対訳文対のアライメントを考慮したサブワード分割

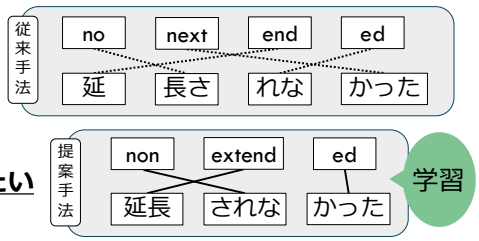
西田 祥人, 二宮 崇, 後藤 功雄（愛媛大学）

目的：対訳関係を考慮した良質なデータセットで機械翻訳をしたい

1. 課題：従来のサブワード分割手法は対訳関係を考慮していない

- 従来のサブワード分割手法は、言語ごとに独立に分割モデルを学習してあり、**対訳関係を考慮していない**
- 先行研究[1]では、EMアルゴリズムを用いて対訳文対のアライメント確率を学習するが、**外部のアライメントツールに依存する**

⇨ **対訳文対のアライメント確率に加え、サブワードアライメント確率の学習もしたい**



[1] 松井ら (2023) バイリンガルサブワード分割のためのEMアルゴリズム. 言語処理学会年次大会.

2. 提案手法：EMアルゴリズムを用いて、対訳文対のサブワードの対応関係を学習

2.1 サブワード分割のための確率モデル

原言語文Xと目的言語文Yの生起確率：

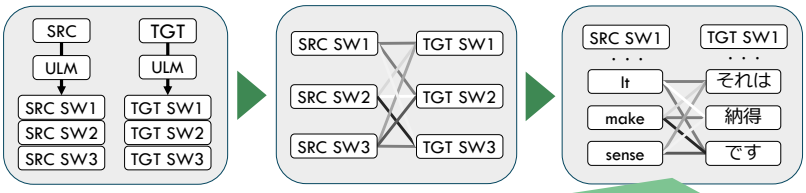
$$P(X,Y) \approx \sum_k \sum_l \sum_{a \in A(x^{(k)}, y^{(l)})} P_U(x^{(k)}) P_U(y^{(l)}) \prod_{u,v \in a} \alpha_{uv}$$

$S(X)$：Xのtop-Kサブワード分割候補（$x^{(1)}, x^{(2)}, \ldots, x^{(K)} \in S(X)$）
$S(Y)$：Yのtop-Lサブワード分割候補（$y^{(1)}, y^{(2)}, \ldots, y^{(L)} \in S(Y)$）
$P_U(x)$：ユニグラム言語モデルが出力するサブワード列の生起確率
$A(x^{(k)}, y^{(l)})$：$x^{(k)}$と$y^{(l)}$のサブワードのアライメントを返す関数
α_{uv}：原言語側サブワードuと目的言語側サブワードvが翻訳（アライメント）関係にある確率（u, vの同時確率$P(u,v)$）

2.2 EMアルゴリズムによるアライメント確率α_{uv}の学習

$$\alpha_{uv}^{\text{new}} = \frac{\sum_n \sum_k \sum_l E_{nkluv}}{\sum_{u'} \sum_{v'} \sum_n \sum_k \sum_l E_{nklu'v'}}$$

$$E_{nkluv} \approx \frac{P_U(x_n^{(k)}) P_U(y_n^{(l)}) \prod_{u \in x_n^{(k)}} \sum_{v \in y_n^{(l)}} \alpha_{uv}^{\text{old}}}{\sum_{k'} \sum_{l'} P_U(x_n^{(k')}) P_U(y_n^{(l')}) \prod_{u \in x_n^{(k')}} \sum_{v \in y_n^{(l')}} \alpha_{uv}^{\text{old}}} C_{nkluv}$$



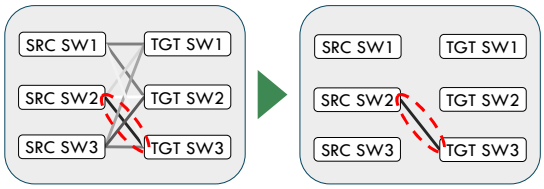
アライメントを繰り返すことで確率が更新

C_{nkluv}：n番目の文のkとlのサブワード文対に
サブワードuとvが同時に存在する回数

2.3 訓練文のサブワード分割

アライメント確率から、**対応関係が最も高いサブワード文対**を取得

$$\hat{k}, \hat{l} = \underset{k,l}{\operatorname{argmax}} P_U(x^{(k)}) P_U(y^{(l)}) \prod_{u,v \in a} \alpha_{uv}$$



2.4 評価文のサブワード分割

目的言語文のない評価文に対してはu, vの同時確率を**目的言語側サブワードで周辺化** → 原言語のサブワード列を取得

u, vの同時確率の周辺化：$\alpha'_u = \sum_{v \in V_{\text{target}}} \alpha_{uv}$

V_{target}：目的言語側のサブワード集合

$$\hat{k} = \underset{k}{\operatorname{argmax}} P_U(x^{(k)}) \prod_{u \in x^{(k)}} \alpha'_u$$

3. 評価実験：提案手法を評価した結果、従来手法と同等以上の性能を達成

3.1 実験設定

- データセット：ASPEC[2]
- ベースライン：SentencePiece[3]
- ユニグラム言語モデル：SentencePiece
- NMTモデル（Transformer）：Fairseq[4]
- サブワードの語彙サイズ
  各言語：16,000
- 分割候補数（K, L）：10
- EMステップ数：5

	訓練用	検証用	評価用
ASPEC-JE	1,000,000	1,790	1,812
ASPEC-JC	672,315	2,090	2,107

3.2 実験結果

		英日	日英	中日	日中
BLEU[5]	ベースライン	27.2	27.0	28.9	35.4
	提案手法	**27.6**	**27.5**	**29.2**	**35.5**
COMET[6]	ベースライン	**0.8882**	0.8182	0.9049	0.8675
	提案手法	0.8880	**0.8195**	**0.9055**	**0.8680**

提案手法の翻訳が改善した例

	分割結果	翻訳結果
正解データ	quilibrious interval disorder	平衡間隔失調
ベースライン	_ qui lib r ious _interval _disorder	巧 妙 な 区間 障害
提案手法	_ qui lib ri ous _interval _disorder	平衡 間隔 障害

[2] Nakazawa et al. (LREC 2016) ASPEC: Asian Scientific Paper Excerpt Corpus.
[3] Kudo & Richardson. (EMNLP 2018) SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing.
[4] Ott et al. (NAACL 2019) fairseq: A Fast, Extensible Toolkit for Sequence Modeling.
[5] Papineni et al. (ACL 2002) Bleu: a Method for Automatic Evaluation of Machine Translation.
[6] Rei et al. (WMT 2022) COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task.