

# Analysis of Vocabulary and Subword Tokenization Settings for Optimal Fine-tuning of MT

A Case Study of In-domain Translation

Javad Pourmostafa   Dimitar Shterionov   Pieter Spronck

Department of Intelligent Systems  
Tilburg University, The Netherlands

RANLP 2025 — September 10 — Varna, Bulgaria

# Introduction

---

# Motivation

- **SW tokenization and vocabulary** crucially affect both training and fine-tuning of MT models.
- **Fine-tuning** adapts models to new data, but:
  - New data introduces **unseen tokens**.
  - Token distributions can **differ from the base domain**.
- The **original SW model** may be less suitable for the new domain.

# Problem Statement

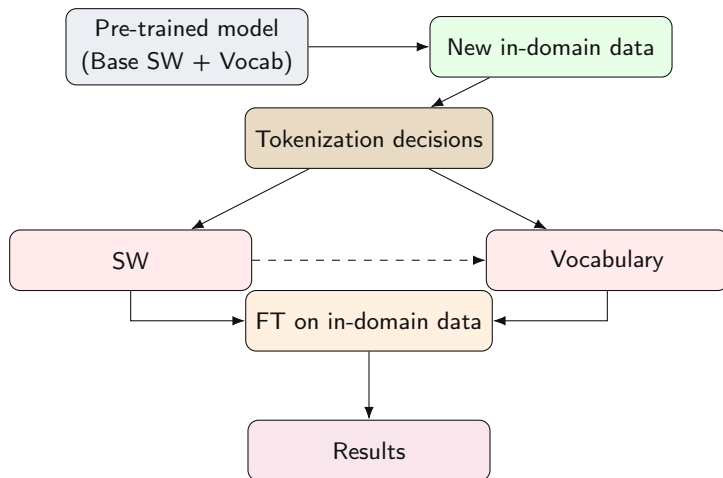
- Prior work: fine-tuning improves MT, but pipelines typically assume **reusing the base tokenizer and vocabulary**.
- **Gap:** the effect of different **SW segmentation** and **vocabulary creation** strategies during FT is not systematically studied.
- We ask:
  - Which SW + vocab configurations give the best **in-domain** performance?
  - How do these choices affect **generalization** to out-of-domain (OOD) data and overall **efficiency**?

Scope: controlled NMT setup, not LLMs, to isolate tokenization effects [2, 3].

# Pipeline

---

# Where SW & Vocabulary sit in the DA pipeline



In this work, we used BPE for subword tokenization.

## Decision Points

---

# Decision points

- **A. SW segmentation:** reuse base BPE; train in-domain BPE; train combined BPE [9, 7].
- **B. Vocabulary:** reuse base vocabulary; expand with base+in-domain; use in-domain vocabulary.

Notation:  $D$  = out-of-domain data,  $E$  = in-domain data.



# Decision points: SW segmentation & Vocabulary

## SW segmentation

- $D_{\text{BPE}}$ : maximum compatibility with base model.
- $E_{\text{BPE}}$ : captures domain morphemes/terms; better for in-domain.
- $(D + E)_{\text{BPE}}$ : compromise across distributions [9, 7].

## Vocabulary

- $|D|$ : safe, no change; misses domain terms.
- $|D + E|$ : extends base with domain tokens.
- $|E|$ : maximal domain capacity; mismatch to base [8].

SW segmentation defines how words are split; the vocabulary source determines which subwords are included in the embeddings.

## Evaluated configurations (C1–C9)

Config	BPE (vocab+FT)	Vocab source	Status
C1	$D_{\text{BPE}}$	$D$	Valid
C2	$D_{\text{BPE}}$	$D+E$	Valid
C3	$D_{\text{BPE}}$	$E$	Weak
C4	$E_{\text{BPE}}$	$D$	Weak
C5	$E_{\text{BPE}}$	$D+E$	Valid
C6	$E_{\text{BPE}}$	$E$	Valid
C7	$(D + E)_{\text{BPE}}$	$D+E$	Valid
C8	$(D + E)_{\text{BPE}}$	$D$	Excluded
C9	$(D + E)_{\text{BPE}}$	$E$	Excluded

**Valid:** consistent SW+vocab sources (C1, C2, C5, C6, C7).

**Weak:** mismatched but tested for comparison (C3, C4).

**Excluded:** severe mismatches (C8, C9).

# Experimental Setup

---

# Data

- **Out-of-domain (OOD,  $D$ ):** WMT18 English–German (En–De) subset  $\sim 12.7\text{M}$  sentence pairs.
- **In-domain ( $E$ ):** Medical En–De  $\sim 248\text{K}$  sentence pairs (cleaned/re-split).
- **Combined ( $D+E$ ):** oversample  $E$  to balance; used for SW/vocab where required [8, 9].

# Model & Training

- OpenNMT-py Transformer: 6e/6d,  $d=512$ , 8 heads, FFN= 2048 [2, 3].
- Noam LR 2.0; warmup 8k; label smoothing 0.1.
- $\leq 200k$  steps; early stopping; batch 10,240 tokens; grad-acc 4.
- BPE merges: 8k/30k/50k by corpus size; src/tgt trained separately [9, 1, 6].

# Design & Metrics

- **Setup:** Base model trained on  $D$ ; fine-tuned on  $E$  under C1–C7 (vary only SW +vocab).
- **Translation metrics:**
  - BLEU (primary, with bootstrap resampling for reliability).
  - TER and chrF2 as secondary metrics for tie-breaking and nuance.
- **Efficiency metrics:**
  - Training time (hours).
  - Carbon emissions via CodeCarbon [7].
- **Decision rule:** rank by BLEU (with stats), resolve close cases with TER/chrF2, and reason about cost with efficiency.

## Results

---

In-domain results (test on  $E$ )

Cfg	BPE	Vocab	BLEU $\uparrow$	TER $\downarrow$	chrF2 $\uparrow$
C1	$D_{\text{BPE}}$	$D$	53.6	49.3	69.4
C2	$D_{\text{BPE}}$	$D+E$	53.4	49.9	69.5
C3	$D_{\text{BPE}}$	$E$	51.7	50.9	68.4
C4	$E_{\text{BPE}}$	$D$	46.6	53.0	64.5
C5	$E_{\text{BPE}}$	$D+E$	53.1	49.7	68.9
C6	$E_{\text{BPE}}$	$E$	<b>54.8</b>	<b>48.9</b>	<b>69.8</b>
C7	$(D+E)_{\text{BPE}}$	$D+E$	53.2	50.1	69.1

**Takeaway:** In-domain BPE+vocab (C6) wins; mismatched segmentation/vocab (C3 and C4) hurts.



# Vocabulary overlap vs. in-domain performance

Cfg	BLEU	SRC Overlap %	New Tokens SRC	New Tokens TGT
C6	<b>54.8</b>	82.84	13,022	13,559
C5	53.1	83.04	14,289	14,157
C7	53.2	<b>97.61</b>	14,300	15,077
C2	53.4	83.04	11,736	11,804
C1	53.6	<b>100.00</b>	0	0

*Note: SRC Overlap % = proportion of source-side vocabulary shared with the base model (higher = fewer new tokens introduced).*

**Takeaway:** C6 performs best by adding many in-domain tokens (low overlap). C1 is most stable (full overlap) but adapts least. Hybrids (C2, C5, C7) sit in-between.

# Out-of-domain & efficiency at a glance

## Out-of-domain (test on $D$ )

Cfg	BPE / Vocab	BLEU	Drop	Drop (%)
Base	$D_{\text{BPE}} / D$	<b>33.9</b>	–	–
C2	$D_{\text{BPE}} / D + E$	15.1	-18.8	-55.5
C7	$(D + E)_{\text{BPE}} / D + E$	15.0	-18.9	-55.8
C1	$D_{\text{BPE}} / D$	13.1	-20.8	-61.4
C6	$E_{\text{BPE}} / E$	7.7	-26.2	-77.3
C5	$E_{\text{BPE}} / D + E$	7.0	-26.9	-79.4

## Efficiency (FT cost)

Cfg	BPE	CO <sub>2</sub> (g)	Time (h)
C6	$E_{\text{BPE}}$	1587.4	09:30
C1	$D_{\text{BPE}}$	1658.7	07:45
C5	$E_{\text{BPE}}$	<b>729.0</b>	<b>03:15</b>
C2	$D_{\text{BPE}}$	1198.7	05:15
C7	$(D + E)_{\text{BPE}}$	<b>543.8</b>	<b>03:08</b>

**Takeaway:** Domain alignment ( $E$ -specific, e.g. C6) boosts in-domain but collapses out-of-domain. Hybrids (C2, C7) preserve OOD and are more efficient (lower CO<sub>2</sub>, shorter training).

## Limits & Next Steps

---

# Limitations & Next steps

## Limitations

- One language pair (En–De) and one domain (medical).
- Only BPE; fixed hyperparameters.
- Automatic metrics only (BLEU, TER, chrF2).

## Next steps

- More domains/pairs; multilingual setups.
- Compare tokenizers (WordPiece, Unigram, LMVR).
- Adaptive vocab selection; test LLM-based MT.
- Add COMET and human evaluation; study HP–tokenization trade-offs.

## Takeaways

---

# Practical recommendations

- **Best in-domain:** train both BPE and vocab on in-domain data.
- **Balanced setup:** use combined base+domain vocab/BPE to keep overlap and OOD strength.
- **Avoid mismatches:** mixing segmentation and vocab sources hurts performance.
- **Plan resources:** introducing many new domain tokens increases FT time and CO<sub>2</sub>; hybrids are faster/greener.










# Благодаря!

Thank you!

<https://github.com/JoyeBright/subword-ft-guide>












# References

-  Papineni, K. et al. (2002). BLEU. *ACL*.
-  Snover, M. et al. (2006). TER. *AMTA*.
-  Popović, M. (2015). chrF. *WMT*.
-  Luong, M.-T. et al. (2015). Attention-based NMT. *EMNLP*.
-  Freitag, M. et al. (2016). Fast DA for NMT. arXiv.
-  Pourmostafa Roshan Sharami, J. et al. (2022). Selecting data for FT NMT.
-  Lim, K. et al. (2018). Subword segmentation for NMT.
-  Sato, M. et al. (2020). Vocabulary adaptation for domain transfer.
-  Sennrich, R. et al. (2016). BPE for rare words. *ACL*.



## References (cont.)

-  Kudo, T., Richardson, J. (2018). SentencePiece. *EMNLP*.
-  Klein, G. et al. (2017). OpenNMT. *ACL*.
-  Vaswani, A. et al. (2017). Attention Is All You Need. *NeurIPS*.
-  Aharoni, R., Goldberg, Y. (2020). Unsupervised domain splits.
-  Koehn, P., Knowles, R. (2017). Six Challenges for NMT. *ACL*.
-  Adlaon, K., Marcos, D. (2024). Optimal BPE merges.
-  Courty, B. et al. (2024). CodeCarbon. *Zenodo*.
-  Wang, X. et al. (2020). Balancing Training for Multilingual NMT. *ACL*.
-  Liu, Y. et al. (2020). Multilingual Denoising Pre-training. *TACL*.