Bilingual Induction and Pseudo Parallel Corpora

Alexander Fraser

LMU Munich

25 June 2022

Building and Using Comparable Corpora (BUCC 2022, Marseille)

Thanks for the invitation!

Outline:

- BUCC and why this topic?
- Main part: using Bilingual Word Embeddings to create a special kind of Pseudo Parallel Corpora
  - Handling out-of-vocabulary words (words that do not occur in the MT training data)
- Time allowing: translating word senses that are rare and even unseen in the training data

Building and Using Comparable Corpora

***Long involvement with core BUCC topics*:**

Parallel sentence extraction (supervised)	Improved Machine Translation Performance via Parallel Sentence Extraction from Comparable Corpora (Munteanu, Fraser, Marcu NAACL 2004)
Parallel sentence extraction (unsupervised)	Unsupervised Parallel Sentence Extraction with Parallel Segment Detection Helps Machine Translation (Hangya, Fraser ACL 2019) and a previous paper
Terminology mining	Combining Bilingual Terminology Mining and Morphological Modeling for Domain Adaptation in SMT (Weller, Fraser, Heid EAMT 2014) and some subsequent work
Bilingual Lexicon Induction (BLI)	Next slide

Building and Using Comparable Corpora

***Long involvement with core BUCC topics (II)*:**

BLI (basic techniques)	Many papers with Viktor Hangya and several others.
BLI (low resource)	Many papers with Hangya, including several papers with Silvia Severini (see presentation later today!).
Transliteration Mining (unsupervised)	Papers with Hassan Sajjad, Helmut Schmid
BLI (transliteration)	Incorporation into BLI with Braune, Severini, Hangya, others.
BLI (applications)	Focus of today's talk

What are Pseudo Parallel Corpora?

- Basic idea: **back-translation.** For instance, MT of a German corpus to English.
- Results in a pseudo parallel corpus consisting of noisy (machine translation output) English, and perfectly fluent and adequate German.
- Often used to incorporate German monolingual corpora into an English to German NMT system.
- Training MT on this works well because Neural Machine Translation is very robust to noise in the input.
- The intution behind back translation is also a key component of *unsupervised* machine translation.
- But in this talk we will introduce a new twist to this that many of you have hopefully not seen before.

Better OOV Translation with Bilingual Terminology Mining

Matthias Huck, Viktor Hangya, Alexander Fraser

LMU Munich

ACL 2019

Motivation

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

Subword segmentation allows for open-vocabulary translation, but out-of-vocabulary words (OOVs) are still often mistranslated.

***Example*:**

src	A coronary **angioplasty** may not be technically possible [. . .]
ref	Eine **Koronarangioplastie** ist wahrscheinlich technisch nicht möglich [. . .]
hyp	Ein **Herzinfarkt** *(heart attack)* ist vielleicht technisch nicht möglich [. . .]

"*OOVs*":

Source language words that weren't observed in the parallel training corpus

Idea: Use BWEs

Can adequate translations of OOV words be learned from additional monolingual corpora?

Bilingual word embeddings (BWEs)

- Represent source and target language words in a joint space
- Higher word vocabulary coverage than the parallel corpus

How to best integrate OOV word translation candidates from the BWE space into the NMT system?

- Cross-lingual nearest neighbors in the BWE space are noisy
- Polysemy: Need to disambiguate – choose amongst multiple options depending on context within sentences

Approach

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

1. **Baseline NMT system**
   - Trained on parallel corpus (subword-segmented)
2. **(Unsupervised) BWEs**
   - Trained on large monolingual data in the two languages
3. **Bilingual terminology mining**
   - Identify test set OOVs & get top-n word translations from BWEs
   - In target-language monolingual data, mine sentences that contain the OOV translation candidates
4. **NMT fine-tuning**
   - Backtranslate the mined target-side sentences, force OOV words to be generated in the backtranslations
   - Fine-tune NMT model on synthetic data (subword-segmented)

Approach

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

1 Baseline NMT system

- Trained on parallel corpus (subword-segmented)

2 (Unsupervised) BWEs

- Trained on large monolingual data in the two languages

3 Bilingual terminology mining

- Identify test set OOVs & get top-n word translations from BWEs
- In target-language monolingual data, mine sentences that contain the OOV translation candidates

4 NMT fine-tuning

- Backtranslate the mined target-side sentences, force OOV words to be generated in the backtranslations
- Fine-tune NMT model on synthetic data (subword-segmented)

Approach

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

1 Baseline NMT system

- Trained on parallel corpus (subword-segmented)

2 (Unsupervised) BWEs

- Trained on large monolingual data in the two languages

3 Bilingual terminology mining

- Identify test set OOVs & get top-n word translations from BWEs
- In target-language monolingual data, mine sentences that contain the OOV translation candidates

4 NMT fine-tuning

- Backtranslate the mined target-side sentences, force OOV words to be generated in the backtranslations
- Fine-tune NMT model on synthetic data (subword-segmented)

Approach

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

1. **Baseline NMT system**
   - Trained on parallel corpus (subword-segmented)
2. **(Unsupervised) BWEs**
   - Trained on large monolingual data in the two languages
3. **Bilingual terminology mining**
   - Identify test set OOVs & get top-n word translations from BWEs
   - In target-language monolingual data, mine sentences that contain the OOV translation candidates
4. **NMT fine-tuning**
   - Backtranslate the mined target-side sentences, force OOV words to be generated in the backtranslations
   - Fine-tune NMT model on synthetic data (subword-segmented)

Bilingual Terminology Mining (1)

src	if you need to take medication for eye health , make sure you take as prescribed and don ’t stop without talking to your GP or **optometrist** .

Top-5 word translations from BWEs for the OOV “optometrist”:

- Gesichtsfeldprüfgerät *(visual field checking device)*
- Augenarzt *(eye doctor)*
- Bildanzeigeverfahren *(image display method)*
- Sehtests *(vision test)*
- Sehtestgerät *(eyesight test device)*

Braune et al. (2018): cosine combined with orthography

Bilingual Terminology Mining (2)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

src	if you need to take medication for eye health , make sure you take as prescribed and don 't stop without talking to your GP or **optometrist** .
top-5	**Gesichtsfeldprüfgerät** | **Augenarzt** | **Bildanzeigeverfahren** | **Sehtests** | **Sehtestgerät**

Mine target-language monolingual sentences with OOV translation candidates:

- kompaktes **Gesichtsfeldprüfgerät** nach Anspruch 2 [. . .]
- bei einer Beeinträchtigung des Sehens oder der Augen während der Behandlung wenden Sie sich bitte umgehend an Ihren **Augenarzt** .
- Bildanzeigeeinheit , **Bildanzeigeverfahren** und Bildanzeigeprogramm
- die Erfordernis eines jährlichen Hör- und **Sehtests**
- die Erfindung betrifft ein Verfahren und ein **Sehtestgerät** zur Ermittlung der Notwendigkeit einer Sehhilfe bei Dunkelheit [. . .]

Backtranslation with Forced OOV Words

mined	bei einer Beeinträchtigung des Sehens oder der Augen während der Behandlung wenden Sie sich bitte umgehend an Ihren **Augenarzt** .

Backtranslation with Forced OOV Words

mined	bei einer Beeinträchtigung des Sehens oder der Augen während der Behandlung wenden Sie sich bitte umgehend an Ihren **OOV** .

Backtranslation with Forced OOV Words

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

mined	bei einer Beeinträchtigung des Sehens oder der Augen während der Behandlung wenden Sie sich bitte umgehend an Ihren **OOV** .
bt	you are turning straight to your **OOV** in the event of interference in the treatment or the eye during the treatment .

Backtranslation with Forced OOV Words

mined	bei einer Beeinträchtigung des Sehens oder der Augen während der Behandlung wenden Sie sich bitte umgehend an Ihren **Augenarzt** .
bt	you are turning straight to your **optometrist** in the event of interference in the treatment or the eye during the treatment .

Backtranslation with Forced OOV Words

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

mined	bei einer Beeinträchtigung des Sehens oder der Augen während der Behandlung wenden Sie sich bitte umgehend an Ihren **Augenarzt** .
bt	you are turning straight to your **optometrist** in the event of interference in the treatment or the eye during the treatment .

mined	die Erfordernis eines jährlichen Hör- und **Sehtests** *(vision test)* .
bt	the requirement for an annual hearing and **optometrist** .

Evaluation: Machine Translation Quality

	BLEU	
	Cochrane	NHS24
baseline	22.4	20.2
with OOV copying	23.4	20.5
fine-tuned with OOV terminology mining	27.2	22.5

Examples: Better OOV Translations

src	[. . .] without talking to your GP or **optometrist**
ref	[. . .] ohne vorherige Rücksprache mit Ihrem Hausarzt oder **Optiker** *(optician)*
base	[. . .] ohne mit Ihrem Arzt oder Ihrem **Arzt** *(physician)* zu sprechen
ours	[. . .] ohne mit Ihrem Arzt oder **Augenarzt** *(eye doctor)* zu sprechen

Examples: Better OOV Translations

src	A coronary **angioplasty** may not be technically possible [. . .]
ref	Eine **Koronarangioplastie** ist wahrscheinlich technisch nicht möglich [. . .]
base	Ein **Herzinfarkt** *(heart attack)* ist vielleicht technisch nicht möglich [. . .]
ours	Eine koronare **Angioplastie** ist möglicherweise nicht technisch möglich [. . .]

Examples: Better OOV Translations

src	regular **nosebleeds**
ref	regelmäßige **Nasenbluten**
base	regelmäßige **Misskredite** *(discredits)*
ours	regelmäßige **Nasenbluten**

Examples: Better OOV Translations

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

src	dizziness or **lightheadedness**
ref	Schwindel oder **Benommenheit**
base	schwindelerregend *(dizzying)* oder **zurückhaltend** *(reluctant)*
ours	Schwindel oder **Schwächegefühl** *(feeling of faintness)*

Examples: Better OOV Translations

src	Four different alpha blockers were tested (**alfuzosin**, **tamsulosin**, **doxazosin** and **silodosin**).
ref	Vier verschiedene Alphablocker wurden getestet (**Alfuzosin**, **Tamsulosin**, **Doxazosin** und **Silodosin**).
base	Vier verschiedene Alphablocker wurden getestet (**alfuzos**, **tasuloin**, **doxasa** und **silodosin**).
ours	Vier unterschiedliche Alphablocker wurden untersucht (**Alfuzosin**, **Tamsulosin**, **Doxazosin** und **Tigecyclin**).

Summary

BWEs help adequately translate vocabulary which isn't present in parallel training data.

- We've presented a simple approach to effectively integrate BWE-suggested OOV word translation candidates into an NMT system
- **Bilingual terminology mining**
  & **backtranslation with forced OOV words**
  & **finetuning**
- Multiple candidates provided from the BWEs that the NMT system can choose from

References I

Braune, F., Hangya, V., Eder, T., and Fraser, A. (2018). Evaluating bilingual word embeddings on the long tail. In *Proc. NAACL-HLT*.

Improving Machine Translation of Rare and Unseen Word Senses

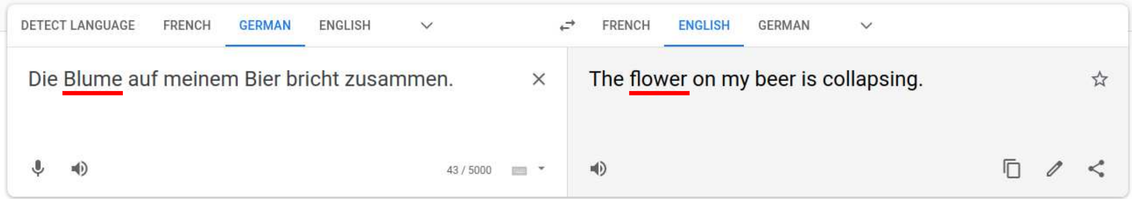
[1]Viktor Hangya, [2]Qianchu Liu, [1]Dario Stojanovski,
[1]Alexander Fraser and [2]Anna Korhonen

[1]Center for Information and Language Processing, LMU Munich, Germany
{hangyav,stojanovski,fraser}@cis.lmu.de

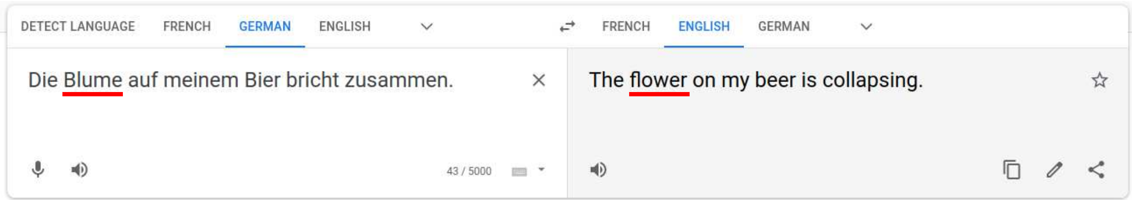
[2]Language Technology Lab, TAL, University of Cambridge, UK
{ql261,alk23}@cam.ac.uk

WMT 2021

Motivation

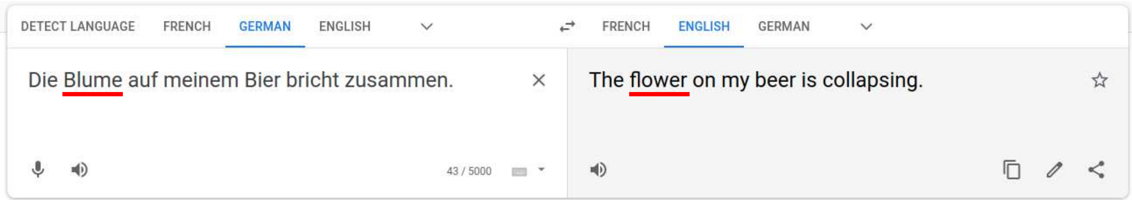

DETECT LANGUAGE
FRENCH
GERMAN
ENGLISH
FRENCH
ENGLISH
GERMAN
Die Blume auf meinem Bier bricht zusammen.
The flower on my beer is collapsing.
43 / 5000

Motivation


DETECT LANGUAGE
FRENCH
GERMAN
ENGLISH
FRENCH
ENGLISH
GERMAN
Die Blume auf meinem Bier bricht zusammen.
The flower on my beer is collapsing.
43 / 5000


rastal

Motivation



- word senses are not uniformly represented in parallel corpora, thus
- the most frequent senses are excessively used
- leading to incomprehensible translations

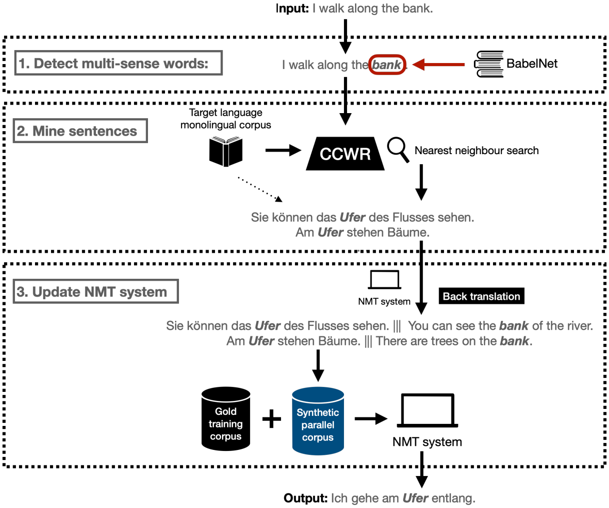
Our Contributions

CMBT (**C**ontextually-**m**ined**B**ack-**T**ranslation)

- Improve translation of multi-sense words
  - especially of rare and missing senses
- We build a synthetic parallel corpus tailored specifically for these senses
  - thus our approach is not limited to the senses contained in parallel corpora
- We show on English-German:
  - significant improvements of rare and missing sense translation
  - while having a low impact on non multi-sense words

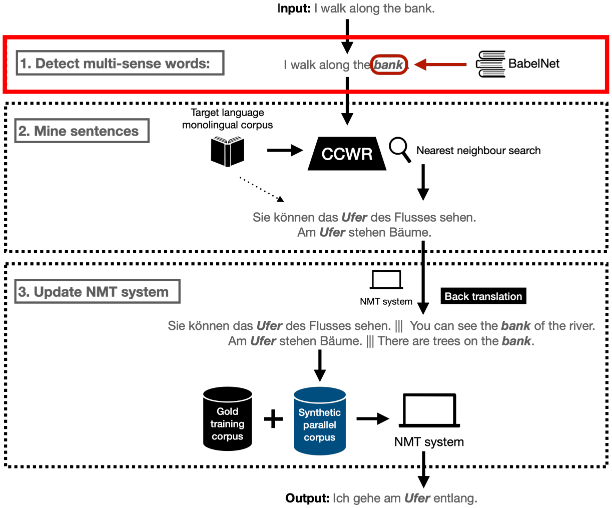
Overview

- We rely on Contextualized Cross-lingual Word Representations (CCWRs)
  - XLM-R (Conneau et al., 2020)
  - trained on large monolingual corpora covering a large set of word senses



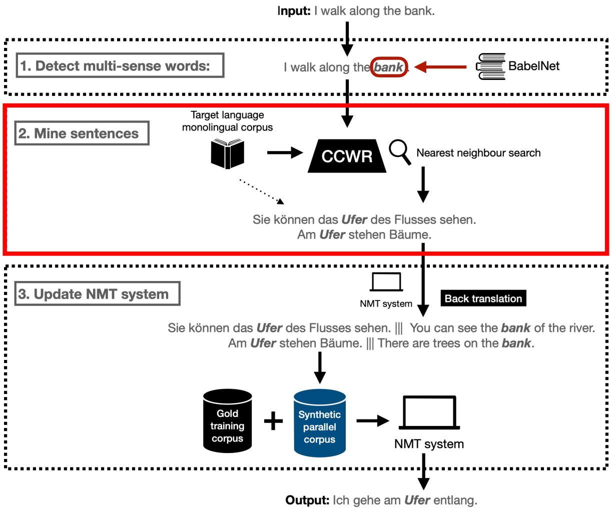
Overview

- Step 1: build a list of multi-sense words



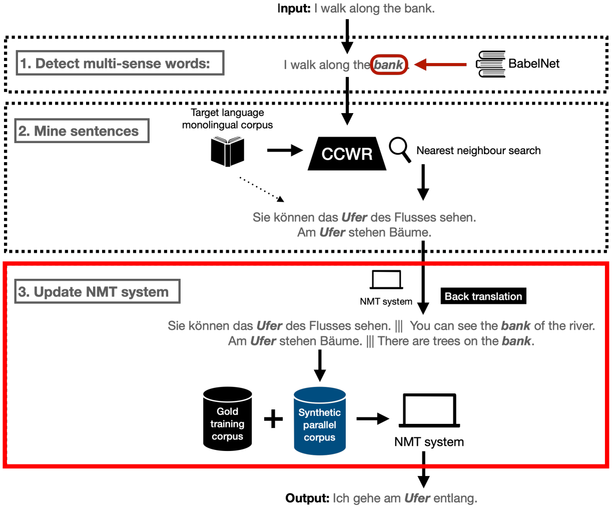
Overview

- Step 2: mine target language sentences

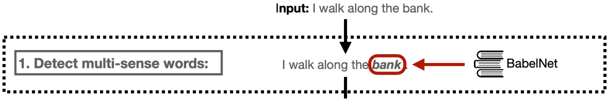


Overview

- Step 3: train an NMT system



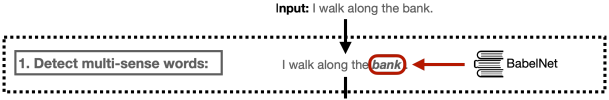
Step 1: Multi-Sense Word Detection



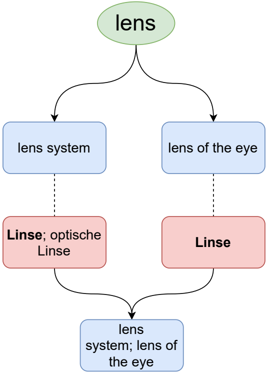
- **Input**:
  - source language corpus to be translated

- Using BabelNet synsets (Navigli and Ponzetto, 2012):
  - if a word is contained in multiple synsets
    $\rightarrow$ multi-sense word

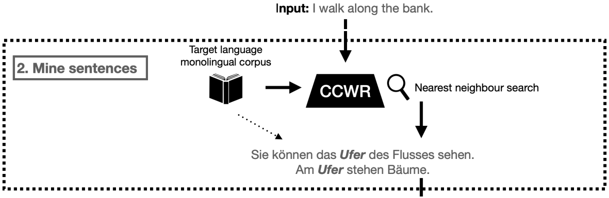
Step 1: Multi-Sense Word Detection



- **Input**:
  - source language corpus to be translated
- Using BabelNet synsets (Navigli and Ponzetto, 2012):
  - if a word is contained in multiple synsets $\rightarrow$ multi-sense word
- **Problem**:
  - BabelNet synsets are too fine grained
  - we merge synsets which have overlapping translations using BabelNet's interlingual links

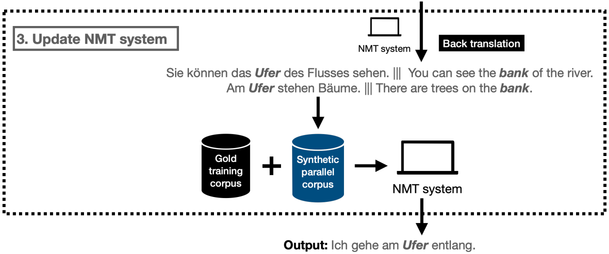


Step 2: Sentence Mining



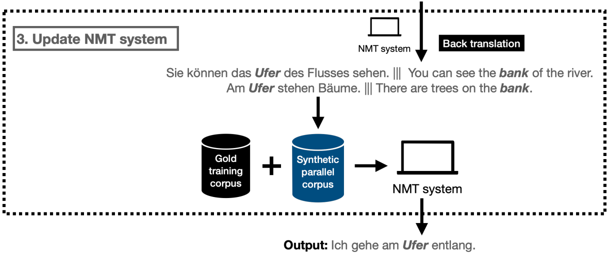
- **Input**:
  - source language sentences containing multi-sense words
  - target language Wikipedia
- Using CCWRs (XLM-R):
  - build contextual representations of words
  - retrieve the top-5 most similar target language word in a sentence for each source word
    - using cosine similarity

Step 3: Back-Translation & NMT Training



- **Input**:
  - mined sentences
- We back translate the mined sentences
- Using gold + the synthetic parallel data we train an NMT system

Step 3: Back-Translation & NMT Training



- **Input**:
  - mined sentences
- We back translate the mined sentences
- Using gold + the synthetic parallel data we train an NMT system
- **Problem**:
  - source multi-sense words might not appear in the translations

Input: *Am* **Ufer** *stehen Bäume.*
Replace: *Am* **[MARK]** *stehen Bäume.*
Translate: *There are trees on the* **[MARK]**.
Restore: *There are trees on the* **bank**.

Experiments

- MuCoW dataset (Raganato et al., 2020)
  - providing gold training and test corpora
  - English→German
- *unseen* set:
  - one sense of each gold multi-sense word is missing from the training data
- *sample-10* set
  - random 10% sample of the training data to have:
  - very rare senses: 0-20% (relative frequency compared to the other senses of a given word)
  - rare senses: 20-40%

Results

set	freq.	system	F_1
unseen	0-0%	baseline	17.14
		BWEs	25.39
		CMBT	**34.80**$\uparrow^{17.66}$
sample-10	0-20%	baseline	35.53
		BWEs	37.70
		CMBT	**47.02**$\uparrow^{11.49}$
	20-40%	baseline	60.98
		BWEs	60.80
		CMBT	**64.49**$\uparrow^{3.51}$

train	freq.	system	F_1
unseen	0-100%	baseline	70.70
		BWEs	71.66
		CMBT	**73.51**$\uparrow^{2.81}$
sample-10	0-100%	baseline	74.58
		BWEs	73.75
		CMBT	**75.86**$\uparrow^{1.28}$

- F_1 scores per frequency bin in the test set
- BWEs: *fastText* embeddings instead of XLM-R (Huck et al., 2019)
- Significant improvements:
    - context is important
    - especially effective at low frequency ranges

- F_1 scores on the complete test set
- CMBT improves overall as well
- Context is important for the mining
    - BWEs decrease F_1 of rare senses

Results

set	freq.	system	F_1
unseen	0-0%	baseline	17.14
		BWEs	25.39
		CmBT	**34.80**$\uparrow^{17.66}$
sample-10	0-20%	baseline	35.53
		BWEs	37.70
		CmBT	**47.02**$\uparrow^{11.49}$
	20-40%	baseline	60.98
		BWEs	60.80
		CmBT	**64.49**$\uparrow^{3.51}$

train	freq.	system	F_1
unseen	0-100%	baseline	70.70
		BWEs	71.66
		CmBT	**73.51**$\uparrow^{2.81}$
sample-10	0-100%	baseline	74.58
		BWEs	73.75
		CmBT	**75.86**$\uparrow^{1.28}$

- F_1 scores per frequency bin in the test set
- BWEs: *fastText* embeddings instead of XLM-R (Huck et al., 2019)
- Significant improvements:
    - context is important
    - especially effective at low frequency ranges

- F_1 scores on the complete test set
- CmBT improves overall as well
- Context is important for the mining
    - BWEs decrease F_1 of rare senses

Results

set	freq.	system	F_1
unseen	0-0%	baseline	17.14
		BWEs	25.39
		CMBT	**34.80**$\uparrow^{17.66}$
sample-10	0-20%	baseline	35.53
		BWEs	37.70
		CMBT	**47.02**$\uparrow^{11.49}$
	20-40%	baseline	60.98
		BWEs	60.80
		CMBT	**64.49**$\uparrow^{3.51}$

train	freq.	system	F_1
unseen	0-100%	baseline	70.70
		BWEs	71.66
		CMBT	**73.51**$\uparrow^{2.81}$
sample-10	0-100%	baseline	74.58
		BWEs	73.75
		CMBT	**75.86**$\uparrow^{1.28}$

- F_1 scores per frequency bin in the test set
- BWEs: *fastText* embeddings instead of XLM-R (Huck et al., 2019)
- Significant improvements:
  - context is important
  - especially effective at low frequency ranges

- F_1 scores on the complete test set
- CMBT improves overall as well
- Context is important for the mining
  - BWEs decrease F_1 of rare senses

Results

set	freq.	system	F_1
unseen	0-0%	baseline	17.14
		BWEs	25.39
		CmBT	**34.80**$\uparrow^{17.66}$
sample-10	0-20%	baseline	35.53
		BWEs	37.70
		CmBT	**47.02**$\uparrow^{11.49}$
	20-40%	baseline	60.98
		BWEs	60.80
		CmBT	**64.49**$\uparrow^{3.51}$

train	freq.	system	F_1
unseen	0-100%	baseline	70.70
		BWEs	71.66
		CmBT	**73.51**$\uparrow^{2.81}$
sample-10	0-100%	baseline	74.58
		BWEs	73.75
		CmBT	**75.86**$\uparrow^{1.28}$

- F_1 scores per frequency bin in the test set
- BWEs: *fastText* embeddings instead of XLM-R (Huck et al., 2019)
- Significant improvements:
  - context is important
  - especially effective at low frequency ranges

- F_1 scores on the complete test set
- CmBT improves overall as well
- Context is important for the mining
  - BWEs decrease F_1 of rare senses

Results

set	freq.	system	F_1
unseen	0-0%	baseline	17.14
		BWEs	25.39
		CmBT	**34.80** $\uparrow^{17.66}$
sample-10	0-20%	baseline	35.53
		BWEs	37.70
		CmBT	**47.02** $\uparrow^{11.49}$
	20-40%	baseline	60.98
		BWEs	60.80
		CmBT	**64.49** $\uparrow^{3.51}$

- F_1 scores per frequency bin in the test set
- BWEs: *fastText* embeddings instead of XLM-R (Huck et al., 2019)
- Significant improvements:
    - context is important
    - especially effective at low frequency ranges

train	freq.	system	F_1
unseen	0-100%	baseline	70.70
		BWEs	71.66
		CmBT	**73.51** $\uparrow^{2.81}$
sample-10	0-100%	baseline	74.58
		BWEs	73.75
		CmBT	**75.86** $\uparrow^{1.28}$

- F_1 scores on the complete test set
- CmBT improves overall as well
- Context is important for the mining
    - BWEs decrease F_1 of rare senses

Results

train	freq.	baseline	BWEs	CmBT
unseen	0-0%	23.0	23.2	**23.3**
	0-100%	25.5	25.6	**25.7**
sample-10	0-20%	22.3	22.3	**22.6**
	20-40%	24.5	24.6	**24.7**
	0-100%	25.0	25.0	**25.1**

- BLEU scores:
  - CmBT improves overall translation
  - only marginally since non multi-sense words are not significantly affected

Results

SRC	*The physician, to whom the soldiers of the **watch** had carried him at the first moment...*
BASE	Der Arzt, zu dem ihn die Soldaten der **Uhr**[timepiece] im ersten Augenblick getragen hatten... ✗
CmBT	Der Arzt, zu dem ihn die Soldaten der **Wache**[guard] im ersten Augenblicke getragen hatten... ✔
REF	*Der Heilkünstler, zu welchem die Soldaten der **Wache** ihn im ersten Augenblicke getragen...*

SRC	*A lover finds his mistress asleep on a mossy **bank**;...*
BASE	Ein Liebhaber findet seine Geliebte schlafend auf einer feuchten **Bank**[bench];... ✔
CmBT	Ein Geliebter findet seine Geliebte schlafend auf einem feuchten **Ufer**[river bank];... ✗
REF	*Ein Liebender findet seine Geliebte auf einer moosigen **Bank** eingeschlafen;...*

- Positive and negative example

Results

SRC	*The physician, to whom the soldiers of the* ***watch*** *had carried him at the first moment...*
BASE	Der Arzt, zu dem ihn die Soldaten der **Uhr**[*timepiece*] im ersten **Augenblick** getragen hatten... ✗
CmBT	Der Arzt, zu dem ihn die Soldaten der **Wache**[*guard*] im ersten **Augenblicke** getragen hatten... ✔
REF	*Der Heilkünstler, zu welchem die Soldaten der* ***Wache*** *ihn im ersten Augenblicke getragen...*
SRC	*A lover finds his mistress asleep on a mossy* ***bank****;...*
BASE	Ein **Liebhaber** findet seine Geliebte schlafend auf **einer** feuchten **Bank**[*bench*];... ✔
CmBT	Ein **Geliebter** findet seine Geliebte schlafend auf **einem** feuchten **Ufer**[*river bank*];... ✗
REF	*Ein Liebender findet seine Geliebte auf einer moosigen* ***Bank*** *eingeschlafen;...*

- Positive and negative example
- Non multi-sense words are kept intact

Conclusions

- CmBT (**C**ontextually-**m**ined**B**ack-**T**ranslation)
  - uses contextualized cross-lingual word embeddings
  - builds synthetic parallel corpus containing missing and rare senses as well
- The resulting NMT system:
  - improves multi-sense word translation
  - especially missing and rare senses!
  - while leaving non multi-sense words intact

Summary

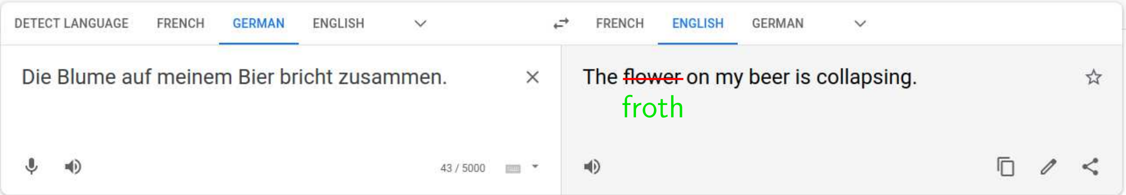
I presented:

- A few words on bilingual induction of sentences and words
- Building a special kind of Pseudo Parallel Corpus for handling out-of-vocabulary words (words that do not occur in the MT training data)
- Translating word senses that are rare and even unseen in the training data

Final words:

- Other forms of Pseudo Parallel Corpora are interesting! For instance, our work on equivalent named entities (Adapting Entities Across Languages and Cultures, Peskov et al 2021).
- Thanks very much to everyone in my team and all co-authors! Also additionally to Matthias and Viktor for slides.
- Advertisement: we are about to announce the very low resource and unsupervised shared task at WMT 22, Upper Sorbian, Lower Sorbian, German, all directions, would be great if you participated!

Thank you!



This project has received funding from the Cambridge LMU Strategic Partnership, the Peterhouse College at University of Cambridge, the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme DASMT (No. 640550), the German Research Foundation (DFG; grant FR 2829/4-1), the ERC Consolidator Grant LEXICAL (No. 648909) and the ERC PoC Grant MultiConvAI (No. 957356).