ニューラル機械翻訳における単語分割器のドメイン適応

榎本大晟　平澤寅庄　金輝燦　岡照晃　小町守
東京都立大学
{enomoto-taisei, hirasawa-tosho, kim-hwichan}@ed.tmu.ac.jp,
{teruaki-oka, komachi}@tmu.ac.jp

概要

fine-tuning（微調整）で in-domain にドメイン適応をする際，一般的にモデルのみが対象であり，単語分割は general-domain で使用した単語分割器をそのまま使用する．既存研究では，タスクやモデルに適した単語分割を行うことにより，モデルの性能が向上することが示されており，ドメインにおいても同様に general-domain の単語分割器が，in-domain において最適とは限らないと考えられる．そこで本研究では，機械翻訳タスクを対象とし，モデルと単語分割器の学習を同時に行う単語分割同時最適化を用いて，単語分割器のドメイン適応を行うことで，翻訳精度の改善を試みる．実験により，単語分割器のドメイン適応が in-domain データの翻訳性能の向上に寄与することを示す．

1　はじめに

単語分割は自然言語処理タスクの精度に影響を与える重要な処理である．これまでの研究から，モデルの性能が向上するような適切な単語分割はタスクやモデルの構造などによって異なることがわかっている [1, 2, 3, 4, 5]．近年では，タスクやモデルに応じて単語分割を自動で最適化する手法が研究されており，その内の 1 つに平岡ら [6] による単語分割同時最適化手法がある．これは，タスクに用いるモデルと単語分割器を同時に End-to-End で学習する手法である．平岡らは，機械翻訳タスクでの単語分割同時最適化の評価実験では in-domain データのみでモデルと単語分割器の学習を行い，翻訳性能の向上を確認している[1)]．

一方，ドメインに特化した機械翻訳システムは高い需要があるが，一般的に機械翻訳モデルの学習には大規模なコーパスが必要であるため，in-domain データのみを用いてモデルを作成することは困難である．この問題を解決するため，ドメイン適応では，大規模な general-domain データで事前学習したモデルを，目的の in-domain データで微調整する（図 1a）．ドメイン適応により，in-domain データが限定的であっても，高い翻訳性能のモデルを学習できる．しかしながら一般的なドメイン適応はタスクのモデルのみを微調整し，単語分割器は general-domain データのものを in-domain データに対してそのまま使用する．前述の通り単語分割はモデルの性能に大きく関与するため，general-domain の単語分割器が in-domain において最適とは限らない．

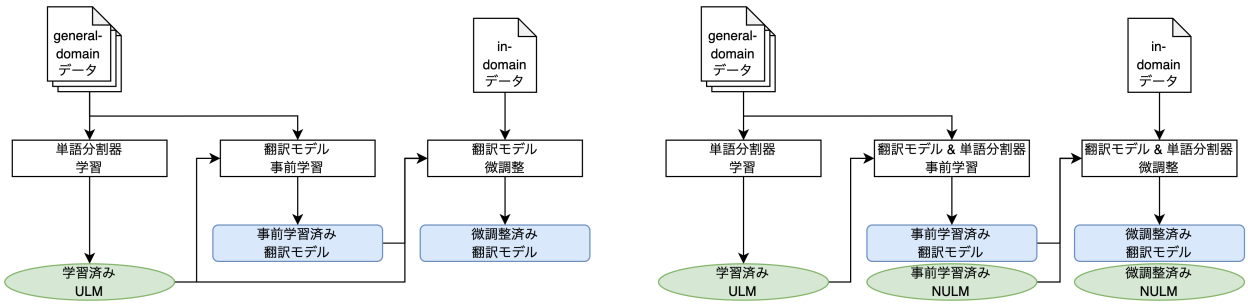
そこで本稿では，平岡らの単語分割同時最適化を用いて，翻訳モデルと単語分割器の両方のドメイン適応を行い，in-domain データにおける翻訳精度の改善に取り組む（図 1b）．提案手法はまず事前学習の際に単語分割同時最適化を適用することで，事前学習に用いる general-domain データに適した単語分割器と翻訳モデルを獲得する．その後，翻訳モデルのみでなく単語分割器も in-domain データに適したものにするために，微調整の際にも単語分割同時最適化を適用する．英日と英独の機械翻訳タスクにおける実験を通して，単語分割器のドメイン適応が，in-domain の翻訳性能向上につながることを示す．

2　関連研究

2.1　サブワード分割

機械翻訳では，未知語の問題を解決するために，単語より小さい単位であるサブワードを扱うことが一般的である．サブワード化を行う代表的な手法にユニグラム言語モデル（ULM）に基づく分割 [7] がある．ULM の語彙内のサブワード w のユニグラム確率 $p(w)$ は，EM アルゴリズムを用いて，与えられた学習データをもとに求める．文 s を単語分割する

1) 平岡らの機械翻訳タスクでの実験では，単語単位に分割した後に単語分割同時最適化をおこなっているため，サブワード単位への分割を学習しているといえる．本研究でも同様にサブワード単位への分割を学習する．



(a) 翻訳モデルのみ微調整を行う従来のドメイン適応.　(b) 翻訳モデルと単語分割器の同時ドメイン適応.

図 1: 機械翻訳タスクにおける翻訳モデルと単語分割器のドメイン適応.

際には，尤度（ユニグラム確率の積）が最大となるようなサブワードの系列 $s' = w_1, ..., w_I$ に変換する．

平岡ら [6] はタスクの学習に用いるモデルに単語分割器を組み合わせ，End-to-End で同時に学習する単語分割同時最適化手法を提案した．単語分割器には，ニューラルネットワークで構成された ULM であるニューラルユニグラム言語モデル（NULM）を用いている．平岡らは，単語分割同時最適化によりモデルの性能の向上を報告した．しかしながら，学習には in-domain データのみを使用しており，単語分割の学習にドメイン適応が有効であるかは明らかになっていない．

2.2 ドメイン適応

ニューラル機械翻訳におけるドメイン適応手法は数多く提案されている．Freitag ら [8] は大規模な対訳データで学習したモデルに対して小規模な in-domain データを用いて追加の学習を行う微調整により，in-domain データでの翻訳性能の向上を報告した．Chu ら [9] は微調整時に general-domain データと目的の in-domain データを組み合わせたデータを使用する混合微調整を提案した．いずれの研究でもドメイン適応によりモデルのパラメータのみを更新しており，単語分割器のようなモデル以外の構成要素に対してドメイン適応は行っていない．単語分割はモデルの性能に影響するため，単語分割器のドメイン適応は in-domain データにおける性能の改善につながると考える．

3 単語分割器のドメイン適応

本研究で単語分割器として用いる NULM と，単語分割同時最適化手法を用いた単語分割器のドメイン適応の手順を概説する．NULM の語彙 V は，ULM や BPE [10] などを用いて初期化を行う．NULM では，語彙 V 内のサブワード w について，単語分散表現 $\boldsymbol{v}_w$ と多層パーセプトロン MLP(·) をもとにしてユニグラム確率 $p(w)$ を次のように計算する．

$$d_w = \mathrm{MLP}(\boldsymbol{v}_w) \tag{1}$$

$$p(w) = \frac{\exp(d_w)}{\sum_{\hat{w} \in V} \exp(d_{\hat{w}})} \tag{2}$$

NULM は後段タスクの損失に基づいてパラメータの更新を行う．

本研究で行う単語分割器のドメイン適応とは，NULM のドメイン適応を指す．本研究で提案する，単語分割同時最適化手法を用いた単語分割器のドメイン適応の手順は以下の通りである．

1. general-domain データを用いて ULM を学習する．
2. general-domain データを用いて翻訳モデルと単語分割器を学習する．単語分割器の語彙の初期化には 1 の ULM を用いる．
3. in-domain データを用いて，2 で得られた翻訳モデルと単語分割器を微調整する．

図 1b は翻訳モデルと単語分割器の同時ドメイン適応の手順を図示したものである．

4 実験

4.1 実験設定

データセット　事前学習に用いる general-domain データには，英日翻訳では JParaCrawl v3.0 [11, 12]，英独翻訳では ParaCrawl v9 [13] を使用した．これらはウェブをクローリングすることで作成された大規模な対訳コーパスである．それぞれ学習データとして，全対訳データから 800 万文対を抽出した．in-domain データには，英日翻訳では IWSLT2017 [14] と ASPEC [15]，英独翻訳では IWSLT2017 と EMEA [16]

を使用した．IWSLT は TED talks，ASPEC は科学技術論文，EMEA は医療機関の PDF ドキュメントから作成された対訳コーパスである．コーパスのドメインの違いによる単語分割器の性質や翻訳性能の変化を確認するために，ASPEC の学習データ数は IWSLT（英日）の学習データ数と等しくなるようにダウンサンプリングを行った．同様に，EMEA の学習データにもダウンサンプリングを行った．先行研究 [6] に倣い，コーパスは日本語には MeCab [17]（IPA 辞書），英語とドイツ語には Moses トークナイザ [18] を適用した後に，SentencePiece [19] を用いて ULM の学習を行う．語彙サイズは 32,000 とした．

学習設定　機械翻訳モデルとして Transformer [20]（base）を使用した．単語分割器のドメイン適応の有効性を検証するために，単語分割器の学習を行わない設定と general-domain データのみで行う設定と in-domain データのみで行う設定をベースラインとして比較する．事前学習では 20 エポックの学習を行い，開発データセットにおいて損失値がもっとも低いモデルを使用した．微調整では 20 エポック，事前学習なしのモデルでは 100 エポックの学習を行い，開発データセットにおいて翻訳性能がもっとも高いモデルを評価に使用した．単語分割器の学習はソース側言語とターゲット側言語で同時に行った．全ての設定でサブワード正則化 [7, 21] を用いて学習を行った．評価には BLEU [22] を用いた．報告する全ての BLEU は 3 つのシードでの平均値である．

4.2　実験結果と考察

表 1 に各 in-domain データセットにおける BLEU スコアを示す．実験結果より，すべてのデータセットにおいて，general-domain データでの事前学習から単語分割同時最適化を行い翻訳モデルのみでなく単語分割器もドメイン適応を行うことで最高性能となることが確認され，単語分割器のドメイン適応の有効性が示された．

英独の IWSLT 以外のデータセットにおいて，in-domain データのみで単語分割同時最適化を行う設定は，単語分割同時最適化を行わない設定の翻訳性能を下回ることが確認された．このことから，in-domain データが小規模であるため，in-domain データのみでは単語分割器の学習が十分にできない場合があると考えられる．これに対し，大規模な general-domain データでの事前学習から単語分割同時最適化を行うことは，単語分割器の過学習や学習

表 1: 各 in-domain データセットにおける各手法の BLEU スコア．事前学習における “✓” は翻訳モデルのドメイン適応を行うことを意味する．単語分割最適化における “✓” は該当のデータで単語分割器の学習を行うことを意味する．“–” は該当のデータでモデル自体の学習を行わないことを意味する．General は general-domain データ，In は in-domain データを示す．

	単語分割最適化		英日		英独	
事前学習	General	In	IWSLT	ASPEC	IWSLT	EMEA
	–		12.06	25.99	22.78	28.42
	–	✓	11.87	25.65	22.92	28.22
✓			14.83	27.51	26.06	35.17
✓	✓		14.94	27.24	26.37	35.19
✓		✓	14.40	27.12	26.23	34.92
✓	✓	✓	**15.16**	**27.68**	**26.58**	**35.52**

不足を回避することができると考えられる．そのため，単語分割器のドメイン適応を行った設定では，モデルや in-domain データに合った単語分割器を獲得でき，全てのデータセットにおいて安定して性能が向上したと考えられる．

5　分析

5.1　単語分割器により翻訳が改善された例

単語分割器のドメイン適応を行うことで，単語分割がどのように変化し，結果として翻訳性能が向上したかを分析する．表 2 にモデルの事前学習を行う設定における，単語分割同時最適化を行わない設定（ULM による単語分割）と，in-domain データのみで行う設定と，general-domain データから行い単語分割器ごとドメイン適応を行う設定のそれぞれの単語分割と翻訳結果の例を示す．“_” は空白記号である．以降では，それぞれの単語分割手法を，SP，In，Ours と記述する．“電子写真プロセス” と訳される “The electrophotographic process” という単語に着目すると，SP と In では “_The / _electro / pho / t / ographic / _process” と分割しているのに対し，Ours では “_The / _electro / photo / graph / ic / _process” と分割していることが確認される．その結果，モデルによる該当箇所の翻訳は，SP では未翻訳であり，In では “電気泳動法” という誤った訳が出力されているのに対し，Ours では “電子写真プロセス” と正しい翻訳が出力されていることが確認される．このことから，単語分割器のドメイン適応を行うことにより，希少単語を妥当なサブワードに分割することが可能になり，その結果として翻訳性能が向上すると考えられる．

表 2: 単語分割器の学習を行わない設定（SP），in-domain データのみで行う設定（In），general-domain データから行い単語分割器のドメイン適応を行う設定（Ours）の単語分割と翻訳の結果．“ ...” は省略を意味する．

(a) 各手法で学習した単語分割器による単語分割．

入力文	**The electrophotographic process** is widely applied ...
SP	**_The / _electro / pho / t / ographic / _process** / _is / _widely / _applied / ...
In	**_The / _electro / pho / t / ographic / _process** / _is / _widely / _applied / ...
Ours	**_The / _electro / photo / graph / ic / _process** / _is / _widely / _applied / ...

(b) 各手法で学習した翻訳モデルによる翻訳結果．各モデルの入力は (a) の単語分割となる．

参照訳	**電子写真プロセス**は，... 広く応用されている。
SP	... / 広く / 応用 / さ / れ / て / いる / 。
In	**電気 / 泳動 / 法** / は / ，/ ... / 広く / 応用 / さ / れ / て / いる / 。
Ours	**電子 / 写真 / プロセス** / は / ，/ ... / 広く / 応用 / さ / れ / て / いる / 。

表 3: 英日翻訳において微調整によりユニグラム確率の増加率が大きいサブワード．

IWSLT		ASPEC	
英	日	英	日
_verifii	TED	ic	ラーゼ
_obsess	プリ	_augment	ED
_sounds	シティ	_defect	_SYN

表 4: in-domain データにおいて微調整により単語分割が変わった文の割合（%）．

	英日		英独	
	IWSLT	ASPEC	IWSLT	EMEA
source（英）	0.98	4.29	0.52	6.68
target（日/独）	0.11	0.18	0.35	6.13

5.2 ドメイン適応後の単語分割器の変化

general-domain データにおいて事前学習を行った単語分割器を in-domain データで微調整することにより，単語分割器がどのように変化するのかについて分析を行う．

ユニグラム確率の増加率が大きいサブワード 表 3 に英日の in-domain データでの微調整後に，ユニグラム確率が大きく増加したサブワードを示す [2]．IWSLT（英日) の日本語側では “TED” [3] のユニグラム確率の増加率が大きいことが確認された．IWSLT は TED talks の字幕から作成されたコーパスであり，学習データ内では “TED” を含む文字列が約 900 回と頻出している．同様に ASPEC の学習データ内には，“magnetic” のような “ic” を接尾辞とする形容詞の単語が頻出しており，その結果，英語側では “ic” のユニグラム確率の増加率が増大した．これらのことから，単語分割器を in-domain データで微調整することにより，in-domain データにおいて重要な役割を持つサブワードのユニグラム確率を増大させていることがわかった．

単語分割が変わった文の割合 表 4 に in-domain データの学習データにおいて，事前学習後の単語分割器と微調整後の単語分割器で単語分割が異なる文の割合を示す．英日では異なる単語分割となる文数は，IWSLT と比較して ASPEC の方が多いことが確認された．これは IWSLT と比較して ASPEC の方が JParaCrawl から離れたドメインデータである（付録 B）ため，微調整によって単語分割器がより変化したのだと考えられる．同様に，英独では EMEA は IWSLT よりも異なる単語分割となる文数が多いことが確認された．この結果は，in-domain データとして使用するコーパスのドメインが特徴的であるほど（general-domain から遠いほど），微調整により単語分割器が大きく変化することを示している．

また，微調整により単語分割が変わった文の割合がもっとも高い EMEA であっても 6.68%と低い割合であり，単語分割器を極端に大きく変化させていないことがわかる．このことから，in-domain データで単語分割器を微調整する際，最小限の調整で翻訳性能を向上させていると考えられる．

6 おわりに

本研究では，ドメイン適応を用いる機械翻訳への単語分割同時最適化の拡張を提案した．提案手法では単語分割器を大規模な general-domain データで事前学習した後に，小規模な in-domain データで微調整する．実験結果より，提案手法が in-domain データの翻訳性能向上に寄与することが確認され，より妥当な単語分割を行うことが可能であることが示唆された．

今後は BERT の事前学習である MLM を行う際に単語分割同時最適化を行い，得られる単語分割器と事前学習済み BERT をさまざまなタスクで微調整した場合に，タスクの性能が向上するかを検証する．

2） 英独については付録 A に掲載する．

3） ここで “_TED” ではなく “TED” である理由は，“_TED” が語彙に登録されていないためである．これは，語彙が JParaCrawl をもとに作成されており，JParaCrawl では “TED” が冒頭となる単語が少なく，“TED” が語中や末尾となる単語が多いためである．

参考文献

[1] Jia Xu, Jianfeng Gao, Kristina Toutanova, and Hermann Ney. Bayesian semi-supervised Chinese word segmentation for statistical machine translation. In **Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)**, pp. 1017–1024, Manchester, UK, August 2008. Coling 2008 Organizing Committee.

[2] Pi-Chuan Chang, Michel Galley, and Christopher D Manning. Optimizing chinese word segmentation for machine translation performance. In **Proceedings of the third workshop on statistical machine translation**, pp. 224–232, 2008.

[3] ThuyLinh Nguyen, Stephan Vogel, and Noah A. Smith. Nonparametric word segmentation for machine translation. In **Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)**, pp. 815–823, Beijing, China, August 2010. Coling 2010 Organizing Committee.

[4] Miguel Domingo, Mercedes Garcıa-Martınez, Alexandre Helle, Francisco Casacuberta, and Manuel Herranz. How much does tokenization affect neural machine translation?, 2018. arXiv:1812.08621.

[5] Tatsuya Hiraoka, Sho Takase, Kei Uchiumi, Atsushi Keyaki, and Naoaki Okazaki. Optimizing word segmentation for downstream task. In **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 1341–1351, Online, November 2020. Association for Computational Linguistics.

[6] Tatsuya Hiraoka, Sho Takase, Kei Uchiumi, Atsushi Keyaki, and Naoaki Okazaki. Joint optimization of tokenization and downstream model. In **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 244–255, Online, August 2021. Association for Computational Linguistics.

[7] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 66–75, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[8] Markus Freitag and Yaser Al-Onaizan. Fast domain adaptation for neural machine translation, 2016. arXiv:1612.06897.

[9] Chenhui Chu, Raj Dabre, and Sadao Kurohashi. An empirical comparison of domain adaptation methods for neural machine translation. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 385–391, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[10] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

[11] Makoto Morishita, Jun Suzuki, and Masaaki Nagata. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 3603–3609, Marseille, France, May 2020. European Language Resources Association.

[12] Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 6704–6710, Marseille, France, June 2022. European Language Resources Association.

[13] Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In **Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks**, pp. 118–119, Dublin, Ireland, August 2019. European Association for Machine Translation.

[14] Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuitho Sudoh, Koichiro Yoshino, and Christian Federmann. Overview of the IWSLT 2017 evaluation campaign. In **Proceedings of the 14th International Workshop on Spoken Language Translation**, pp. 2–14, 2017.

[15] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. ASPEC: Asian scientific paper excerpt corpus. In **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**, pp. 2204–2208, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[16] Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)**, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[17] Taku Kudo. MeCab: Yet another part-of-speech and morphological analyzer, 2006. http://taku910.github.io/mecab/.

[18] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In **Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions**, pp. 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[19] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[21] Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. BPE-dropout: Simple and effective subword regularization. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 1882–1892, Online, July 2020. Association for Computational Linguistics.

[22] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[23] Roee Aharoni and Yoav Goldberg. Unsupervised domain clusters in pretrained language models. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7747–7763, Online, July 2020. Association for Computational Linguistics.

付録

A　英独においてユニグラム確率の増加率が大きいサブワード

表 5 に英独の in-domain データでの微調整後に，ユニグラム確率が大きく増加したサブワードを示す．EMEA のドイツ語側では “kin” のユニグラム確率の増加率が大きいことが確認された．EMEA の学習データ内では “pharmakokinetik” という医学分野特有の単語が約 1,500 回と頻出しており，この単語は “pharmako / kin / etik” と分割されることが意味的に妥当であると考えられる．また，“Interleukin” や “Hodgkin” のように “kin” で終わる医学系の単語も頻出している．これらのことから，EMEA のドイツ語側では “kin” が重要な役割を持つサブワードであると考えられる．英語側では “g” のユニグラム確率の増加率が大きいことが確認された．EMEA は医療ドメインであり，薬品などの質量について記述した文が多くある．そのため，質量の単位である “g”，“mg”，“ng” などがコーパス内では頻出している．

これらのことから英日と同様に，英独においても in-domain データにおいて重要な役割を持つサブワードのユニグラム確率を増大させていることがわかった．

表 5: 英独翻訳において微調整によりユニグラム確率の増加率が大きいサブワード．

IWSLT		EMEA	
英	独	英	独
_boost	_Sch	g	kin
_sup	liz	_mugg	tro
ory	rie	ara	ati

B　コーパスのドメインの離れ具合

使用したコーパスのドメインの離れ具合について報告する．Aharoni ら [23] に倣い，各コーパスのソース側の文に対して，事前学習済み BERT の隠れ層のベクトルを取得し，PCA を用いて 2 次元可視化を行う．図 2，3 に各言語対に使用したコーパスの 2 次元可視化の結果を示す．図 2 から IWSLT のデータの多くが JParaCrawl のデータに重なっているのに対し，ASPEC のデータの多くは JParaCrawl のデータに重なっていないことが確認された．このことから，IWSLT と比較して ASPEC の方が JParaCrawl から離れたドメインデータであると考えられる．同様に図 3 から，IWSLT と比較して EMEA の方が ParaCrawl から離れたドメインデータであると考えられる．

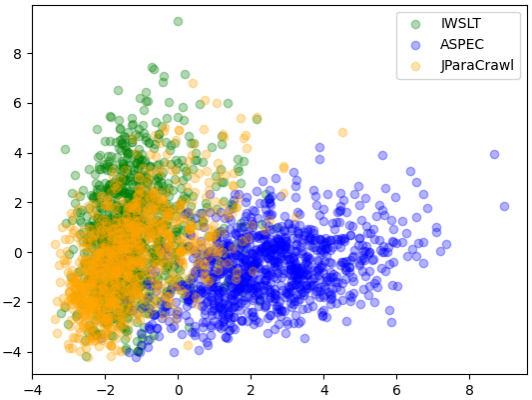


図 2: PCA を用いた英日データセットの BERT 隠れ層の 2 次元可視化．

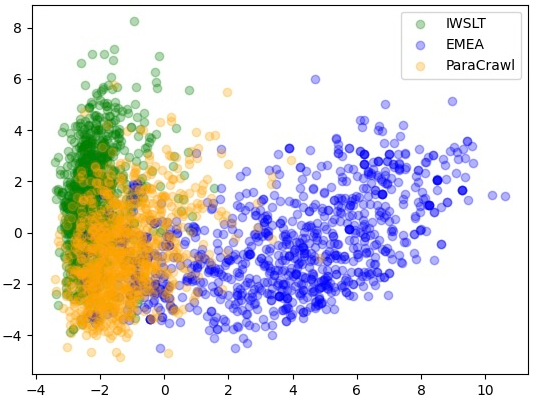


図 3: PCA を用いた英独データセットの BERT 隠れ層の 2 次元可視化．