Adapting Pre-trained Language Models to Low-Resource Text Simplification: The Path Matters

Cristina Garbacea[1], **Qiaozhu Mei**[1,2]

[1]Department of EECS, [2]School of Information, University of Michigan, Ann Arbor, MI, USA

Text Simplification Matters for Real People in Real Life!

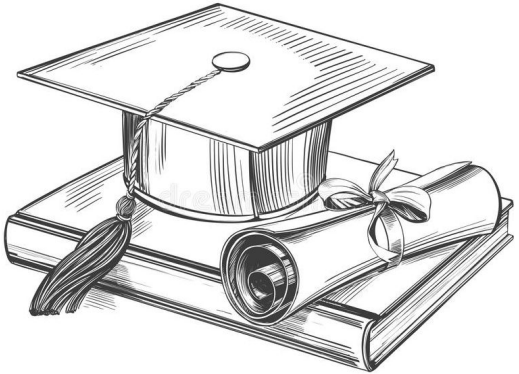
- Text simplification aims to transform complex/ specialized content into simpler and more accessible text
- Improves the ***fairness*** and ***transparency*** of information systems

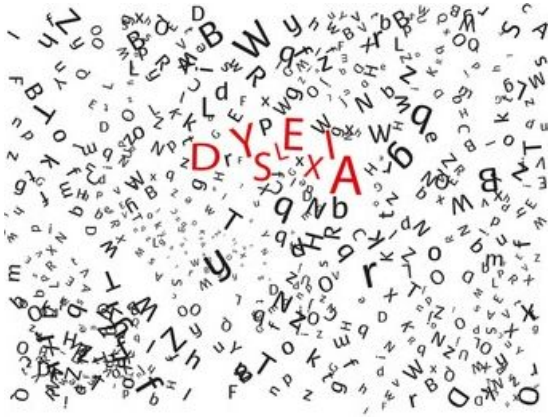
Despite its difficulty, text simplification provides equitable information services to the broad population!

Text Simplification Matters for Real People in Real Life!

Healthcare **Education** **Kids**

Text Simplification Matters for Real People in Real Life!





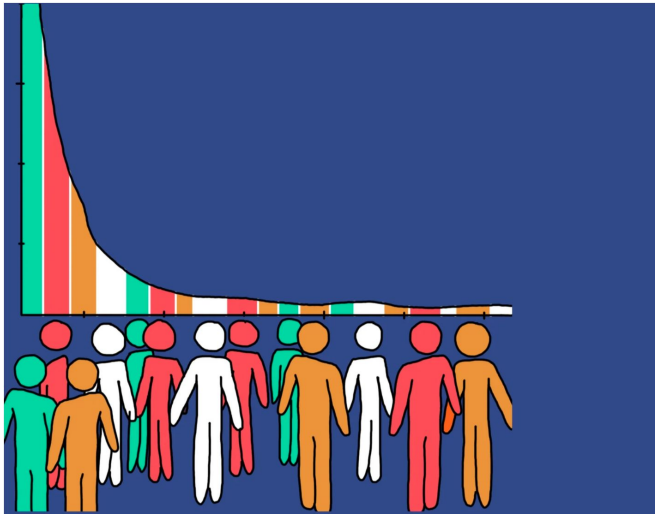
Non-native speakers **Low education** **Medical conditions**

Low-Resource Text Simplification

Text simplification is a low-resource setting:

- Abundant training examples do not exist
  - few domains with limited amount of parallel aligned complex-simple sentences

- Data labeling is costly
  - running user studies is costly and time-consuming
  - annotators need proper and clear instructions



NLP for Low-Resource Settings

- **Transfer Learning:**
  - source and target domains consist of the same feature space (Day, 2017)
  - sufficient amounts of in-domain data for the target task (Chen et al, 2019)
  - ***Issues***: model overfitting, catastrophic forgetting, negative transfer across tasks (Xu et al, 2020), (Thompson et al, 2019), (Kirkpatrick, 2017)

- **Meta-Learning:**
  - Promising general learning strategy suitable for few-shot learning and cross-domain generalization (Li et al, 2018), (Wang et al, 2020)
  - applicable to resource constrained problems where there is a distribution of tasks
  - ***Metric Learning*** (Vinyals et al, 2016), (Snell et al, 2017), ***Memory Networks*** (Santoro et al, 2016),(Oreshkin et al, 2018), ***Gradient based*** (Finn et al, 2017), (Zhang et al, 2018)

NLP for Low-Resource Settings

- **Transfer Learning:**
  - source and target domains consist of the same feature space (Day, 2017)
  - sufficient amounts of in-domain data for the target task (Chen et al, 2019)
  - ***Issues***: model overfitting, catastrophic forgetting, negative transfer across tasks (Xu et al, 2020), (Thompson et al, 2019), (Kirkpatrick, 2017)

- **Meta-Learning:**
  - Promising general learning strategy suitable for few-shot learning and cross-domain generalization (Li et al, 2018), (Wang et al, 2020)
  - applicable to resource constrained problems where there is a distribution of tasks
  - ***Metric Learning*** (Vinyals et al, 2016), (Snell et al, 2017), ***Memory Networks*** (Santoro et al, 2016),(Oreshkin et al, 2018), ***Gradient based*** (Finn et al, 2017), (Zhang et al, 2018)

NLP for Low-Resource Settings

- **Transfer Learning:**
  - source and target domains consist of the same feature space (Day, 2017)
  - sufficient amounts of in-domain data for the target task (Chen et al, 2019)
  - ***Issues***: model overfitting, catastrophic forgetting, negative transfer across tasks (Xu et al, 2020), (Thompson et al, 2019), (Kirkpatrick, 2017)

- **Meta-Learning:**
  - Promising general learning strategy suitable for few-shot learning and cross-domain generalization (Li et al, 2018), (Wang et al, 2020)
  - applicable to resource constrained problems where there is a distribution of tasks
  - ***Metric Learning*** (Vinyals et al, 2016), (Snell et al, 2017), ***Memory Networks*** (Santoro et al, 2016),(Oreshkin et al, 2018), ***Gradient based*** (Finn et al, 2017), (Zhang et al, 2018)

Adapting Pre-trained Language Models to Low-Resource Text Simplification

Research Questions:

RQ1: Can we learn how to quickly adapt pre-trained language models to new tasks and domains with few training examples in the context of text simplification?

Adapting Pre-trained Language Models to Low-Resource Text Simplification

Research Questions:

RQ1: Can we learn how to quickly adapt pre-trained language models to new tasks and domains with few training examples in the context of text simplification?

RQ2: Can we combine the advantages of task and domain adaptation?

RQ3: Can consecutive stages of task and domain adaptation improve one-stage adaptation?

RQ4: Which is the ideal order in which to perform adaptation?

Adapting Pre-trained Language Models to Low-Resource Text Simplification

Proposed Approach:

- Frame the problem of text simplification from a ***task and domain adaptation perspective*** in diverse domains, including news and scientific articles

- Consider complex-simple parallel aligned examples as samples drawn from a distribution of text generation tasks with ***varying constraints on the level of text complexity and readability***

Adapting Pre-trained Language Models to Low-Resource Text Simplification

Methods:

a) **(Domain Adaptation) Transfer Learning:** fine-tunes a general purpose language model to the new domains of text simplification with limited in-domain data

b) **(Task Adaptation) Gradient based meta-learning:**
simulate many domain adaptation tasks to learn model parameters that can generalize to new tasks with few examples

Related Work (I)

- **Machine translation:**
  - knowledge extracted from high-resource language pairs is leveraged to adapt MT systems to low-resource languages (Gut et al, 2018), (Sharaf et al, 2020)

- **Visual Classification Benchmarks:**
  - meta-learning approaches struggle on OOD tasks, overall performance is inferior to transfer learning (Dumoulin et al, 2021); having sufficient heterogeneous tasks is a critical for meta-model training (Kang et al, 2018)

- **Text Classification:**
  - combining domain and task adaptive pre-training improves performance (Gururangan et al, 2020), 4 domains only

Related Work (II)

- **Similarity between source and target** (task and domains) is an important predictive factor of successful adaptation (Vu et al, 2020)

- **Meta-parameterized pre-training** (Raghu et al, 2021):
  meta-learn the pre-training hyper-parameters

- **Meta-finetuning** (Wang et al, 2020):
  improve fine-tuning by meta-learning class prototypes and domain invariant representations; minimizing within-class feature variation is critical for robust few-shot performance on complex tasks (Goldblum et al, 2020)

Experimental Setup (I)

Goal: investigate whether transfer learning or meta-learning is a suitable adaptation strategy when there is a distribution of low-resource text simplification tasks/domains

(Domain Adaptation) Transfer Learning:
- *Fine-tune T5* (Raffel et al, 2020) small (60 million parameters)

(Task Adaptation) Gradient based meta-learning:
- *MAML* (Finn et al, 2017), *Reptile* (Nichols et al, 2018)

Experimental Setup (II)

1) Direct Task Adaptation

2) Direct Domain Adaptation

3) Adding an intermediate stop between source and target

4) Use a “pseudo-stop” based on the target task/ domain

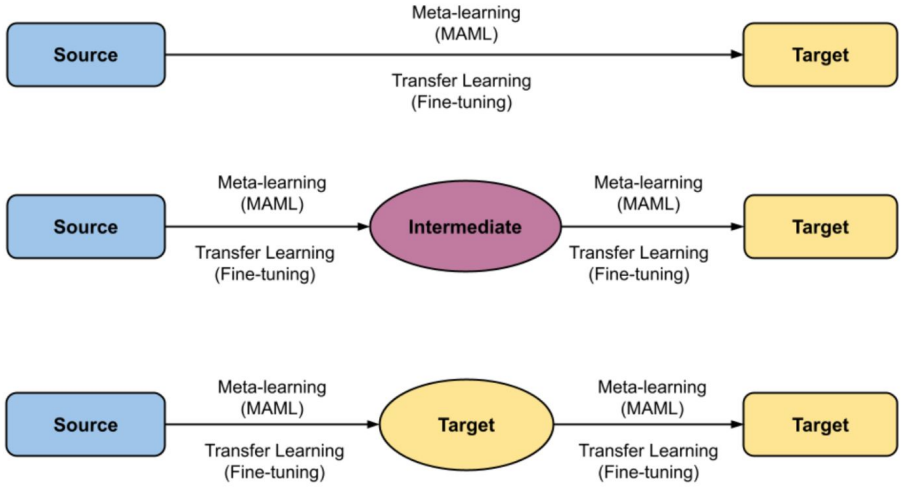


Figure 1: Adapting a pre-trained language model (source) to low-resource text simplification (target).

Experimental Setup (III)

- **Datasets:** different domains and application scenarios of text simplification

  a) **News Simplification**:
  Newsela (Xu et al, 2015) - news articles simplified by professional news editors

  b) **Scientific Press Release:**
  Biendata - research papers from various disciplines matched with press releases

  c) **Auxiliary**:
  WikiLarge (Zhang and Lapata, 2017), *WikiSmall* (Zhu et al, 2010)
  parallel aligned Wikipedia - Simple Wikipedia sentence pairs

Experimental Setup (IV)

- **Newsela** splits by complexity level

Table 6: Newsela splits according to complexity level. 0 denotes the most complex level, and 4 represents the simplest.

Complexity level	Sentence pairs	TRAIN (70%)	DEV (15%)	TEST (15%)	
0 - 1	16,611	11,627	2,492	2,492	META-TRAIN
0 - 2	20,122	14,086	3,018	3,018	
0 - 3	19,891	13,923	2,984	2,984	
1 - 2	12,888	9,022	1,933	1,933	
1 - 3	13,296	9,308	1,994	1,994	
2 - 3	12,146	8,502	1,822	1,822	
2 - 4	9,780	6,846	1,467	1,467	META-DEV
3 - 4	10,185	7,129	1,528	1,528	
0 - 4	16,086	11,260	2,413	2,413	META-TEST
1 - 4	10,577	7,403	1,587	1,587	

Experimental Setup (V)

- **Biendata** splits by scientific domain

Table 8: Biendata splits according to scientific domain.

Scientific Domain	Sentence pairs	TRAIN (50%)	DEV (25%)	TEST (25%)	
Medicine	7,993	3,997	1,998	1,998	META-TRAIN
Biology	10,040	5,020	2,510	2,510	
Internal Medicine	1,095	547	274	274	
Psychology	3,367	1,683	842	842	
Chemistry	1,516	758	379	379	
Cancer Research	1,044	522	261	261	
Neuroscience	1,411	705	353	353	
Virology	1,106	554	276	276	
Pediatrics	812	406	203	203	
Disease	582	292	145	145	
Immunology	2,281	1,141	570	570	META-DEV
Genetics	2,151	1,075	538	538	
Social Psychology	1,090	546	272	272	
Surgery	1,261	631	315	315	
Psychiatry	1,045	523	261	261	
Cognition	662	330	166	166	
Demography	992	496	248	248	
Climate Change	847	423	212	212	
Zoology	645	323	161	161	
Endocrinology	1,582	790	396	396	META-TEST
Cell Biology	2,154	1,076	539	539	
Molecular Biology	904	452	226	226	
Biochemistry	640	320	160	160	
Physical Therapy	1,189	595	297	297	
Nanotechnology	378	188	95	95	
Gerontology	649	325	162	162	
Computer Science	739	369	185	185	
Physics	1,108	554	277	277	
Materials Science	967	483	242	242	
Ecology	2,869	1,435	717	717	
Geography	658	330	164	164	
Economics	384	192	96	96	

Experimental Setup (VI)

- **WikiLarge** random splits

Table 7: WikiLarge random splits.

Subset	Sentence pairs	TRAIN (50%)	DEV (25%)	TEST (25%)	
Wikipedia 0	20,000	10,000	5,000	5,000	META-TRAIN
Wikipedia 1	20,000	10,000	5,000	5,000	
Wikipedia 2	20,000	10,000	5,000	5,000	
Wikipedia 3	20,000	10,000	5,000	5,000	
Wikipedia 4	20,000	10,000	5,000	5,000	
Wikipedia 5	20,000	10,000	5,000	5,000	
Wikipedia 6	20,000	10,000	5,000	5,000	
Wikipedia 7	20,000	10,000	5,000	5,000	
Wikipedia 8	20,000	10,000	5,000	5,000	
Wikipedia 9	20,000	10,000	5,000	5,000	
Wikipedia 10	20,000	10,000	5,000	5,000	META-DEV
Wikipedia 11	20,000	10,000	5,000	5,000	
Wikipedia 12	20,000	10,000	5,000	5,000	
Wikipedia 13	20,000	10,000	5,000	5,000	
Wikipedia 14	20,000	10,000	5,000	5,000	
Wikipedia 15	20,000	10,000	5,000	5,000	META-TEST
Wikipedia 16	20,000	10,000	5,000	5,000	
Wikipedia 17	20,000	10,000	5,000	5,000	
Wikipedia 18	20,000	10,000	5,000	5,000	

Experimental Setup (VII)

- **Evaluation Metrics:**

SARI (Xu et al, 2016): output vs. source and reference simplifications
BLEU (Papineni et al, 2002): output vs. reference simplifications
FKGL (Kincaid et al, 1975): readability
MoverScore (Zhao et al, 2019): semantic distance
MAUVE (Pillutla et al, 2021): KLdiv(output, reference) distributions
BARTScore (Yuan et al, 2021): evaluate text generation as text generation
- ***Faithfulness*** (source → hypotheses),
- ***Precision*** (reference → hypotheses),
- ***Recall*** (hypotheses → reference),
- ***F1***(reference ←→ hypotheses)

Experimental Setup (VIII)

Baseline (No Adaptation):

1) *Transformer* (Vaswani et al, 2017) trained directly on Newsela and Biendata

Additional Baselines (Pre-trained Text Simplification models):

2) *ACCESS* (Martin et al, 2020): controllable sequence-to-sequence simplification model reported highest performance on WikiLarge

3) *DMLMTL* (Guo et al, 2018): Dynamic Multi-Level Multi-Task Learning for Sentence Simplification reported the highest performance on Newsela

Results - Baseline (No Adaptation)

Transformer (Vaswani et al, 2017):
- performance is suboptimal compared to adaptation-based methods
- for Biendata, the model over-simplifies scientific content (FKGL lower than ground-truth)
- results on WikiLarge are better than Newsela and Biendata, i.e. when training data is abundant, neural text simplification performs well without any adaptation

ACCESS (Martin et al, 2020) & DMLMTL (Guo et al, 2018):
- results on Newsela and WikiLarge are consistent with the literature
- performance of both pre-trained models degrades significantly on Biendata
- critical need for task/domain adaptation
- Transformer model directly trained on each dataset outperforms both pre-trained models

Results - One Stage Adaptation

- ***Direct Task Adaptation (Meta-Learning)***
  - Two source language models: T5 and Transformer trained on WikiLarge
  - MAML T5 > MAML Wiki, Reptile Wiki: ***adapting from a powerful pre-trained language model outperforms training a model directly from limited resource***

- ***Direct Domain Adaptation (Fine-tuning)***
  - using T5 as source > Transformer trained on WikiLarge
  - domain adaptation outperforms task adaptation (either MAML or Reptile)
  - ***benefit is larger on OOD scientific press release tasks and domains*** (Biendata)

Results - Two Stage Adaptation

RQ2: Can we combine the advantages of task and domain adaptation?

RQ3: Can consecutive stages of task and domain adaptation improve one-stage adaptation?

RQ4: Which is the ideal order in which to perform adaptation?

- **Intermediate dataset available**
- **No intermediate dataset available**

Results - Two Stage Adaptation

- **Intermediate dataset available (WikiLarge)**

  - first adapt pre-trained T5 to WikiLarge, then continue to adapt the resulting model to Newsela or Biendata; explore combinations of task and domain adaptation at each stage

  - most promising strategy: ***adapt to new tasks*** (through MAML or Reptile) in first stage, ***then continue adapting to new domains*** (through fine-tuning) in the second stage

  - critical to do ***domain adaptation in the final stage***, suggesting that the difference over data distributions is more critical than the difference over tasks in our scenario

  - Repeating the same type of (task/domain) adaptation in both stages is less effective than task + domain; ***two-stage task and domain adaptation is better than one-stage***

Results - Two Stage Adaptation

- **No Intermediate dataset available (use Target as Intermediate)**
  (Source → Target) → Target: (pre-trained T5 → Newsela/ Biendata) → Newsela/ Biendata

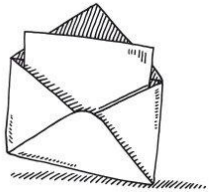
  - ***task adaptation + domain adaptation remains the best adaptation strategy***

  - higher benefit on OOD tasks and domains from Biendata

  - ***using the target dataset directly for intermediate adaptation overcomes the need for extra data*** (slightly lower but comparable results to relying on Wikipedia)

  - results remain consistent when using either MAML or Reptile for task adaptation

Main Takeaways



- Frame the task of text simplification from a task and domain adaptation perspective
- **Direct Adaptation:** fine-tuning pre-trained language models outperforms metalearning models for the task of low-resource text simplification
- **Decomposing the adaptation process into multiple steps**:
  a) *Intermediate dataset available:*
  Adapt to new tasks first, then continue adapting to new domains!

  b) *No intermediate dataset available:*
  build a pseudo-stop based on the target task/ domain itself
  Task + domain adaptation remains the best performing adaptation strategy!

  Coupling of task and domain adaptation is beneficial!