A Lifelong Multilingual Multi-granularity Semantic Alignment Approach via Maximum Co-occurrence Probability

Xin Liu[1], Hongwei Sun[2], Shaojie Dai[3,1], Bo Lv[3,1], Youcheng Pan[1], Hui Wang[1,*] and Yue Yu[1,*]

[1] Peng Cheng Laboratory, [2] Mudanjiang Normal University,

[3] University of Chinese Academy of Sciences

Content

Introduction

Background & Problem

1) The alignment between languages is the key message for machine translation, and it encourages the models to learn the correlation of different languages and to achieve multilingual interaction.

2) Although most models could learn accurate multilingual alignment information through the parallel corpora, they are usually limited by the insufficient scale of the parallel corpora and thus cannot learn sufficiently.

3) Therefore, the cross-lingual pre-training methods are proposed, which mask and predict tokens in a multilingual text to generalize diverse multilingual information.

4) However, due to the lack of sufficient aligned multilingual resources in the pre-training process, these methods may not fully explore the multilingual correlation of masked tokens, resulting in the limitation of multilingual information interaction.

Introduction

- Solution

1) This paper proposes a lifelong multilingual multi-granularity semantic alignment approach via maximum co-occurrence probability in noisy parallel data and uses it to build a semantic alignment resource.

1) The approach collects a group of noisy pairs that contain the same linguistic unit in one language and computes the co-occurrence probability of one candidate linguistic unit in the other languages. The co-occurrence probability is the probability that one linguistic unit appears in all sentences, so one candidate linguistic unit will have a higher probability if it occurs in most sentences.

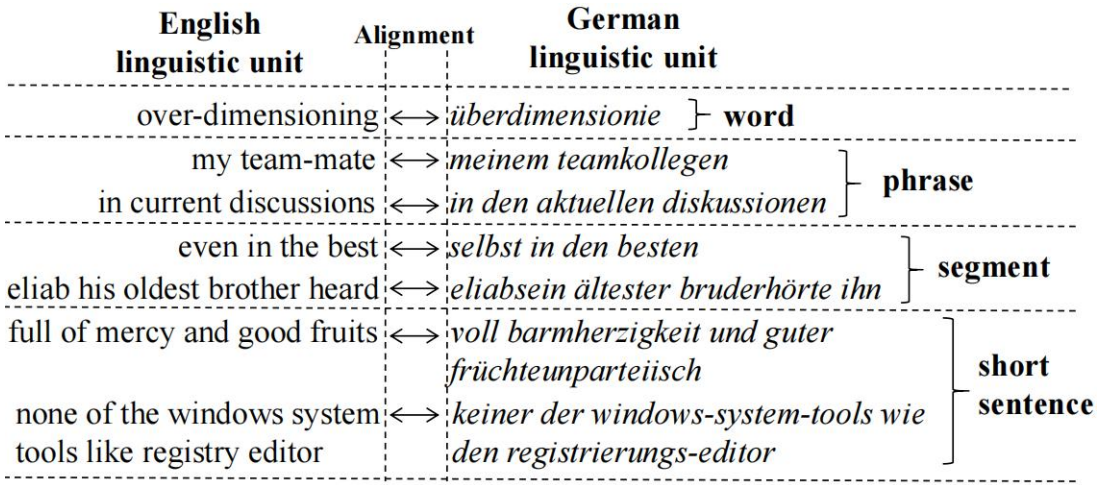

Figure 1: Illustration of English-German multi-granularity alignment linguistic units in the resource built by this approach.

Introduction

■ Contributions

1) proposing a lifelong multilingual multi-granularity semantic alignment approach that only relies on the co-occurrence constraints in the multilingual noisy data, and can identify massive semantically aligned linguistic units at various granularity through the maximum occurrence probability continuously and unsupervised;

2) releasing a version of the **l**ifelong **m**ultilingual **m**ulti-**g**ranularity **s**emantic **a**lignment **r**esource (called LM^2_gSAR). In this version, LM^2_gSAR supports the multilingual alignment between seven languages, namely English (en), Czech (cs), German (de), Russian (ru), Romanian (ro), Hindi (hi) and Turkish (tr). Meanwhile, it also supports the continuous expansion in scale, language coverage, and granularity.

3) conducting exhaustive experiments on the aligner comparisons and the bi-direction translation tasks between English and the above six languages. Compared to the other aligners, the approach shows higher alignment accuracy. The models using LM^2_gSAR have shown significant improvements in almost all translation directions. In addition, we perform objective analysis and discussion as strong evidence of the value and significance of this work.

Source code is released at *https://github.com/Gdls/MCoPSA*

Related Works

Weakness & Advantages

1) The mainstream pre-training methods rely on different mechanisms, techniques, or tools to learn multilingual alignment information. These mechanisms, techniques, or tools have boosted the capabilities of these methods on generalizing alignment information, but due to the absence of accurate and sufficient alignment resources, there is still a lot of room for improving their capabilities.

2) The other methods propose to use the common multilingual alignment resources, which are usually the parallel corpora, coming from the public releases, web mining, or competitions. Although these resources could provide multilingual alignment information, the scale or linguistic diversity may not meet the need for the pre-training methods.

Compared to the previous methods and resources, the proposed approach considers the scale, diversity, and other linguistic properties. Meanwhile, the resource provides more specific and sufficient semantic alignment information than those alignment techniques in the pre-training methods.

Method Description

- The **m**aximum **c**o-**o**ccurrence **p**robability based **s**emantic **a**lignment algorithm(MCoPSA)

Core idea:

1) There is a group of translated pairs from noisy data, and each pair consists of sentences in two languages.
2) A linguistic unit of one language exists in all sentences in the group, and the algorithm calculates the co-occurrence probability of each candidate linguistic unit in all sentences in the other language.
3) The co-occurrence probability means the probability of one candidate linguistic unit appearing in all the sentences of the group.
4) Then, the algorithm selects the candidate with the maximum co-occurrence probability as the aligned linguistic unit.

Algorithm 1 The maximum co-occurrence probability based semantic alignment algorithm.

procedure *MCoPSA*(u_l, $\mathfrak{D}_{\mathfrak{N}}$)
  Assert $u_l \in X$ and $u_l^X = u_l$;
  Initialize $G(u_l^X) = (p_0, ..., p_n)$ from $\mathfrak{D}_{\mathfrak{N}}$;
  Initialize lists L = [], uL = [], dict D = {}
  Select $t_0^{\mathcal{Y}}, t_1^{\mathcal{Y}}$ from $G(u_l^X)$;
  $G(u_l^X)$ = $G(u_l^X)$ - p_0 - p_1;
  L.append($t_0^{\mathcal{Y}}$, $t_1^{\mathcal{Y}}$);
  uL.extend(*CSFunc*($t_0^{\mathcal{Y}}$, $t_1^{\mathcal{Y}}$));
  Update *cnt*(uL);
  for ${u_l}_i^Y \in uL$ **do**
    $D[{u_l}_i^Y]$ = *cnt*[${u_l}_i^Y$]/*len*(L);
  end for
  while $G(u_l^X)$ is not Ø **do**
    Select $t_i^{\mathcal{Y}}$ from $G(u_l^X)$;
    $G(u_l^X)$ = $G(u_l^X)$ - p_i;
    for $t_j^{\mathcal{Y}}$ in L **do**
      uL.extend(*CSFunc*($t_i^{\mathcal{Y}}$, $t_j^{\mathcal{Y}}$));
      Update *cnt*(uL);
    end for
    L.gappend($t_j^{\mathcal{Y}}$);
    for ${u_l}_k^Y \in uL$ **do**
      $D[{u_l}_k^Y]$ = *cnt*[${u_l}_k^Y$]/*len*(L);
    end for
  end while
  ${u_l}^Y$ = *maxProb*(D,ϱ);
  Return (u_l^X, ${u_l}^Y$);
end procedure

Method Description

- The **m**aximum **c**o-**o**ccurrence **p**robability based **s**emantic **a**lignment algorithm(MCoPSA)

Core idea:

1) There is a group of translated pairs from noisy data, and each pair consists of sentences in two languages.
2) A linguistic unit of one language exists in all sentences in the group, and the algorithm calculates the co-occurrence probability of each candidate linguistic unit in all sentences in the other language.
3) The co-occurrence probability means the probability of one candidate linguistic unit appearing in all the sentences of the group.
4) Then, the algorithm selects the candidate with the maximum co-occurrence probability as the aligned linguistic unit.

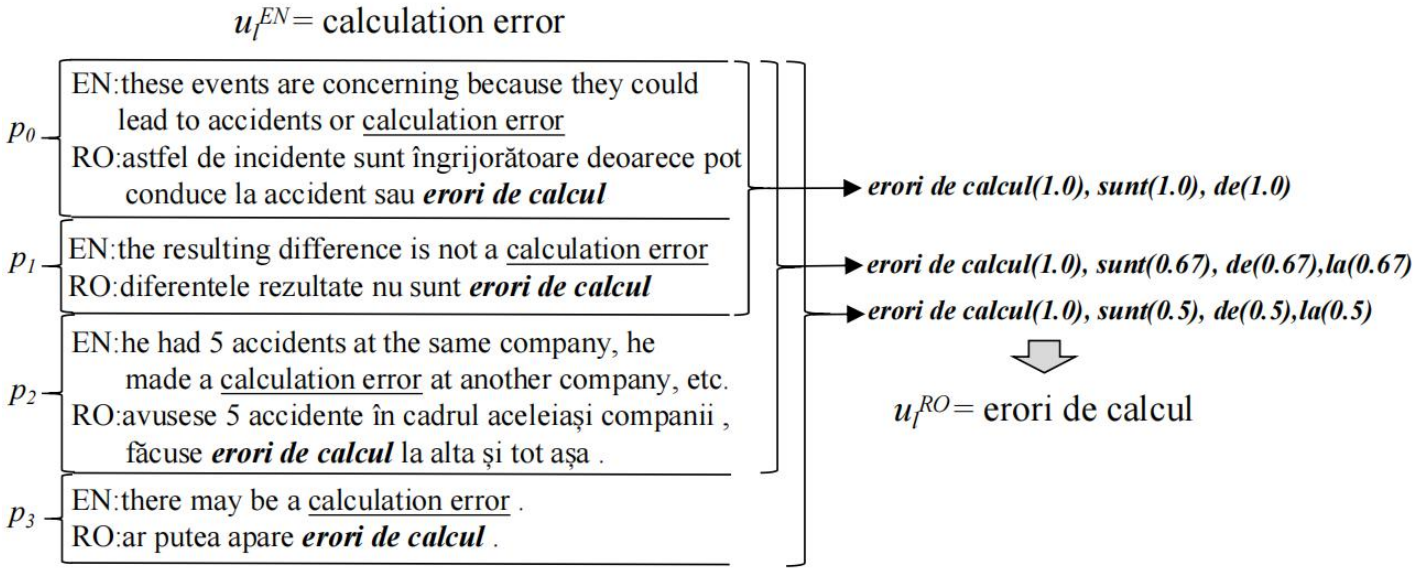

Figure 2: An example to illustrate the processing of Algorithm 1. The underlined part in the English sentences is the input linguistic unit, and the bold and italic part in the Romanian sentences is the output(aligned linguistic unit). The float in brackets is the co-occurrence probability of each candidate unit in the current step.

Method Description

- The **l**ifelong **m**ultilingual **m**ulti-**g**ranularity **s**emantic **a**lignment **r**esource (LM^2_gSAR).

1) LM^2_gSAR supports seven languages, namely English, Czech, German, Russian, Romanian, Hindi and Turkish. It is built on the noisy bilingual data from published CCMatrix v1 between English and the other six languages.

2) The statistics on LM^2_gSAR:

Table 1: The scale of the bilingual data from published CCMatrix v1 used in the MCoPSA algorithm for building LM^2_gSAR.

Languages	Volume	PinND (%)
en-de	21.5M	8.7
en-ru	20.6M	14.7
en-cs	15.6M	27.7
en-ro	15.1M	27.2
en-tr	14.2M	29.2
en-hi	5.5M	36.4

Table 2: The statistics on the scale of the aligned linguistic units in LM^2_gSAR between seven languages.

	en	cs	de	ru	ro	tr
cs	3.30M	-				
de	2.53M	0.33M	-			
ru	3.04M	0.19M	0.13M	-		
ro	3.11M	0.53M	0.30M	0.17M	-	
tr	1.95M	0.36M	0.21M	0.11M	0.34M	-
hi	1.01M	0.19M	0.13M	0.33M	0.20M	0.18M

Table 3: The statistics on the granularity distribution (%) of the aligned linguistic units in LM^2_gSAR based on the en-XX alignment.

Aligment	Multi-granurality(%)				
	w-1	p-2	p-3	s-4	ss-5+
en-cs	3	21	44	31	1
en-de	4	23	42	30	1
en-ru	4	20	44	31	1
en-ro	3	19	42	34	2
en-tr	3	30	44	21	2
en-hi	4	28	43	24	1
Avg	3.5	23.5	43.2	28.5	1.3

Experiments and Results

- Experimental Datasets

The aligner comparison experiment

1) we andomly collected a group of en-XX corpora and selected the top 500 language units in each en-XX corpus based on the term frequency and inverse document frequency.

2) we recruited some language experts with English and XX backgrounds, and for each pair, they manually annotated the golden alignment of the English language unit in XX sentences, which serves as the evaluation test set.

The machine translation experiment

1) the WMT datasets including twelve translation directions as the evaluation benchmarks, namely en-de (4.5M), en-ru (1.1M), and en-hi (32K) in WMT14, en-ro (0.6M) in WMT16, en-tr (0.2M) in WMT17, and en_x0002_cs (11M) in WMT18.

- Evaluation Metrics

1) The Alignment Error Rates (AER)

2) BLEU score (sacreBLEU)

Experiments and Results

- Baseline and Comparison Methods

  In the aligner comparison experiment

  1) GIZA++ is an extension of the program GIZA (part of the SMT toolkit EGYPT).

  2) Fast-align is a simple, fast, unsupervised word aligner.

  3) Awesome-align is a tool that can extract word alignments from multilingual BERT.

  In the machine translation experiment

  1) mBART is one of the first methods for pre-training a complete sequence-to-sequence model by denoising full texts in multiple languages.

  2) M2M-100 is a Many-to-Many multilingual translation model that can translate directly between any pair of 100 languages.

  3) mT5 is pre-trained on a new Common Crawl-based dataset covering 101 languages and has shown SOTA performance on many multilingual benchmarks.

Experiments and Results

Experimental setup

- Tokenizer: MBartTokenizer, M2M100Tokenizer, and T5Tokenizer, respectively;
- Training batch size: 4~16;
- Max sequence length: 1024;
- Beam size: 5;
- Optimizer: AdamW, learning rate: 5e-5;
- More details are listed in the paper.

Training strategy in machine translation experiment

1) LM^2_gSAR-based pre-training: we first apply the alignment substitution technique (AST) with LM^2_gSAR to prepare the pre-training corpus. The monolingual sentences used to construct the alignment substitution with LM^2_gSAR come from the corresponding bilingual training data. A baseline model is pre-trained with the corpus.

2) Fine-tuning: the pre-trained model in the previous step is fine-tuned with the training data.

3) Evaluation: the trained model is evaluated on the test set.

Experiments and Results

■ Results on the aligner comparison experiment

Table 5: The AER scores on each en-XX test set of the MCoPSA, GIZA++, FastAlign (Fa-Align), and Awesome-Align (Aw-Align).

	GIZA++	Fa-Align	Aw-Align	MCoPSA
en-cs	48.4	42.1	30.8	**19.8**
en-de	61.2	58.7	29.0	**17.2**
en-hi	67.2	66.0	30.3	**16.2**
en-ro	53.4	50.2	24.4	**16.4**
en-ru	50.1	46.1	21.8	**15.2**
en-tr	63.2	72.6	30.3	**12.4**
Avgs	57.3	55.9	24.4	**16.2**

Table 6: The AER score of each method at word, phrase, and segment granularity of the linguistic unit on the en-ru test set.

	word	phrase	segment	Avgs
GIZA++	52.2	48.4	58.1	52.9
Fa-Align	54.5	47.4	56.1	52.7
Aw-Align	**14.1**	24.1	20.2	19.5
MCoPSA	25.4	**13.1**	**13.4**	**17.3**

1) The performance difference between MCoPSA and the others indicates that the alignments by MCoPSA are of better quality and more promising for the pre-training stage in real world scenarios.

2) The proposed McoPSA show a much better performance on phrase and segment-level, and this may be the main reason why the proposed McoPSA achieves the best results

Experiments and Results

- Results on the machine translation experiment

Table 7: The BLEU scores of the baseline models under different training strategies on the test sets of each WMT benchmark.

Benchmark	mBART		M2M100		mT5		Avg△
	→	←	→	←	→	←	
en-cs^{H}	16.6 17.5$^{\uparrow}$	28.9 **30.5(+1.6)**$^{\uparrow}$	17.2 17.7$^{\uparrow}$	30.1 29.4	18.1 **18.5(+0.4)**$^{\uparrow}$	27.9 29.7$^{\uparrow}$	+0.8
en-de^{M}	26.5 **27.3(+0.8)**$^{\uparrow}$	32.5 **33.0(+0.5)**$^{\uparrow}$	26.9 26.9	31.6 32.2$^{\uparrow}$	24.4 24.6$^{\uparrow}$	28.9 29.6$^{\uparrow}$	+0.5
en-ru^{M}	33.0 **34.1(+1.1)**$^{\uparrow}$	32.2 **32.6(+0.4)**$^{\uparrow}$	33.0 33.1$^{\uparrow}$	32.4 32.4	26.8 26.9$^{\uparrow}$	26.8 27.0$^{\uparrow}$	+0.4
en-ro^{L}	25.6 **26.8(+1.2)**$^{\uparrow}$	36.2 **37.0(+0.8)**$^{\uparrow}$	26.4 26.3	36.7 36.6	22.8 22.9$^{\uparrow}$	32.9 33.2$^{\uparrow}$	+0.4
en-tr^{L}	19.1 **20.0(+0.9)**$^{\uparrow}$	22.7 23.1$^{\uparrow}$	19.8 19.8	22.7 **23.3(+0.6)**$^{\uparrow}$	13.1 13.3$^{\uparrow}$	17.2 16.8	+0.3
en-hi^{El}	0.9 2.3$^{\uparrow}$	1.1 2.1$^{\uparrow}$	10.3 **13.1(+2.8)**$^{\uparrow}$	13.6 **14.5(+0.9)**$^{\uparrow}$	0.2 0.3$^{\uparrow}$	0.1 0.3$^{\uparrow}$	+1.1

1) Each model with the LM^{2}$_{g}$SAR pre-training and fine-tuning shows better performances than the only fine-tuning one on all benchmarks, with an average of 0.3~1.1 BLEU improvement in the six benchmarks.

2) In the significant test, the improvement of the models that report the best BLEU scores in both directions is significant.

Discussion

■ Quality control in MCoPSA agorithm

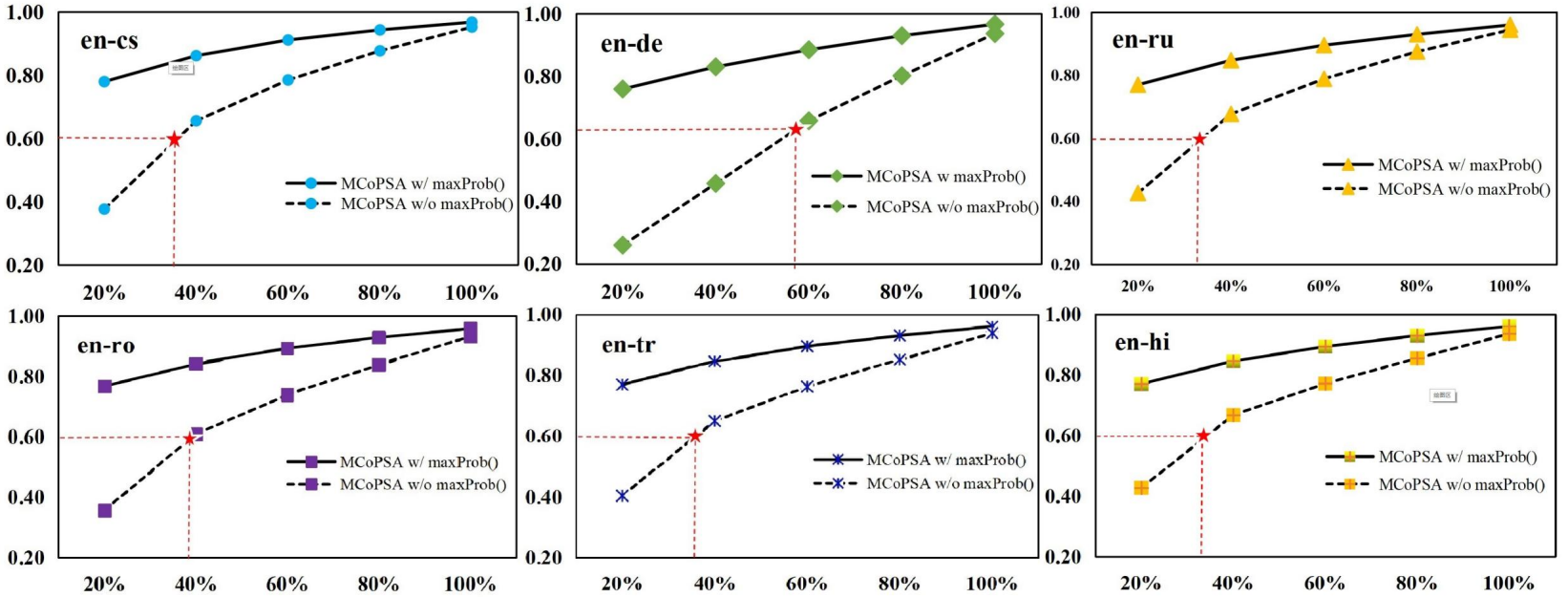

Figure 3: The average LaBSE score curve of each fold on the en-XX languages. The X-axis is thedistribution interval of each fold, and the Y-axis is the LaBSE score.

- collected the aligned linguistic units based on "MCoPSA w/o maxProb(·)" in en-XX languages.
- used LaBSE to compute the similarity scores and ranked them in ascending order.
- divided the ranked units into five folds, and each fold contains 20% of the whole units.
- averaged the LaBSE scores in each fold.

Discussion

- Lifelong property of the approach

The term "lifelong" refers to the sustainability and extensibility of the proposed approach, which is mainly reflected in the language extension and continuous scale expansion of the resource.

1) Language extension: the approach can easily extend new languages into LM^2_gSAR through their bilingual data, and make connections between the new language and other languages to improve linguistic diversity.

2) Scale expansion: with the expansion of the parallel data, the approach can expand the scale of the resource, and supplement more linguistic units to perfect its resource.

Conclusion

■ In this work

1) We proposed a lifelong multilingual multi-granularity semantic alignment approach via maximum co-occurrence probability in the noisy parallel data.

1) We released a version of its corresponding resource.

2) We also conducted experiments to prove the ability of the MCoPSA algorithm compared to the traditional aligners and elaborate on how to use the resource to prove its effectiveness in machine translation tasks.

■ In the future

1) Continue to optimize the approach from the quality and linguistic diversity.

2) Continue to release more versions of the resource with the optimized approach to support more languages and provide a bigger scale.

3) Continue to explore the strategies for utilizing the resource to contribute to the pre-training methods.

That is all, thanks!