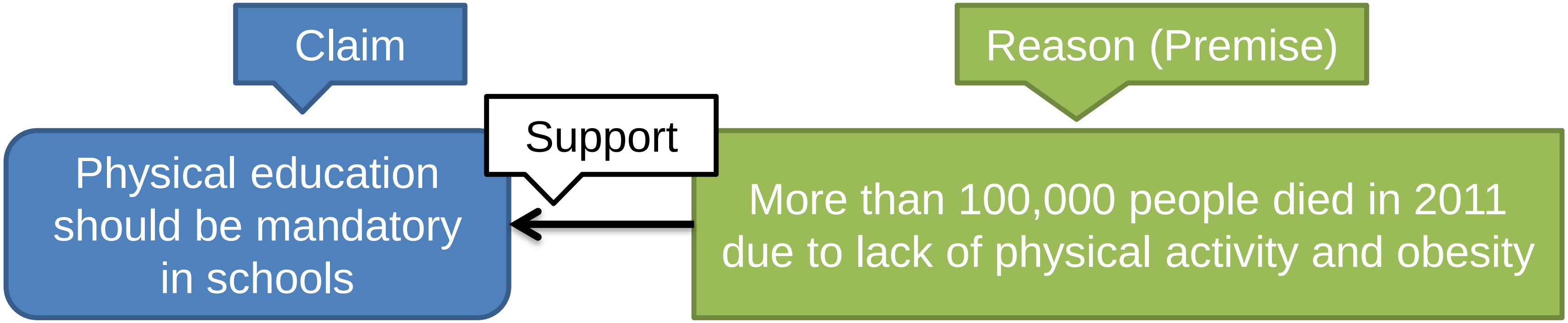
http://xkcd.com/386/

In fact, the bridge in between is a fundamental research question itself!

van Eemeren, F. H., Garssen, B., Krabbe, E. C. W., Snoeck Henkemans, a. F., Verheij, B., & Wagemans, J. H. M. (2014). Handbook of Argumentation Theory. Dordrecht: Springer Netherlands

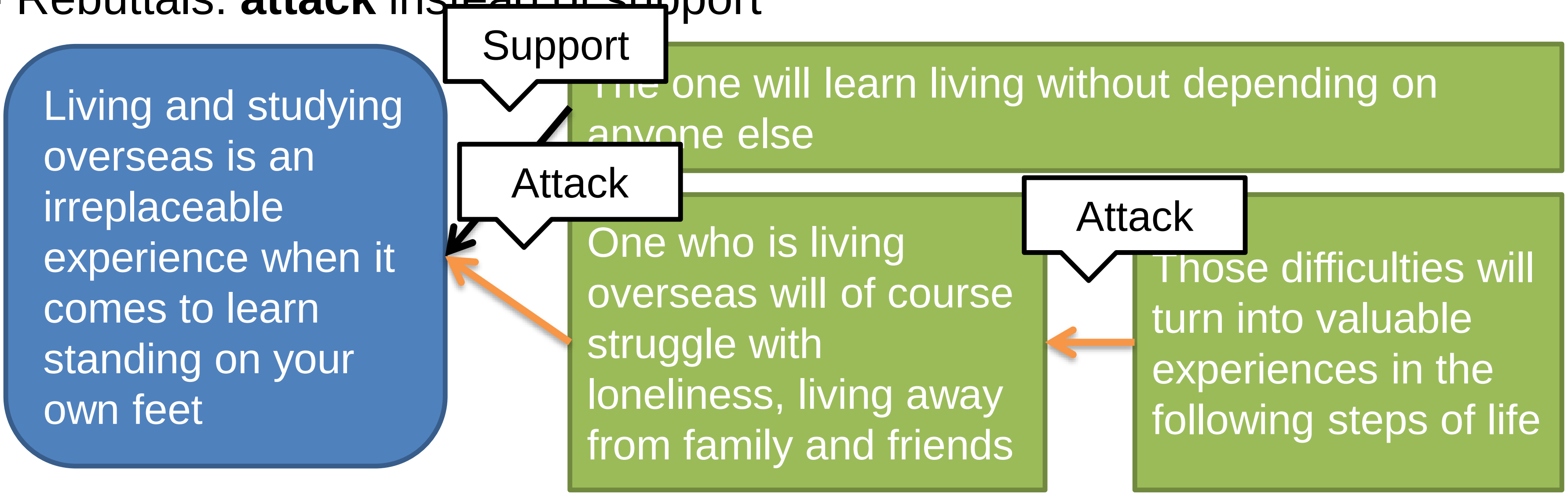
A Simple Argument

- An argument is a **claim**, supported by **reasons**, intended to persuade



A More Complex Argument Structure

- Rebuttals: **attack** instead of support



[…] Second, living and studying overseas is an irreplaceable experience when it comes to learn standing on your own feet. One who is living overseas will of course struggle with loneliness, living away from family and friends but those difficulties will turn into valuable experiences in the following steps of life. Moreover, the one will learn living without depending on anyone else. […]

Outline

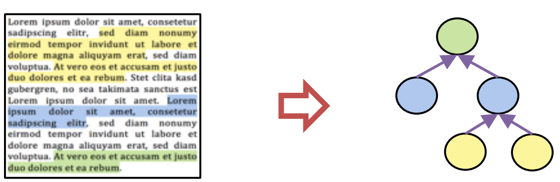
TECHNISCHE UNIVERSITÄT DARMSTADT

- Cross-topic Argument Mining from Heterogeneous Sources
  Stab, C., Miller, T., Schiller, B., Rai, P., Gurevych, I. (2018). https://arxiv.org/abs/1802.05758

ArgumenText

- Document-Level Stance Classification for Fake News Detection
  Hanselowski, A., Schiller, B., Caspelherr, F., Avinesh PVS, Chaudhuri, D., Gurevych, I. (2018). COLING 2018, to appear.

- Cross-Lingual Argumentation Mining: Machine Translation (and a bit of Projection) is All You Need
  Eger, S., Daxenberger, J., Stab, C., & Gurevych, I. (2018). COLING 2018, to appear.

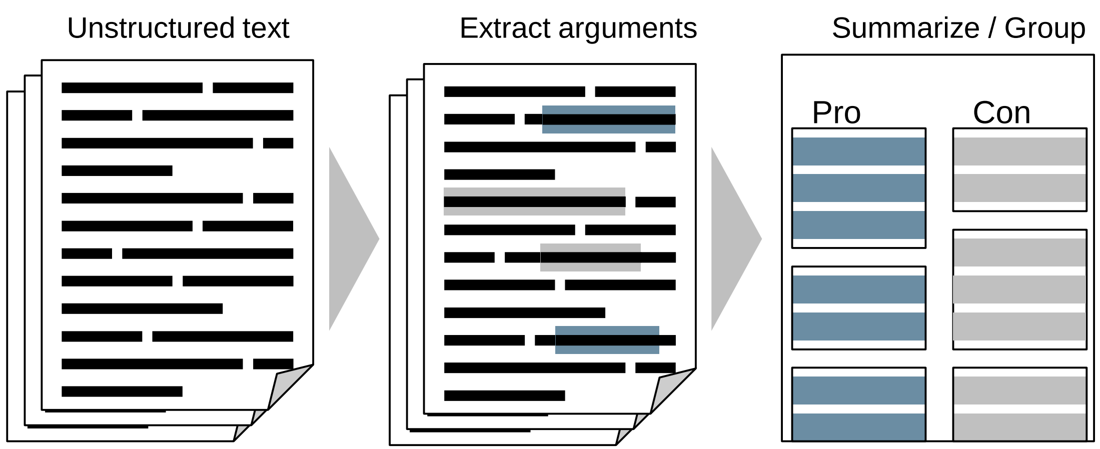


UKP

Goals and Challenges

Goals

- Mine arguments for a given **topic** from **arbitrary** Web sources



Challenges

- How to deal with different text types / genres / writing styles?
- How can we scale the annotation of arguments to arbitrary texts?
- How to generalize argument mining to different topics?

Annotation Scheme and Examples

Requirements

- General enough for use on various text types
- Simple enough to be applied by untrained annotators

Annotation scheme

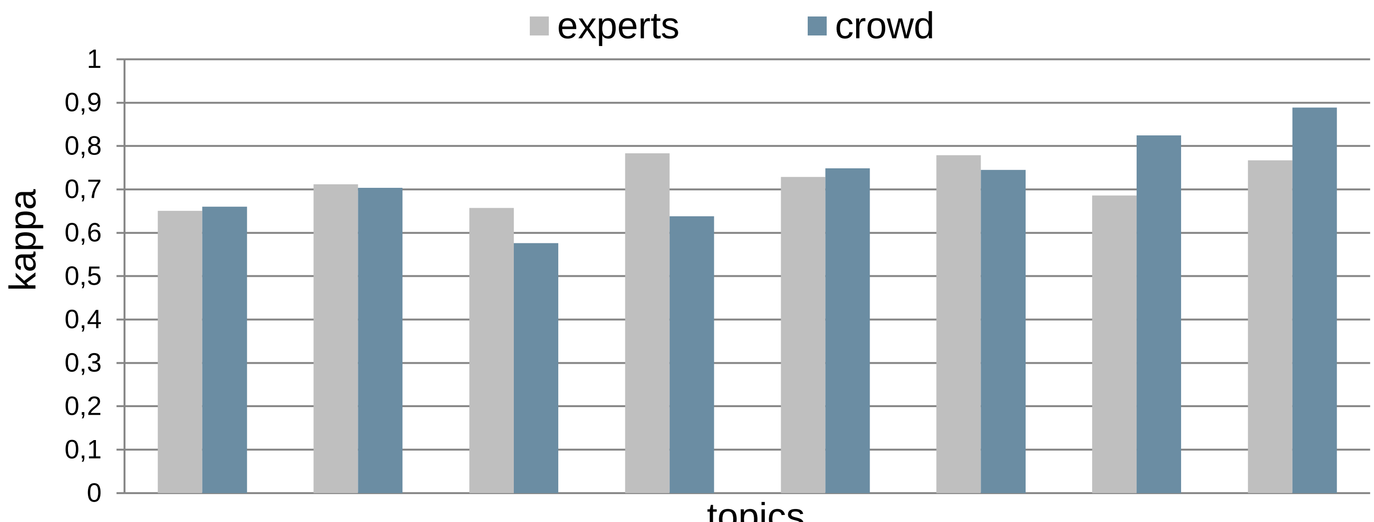
- A span of text expressing evidence supporting or opposing the given topic
- Labels: (1) Supporting argument, (2) Attacking Argument, (3) No Argument

Examples

Topic	Sentence	Label
nuclear energy	Nuclear fission is the process that is used in nuclear reactors to produce high amount of energy using element called uranium.	No Argument
nuclear energy	It has been determined that the amount of greenhouse gases have decreased by almost half because of the prevalence in the utilization of nuclear power.	Supporting Argument

Crowdsourcing Large Dataset

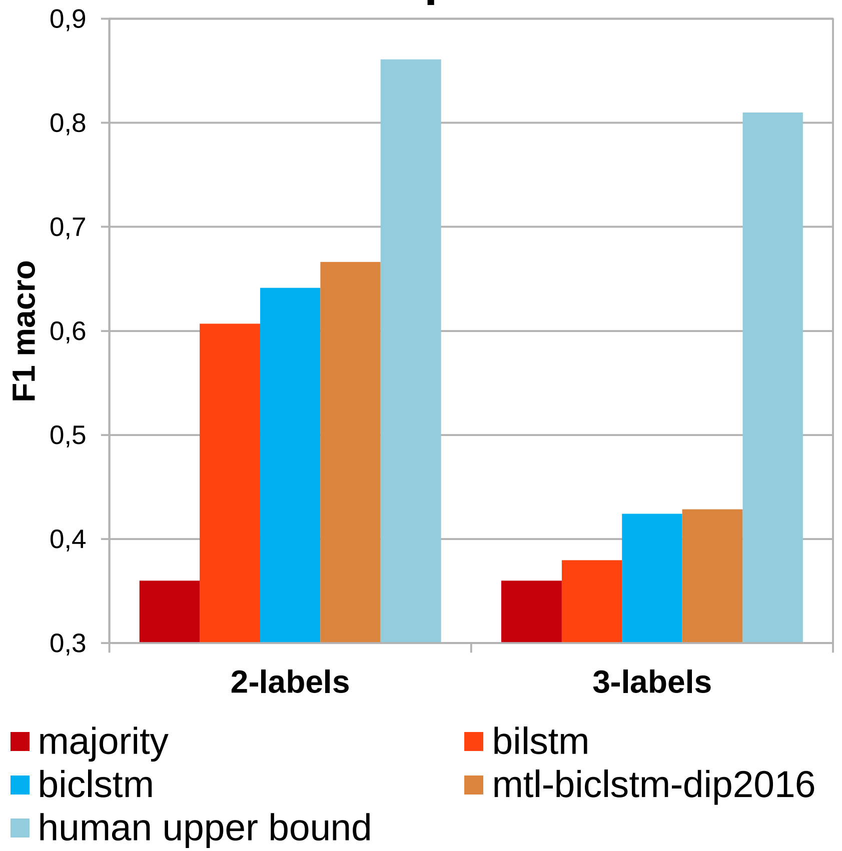
- **Data**
  - Web documents retrieved using Google Search API
- **Resulting corpus**
  - High quality annotations using crowdsourcing (κ=.723)
  - Process is scalable: **40 domains** in less than a week (with 750 workers)
  - Corpus size: **25k+ instances** of high quality annotations



Experiments

- Modified LSTM-cell: **(Bi)CLSTM**
  - Integrates topic directly into LSTM-cell
- **Shared-private[1] multi-task learning** model (mtl-biclstm-dip2016)
  - Combines BiCLSTM with multi-task learning
  - Leverages DIP2016 corpus[2] to learn topic-relevance
- **Results**
  - Our models substantially increase recall of arguments
  - Outperform vanilla BiLSTM model by more than 5% F1 macro

Cross-topic results



[1] (Liu et al., 2017)
[2] (Habernal et al., 2016)

Take Aways

- **Use Cases**
  - Customer feedback analysis, online journalism, educational applications

    http://www.argumentsearch.com/

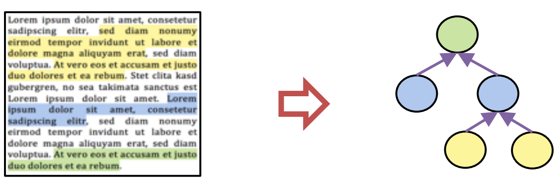
- **Research Findings**
  - Our annotation scheme is applicable to arbitrary Web texts
  - Training data can be reliably created using crowdsourcing
  - Topic-integrating models (e.g. CLSTM) generalize better to unknown topics than common deep learning approaches
  - Leveraging information of datasets from similar tasks can further improve the classification of arguments (e.g. mtl-biclstm-dip2016)

Outline

- Cross-topic Argument Mining from Heterogeneous Sources
  Stab, C., Miller, T., Schiller, B., Rai, P., Gurevych, I. (2018). https://arxiv.org/abs/1802.05758

- Document-Level Stance Classification for Fake News Detection
  Hanselowski, A., Schiller, B., Caspelherr, F., Avinesh PVS, Chaudhuri, D., Gurevych, I. (2018). COLING 2018, to appear.

- Cross-Lingual Argumentation Mining: Machine Translation (and a bit of Projection) is All You Need
  Eger, S., Daxenberger, J., Stab, C., & Gurevych, I. (2018). COLING 2018, to appear.



The Fake News Challenge (FNC)

- Stance detection – determining the **relative perspective** of text **T** to target text or entity **E**

Text → Target Entity

- …which is a **support** or **attack** relation between arguments in abstract argumentation frameworks

Hundreds of Palestinians flee floods in Gaza as Israel opens dams

Disagree

[..] ‘southern Israel does not have any dams,’ said a statement from COGAT. [..]

Agree

Hundreds of Palestinians were evacuated after Israel opened the gates of several dams on the border with the Gaza strip and flooded at least 80 households. Israel has denied the claim as “entirely false”. [..]

(discuss)

(unrelated)

Pomerleau, D. & Rao, D. 2017. The fake news challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news. www.fakenewschallenge.org

Problem Solved?

- UKP Lab: **81.97** FNC-score in 2017 (2nd rank, 1st rank **82.02**)
- The FNC-score is problematic, it neglects the skewed class distribution
  - Getting only unrelated-related correct and predicting for "discuss" class yields 83.3 FNC-score: enough to win the FNC!

Re-assess the models with better metrics

- “Plain old” macro F1 score

Annotation studies on the original data

- Very challenging: 0.754 macro F1, 0.218 Fleiss’ kappa (on related classes)

Generalizing to another dataset

- Stance of newswire arguments, additional 18k instances

Top-Scoring FNC Systems on FNC corpus

- "Talos Intelligence Model" – deep CNN combined with gradient-boosted decision trees S. Baird et al. 2017. Talos targets disinformation with fake news challenge victory.
- ATHENE – ensemble of five MLP with 6 hidden layers + handcrafted features (UKP)
- "UCL Model" – multi-layer perceptron with bag-of-words features B. Riedel et al. 2017. A simple but tough-to-beat baseline for the fake news challenge stance detection task.
  - Results of the Fake News Challenge based on the **F1 metric**:

Model	Overall	Agree	Disagree	Discuss	Unrelated
Majority vote	21.0	0.0	0.0	0.0	83.9
Talos model	58.2	**53.9**	3.5	76.0	99.4
UCL model	58.3	47.9	11.4	74.7	98.9
ATHENE	**60.4**	48.7	**15.1**	**78.0**	**99.6**
Human UB	*75.4*	*58.8*	*66.7*	*76.5*	*99.7*

Top-Scoring FNC Systems on ARC Corpus

TECHNISCHE UNIVERSITÄT DARMSTADT

- Corpus – Argument Reasoning Comprehension [Habernal et al.] modified
- featMLP – ATHENE with an improved feature set (Lexical features that best performed in an ablation study)
- stackLSTM – A stacked Long Short-Term Memory Network combined with the improved feature set

Model	Overall	Agree	Disagree	Discuss	Unrelated
Majority vote	21.4	0.0	0.0	0.0	85.7
Talos model	**57.3**	**59.3**	**59.8**	16.0	**94.4**
UCL model	51.9	51.7	50.3	12.1	93.2
ATHENE	54.8	51.6	48.2	19.0	93.3
featMLP	52.6	52.6	50.6	14.4	93.4
stackLSTM	52.4	45.1	51.8	**19.4**	93.5
Human UB	*77.3*	*71.0*	*85.7*	*57.1*	*95.4*

UKP

Error Analysis for Top-Scoring FNC Systems

Observations:

- Models exploit lexical overlap between the two texts for classification
  →Lexical cues are important: “reports”, “said”, “false”, “hoax”, ..
- The models **fail when**:
  - Semantic relations between words need to be taken into account
    → Synonymy, Hyponymy, Entailment, …
  - Complex disagreement cases are encountered
    → Disagreement is often expressed in complex terms:
    e.g.: “If the bizarre story about ... sounded outlandish, that’s because it was”
  - Understanding of propositional content in general is required

Take Aways

- **What did we do?**
  - Revisited the problem setting
  - Introduced a new dataset
  - Tested high performing models

- **What are future challenges?**
  - Error analysis: models exploit similarity between the headline and the article body in terms of lexical overlap
  - Lexical cues, such as ”reports”, ”said”, ”false”, ”hoax” are important
  - Systems fail on semantic relations between words
    - Complex negation instances
    - Understanding of propositional content

Outline

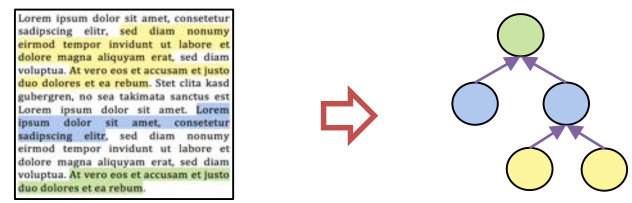
- Cross-topic Argument Mining from Heterogeneous Sources
  Stab, C., Miller, T., Schiller, B., Rai, P., Gurevych, I. (2018). https://arxiv.org/abs/1802.05758

- Document-Level Stance Classification for Fake News Detection
  Hanselowski, A., Schiller, B., Caspelherr, F., Avinesh PVS, Chaudhuri, D., Gurevych, I. (2018). COLING 2018, to appear.

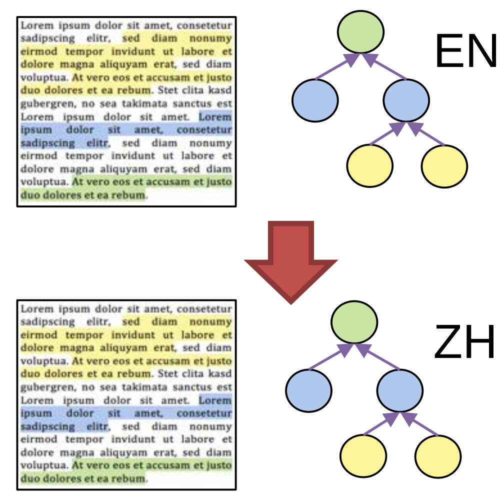
- Cross-Lingual Argumentation Mining: Machine Translation (and a bit of Projection) is All You Need
  Eger, S., Daxenberger, J., Stab, C., & Gurevych, I. (2018). COLING 2018, to appear.

Why cross-lingual NLP approach?

- Creating **annotated resources** for argumentation mining is expensive
  - Low agreement without training
  - Complex discourse comprehension (“disentangling thoughts”)



- Going across languages – annotation efforts grow with the number of languages
  - **Not feasible!**



- **Cross-lingual transfer** becomes critical: transferring from a high-resource language with labeled data to other languages

Experiments

- The task: Argument component extraction (Major claim, Claim, Premise)
- Data
  - Bilingual "**Microtexts**" (Peldszus & Stede, 2015), EN-DE
  - Monolingual "**Chinese Review Corpus**" (Li et al. 2017), ZH
  - Monolingual "**Persuasive Essays**" (Stab and Gurevych, 2017), EN
  - Parallel: Translated "Persuasive Essays" into DE (human) and DE, FR, ES, ZH (machine translation)

•Andreas Peldszus and Manfred Stede. 2015. An annotated corpus of argumentative microtexts. In Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation. Lisbon, Portugal, pages 801–815.
•Mengxue Li, Shiqiang Geng, Yang Gao, Shuhua Peng, Haijing Liu, and Hao Wang. 2017. Crowdsourcing Argumentation Structures in Chinese Hotel Reviews. In Proceedings of the 2017 IEEE International Conference on Systems, Man, and Cybernetics. Banff, Canada, pages 87–92.
•Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. Computational Linguistics 43(3):619–659.

Experimental Setup

Since [*it killed many marine lives*] [**tourism has threatened nature**]

[**Tourismus bedroht die Natur**] weil [*durch ihn viele Tiere sterben*]

Experimental Setup

Since [*it killed many marine lives*] [**tourism has threatened nature**]

[**Tourismus bedroht die Natur**] weil [*durch ihn viele Tiere sterben*]

- We adapt two popular approaches
  - **(1) Direct Transfer (operates on source language with gold labels)**
    Directly apply a model trained on *shared representations* (bilingual word embeddings) to the target language
  - **(2) Projection (operates on the target language with noisy labels)**
    Project annotations from source to target language on parallel data and train a system on the target language
    Need to adapt the projection algorithm to handle spans rather than individual tokens as in POS and NER

Experimental Setup

Since [*it killed many marine lives*] [**tourism has threatened nature**]

[**Tourismus bedroht die Natur**] weil [*durch ihn viele Tiere sterben*]

- We adapt two popular approaches
  - **(1) Direct Transfer (operates on source language with gold labels)**
    Directly apply a model trained on *shared representations* (bilingual word embeddings) to the target language
  - **(2) Projection (operates on the target language with noisy labels)**
    Project annotations from source to target language on parallel data and train a system on the target language
    Need to adapt the projection algorithm to handle spans rather than individual tokens as in POS and NER

- For both (1)+(2) need to take a **neural** model that can capture **long-range dependencies** for Argument Mining (i.e. can't use an HMM)

Experiments and Findings

1. **Microtexts dataset** is “too easy”
   - Transfer works well because arguments mostly depend on punctuation
2. **Chinese Hotel Reviews** ↔ **Persuasive Essays** is too difficult
   - Domain differences do not allow for direct cross-lingual transfer, with neither of the two approaches considered (worse than random baseline)

	$CRC \leftrightarrow PE_{EN}$				$MTX_{EN} \leftrightarrow MTX_{DE}$			
Model	ZH→ZH	ZH→EN	EN→EN	EN→ZH	EN→EN	EN→DE	DE→DE	DE→EN
BLCRF+Char	**46.31**	14.01	68.87	9.50	**73.12**	67.03	**73.41**	**66.62**
BLCRF	44.95	16.52	69.27	12.60	72.15	**69.46**	72.52	63.71
Baseline	18.	**17.**	20.	**17.**	45.	46.	50.	50.

Table 5: Direct transfer results for CRC and MTX. Scores are macro-F1. Embeddings are BISKIP-100.

Experiments and Findings

3. For **Persuasive Essays dataset** in parallel versions:
   - Projection works considerably better than Direct Transfer
   - Projection works very well independent of whether we use machine or human translated parallel data. In both cases, we almost reach the in-language upper bound

In-Language

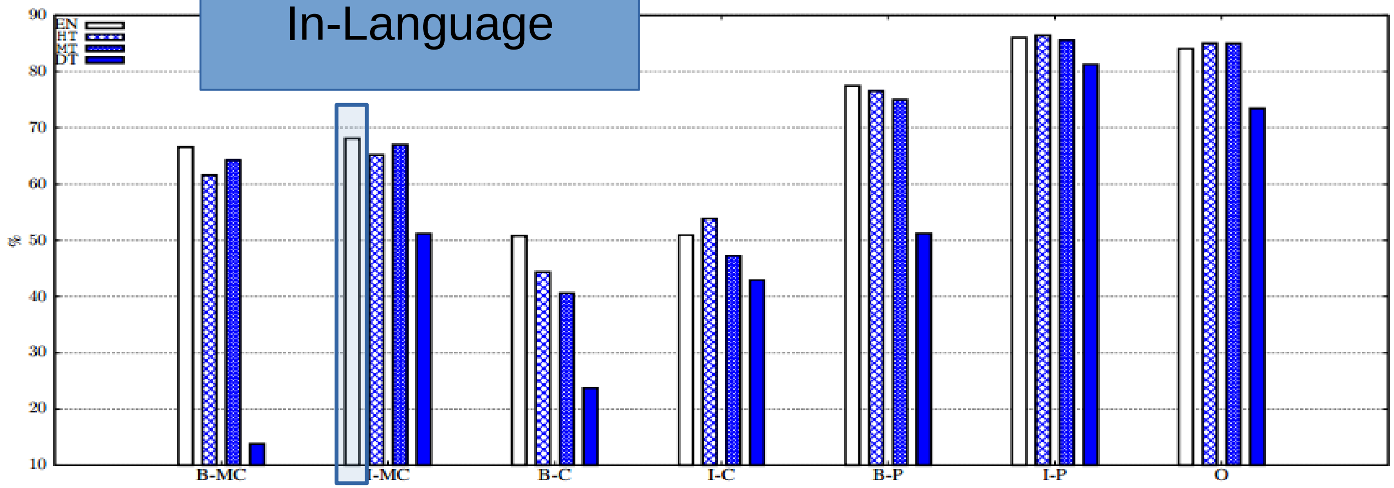


Figure 1: Individual F1-scores for four indicated systems and seven labels. All transfer systems are from PE_{DE} to PE_{EN}; EN is in-language. DT stands for Direct Transfer. HT/MT are projection-based approaches. Embeddings are BISKIP-100.

Experiments and Findings

3. For **Persuasive Essays dataset** in parallel versions:
   - Projection works considerably better than Direct Transfer
   - Projection works very well independent of whether we use machine or human translated parallel data. In both cases, we almost reach the in-language upper bound

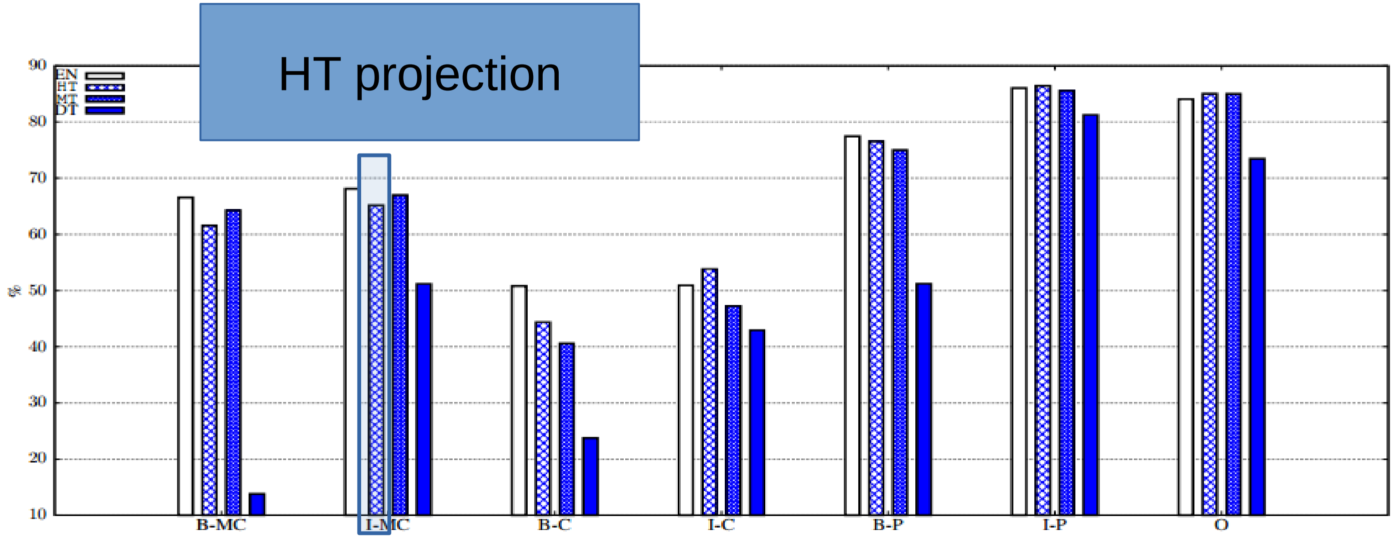


Figure 1: Individual F1-scores for four indicated systems and seven labels. All transfer systems are from PE_{DE} to PE_{EN}; EN is in-language. DT stands for Direct Transfer. HT/MT are projection-based approaches. Embeddings are BISKIP-100.

Experiments and Findings

TECHNISCHE UNIVERSITÄT DARMSTADT

3. For **Persuasive Essays dataset** in parallel versions:
   - Projection works considerably better than Direct Transfer
   - Projection works very well independent of whether we use machine or human translated parallel data. In both cases, we almost reach the in-language upper bound

MT projection

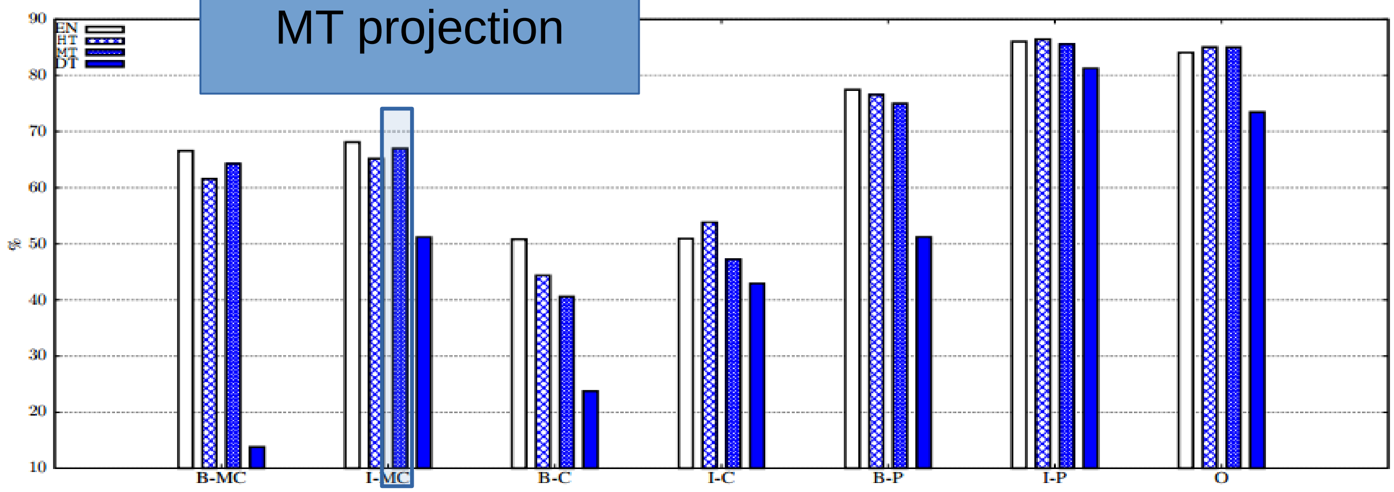


Figure 1: Individual F1-scores for four indicated systems and seven labels. All transfer systems are from PE_{DE} to PE_{EN}; EN is in-language. DT stands for Direct Transfer. HT/MT are projection-based approaches. Embeddings are BISKIP-100.

UKP

Experiments and Findings

3. For **Persuasive Essays dataset** in parallel versions:
   - Projection works considerably better than Direct Transfer
   - Projection works very well independent of whether we use machine or human translated parallel data. In both cases, we almost reach the in-language upper bound

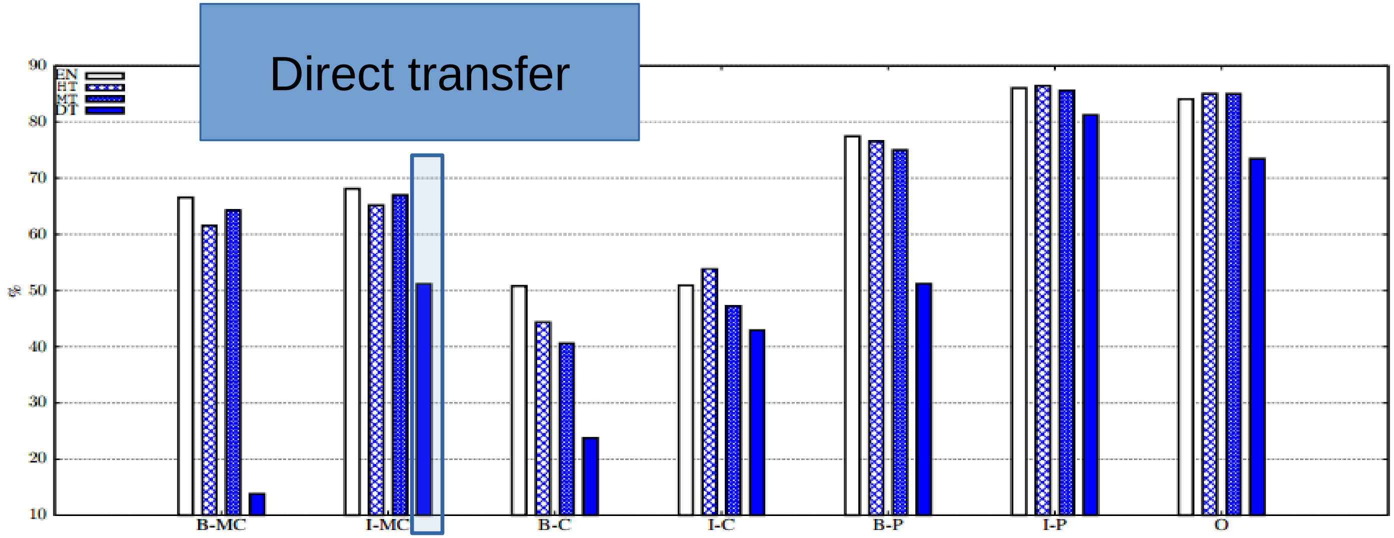


Figure 1: Individual F1-scores for four indicated systems and seven labels. All transfer systems are from PE_{DE} to PE_{EN}; EN is in-language. DT stands for Direct Transfer. HT/MT are projection-based approaches. Embeddings are BISKIP-100.

Experiments and Findings

TECHNISCHE UNIVERSITÄT DARMSTADT

3. For **Persuasive Essays dataset** in parallel versions:
    - Both direct transfer + annotation projection make errors at beginnings of components
    - Direct transfer: Due to "OOV" problem
    - Projection: Due to misalignments
    - Direct transfer makes a lot more errors
    - Almost no difference between HT and MT projection

- "When we have **no domain gap**, all we need is (very cheap) machine translation and (naive) projection"
- Code and data will be here: https://github.com/UKPLab/coling2018-xling_argument_mining

UKP

Conclusions

- Computational Argumentation has **great impact** on a large number of important downstream tasks
- **Cross-topic argument search** works well
- **Stance identification** is not yet solved
- Methods work O.K.-ish in a single-domain and cross-language, but **cross-domain is a big challenge**
- And real **inference and reasoning** is hard

- **Further research?**
  - Integrating knowledge and common-sense reasoning with neural networks
  - Pragmatic and social dimensions of argumentation
  - A vast number of open research problems

Conclusions

Contact

Iryna Gurevych
Technische Universität Darmstadt
Ubiquitous Knowledge Processing Lab

Hochschulstr. 10, 64289 Darmstadt, Germany
+49 (0)6151 16–25293
+49 (0)6151 16–25295
gurevych (at) ukp.informatik.tu-darmstadt.de

tps://pixabay.com/en/hand-leave-pen-paper-thank-you-226358/ (CC0 Public Domain)