Overlap-based Vocabulary Generation Improves Cross-lingual Transfer Among Related Languages

Vaidehi Patil
Indian Institute of Technology Bombay

Partha Talukdar
Google Research, India

Sunita Sarawagi
Indian Institute of Technology Bombay
• Multilingual Models form the core of many NLP tasks

  E.g. theorem proving, solving reading comprehension
Multilingual Models: An overview

- Multilingual Models form the core of many NLP tasks
  
  E.g. theorem proving, solving reading comprehension

  ○ Labeled data for such tasks is scarce
Multilingual Models: An overview

- Multilingual Models form the core of many NLP tasks
  
  E.g. theorem proving, solving reading comprehension
  
  ○ Labeled data for such tasks is scarce

Top 100 languages with the largest Wikipedias
Multilingual Models: An overview

- Multilingual Models form the core of many NLP tasks
  
  E.g. theorem proving, solving reading comprehension

  - Labeled data for such tasks is scarce

- Multilingual models have been effective for cross-lingual transfer
  
  - when there is sufficient LRL unlabeled corpus (Wu and Dredze, 2020)
Multilingual Models: An overview

- Multilingual Models form the core of many NLP tasks
  - E.g. theorem proving, solving reading comprehension
    - Labeled data for such tasks is scarce

- Multilingual models have been effective for cross-lingual transfer
  - when there is sufficient LRL unlabeled corpus (Wu and Dredze, 2020)

If languages belong to the same family, what more can be done to improve cross-lingual transfer?
Multilingual Models: An overview

\[ \mathcal{D}_L \]

... English ...
Multilingual Models: An overview

\[ \mathcal{D}_L \]
- English
- German
- Dutch
- Frisian
Multilingual Models: An overview

Vocabulary generation (BPE)

\[ \mathcal{D}_L \]

English
German
Dutch
Frisian
Multilingual Models: An overview

Multilingual Model: (MLM pre-training)

Vocabulary generation (BPE)

D_L
English

D_L
German

D_L
Dutch

D_L
Frisian
Multilingual Models: An overview

Task-specific labeled data: en

Multilingual Model: (MLM pre-training)

Vocabulary generation (BPE)

$\mathcal{D}_L$ English
$\mathcal{D}_L$ German
$\mathcal{D}_L$ Dutch
$\mathcal{D}_L$ Frisian
Multilingual Models: An overview

- Task-specific labeled data: en
- Cross-lingual transfer to de, nl, fy
- Shared representations

Multilingual Model: (MLM pre-training)

- Vocabulary generation (BPE)

- \( \mathcal{D}_L \) English
- \( \mathcal{D}_L \) German
- \( \mathcal{D}_L \) Dutch
- \( \mathcal{D}_L \) Frisian
Multilingual Vocabulary: An overview

- Vocabulary is generated using
  - Byte Pair Encoding (BPE) (Sennrich et al., 2016).
  - WordPiece (Schuster and Nakajima, 2012; Wu et al., 2016)
  - SentencePiece (Kudo and Richardson, 2018)
Multilingual Vocabulary: An overview

- Vocabulary is generated using
  - Byte Pair Encoding (BPE) (Sennrich et al., 2016).
  - WordPiece (Schuster and Nakajima, 2012; Wu et al., 2016)
  - SentencePiece (Kudo and Richardson, 2018)

- Given a desired vocabulary size, these algorithms
  - Select an inventory of subwords that compactly represent the data
  - Look at subword frequencies in the combined multilingual corpus
Multilingual Vocabulary: An overview

- Vocabulary is generated using
  - Byte Pair Encoding (BPE) (Sennrich et al., 2016).
  - WordPiece (Schuster and Nakajima, 2012; Wu et al., 2016)
  - SentencePiece (Kudo and Richardson, 2018)

- Given a desired vocabulary size, these algorithms
  - Select an inventory of subwords that compactly represent the data
  - Look at subword frequencies in the combined multilingual corpus
    - Learn suboptimal decompositions for LRLs
Multilingual Vocabulary: An overview

- Vocabulary is generated using
  - Byte Pair Encoding (BPE) (Sennrich et al., 2016).
  - WordPiece (Schuster and Nakajima, 2012; Wu et al., 2016)
  - SentencePiece (Kudo and Richardson, 2018)

- Given a desired vocabulary size, these algorithms
  - Select an inventory of subwords that compactly represent the data
  - Look at subword frequencies in the combined multilingual corpus
    - Learn suboptimal decompositions for LRLs

If languages belong to the same family, what more can be done while generating vocabulary for supervision transfer from HRL to LRL?
Lexical Overlap

<table>
<thead>
<tr>
<th>Family</th>
<th>Hindi</th>
<th>Marathi</th>
<th>Punjabi</th>
<th>Gujarati</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indo-Aryan</td>
<td>Vaapariyo</td>
<td>Vaapartat</td>
<td>Vaaparan</td>
<td>Vaaparvana</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Family</th>
<th>English</th>
<th>German</th>
<th>Dutch</th>
<th>Western Frisian</th>
</tr>
</thead>
<tbody>
<tr>
<td>West-Germanic</td>
<td>Category</td>
<td>Kategorie</td>
<td>Categorie</td>
<td>Kategorie</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Family</th>
<th>French</th>
<th>Spanish</th>
<th>Portuguese</th>
<th>Italian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Romance</td>
<td>Association</td>
<td>Associacion</td>
<td>Associacao</td>
<td>Associazione</td>
</tr>
</tbody>
</table>

Lexically overlapping tokens with similar meanings across four languages in each of three families
Lexical Overlap

<table>
<thead>
<tr>
<th>Language Family</th>
<th>Hindi: Vaapariyo, Marathi: Vaapartat, Punjabi: Vaaparan, Gujarati: Vaaparvana</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indo-Aryan</td>
<td>English: Category, German: Kategorie, Dutch: Categorie, Western Frisian: Kategory</td>
</tr>
</tbody>
</table>

Lexically overlapping tokens with similar meanings across four languages in each of three families

How can relatedness help improve cross-lingual transfer?
Main takeaways

- Oversampling is less effective than exploiting token-overlap zero-shot transfer in related languages setting
Main takeaways

- Oversampling less effective than exploiting token-overlap zero-shot transfer in related languages setting
- Token overlap matters (unlike K et al., 2020) under two settings:
  - Languages are sufficiently related
Main takeaways

- Oversampling less effective than exploiting token-overlap zero-shot transfer in related languages setting
- Token overlap matters (unlike K et al., 2020) under two settings:
  - Languages are sufficiently related
  - LRL is resource-poor even in the amount of unlabeled data
PROBLEM: Resource Scarcity
PROBLEM: Resource Scarcity

OPPORTUNITY: Relatedness between Languages
PROBLEM: Resource Scarcity

OPPORTUNITY: Relatedness between Languages

○ Many languages are part of the same language family (e.g. Indo-Aryan, Germanic, Romance etc)
PROBLEM: Resource Scarcity

OPPORTUNITY: Relatedness between Languages

- Many languages are part of the same language family (e.g. Indo-Aryan, Germanic, Romance etc)
- Languages of same family have lexically overlapping words with similar meanings even when languages are of different scripts
PROBLEM: Resource Scarcity

OPPORTUNITY: Relatedness between Languages

- Many languages are part of the same language family (e.g. Indo-Aryan, Germanic, Romance etc.)
- Languages of same family have lexically overlapping words with similar meanings even when languages are of different scripts

Can we take advantage of this relatedness to overcome the barriers of resource scarcity?
Example of BPE in action

Related Languages
Example of BPE in action
Example of BPE in action

Related Languages

English
German
Dutch
Frisian

words
Example of BPE in action

Related Languages
- English
- German
- Dutch
- Frisian

words
- University
- versity
- Universitaten
- Universiteit
- Universiteiten
Example of BPE in action

Related Languages
- English
- German
- Dutch
- Frisian

Frequency

words
- University
- versity
- Universitaten
- Universiteit
- Universiteiten
Example of BPE in action

- English: University (10)
- German: Universitäten (6)
- Dutch: Universiteit (2)
- Frisian: Universiteiten (2)

Related Languages:
- English
- German
- Dutch
- Frisian
Example of BPE in action

Related Languages
- English
- German
- Dutch
- Frisian

Frequency
- 10: University
- 6: versity
- 2: Universitaten
- 1: Universiteit
- 1: Universiteiten

Starting vocab
Example of BPE in action

- Related Languages:
  - English
  - German
  - Dutch
  - Frisian

- Frequency:
  - University: 10
  - versity: 6
  - Universitaten: 2
  - Universiteit: 1
  - Universiteiten: 1

- Starting vocab:
  - Uni
  - versit
  - y</w>
Example of BPE in action

Related Languages:
- English
- German
- Dutch
- Frisian

Frequency:
- University: 10
- versity: 6
- Universitaten: 2
- Universiteit: 1
- Universiteiten: 1

Starting vocab:
- Uni
- versit

Candidate Vocabulary tokens:
- Universit
- versity</w>

Example of BPE in action:
Example of BPE in action

Related Languages
- English
- German
- Dutch
- Frisian

Frequency
- 10
- 6
- 2
- 1

Starting vocab
- University
- versity
- Universitaten
- Universiteit
- Universiteiten

Candidate Vocabulary tokens
- Uni
- versit
- versity</w>
- Universit
- versity</w>
Example of BPE in action

Related Languages
- English
- German
- Dutch
- Frisian

Frequency
- 10
- 6
- 2
- 1
- 1

words
- University
- versity
- Universitaten
- Universiteit
- Universiteiten

Starting vocab
- Uni
- versit
- y</w>
- versity</w>

Candidate Vocabulary tokens
- Universit
- versity</w>

BPE cost function
10+2+1+1
Example of BPE in action

<table>
<thead>
<tr>
<th>Related Languages</th>
<th>Frequency</th>
<th>words</th>
<th>Starting vocab</th>
<th>Candidate Vocabulary tokens</th>
<th>BPE cost function</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>10</td>
<td>University</td>
<td>Uni</td>
<td>Universit</td>
<td>10+2+1+1</td>
</tr>
<tr>
<td>German</td>
<td>6</td>
<td>versity</td>
<td>versit</td>
<td>versity&lt;/w&gt;</td>
<td>10+6</td>
</tr>
<tr>
<td>Dutch</td>
<td>2</td>
<td>Universitat</td>
<td>Universit</td>
<td>versity&lt;/w&gt;</td>
<td></td>
</tr>
<tr>
<td>Frisian</td>
<td>1</td>
<td>Universiteiten</td>
<td>y&lt;/w&gt;</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example of BPE in action

Related Languages

English

German

Dutch

Frisian

Frequency

10

6

2

1

1

words

University

versity

Universitaten

Universiteit

Universiteiten

Starting vocab

Uni

versit

versity</w>

Candidate Vocabulary tokens

Universit

BPE cost function

10+2+1+1

10+6
OBPE

- OBPE
  - prefers vocabulary units which are shared across multiple languages
  - encodes the input corpora compactly
OBPE

- OBPE
  - prefers vocabulary units which are shared across multiple languages
  - encodes the input corpora compactly

- The objective function that governs the candidate token to be added to the vocabulary at every iteration comprises of two terms:
OBPE

- OBPE
  - prefers vocabulary units which are shared across multiple languages
  - encodes the input corpora compactly
- The objective function that governs the candidate token to be added to the vocabulary at every iteration comprises of two terms:
  - The first term compactly represents the total corpus, as in BPE
OBPE

- OBPE
  - prefers vocabulary units which are shared across multiple languages
  - encodes the input corpora compactly
- The objective function that governs the candidate token to be added to the vocabulary at every iteration comprises of two terms:
  - The first term compactly represents the total corpus, as in BPE
  - The second term additionally biases towards vocabulary with greater overlap of each LRL to one HRL
OBPE

- OBPE quantifies overlap between two languages’ encoding as a generalized mean function

\[
\text{overlap}(L_i, L_h, S) = \sum_{k \in S} \left( \frac{f_{ki}^p + f_{kh}^p}{2} \right)^{\frac{1}{p}}, \quad p \leq 1
\]
OBPE quantifies overlap between two languages’ encoding as a generalized mean function:

$$\text{overlap}(L_i, L_h, S) = \sum_{k \in S} \left( \frac{f_{ki}^p + f_{kh}^p}{2} \right)^{\frac{1}{p}}, \quad p \leq 1$$

The greedy version of the objective that controls the candidate vocabulary item to be inducted in each iteration of OBPE:

$$\mathcal{V} = \mathcal{V} \cup \arg\max_{k=[u,v]: u, v \in \mathcal{V}} (1 - \alpha) \sum_j f_{kj} + \alpha \sum_{i \in \mathcal{L}_{LRL}} \max_{h \in \mathcal{L}_{HRL}} \left( \frac{f_{ki}^p + f_{kh}^p}{2} \right)^{\frac{1}{p}}$$
Example of OBPE in action

- **Related Languages**
  - English
  - German
  - Dutch
  - Frisian

- **Frequency**
  - 10
  - 6
  - 2
  - 1

- **words**
  - University
  - versity
  - Universitaten
  - Universiteit
  - Universiteiten

- **Starting vocab**
  - Uni
  - versit
  - y</w>

- **Candidate Vocabulary tokens**
  - Universit
  - versity</w>
Example of OBPE in action

- **Related Languages**: English, German, Dutch, Frisian
- **Frequency**:
  - University: 10
  - versity: 6
  - Universitaten: 2
  - Universiteit: 1
  - Universiteiten: 1
- **Starting vocab**: Uni, Universit
- **Candidate Vocabulary tokens**: Universit, versit, versity<w/>
- **OBPE cost function**
Example of OBPE in action

**Related Languages**
- English
- German
- Dutch
- Frisian

**Frequency**
- 10

**Words**
- University
- versity
- Universitaten
- Universiteit
- Universiteiten

**Starting vocab**
- Uni
- versit
- y</w>

**Candidate Vocabulary tokens**
- Universit
- versity</w>

**OBPE cost function**
\[ 0.5(10+2+1+1) + \]
Example of OBPE in action

- **Related Languages**
  - English
  - German
  - Dutch
  - Frisian

- **Frequency**
  - 10
  - 6
  - 2
  - 1

- **Starting vocab**
  - University
  - versity
  - Universitaten
  - Universiteit
  - Universiteiten

- **Candidate Vocabulary tokens**
  - Uni
  - versit
  - versity</w>
  - Universit

- **OBPE cost function**
  - 0.5(10+2+1+1) + 0.5(2+1)
Example of OBPE in action

Related Languages

- English
- German
- Dutch
- Frisian

Frequency

- University: 10
- versity: 6
- Universitaten: 2
- Universiteit: 1
- Universiteiten: 1

words

- University
- versity
- Universitaten
- Universiteit
- Universiteiten

Starting vocab

- Uni
- versit
- y</w>

Candidate Vocabulary tokens

- Universit
- versity</w>

OBPE cost function

0.5(10+2+1+1) + 0.5(2+1)
Example of OBPE in action

**Related Languages**
- English
- German
- Dutch
- Frisian

**Frequency**
- 10
- 6
- 2
- 1

**words**
- University
- versity
- Universitaten
- Universiteit
- Universiteiten

**Starting vocab**
- Uni
- versit
- y</w>

**Candidate Vocabulary tokens**
- Universit
- versity</w>

**OBPE cost function**

$$0.5(10+2+1+1) + 0.5(2+1+1)$$
Example of OBPE in action

**Related Languages**
- English
- German
- Dutch
- Frisian

**Frequency**
- 10
- 6
- 2
- 1

**Words**
- University
- versity
- Universitaten
- Universiteit
- Universiteiten

**Starting vocab**
- Uni
- versit
- university

**Candidate Vocabulary tokens**
- Universit
- versity</w>

**OBPE cost function**
- $0.5(10+2+1+1) + 0.5(2+1+1)$
- $0.5(10+6)$
Example of OBPE in action

- **Related Languages**
  - English
  - German
  - Dutch
  - Frisian

- **Frequency**
  - 10 for **University**
  - 6 for **versity**
  - 2 for **Universitaten**
  - 1 for **Universiteit**
  - 1 for **Universiteiten**

- **Starting vocab**
  - **Uni**
  - **versit**
  - **versity</w>**
  - **y</w>**

- **Candidate Vocabulary tokens**
  - Universit

- **OBPE cost function**
  - $0.5(10+2+1+1) + 0.5(2+1+1) = 0.5(10+6)$
Twelve Languages simulated as HRLs and LRLs across with two different corpus distribution: Balanced and Skewed
Number of documents in languages simulated as LRLs is 20K

<table>
<thead>
<tr>
<th>Family</th>
<th>HRL</th>
<th>LRLs</th>
<th>Number of HRL Docs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Balanced</td>
<td>Skewed</td>
</tr>
<tr>
<td>West Germanic</td>
<td>English (en)</td>
<td>German (de), Dutch (nl), Western Frisian (fy)</td>
<td>0.16M</td>
</tr>
<tr>
<td>Romance</td>
<td>French (fr)</td>
<td>Spanish (es), Portuguese (pt), Italian (it)</td>
<td>0.16M</td>
</tr>
<tr>
<td>Indo-Aryan</td>
<td>Hindi (hi)</td>
<td>Marathi (mr), Punjabi (pa), Gujarati (gu)</td>
<td>0.16M</td>
</tr>
</tbody>
</table>
### Question

Is OBPE more effective than BPE for zero-shot transfer?

### Balanced setting

<table>
<thead>
<tr>
<th>Method</th>
<th>LRL Performance(↑)</th>
<th>HRL Performance(↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NER</td>
<td>TC</td>
</tr>
<tr>
<td>BPE</td>
<td>64.48</td>
<td>65.52</td>
</tr>
<tr>
<td>BPE-dp</td>
<td>63.92</td>
<td>64.15</td>
</tr>
<tr>
<td>CV</td>
<td>59.58</td>
<td>61.91</td>
</tr>
<tr>
<td>TokComp</td>
<td>63.79</td>
<td>65.77</td>
</tr>
<tr>
<td>OBPE</td>
<td><strong>65.72</strong></td>
<td><strong>68.02</strong></td>
</tr>
</tbody>
</table>

Zero-shot LRL accuracy improves compared to the baselines across all four tasks.
Is OBPE more effective than BPE for zeroshot transfer?

<table>
<thead>
<tr>
<th>Method</th>
<th>LRL Performance(↑)</th>
<th>HRL Performance(↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NER</td>
<td>TC</td>
</tr>
<tr>
<td>BPE</td>
<td>52.91</td>
<td>51.68</td>
</tr>
<tr>
<td>CV</td>
<td>52.73</td>
<td>54.40</td>
</tr>
<tr>
<td>OBPE</td>
<td><strong>55.09</strong></td>
<td><strong>55.37</strong></td>
</tr>
</tbody>
</table>

Zero-shot LRL accuracy improves compared to the baselines across all four tasks.
Is OBPE more effective than BPE for zero-shot transfer?

<table>
<thead>
<tr>
<th>Method</th>
<th>LRL Performance(↑)</th>
<th>HRL Performance(↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NER</td>
<td>TC</td>
</tr>
<tr>
<td>BPE</td>
<td>64.5</td>
<td>65.5</td>
</tr>
<tr>
<td>+overSample</td>
<td>64.4</td>
<td>67.6</td>
</tr>
<tr>
<td>OBPE</td>
<td>65.7</td>
<td>68.0</td>
</tr>
<tr>
<td>+overSample</td>
<td>64.6</td>
<td>67.9</td>
</tr>
</tbody>
</table>

Even though BPE_overSamp improves LRL performance somewhat, it causes HRL performance to drop.

Balanced setting
Is OBPE more effective than BPE for zeroshot transfer?

### Balanced setting

<table>
<thead>
<tr>
<th>Method</th>
<th>LRL Performance (↑)</th>
<th>HRL Performance (↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NER</td>
<td>TC</td>
</tr>
<tr>
<td>BPE</td>
<td>64.5</td>
<td>65.5</td>
</tr>
<tr>
<td>+overSample</td>
<td>64.4</td>
<td>67.6</td>
</tr>
<tr>
<td>OBPE</td>
<td>65.7</td>
<td>68.0</td>
</tr>
<tr>
<td>+overSample</td>
<td>64.6</td>
<td>67.9</td>
</tr>
</tbody>
</table>

OBPE with default sampling is best for both LRLs and HRLs.
<table>
<thead>
<tr>
<th>Method</th>
<th>LRL Performance(↑)</th>
<th>HRL Performance(↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NER</td>
<td>TC</td>
</tr>
<tr>
<td>BPE</td>
<td>64.5</td>
<td>65.5</td>
</tr>
<tr>
<td>+overSample</td>
<td>64.4</td>
<td>67.6</td>
</tr>
<tr>
<td>OBPE</td>
<td>65.7</td>
<td>68.0</td>
</tr>
<tr>
<td>+overSample</td>
<td>64.6</td>
<td>67.9</td>
</tr>
</tbody>
</table>

Is OBPE more effective than BPE for zeroshot transfer?

Balanced setting

OBPE_overSampled is better than BPE_overSampled
How does increased LRL representation in the vocabulary impact accuracy?
How does increased LRL representation in the vocabulary impact accuracy?

OBPE increases LRL representation by favoring overlapping tokens.
How does increased LRL representation in the vocabulary impact accuracy?

OBPE increases fraction of tokens shared across related languages, Oversampling decreases fraction of shared tokens.
How does increased LRL representation in the vocabulary impact accuracy?

OBPE increases HRL representation. Oversampling decreases HRL representation.
What is the effect of token overlap on overall accuracy?

Increased gains in LRL accuracy as we go from no overlap to full overlap on all three tasks.
What is the effect of token overlap on overall accuracy?

<table>
<thead>
<tr>
<th>Task</th>
<th>More related</th>
<th>Less related</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High (hi: 110K)</td>
<td>Low (mr: 20K)</td>
</tr>
<tr>
<td>NER</td>
<td>-12.2</td>
<td>-41.6</td>
</tr>
<tr>
<td>TC</td>
<td>-2.7</td>
<td>-41.3</td>
</tr>
<tr>
<td>POS</td>
<td>-6.6</td>
<td>-7.8</td>
</tr>
</tbody>
</table>

Drop in Accuracy of Zero-shot transfer when we synthetically reduce token overlap to zero
What is the effect of token overlap on overall accuracy?

<table>
<thead>
<tr>
<th>Task</th>
<th>High (hi: 110K)</th>
<th>Low (mr: 20K)</th>
<th>Task</th>
<th>High (en: 1GB)</th>
<th>Low (es: 20K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NER</td>
<td>-12.2</td>
<td>-41.6</td>
<td>NER</td>
<td>-1.4</td>
<td>-11.7</td>
</tr>
<tr>
<td>TC</td>
<td>-2.7</td>
<td>-41.3</td>
<td>XNLI</td>
<td>-1.3</td>
<td>-1.3</td>
</tr>
<tr>
<td>POS</td>
<td>-6.6</td>
<td>-7.8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Token overlap is important for related languages and its benefit is higher in the low resource setting.
Conclusion

- OBPE exploits language relatedness along lexical overlap
Conclusion

- OBPE exploits language relatedness along lexical overlap
- OBPE vocabulary maximizes overlap across related languages to create more pathways for cross-lingual supervision transfer
OBPE exploits language relatedness along lexical overlap

OBPE vocabulary maximizes overlap across related languages to create more pathways for cross-lingual supervision transfer

Exploiting language relatedness results in an overall more effective vocabulary compared to oversampling
Conclusion

- OBPE exploits language relatedness along lexical overlap

- OBPE vocabulary maximizes overlap across related languages to create more pathways for cross-lingual supervision transfer

- Exploiting language relatedness results in an overall more effective vocabulary compared to oversampling

- Token overlap is important in a low resource, related-language setting
Thank You

Github : [https://github.com/Vaidehi99/OBPE](https://github.com/Vaidehi99/OBPE)
Contact : vaidehipatil16@gmail.com, partha@google.com, sunita@iitb.ac.in