HABLex: Human Annotated Bilingual Lexicons for Experiments in Machine Translation

Brian Thompson,* Rebecca Knowles,* Xuan Zhang,* Huda Khayrallah, Kevin Duh, Philipp Koehn

Johns Hopkins University

Overview

What is in the HABLex dataset?

- Human-generated *alignments* of words and phrases.
- Development and test set.

When to use the HABLex dataset?

Benchmarking methods for *bilingual lexicon integration* into neural machine translation.

Why is bilingual lexicon integration desirable?

- high-tech vocabulary
- user requirement
- low resource
- improve rare word translation

What are the challenges of bilingual lexicon integration?

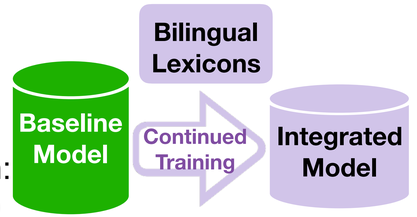
- Arbitrary dictionaries have problems:
  e.g. overlap entries, ineffective
- Hard to evaluate only based on BLEU.

In need of bilingual lexicons *tailored to dev and test set*.

Methods

1. Continued Training (CT)

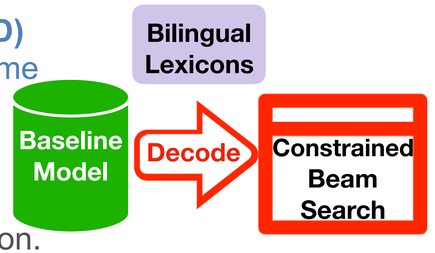
- Incorporation at training time



- Standard CT
- Elastic Weight Consolidation: (EWC; Thompson et al., 2019)
  Train a neural network to learn a new task without catastrophic forgetting.

2. Constrained Decoding (CD)

- Incorporation at inference time

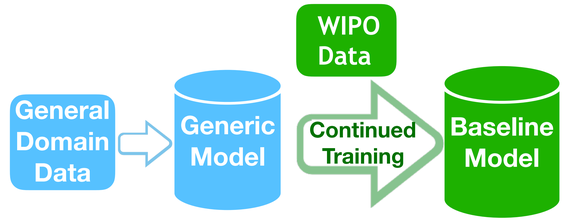


Dynamic Beam Allocation
(Post and Vilar, 2018)
Limitation: work on lexical constraints with one translation.

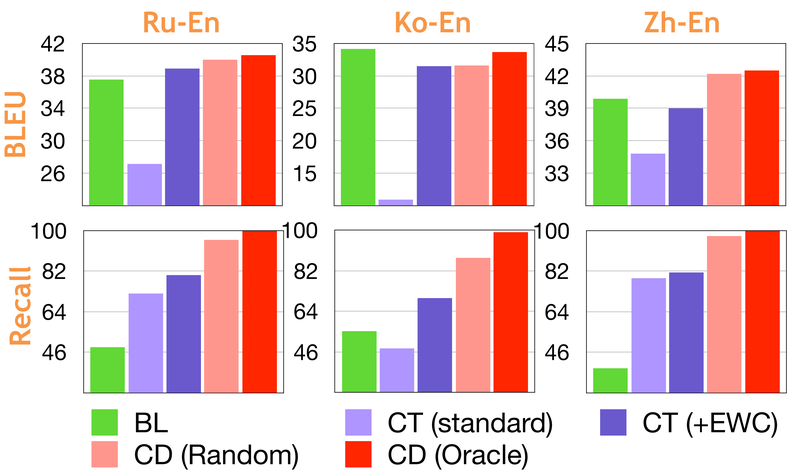
- Oracle choice: use the right lexical translation
- Random choice: pick one random lexical translation

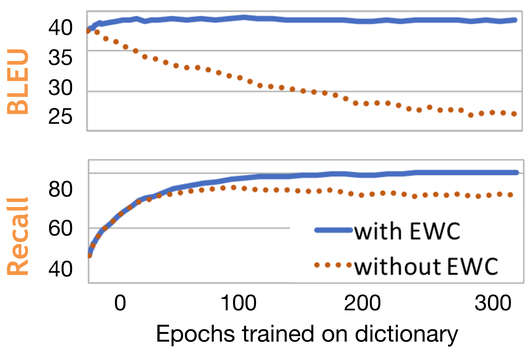
Experiments and Results

Baseline system: (domain-adapted) First train on general domain data, then fine-tune on Patent data.



Recall: Percentage of the time the system output contains the correct lexicon translation.





HABLex Dataset

SCAN ME

http://www.cs.jhu.edu/~kevinduh/a/hablex2019

HABLex Dataset

source 本 发明 用于 板材 软膜 成形 。
alignment
reference The present invention is used for **flexible die** forming a plate .
lexical entry **软膜 ⟷ flexible die**

Two-step process:

1. Identifying *rare words* on the source side of the test and development sets.
2. Human annotators correcting or validating automatic alignments of the identified words.

Domain: Patent
Corpus:
World Intellectual Property Organization (WIPO) COPPA-V2
Language Pairs:
Russian -> English, Korean -> English, Chinese -> English

	Development		Test	
	Entries	Sentences	Entries	Sentences
Ru	9040	2412	8001	2142
Ko	5593	1744	5595	1756
Zh	1773	885	2289	1025

* These authors contributed equally to this work.