Unsupervised Extraction of Partial Translations for Neural Machine Translation

Benjamin Marie and Atsushi Fujita
{bmarie,atsushi.fujita}@nict.go.jp
National Institute of Information and Communications Technology (NICT), Kyoto, Japan

Key Points

Assumptions

- Partial translations can be extracted from monolingual data
- Prior translation knowledge can bias newly extracted knowledge

Method

1. Extraction: only using an unsupervised phrase table
2. Post-processing: masking untranslated tokens
   - Application: training data for NMT

Using partial translations in NMT

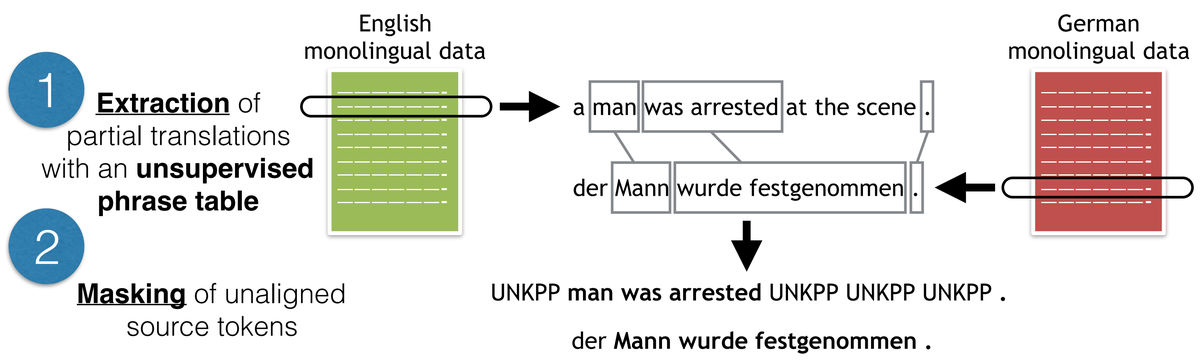
- Better translation quality for low-resource settings
- Complementary to back-translations as NMT training data

Motivation

Generation vs. extraction of synthetic bilingual data

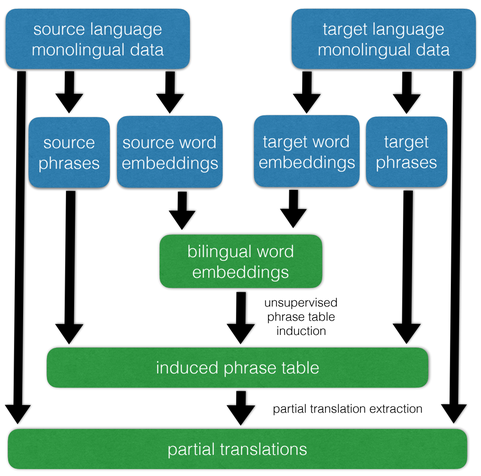
- Generation of back-translations (Sennrich+, 16) : synthetic bilingual data with a synthetic source side
- Extraction of partial translations: synthetic bilingual data in which both sides are fluent sentences

Example of a Partial Translation



Approach: the Whole Framework

1. Step by step partial translation extraction



- Induce a phrase table from monolingual data (Marie+, 17; Artetxe+, 18; Lample+, 18)
- Make a set of sentence pairs through the Cartesian product of the sets of sentences in source and target monolingual data
- Find the best target sentence for each source sentence
  - Filter sentence pairs given their coverage by the phrase table
  - Score the remaining sentence pairs with phrase-based forced decoding (PBFD) (Zhang+, 17)
- Keep the n best sentence pairs with the highest PBFD score

2. For training NMT

- Mask unaligned tokens found by PBFD on the source side of partial translations
- Resource-rich scenario: mix the resulting partial translations with back-translations and the original bilingual data
- Low-resource scenario: mix the resulting partial translations with the small original bilingual data

Evaluation: Experiments in NMT

Data

- Language pairs: English-German, English-Turkish, Bengali-Malay
- Bilingual training data (low-resource conditions):
  - en-de and en-tr: 100k sentence pairs from WMT18
  - bn-ms: 18k sentence pairs from ALT project (Riza+, 16)
- Monolingual data for phrase table induction and partial translations extraction:
  - Resouch-rich languages: 239M, 237M, and 104M lines, for English, German, and Turkish
  - Low-resource languages: 5.3M and 4.6M lines for Bengali and Malay

NMT systems

- baseline: trained only using the original bilingual training data
- `backtr`: same data as baseline mixed with 100k back-translations
- `partial`: same data as baseline mixed with 100k partial translations

Results

Training data	en→de	en→tr	bn→ms
baseline	7.1	9.3	**6.1**
`backtr`	9.1	11.4	5.4
`partial`	9.9	10.4	5.5
`backtr+partial`	**11.5**	**11.6**	4.5

- `backtr` and `partial` are complementary (best configuration)
- Partial translations can be more useful than back-translations (en-de)
- For bn-ms (low-resource), both `backtr` and `partial` fail

Analysis

Phrase table

Training data	en→de	bn→ms
baseline	7.1	6.1
unsupervised	**9.9**	5.5
surpervised	9.4	**6.3**

- More useful partial translations with an unsupervised phrase table

Masking vs. dropping

Training data	en→de	en→tr
baseline	7.1	9.3
`original`	6.2	7.7
`dropped`	8.8	10.0
`partial`	**9.9**	**10.4**

- Masking unaligned tokens performs better than dropping them

With large parallel data

para.	`backtr`	`partial`	en→de	en→tr
all			26.2	13.6
all	1M		27.7	18.6
all		1M	26.4	14.7
all	1M	1M	**28.2**	**19.0**

- Back-translations are more useful
- `backtr` and `partial` are complementary