Towards a Broad Coverage Named Entity Resource: A Data-Efficient Approach for Many Diverse Languages

Silvia Severini, Ayyoob Imani,
Philipp Dufter, Hinrich Schütze

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

LREC 2022 : 13th Conference on Language Resources and Evaluation

Outline

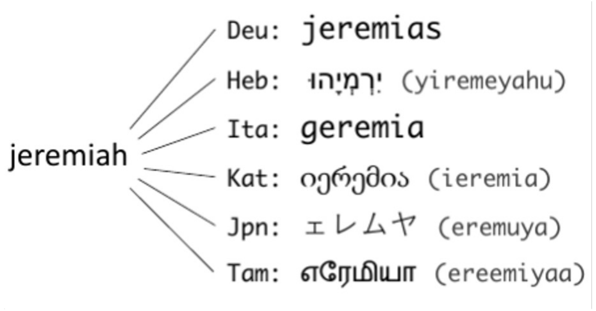
1. Introduction
2. Method
3. Evaluation and Analysis
4. Use cases
5. Resource
6. Conclusion

Outline

1. Introduction
2. Method
3. Evaluation and Analysis
4. Use cases
5. Resource
6. Conclusion

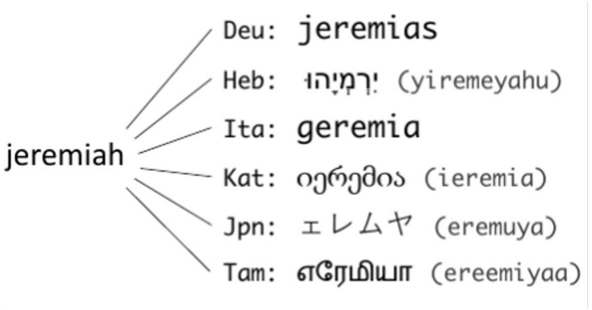
Introduction

- Named entities (NEs):
  - Crucial for monolingual and cross-lingual NLP tasks
  - Multilingual NE lexicons are not available for many low-resource languages



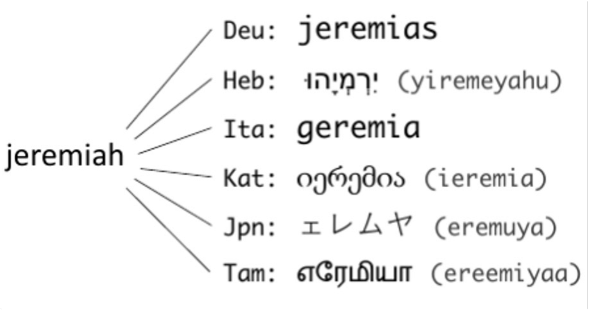
Introduction

- Named entities (NEs):
    - Crucial for monolingual and cross-lingual NLP tasks
    - Multilingual NE lexicons are not available for many low-resource languages
- **Goal**: create a MNE resource for low-resource languages



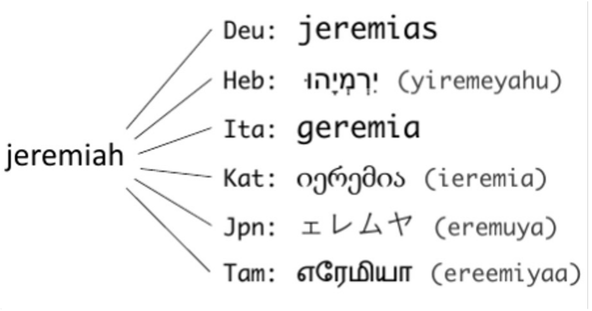
Introduction

- Named entities (NEs):
  - Crucial for monolingual and cross-lingual NLP tasks
  - Multilingual NE lexicons are not available for many low-resource languages
- **Goal**: create a MNE resource for low-resource languages
- Approach:
  - We use the corpus that has the best coverage of low-resource languages: Parallel Bible Corpus (PBC)



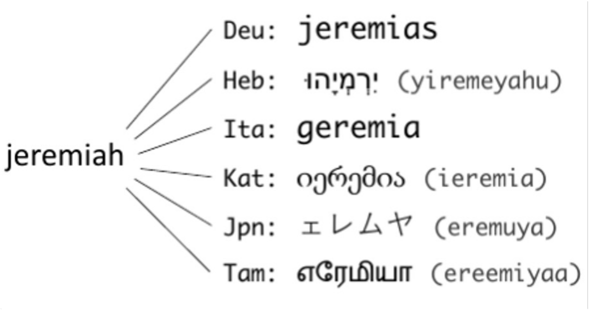
Introduction

- Named entities (NEs):
  - Crucial for monolingual and cross-lingual NLP tasks
  - Multilingual NE lexicons are not available for many low-resource languages
- **Goal**: create a MNE resource for low-resource languages
- Approach:
  - We use the corpus that has the best coverage of low-resource languages: Parallel Bible Corpus (PBC)
  - For most languages no other resource available (no named entity recognizer, no annotated data, no pretrained LMs)



Introduction

- Named entities (NEs):
  - Crucial for monolingual and cross-lingual NLP tasks
  - Multilingual NE lexicons are not available for many low-resource languages
- **Goal**: create a MNE resource for low-resource languages
- Approach:
  - We use the corpus that has the best coverage of low-resource languages: Parallel Bible Corpus (PBC)
  - For most languages no other resource available (no named entity recognizer, no annotated data, no pretrained LMs)
  - **Our method creates a very broad-coverage NE resource based on parallel text only**



Contribution

- We present **CLC-BN**, a method that first identifies named entity correspondences in a parallel corpus and then learns a neural transliteration model from them

Contribution

- We present **CLC-BN**, a method that first identifies named entity correspondences in a parallel corpus and then learns a neural transliteration model from them

- We annotate a set of NEs to evaluate CLC-BN's performance on **13** languages through crowdsourcing and show a performance increase in comparison to prior work
  - We release the gold annotated sets as a resource for future work

Contribution

- We present **CLC-BN**, a method that first identifies named entity correspondences in a parallel corpus and then learns a neural transliteration model from them
- We annotate a set of NEs to evaluate CLC-BN's performance on **13** languages through crowdsourcing and show a performance increase in comparison to prior work
  - We release the gold annotated sets as a resource for future work
- Using CLC-BN, we create and release a named entity resource for **1340** languages

Outline

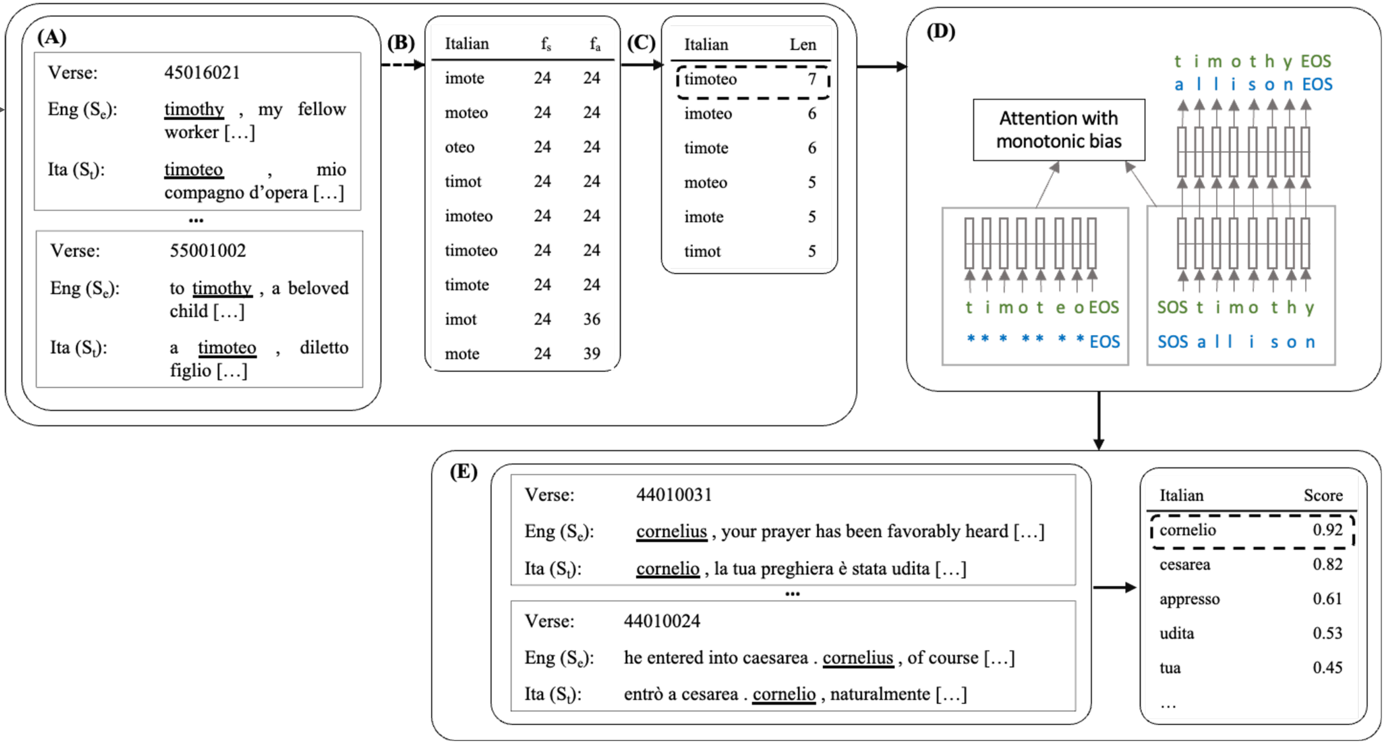
1. Introduction
2. Method
3. Evaluation and Analysis
4. Use cases
5. Resource
6. Conclusion

Outline

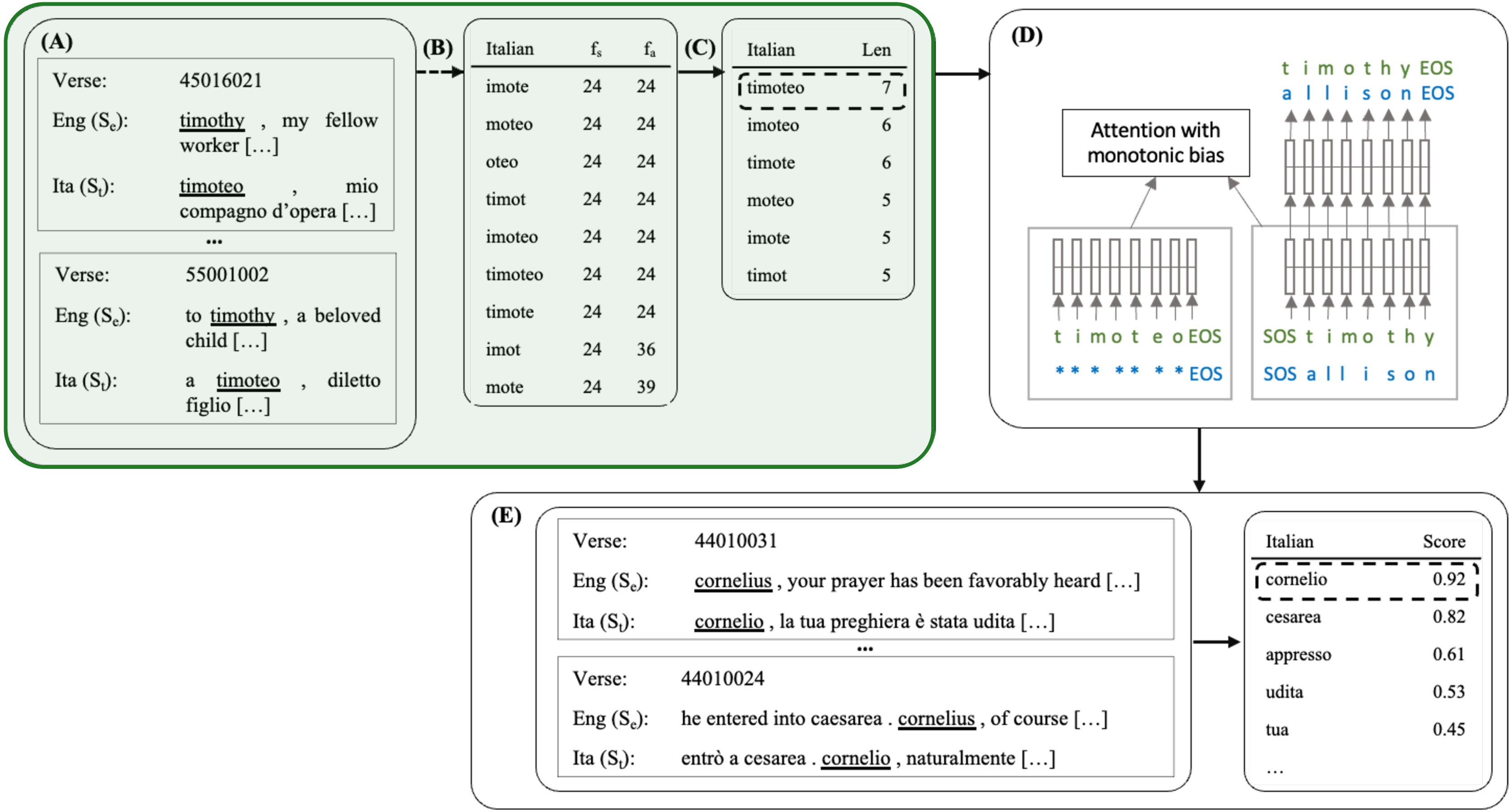
CLC-BN: CLC-Bootstraping + Neural transliteration

English NE:

Timothy



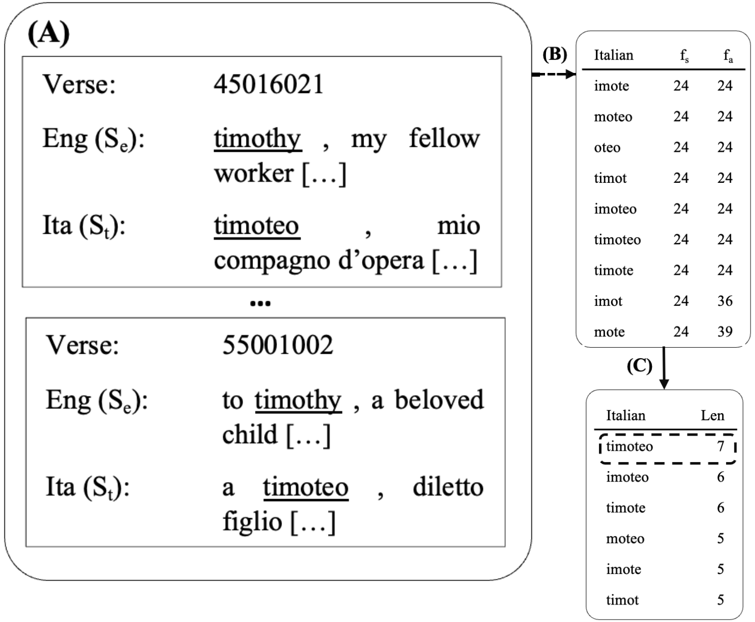
CLC-B: extract character-level correspondences



CLC-B: extract character-level correspondences

A. Extract the parallel subcorpus that contains **Timothy**

English NE: **Timothy**



CLC-B: extract character-level correspondences

A. Extract the parallel subcorpus that contains **Timothy**

B. For all character n-grams in the target corpus, determine f_s and f_a . Discard n-grams with f_a>50

English NE: **Timothy**

(A)

Verse:	45016021
Eng (S_e):	timothy , my fellow worker […]
Ita (S_t):	timoteo , mio compagno d'opera […]

…

Verse:	55001002
Eng (S_e):	to timothy , a beloved child […]
Ita (S_t):	a timoteo , diletto figlio […]

Italian	f_s	f_a
imote	24	24
moteo	24	24
oteo	24	24
timot	24	24
imoteo	24	24
timoteo	24	24
timote	24	24
imot	24	36
mote	24	39

Italian	Len
timoteo	7
imoteo	6
timote	6
moteo	5
imote	5
timot	5

CLC-B: extract character-level correspondences

A. Extract the parallel subcorpus that contains **Timothy**

B. For all character n-grams in the target corpus, determine f_s and f_a . Discard n-grams with f_a>50

C. Filter the remaining n-grams:

  a. Keep n-grams with the highest f_s

  b. Keep n-grams with the minimum absolute difference between f_s and f_a

  c. Return the n-gram with the smallest length difference

English NE: **Timothy**

(A)

Verse:	45016021
Eng (S_e):	timothy , my fellow worker […]
Ita (S_t):	timoteo , mio compagno d'opera […]

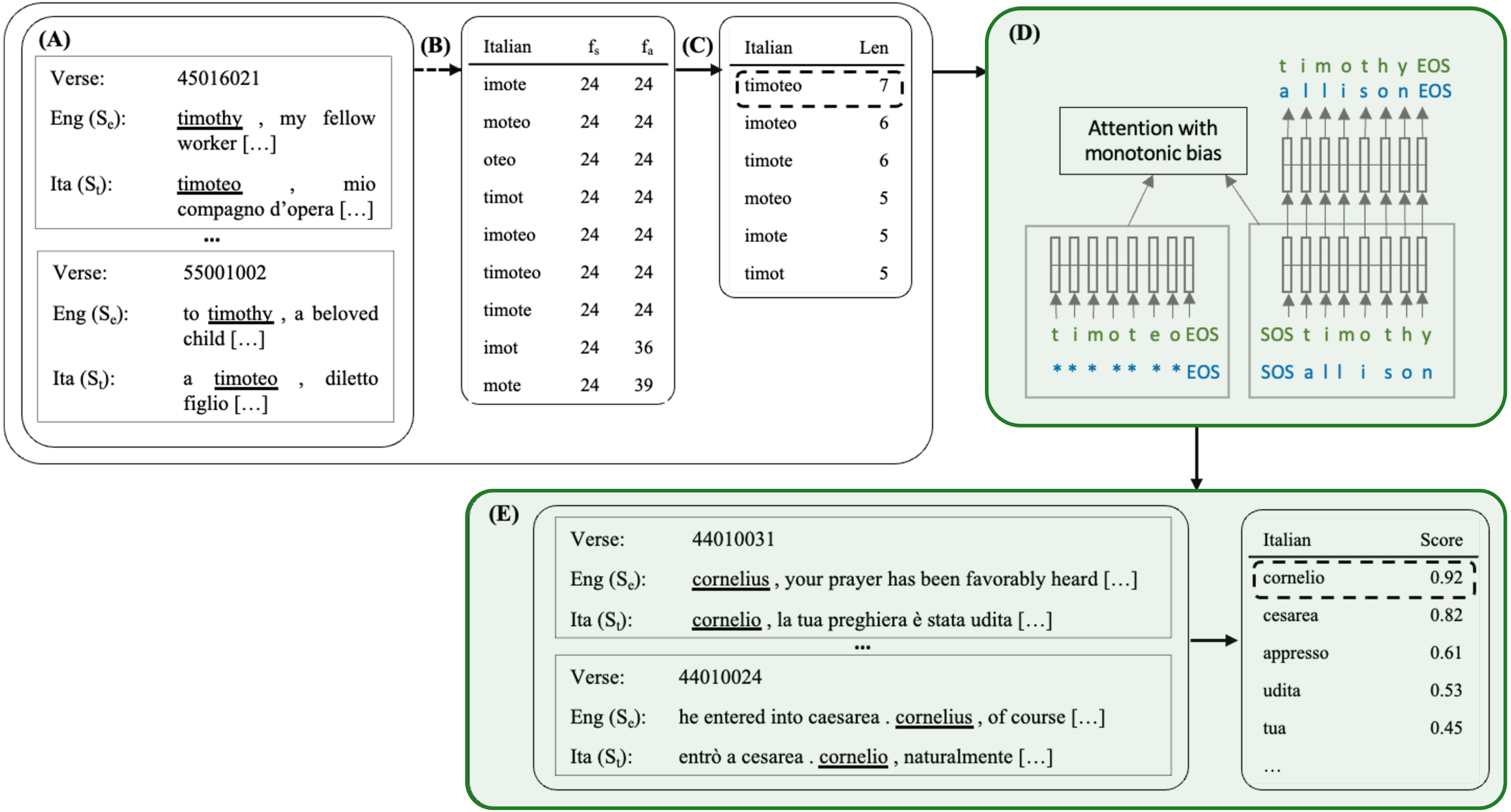
…

Verse:	55001002
Eng (S_e):	to timothy , a beloved child […]
Ita (S_t):	a timoteo , diletto figlio […]

Italian	f_s	f_a
imote	24	24
moteo	24	24
oteo	24	24
timot	24	24
imoteo	24	24
timoteo	24	24
timote	24	24
imot	24	36
mote	24	39

Italian	Len
timoteo	7
imoteo	6
timote	6
moteo	5
imote	5
timot	5

CLC-BN: Neural Transliteration model



CLC-BN: Neural Transliteration model

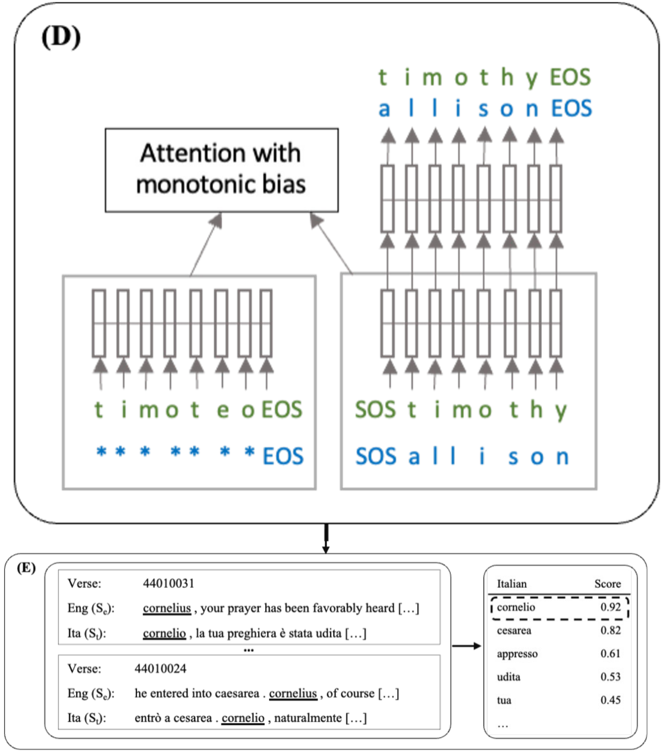
- Goal: mine pairs with a neural Seq2seq model

CLC-BN: Neural Transliteration model

- Goal: mine pairs with a neural Seq2seq model

Model (D):

- Character-level Bi-GRU (Target-to-Source)
- English data augmentation with Wikipedia dump and Flair POS tagger
- Monotonic bias



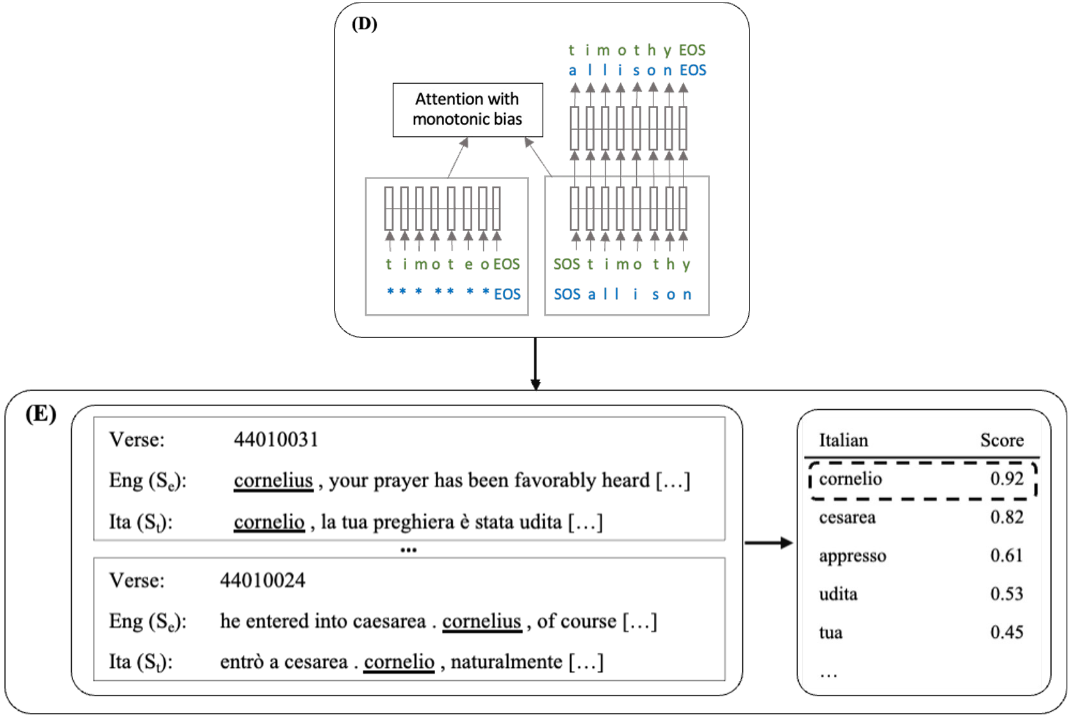
CLC-BN: Neural Transliteration model

- Goal: mine pairs with a neural Seq2seq model

Model (D):

- Character-level Bi-GRU (Target-to-Source)
- English data augmentation with Wikipedia dump and Flair POS tagger
- Monotonic bias

Candidate words: all words in the parallel target verses in which the English word appears



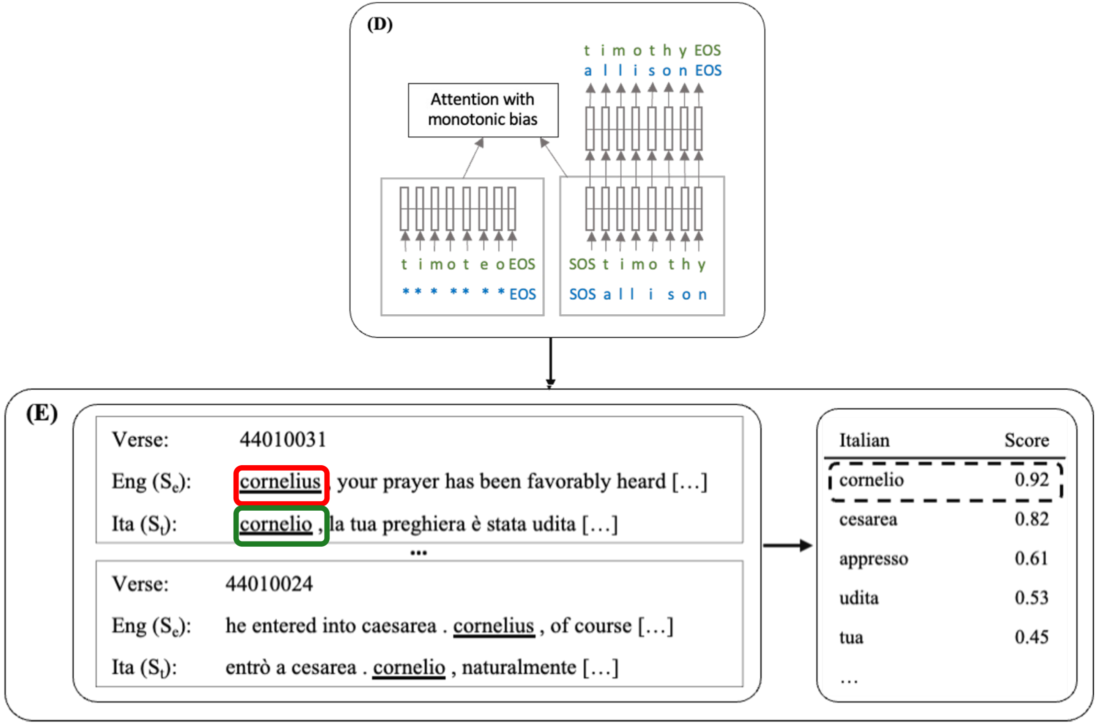
CLC-BN: Neural Transliteration model

- Goal: mine pairs with a neural Seq2seq model

Model (D):

- Character-level Bi-GRU (Target-to-Source)
- English data augmentation with Wikipedia dump and Flair POS tagger
- Monotonic bias

Candidate words: all words in the parallel target verses in which the English word appears



CLC-BN: Neural Transliteration model

- Goal: mine pairs with a neural Seq2seq model

Model (D):

- Character-level Bi-GRU (Target-to-Source)
- English data augmentation with Wikipedia dump and Flair POS tagger
- Monotonic bias

Candidate words: all words in the parallel target verses in which the English word appears

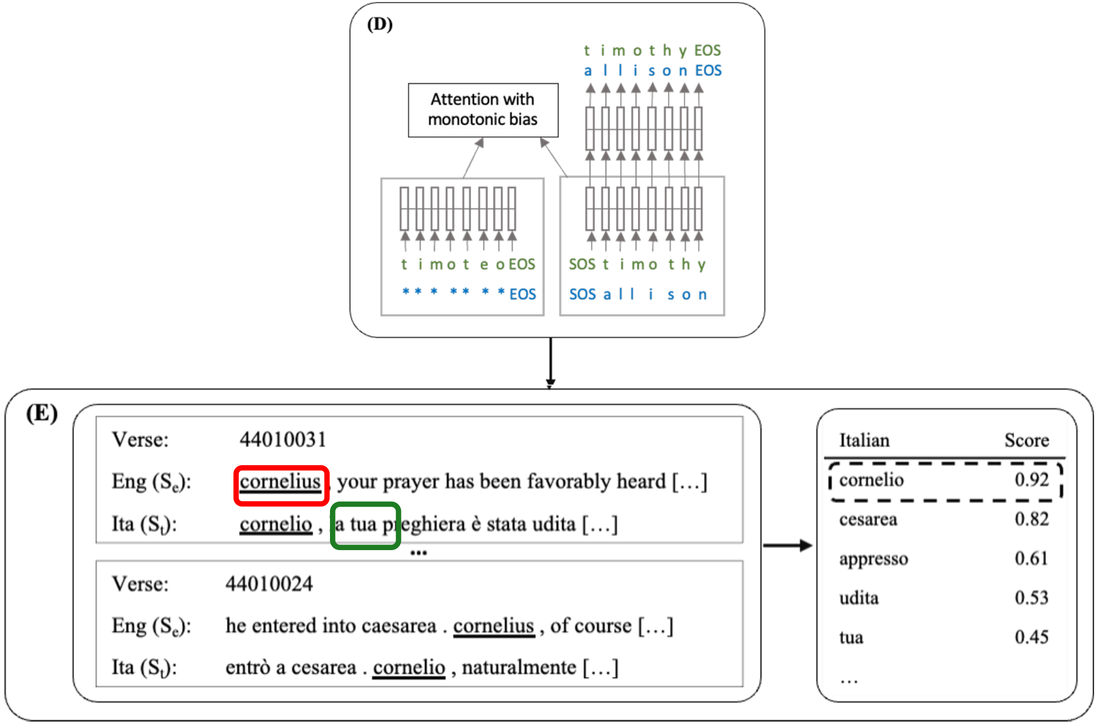
CLC-BN: Neural Transliteration model

- Goal: mine pairs with a neural Seq2seq model

Model (D):

- Character-level Bi-GRU (Target-to-Source)
- English data augmentation with Wikipedia dump and Flair POS tagger
- Monotonic bias

Candidate words: all words in the parallel target verses in which the English word appears



Outline

1. Introduction
2. Method
3. Evaluation and Analysis
4. Use cases
5. Resource
6. Conclusion

Outline

1. Introduction
2. Method
3. Evaluation and Analysis
4. Use cases
5. Resource
6. Conclusion

Experimental setup

- Parallel Bible Corpus (PBC)
- Evaluation over 13 languages with different scripts, resource availabilities, and language families
- Silver evaluation using the Google translation API
- Gold human evaluation through crowd-sourcing

	Lang	ISO	# verses	# parallel
low-resource languages	Arabic	Arb	31173	31062
	Finnish	Fin	31167	31061
	Greek	Ell	31183	31062
	Russian	Rus	31173	31062
	Spanish	Spa	31167	31062
	Swedish	Swe	31167	31062
	Zulu	Zul	31167	31062
lowest-resource languages	Hebrew	Heb	7952	7917
	Hindi	Hin	7952	7917
	Kannada	Kan	7952	7917
	Korean	Kor	7913	7869
	Georgian	Kat	4904	4844
	Tamil	Tam	7942	7917

Gold human evaluation - baseline

Low-resource setting

	Arb	Ell	Fin	Spa	Swe	Rus	Zul	AVG
Wu et al. (2018)	70.0	80.0	90.0	91.7	88.3	72.9	84.8	82.5
CLC-B	56.7	45.0	50.0	48.3	48.3	57.6	74.6	54.4
CLC-BN	<u>81.7</u>	<u>91.7</u>	<u>93.3</u>	<u>96.7</u>	<u>91.7</u>	<u>84.8</u>	<u>93.2</u>	<u>90.4</u>

Lowest-resource setting

	Heb	Hin	Kan	Kat	Kor	Tam	AVG
Wu et al. (2018)	62.5	76.3*	61.7	70.0	54.2	66.1*	65.1
CLC-B	51.8	39.0*	48.3	45.0	37.3	47.5*	44.8
CLC-BN	71.4	<u>94.9</u>*	<u>93.3</u>	<u>88.3</u>	<u>78.0</u>	<u>91.5</u>*	<u>86.2</u>

Gold human evaluation - baseline

Low-resource setting

	Arb	Ell	Fin	Spa	Swe	Rus	Zul	AVG
Wu et al. (2018)	70.0	80.0	90.0	91.7	88.3	72.9	84.8	82.5
CLC-B	56.7	45.0	50.0	48.3	48.3	57.6	74.6	54.4
CLC-BN	81.7	91.7	93.3	96.7	91.7	84.8	93.2	90.4

Lowest-resource setting

	Heb	Hin	Kan	Kat	Kor	Tam	AVG
Wu et al. (2018)	62.5	76.3∗	61.7	70.0	54.2	66.1∗	65.1
CLC-B	51.8	39.0∗	48.3	45.0	37.3	47.5∗	44.8
CLC-BN	71.4	94.9∗	93.3	88.3	78.0	91.5∗	86.2

Gold human evaluation - baseline

Low-resource setting

	Arb	Ell	Fin	Spa	Swe	Rus	Zul	AVG
Wu et al. (2018)	70.0	80.0	90.0	91.7	88.3	72.9	84.8	82.5
CLC-B	56.7	45.0	50.0	48.3	48.3	57.6	74.6	54.4
CLC-BN	81.7	91.7	93.3	96.7	91.7	84.8	93.2	90.4

Lowest-resource setting

	Heb	Hin	Kan	Kat	Kor	Tam	AVG
Wu et al. (2018)	62.5	76.3*	61.7	70.0	54.2	66.1*	65.1
CLC-B	51.8	39.0*	48.3	45.0	37.3	47.5*	44.8
CLC-BN	71.4	94.9*	93.3	88.3	78.0	91.5*	86.2

Gold human evaluation - word alignment

Low-resource setting

	Arb	Ell	Fin	Spa	Swe	Rus	Zul	AVG
Östling et al. (2016)	61.7	88.3	76.7	86.7	85.0	83.1	86.4	81.1
Sabet et al. (2020)	20.0	40.0	60.0	45.0	50.0	45.8	25.4	40.9
CLC-B	56.7	45.0	50.0	48.3	48.3	57.6	74.6	54.4
CLC-BN	81.7	91.7	93.3	96.7	91.7	84.8	93.2	90.4

Lowest-resource setting

	Heb	Hin	Kan	Kat	Kor	Tam	AVG
Östling et al. (2016)	83.9	69.5*	38.3	68.3	33.9	35.6*	57.5
Sabet et al. (2020)	23.2	47.5*	46.7	20.0	40.0	47.5*	37.5
CLC-B	51.8	39.0*	48.3	45.0	37.3	47.5*	44.8
CLC-BN	71.4	94.9*	93.3	88.3	78.0	91.5*	86.2

Gold human evaluation - word alignment

Low-resource setting

	Arb	Ell	Fin	Spa	Swe	Rus	Zul	AVG
Östling et al. (2016)	61.7	88.3	76.7	86.7	85.0	83.1	86.4	81.1
Sabet et al. (2020)	20.0	40.0	60.0	45.0	50.0	45.8	25.4	40.9
CLC-B	56.7	45.0	50.0	48.3	48.3	57.6	74.6	54.4
CLC-BN	81.7	91.7	93.3	96.7	91.7	84.8	93.2	90.4

Lowest-resource setting

	Heb	Hin	Kan	Kat	Kor	Tam	AVG
Östling et al. (2016)	83.9	69.5∗	38.3	68.3	33.9	35.6∗	57.5
Sabet et al. (2020)	23.2	47.5∗	46.7	20.0	40.0	47.5∗	37.5
CLC-B	51.8	39.0∗	48.3	45.0	37.3	47.5∗	44.8
CLC-BN	71.4	94.9∗	93.3	88.3	78.0	91.5∗	86.2

Error Analysis

#	English	Arabic	Finnish	Greek	Hebrew	Kannada	Russian	Tamil
28	elijah	alalihaau	eliaa	elia	veaeliyahu	eliiyanaagali	elisei	eliyaavaa
12	titus	tiytusa	titus	titos	titos	titanannu	titu	tiittuvin
8	elizabeth	aaliysaabaata	elisabet	elisabet	elisheva	elisabeet	elizaveta	elicapet
3	miletus	miyliytusa	miletokseen	mileto	lemilitos	mileetakke	mileta	mileettu
2	rufus	ruwfusa	rufuksen	roufo	vishelom	uphaniguu	rufa	ruupuvukkum
2	hermes	wahirmisa	hermeeksi	epairne	heremes	meeyaniguu	germes	ermee

Error Analysis

#	English	Arabic	Finnish	Greek	Hebrew	Kannada	Russian	Tamil
28	elijah	alalihaau	eliaa	elia	veaeliyahu	eliiyanaagali	elisei	eliyaavaa
12	titus	tiytusa	titus	titos	titos	titanannu	titu	tiittuvin
8	elizabeth	aaliysaabaata	elisabet	elisabet	elisheva	elisabeet	elizaveta	elicapet
3	miletus	miyliytusa	miletokseen	mileto	lemilitos	mileetakke	mileta	mileettu
2	rufus	ruwfusa	rufuksen	roufo	vishelom	uphaniguu	rufa	ruupuvukkum
2	hermes	wahirmisa	hermeeksi	epairne	heremes	meeyaniguu	germes	ermee

Outline

1. Introduction
2. Method
3. Evaluation and Analysis
4. Use cases
5. Resource
6. Conclusion

Outline

1. Introduction
2. Method
3. Evaluation and Analysis
4. Use cases
5. Resource
6. Conclusion

Use cases

- Transliteration

Use cases

- Transliteration
- Extending existing multilingual resources (i.e., BabelNet)

Lang.	CLC-BN	Babel	New NEs %
Arb	977	683	30.1
Fin	979	647	33.9
Ell	979	658	32.8
Rus	485	449	7.4
Spa	979	784	19.9
Swe	979	684	30.1
Zul	979	471	51.9
Heb	467	413	11.6
Hin	467	334	28.5
Kan	467	299	36.0
Kor	467	386	17.3
Kat	368	271	26.4
Tam	433	318	26.6
Jpn	979	715	27.0
Zho	979	698	28.7
Tha	467	337	27.8
AVG.	715	509	27.2

Use cases

- Transliteration
- Extending existing multilingual resources (i.e., BabelNet)

Lang.	CLC-BN	Babel	New NEs %
Arb	977	683	30.1
Fin	979	647	33.9
Ell	979	658	32.8
Rus	485	449	7.4
Spa	979	784	19.9
Swe	979	684	30.1
Zul	979	471	51.9
Heb	467	413	11.6
Hin	467	334	28.5
Kan	467	299	36.0
Kor	467	386	17.3
Kat	368	271	26.4
Tam	433	318	26.6
Jpn	979	715	27.0
Zho	979	698	28.7
Tha	467	337	27.8
AVG.	715	509	27.2

Use cases

- Transliteration
- Extending existing multilingual resources (i.e., BabelNet)
- Cross-lingual mapping of word embeddings
  - VecMap
  - Bilingual Lexicon Induction (MUSE)

	Eng-Jpn	Eng-Tam	Eng-Zho
Unsupervised	0.0	0.0	0.0
Semisupervised	30.43	14.4	30.1

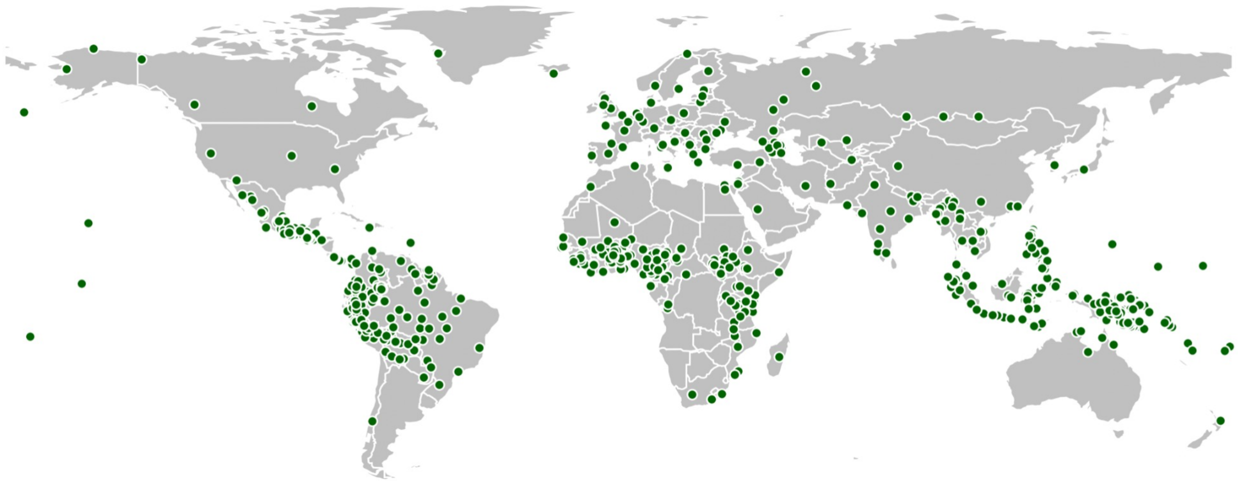
Outline

1. Introduction
2. Method
3. Evaluation and Analysis
4. Use cases
5. Resource
6. Conclusion

Outline

Resource

- 1340 languages, 1134 of which are lowest-resource, average of 503 NEs per language
- Best represented language families: Austronesian, Niger-Congo and Indo-European
- We cover all major areas of linguistic diversity (e.g., Amazonian, African, and Papua New Guinea)
- Our NEs resource is freely available at http://cistern.cis.lmu.de/ne_bible/



Geographical distribution of some languages in the PBC (Mayer et al., 2014) and our resource

Example: English – Italian resource

English	Italian
alexander	alessandro
deborah	debora
egypt	egitto
jahaziel	iahaziel
lucius	lucio
philadelphia	filadelfia
rachel	rachele
tiberius	tiberio

Example: English – Italian resource

English	Italian
alexander	alessandro
deborah	debora
egypt	egitto
jahaziel	iahaziel
lucius	lucio
philadelphia	filadelfia
rachel	rachele
tiberius	tiberio

Outline

1. Introduction
2. Method
3. Evaluation and Analysis
4. Use cases
5. Resource
6. Conclusion

Outline

Conclusion

- We presented CLC-BN, a new method that identifies NE correspondences using co-occurrence statistics and a neural transliteration model

Conclusion

- We presented CLC-BN, a new method that identifies NE correspondences using co-occurrence statistics and a neural transliteration model

- We showed that it outperforms prior work on human-annotated gold data

Conclusion

- We presented CLC-BN, a new method that identifies NE correspondences using co-occurrence statistics and a neural transliteration model
- We showed that it outperforms prior work on human-annotated gold data
- We illustrated its utility for knowledge graph augmentation and Bilingual Lexicon Induction

Conclusion

- We presented CLC-BN, a new method that identifies NE correspondences using co-occurrence statistics and a neural transliteration model
- We showed that it outperforms prior work on human-annotated gold data
- We illustrated its utility for knowledge graph augmentation and Bilingual Lexicon Induction
- We publish a new NE resource for 1340 languages by applying CLC-BN to the Parallel Bible Corpus

References (selected)

- Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019). Flair: An easy- to-use framework for state-of-the-art nlp. In NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pages 54–59.
- Mayer, Thomas and Cysouw, Michael. (2014). Creating a massively parallel bible corpus. Navigli, Roberto and Ponzetto, Simone Paolo. (2012).
- Wu, Winston and Vyas, Nidhi and Yarowsky, David. (2018). Creating a translation matrix of the Bible's names across 591 languages.
- Östling, R., Tiedemann, J., et al. (2016). Efficient word alignment with markov chain monte carlo. The Prague Bulletin of Mathematical Linguistics.
- Sabet, M. J., Dufter, P., Yvon, F., and Schütze, H. (2020). Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1627–1643.
- BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Elsevier.
- Alexis Conneau and Guillaume Lample and Marc'Aurelio Ranzato and Ludovic Denoyer and Herv'e J'egou. (2018). Word Translation Without Parallel Data.
- https://dumps.wikimedia.org/ (01.04.2020)
- https://cloud.google.com/translate
- https://github.com/artetxem/vecmap

Thank you!

Silvia Severini
Oettingenstraße 67 - 80538 Munich - Germany
silvia@cis.uni-muenchen.de - https://silviaseverini.github.io/