Argument Mining Summer 2023

Cross-lingual Argumentation Mining: Machine Translation(a bit of projection) is all you Need!

Steffen Eger, Johannes Daxenberger,Christian Stab, and Iryna Gurevych

Akshat Gupta

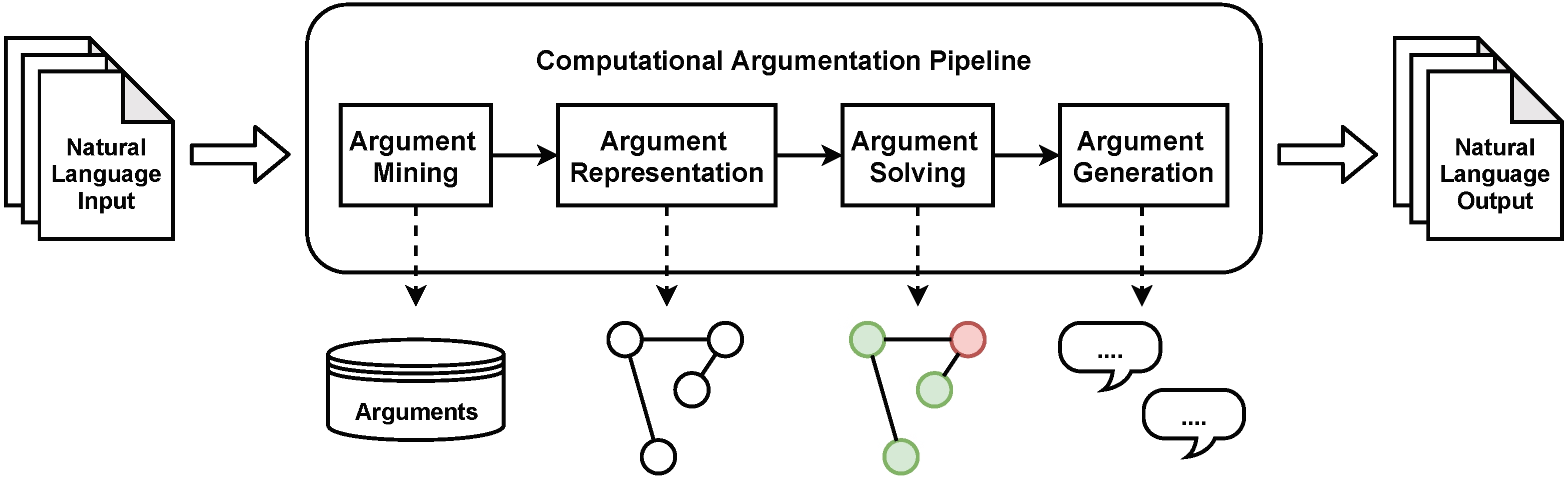
“Machine translation will only displace those humans who translate like a machine”

-Arle Richard Lommel

What is Argumentation Mining?

Automatic identification and extraction of the structure of inference and reasoning expressed as arguments.

Argumentation Mining Pipeline



Source: https://www.mdpi.com/2076-3417/11/15/7160

Motivation

- Current argument mining methods work in a single language
- They are not adequate for cross-lingual argument mining
- Lack of complexity
- Lack of parallel datasets

Table of Contents

Previous Research

- Argument unit segmentation
- Identification of argument components
- Recognizing argumentation discourse relations
- Extracting argument components
- Most of the research is on English data

Previous Research

- Cross-lingual sequence tagging in POS and NER

- Two major approaches are:
  - Projection
  - Direct Transfer

- In POS and NER, the label depends on token + context

- In argument mining, token, and context is absent

- So, methods in POS and NER fail in argument mining

Dataset

- Microtexts(MTX) by Peldszus and Stede (2015)
- Chinese Review Corpus(CRC) by Li et al. (2017)
- A Large-Scale Parallel Dataset of Persuasive Essays(PE)

Dataset Statistics

Name	Docum.	Tokens	Sentences	Major Cl.	Cl.	Prem.	Genre	Lang.
MTX	112	8,865 (en)	449	-	112	464	short texts	en, de
CRC	315	21,858	957	135	1,415	684	reviews	zh [en]
PE	402	148,186 (en)	7,141	751	1,506	3,832	persuasive essays	en [de, fr, es, zh]

Table 1: Statistics for datasets used in this work. Languages in brackets added by the current work.

Dataset Sample

Orig-EN	In the end , I think [**any great success need great work not great luck**] , even though [luck is one factor in reaching goal] but [*its impact is extraneous and we must not reckon on luck in our plans*] .
HT-DE-HumanAnno	Schließlich denke ich , dass [**jeder große Erfolg auf harter Arbeit statt Glück beruht**] , obwohl [Glück ein Faktor in der Erreichung des Ziels ist] , jedoch [*ist dessen Auswirkung unwesentlich und wir sollten uns nicht in unseren Projekten auf unser Glück verlassen*] .
HT-DE-ProjAnno	Schließlich denke ich , dass [**jeder große Erfolg auf harter Arbeit statt Glück beruht , obwohl Glück**] [ein Faktor in der Erreichung des Ziels ist] , jedoch [*ist dessen Auswirkung unwesentlich und wir sollten uns nicht in unseren Projekten auf unser Glück verlassen*] .
MT-DE-ProjAnno	Am Ende denke ich , dass [**jeder große Erfolg große Arbeit erfordert , nicht viel Glück**] , auch wenn [Glück ein Faktor beim Erreichen des Ziels] [*ist , aber seine Auswirkungen sind irrelevant und wir dürfen nicht mit Glück in unseren Plänen rechnen*] .
MT-ES-ProjAnno	Al final , creo que [**cualquier gran éxito requiere un gran trabajo y no mucha suerte**] , aunque la [suerte es un factor para alcanzar el objetivo] , pero [*su impacto es extraño y no debemos tener en cuenta la suerte en nuestros planes*] .
MT-FR-ProjAnno	En fin de compte , je pense que [**tout grand succès a besoin d' un bon travail , pas de chance**] , même si la [chance est un facteur d' atteinte de l' objectif] , mais [*son impact est étranger et nous ne devons pas compter sur la chance dans nos plans*] .
MT-ZH-ProjAnno	最后, 我认为[**任何伟大的成功都需要伟大的工作, 而不是运气好**] , 即使 [运气是达成目标的一个因素] , [但其影响是无关紧要的, 我们不能算计划中的运气] 。

Table 2: Human-annotated English sentence in the PE dataset as well as translations with human-created and projected annotations. Major claims in bold, claims underlined, premises in italics. HT/MT =human/machine translation.

Let's talk about data

Microtexts(MTX)

- By Peldszus and Stede in 2015
- 112 German shorts texts
- Written in response to a questions
- Phrased like "Should one do X"
- Annotated according to Freemans' theory of argumentation macro-structure
- Each text has one claim and several premises
- No "O" token and no major claims
- English translation of German sentences

Chinese Review Corpus(CRC)

- By Li et al. in 2017
- Large-scale argument dataset in Chinese
- Annotations are on the component level according to the claim-premise scheme
- Consist of hotel reviews from tripadvisor.com
- Four component types: major claim, claim, premise, and,
- Premise supporting an implicit claim
- But using only the top 3 components in research
- Used only Easy Review Corpus from CRC.

Persuasive Essays(PE)

- By Stab and Gurevych in 2017
- Essays are written on essayforum.com on conversational topics
- Human-translated german data for 402 essays with annotations
- Google Translate in German, French, Spanish, and Chinese
- Main focus is on EN <> DE and EN<>ZH

Approaches

- Two main approaches:
    - Direct Transfer
    - Projection

Direct Transfer

- A system trained on bilingual embeddings from scratch
- For EN <_> DE, BIVCD Embeddings were used with the BISKP model on 2 million aligned sentences, from europarl corpus
- BIVCD concatenates bilingual aligned sentences
- Used word2vec skip-gram model
- For EN <_> ZA, we trained the same model on the UN corpus with 11 million parallel sentences
- 100 and 200-dimensional embeddings were trained

Projection

- The problem of token-level argument mining
- Input is humans label L1 data and align it with parallel L2 data using fast-align
- Each argument in L1 of type a consists of MajorClaim, Claim, Premise
- Label all words in L2 sentences between scope with type a, using the BIO structure
- Projections of the words which do not align are ignored.

Experiments

- Token-level sequence tagging is performed with BIO labels
- Claim, Premise, and MajorClaim are the main classes
- Used 2 bi-directional LSTM layer with CRF layer of 100 units to get word embedding
- Another bi-directional LSTM model with 50 units is used to learn the character level representation

Experiments

- Concatenate both word embedding and character representation and call this BLCRF + char

- Training is done for 50 epochs

- The F1 score is used for evaluation

Baseline

- Choose majority label in test data and performance is poor on token-level
- Split the dataset by sentences and compute the probability distribution of how likely each argument appears in the sentence
- Label all the tokens in the sentence with the argument component type with BIO structure
- Label the last token with an "O" label in PE and CRC

Train/Dev/Test Split

- For the PE corpus, 286 essays were in the train and 80 in test data with 106k and 29k tokens respectively.
- 36 essays with 12k tokens which is10% of training is used as a dev set
- Average score of 5 random initialization is reported in results

Train/Dev/Test Split

- For CRC, we perform 5-fold cross-validation
- Training has 15k tokens, dev has 2k and test has 4k
- For MTX, 5-fold cross-validation
- The train has 6k tokens, dev has 500, and the test has 1500 tokens

Results

Model	Embedding Type	EN→EN	EN→DE	DE→DE	DE→EN
BLCRF+Char	BIVCD-100	68.87	41.89	65.22	49.91
	BIVCD-200	$\boxed{\textbf{70.51}}$	39.87	$\boxed{\textbf{65.92}}$	49.52
	BISKIP-100	69.27	37.01	63.33	48.23
BLCRF	BIVCD-100	69.27	49.70	65.90	50.14
	BISKIP-100	69.15	$\boxed{\textbf{49.76}}$	64.92	$\boxed{\textbf{50.28}}$
Baseline		20.	20.	20.	20.

Table 4: Direct transfer results for $PE_{EN} \leftrightarrow PE_{DE}$. Scores are macro-F1.

Results: For Direct Transfer

- English language results are above **69% macro F1**
- German in-language results are **4-5%** below English
- Reason: German is more complex than English
- Drop > **40%** for the direction of **EN→DE** and less for **DE→EN**
- Reason: Due to a discrepancy between train and test distribution
- **EN→ DE** performance increases from **40% to 50%** when disabling character information

Results

Model	CRC↔PE_{EN} ZH→ZH	ZH→EN	EN→EN	EN→ZH	MTX_{EN}↔MTX_{DE} EN→EN	EN→DE	DE→DE	DE→EN
BLCRF+Char	**46.31**	14.01	**69.27**	9.50	**73.12**	67.03	**73.41**	**66.62**
BLCRF	44.95	16.52	69.15	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.

Results: For Direct Transfer

- Reason: Diverging German character sequences
- Language CRC results are lower than language PE(**46% vs 69% for PE**)
- Reason: Due to the domain gap between student essays and reviews
- For MTX, the smallest dataset yields the highest F1 scores
- Language drop is small
- Reason: Arguments are separated by punctuation symbols which is easy to learn

Error Analysis and Discussion

- A major source of incorrect classification of tokens labeled as "B"
- The "blurring effect" at test time makes the detection of exact arguments difficult
- German and English have more than 97% cosine similarity in BISKIP-100d
- Apart from the semantic shift, direct transfer also faces syntactic shift
- Language adaption between CRC and PE corpus is difficult because argumentation units are very different

Projection

	EN→DE			DE→EN		
	HT	MT	In-Lang.	HT	MT	In-Lang.
BLCRF+Char	63.67	**64.00**	63.33	**67.57**	66.39	69.27
BLCRF	61.18	63.34	64.92	64.87	64.68	69.15

Table 6: Projection on HT/MT translations, evaluated on human-created test data. Scores are macro-F1. Embeddings are BISKIP-100.

HT Projection

- For PE English and parallel German HT data, sores improved from **49.76%** to **63.67%** which is a **30%** increase for English to German.

- In German to English also, improvement is **30%** compared to direct transfer

Projection

	EN→DE			DE→EN		
	HT	MT	In-Lang.	HT	MT	In-Lang.
BLCRF+Char	63.67	**64.00**	63.33	**67.57**	66.39	69.27
BLCRF	61.18	63.34	64.92	64.87	64.68	69.15

Table 6: Projection on HT/MT translations, evaluated on human-created test data. Scores are macro-F1. Embeddings are BISKIP-100.

MT Projection

- For the PE dataset, English-to-German, results are better than German to English
- Overall, machine translations are as good as human translations
- For CRC, using MT, the f1 score improves from **16.52% to 23.15%**

Other Languages

- MT translation of PE in French, Spanish, and Chinese
- No human test data so evaluation is done on machine translations and projected annotations.
- For BLCRF + char model, performance scores were **62.45%, 65.92%, and 59.20%** for French, Spanish, and Chinese
- For German and English, scores were **63.20%** and **61.45%**
- For CRC, with BLCRF + char obtains **47.92%**

Error Analysis

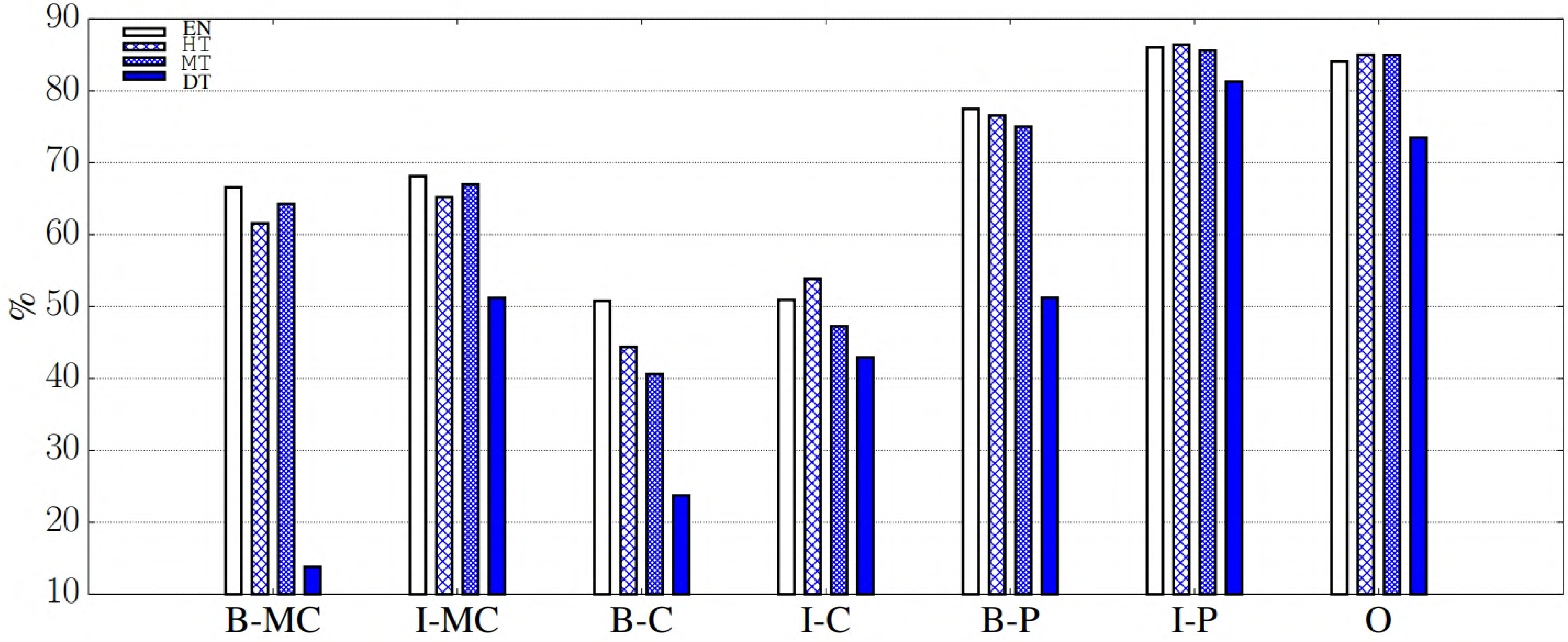


Figure 1: Individual F1-scores for four indicated systems and seven labels. All transfer systems are from PE_{DE} to PE_{EN}; EN is in-language. DT stands for Direct Transfer. `HT`/`MT` are projection-based approaches. Embeddings are BISKIP-100. Systems are BLCRF+Char.

Error Analysis

- The main bottleneck is the quality of cross-lingual projections
- Algorithm projections match is 97.24% for en→de
- The most mismatch is between "B" and "I" with the "O" category

Conclusion

- Currently, available datasets for AM are not adequate for cross-lingual AM
- Created human and machine translations of AM dataset
- Machine Translation and projection work better than direct transfer
- Translation and projected labels can create data in multiple languages, eliminating OOV and ordering problems.
- Machine translation in combination with projection performs on the level of in-language upper-bound results.

Thank you
Questions?