Better OOV Translation with Bilingual Terminology Mining

Matthias Huck, Viktor Hangya, Alexander Fraser

LMU Munich

31 July 2019

Motivation

LMU 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

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

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)

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

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

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

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

	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

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

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

Examples: Better OOV Translations

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

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

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

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

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

Questions?

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Thank you for your attention

Matthias Huck

mhuck@cis.lmu.de

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement № 640550).

References I

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

Instrinsic Evaluation: Bilingual Lexicon Induction

		Acc_1	Acc_5
freq	Braune et al. (2018)	38.6	47.4
	EU+UFAL+orth	25.9	40.6
rare	Braune et al. (2018)	26.3	28.2
	EU+UFAL+orth	17.5	28.8

Table: Medical bilingual lexicon induction results showing the quality of the BWE based dictionaries using top-1 and top-5 translations.

Instrinsic Evaluation: Bilingual Terminology Mining

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

	UFAL			UFAL+orth		
n	$P@n$	$R@n$	$F_1@n$	$P@n$	$R@n$	$F_1@n$
1	58.19	13.58	22.02	58.13	25.28	35.24
5	44.46	26.10	32.89	50.05	43.82	46.73
10	35.80	29.84	32.55	41.04	47.64	44.09
20	29.54	33.58	31.43	34.43	50.16	40.83

	EU+UFAL			EU+UFAL+orth		
n	$P@n$	$R@n$	$F_1@n$	$P@n$	$R@n$	$F_1@n$
1	68.65	37.56	48.55	69.59	41.87	52.28
5	54.33	48.46	51.22	51.13	51.71	51.41
10	42.94	53.41	47.61	44.45	56.34	49.70
20	36.42	58.78	44.98	37.42	61.30	46.47

Table: Mining quality with top-n OOV word translation candidates.