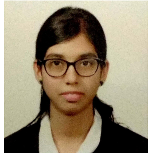
Overlap-based Vocabulary Generation Improves Cross-lingual Transfer Among Related Languages



Vaidehi Patil

Indian Institute of Technology Bombay

Partha Talukdar

Google Research, India

Sunita Sarawagi

Indian Institute of Technology Bombay

Multilingual Models: An overview

- 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

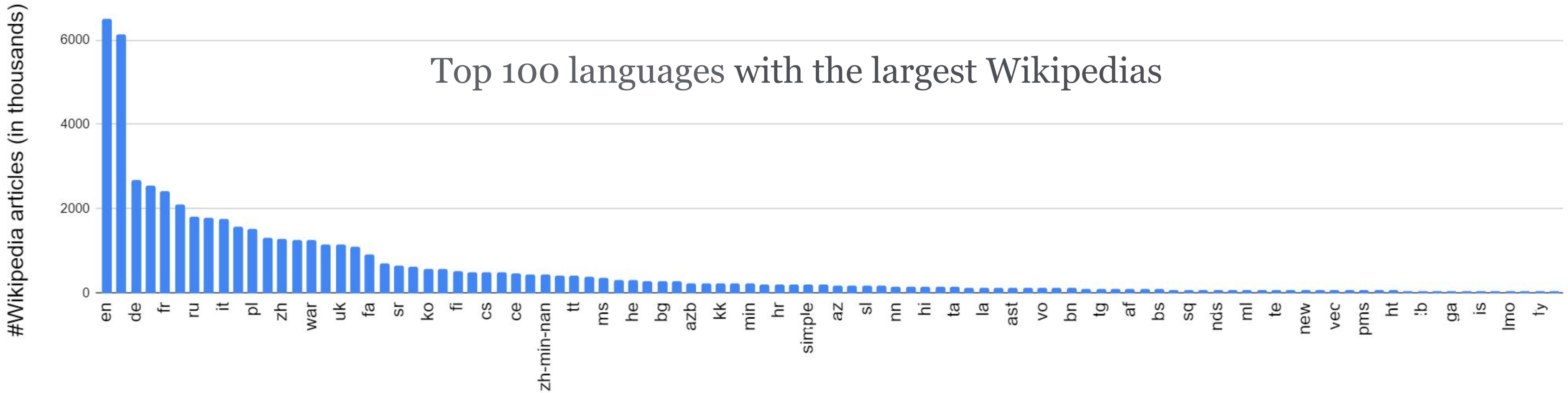
    - Labeled data for such tasks is scarce

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

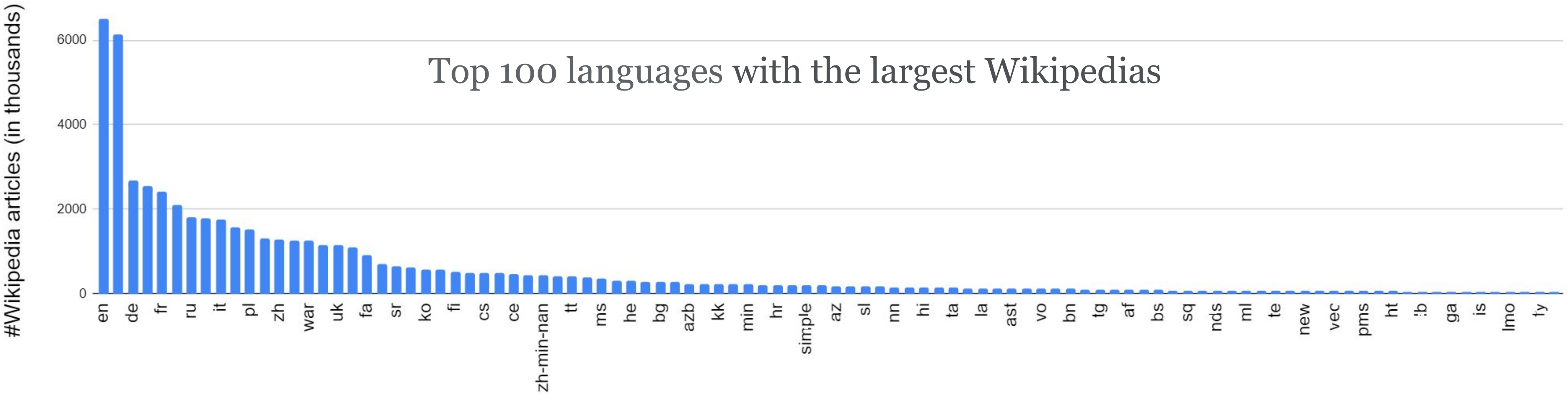


Multilingual Models: An overview

- 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)

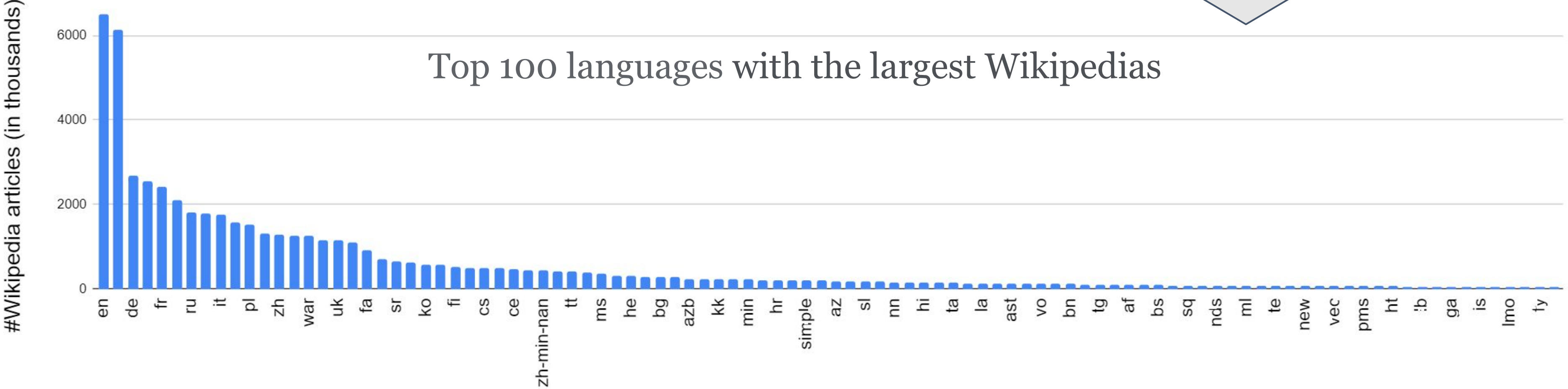
Multilingual Models: An overview

- Multilingual Models form the core of many NLP tasks

    E.g. theorem proving, solving reading comprehension

    - Labeled data for such tasks is scarce

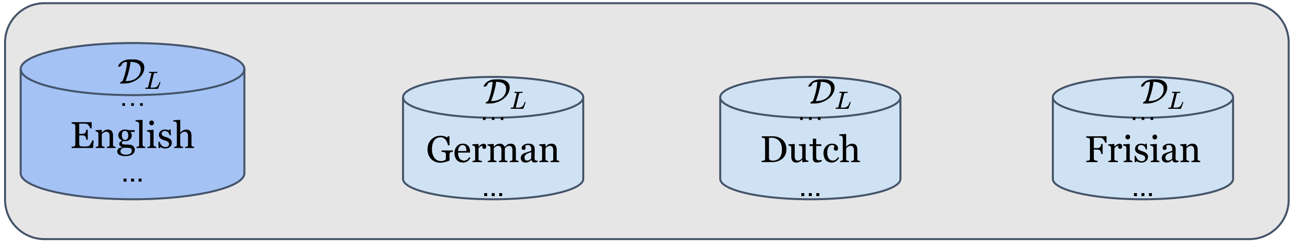
If languages belong to the same family, what more can be done to improve cross-lingual transfer?



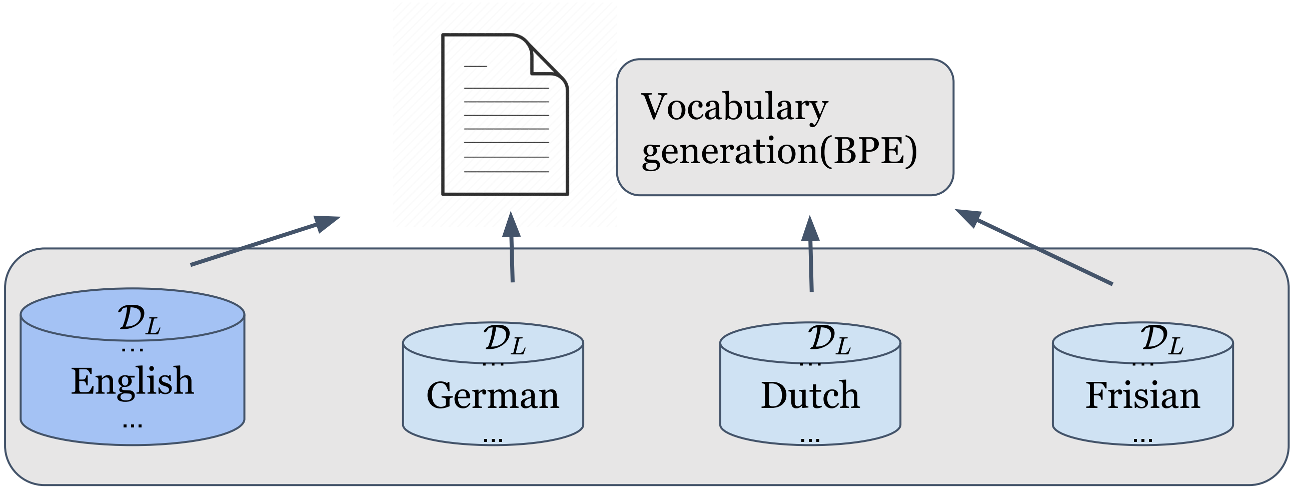
- Multilingual models have been effective for cross-lingual transfer
    - when there is sufficient LRL unlabeled corpus (Wu and Dredze, 2020)

Multilingual Models: An overview

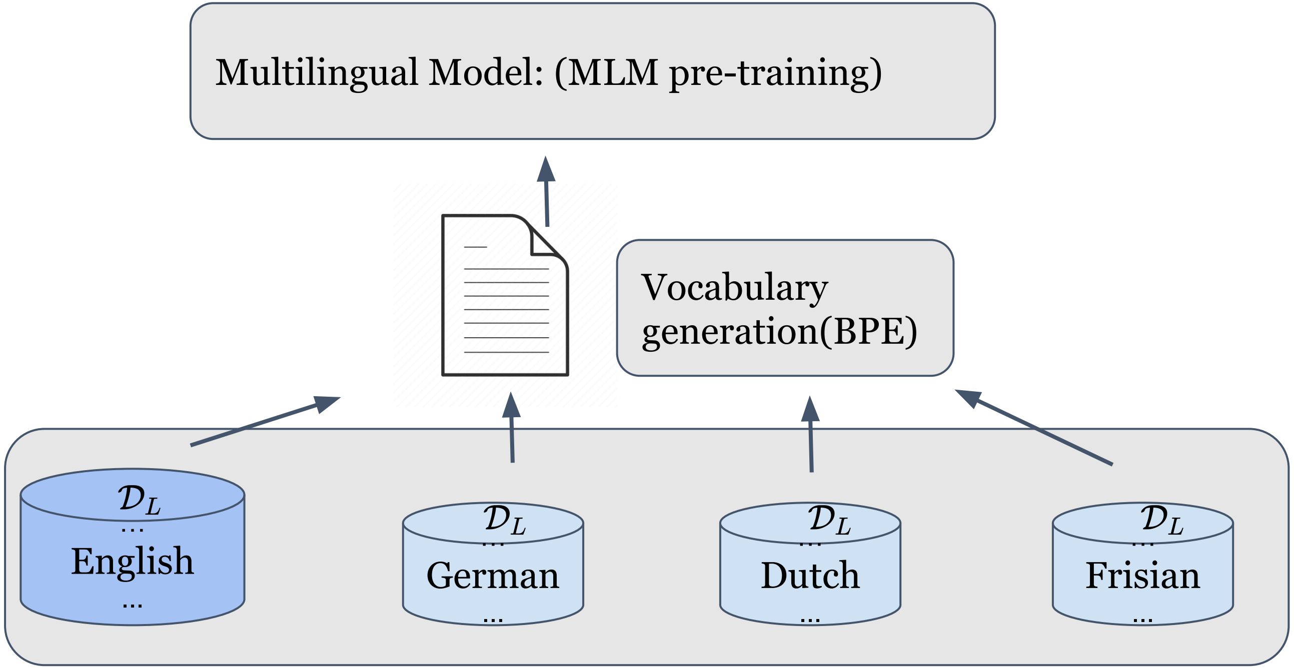
Multilingual Models: An overview



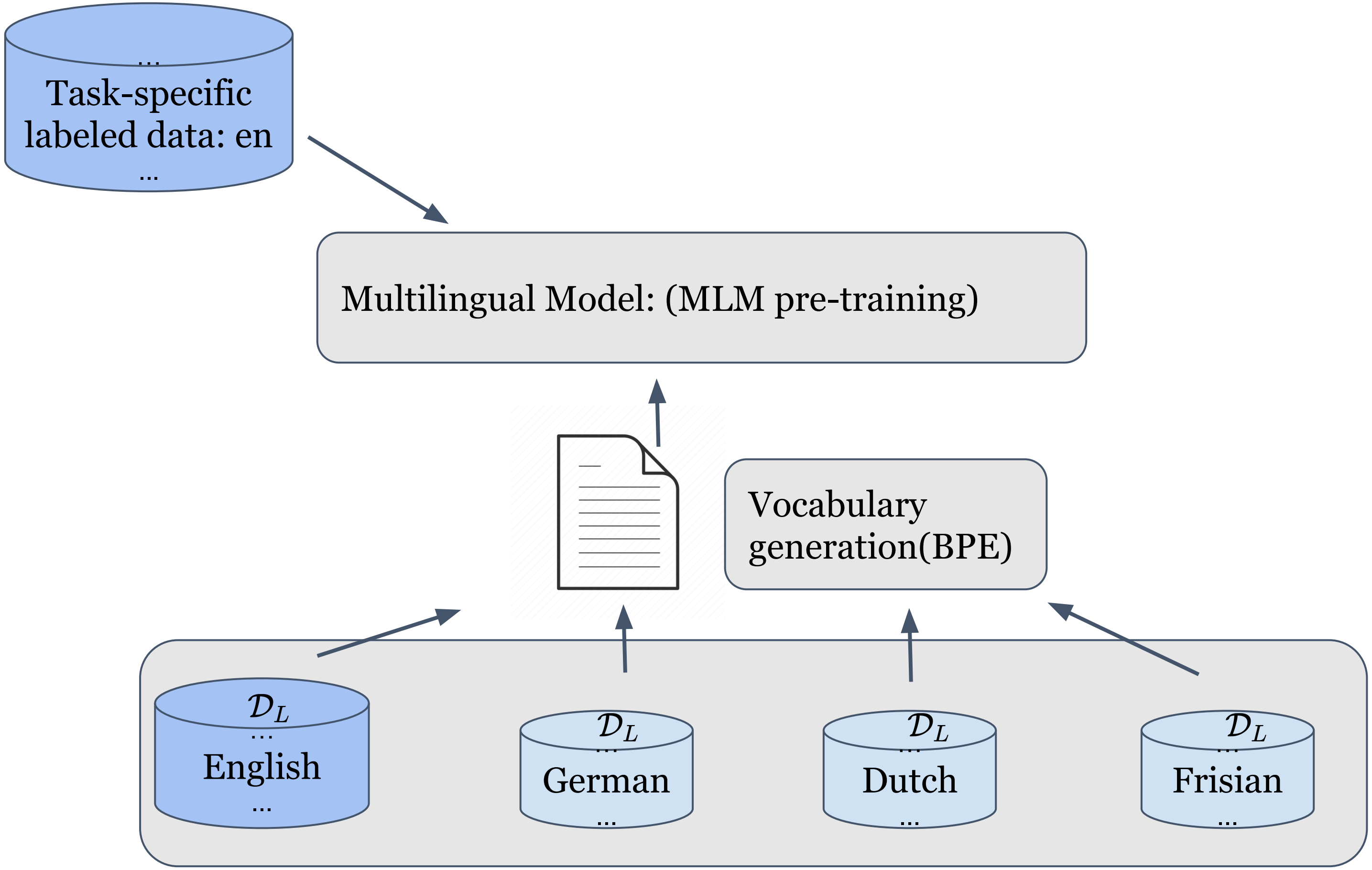
Multilingual Models: An overview



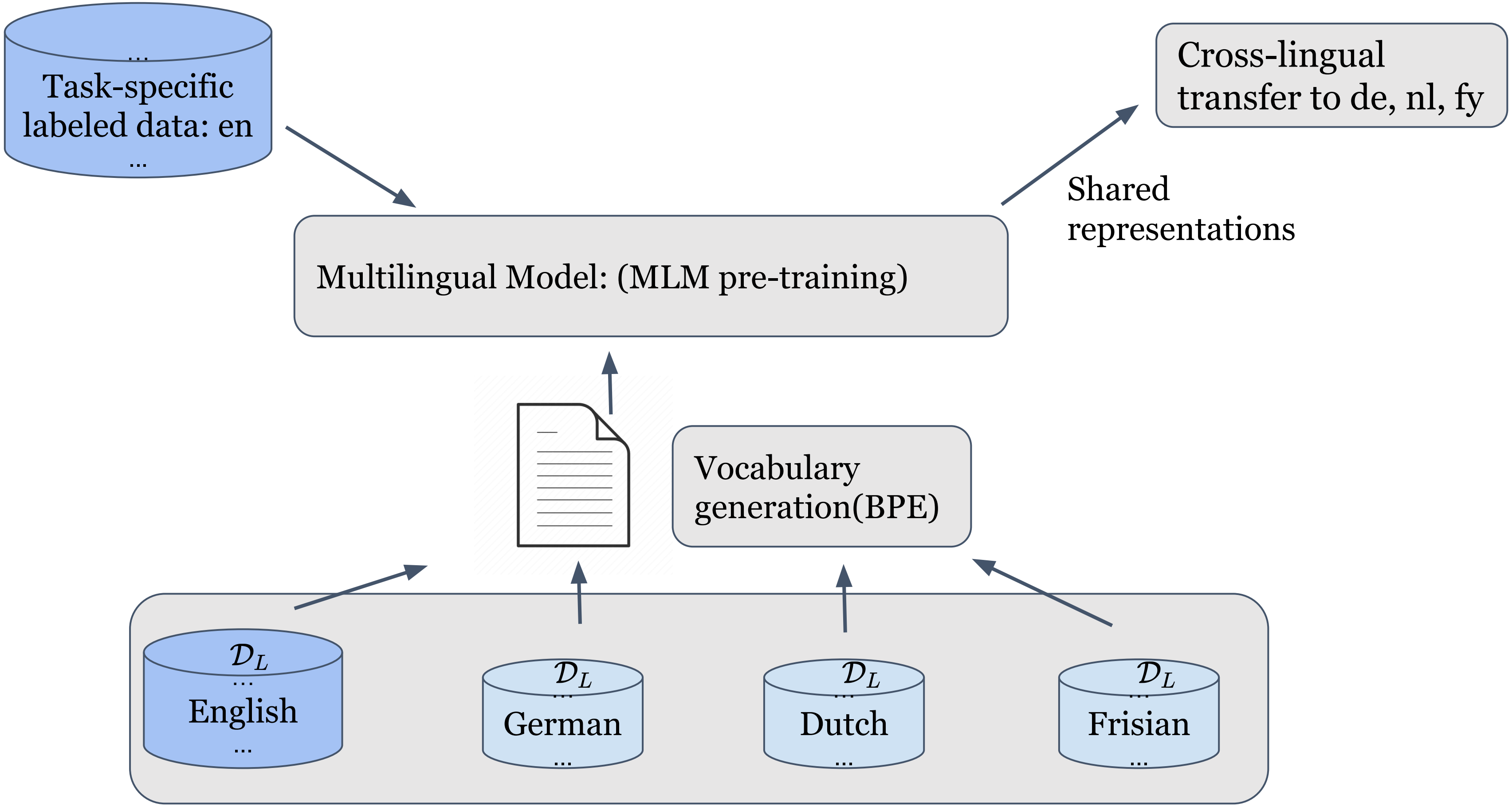
Multilingual Models: An overview



Multilingual Models: An overview



Multilingual Models: An overview



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

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

- 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

Lexical Overlap

Indo-Aryan	Hindi: Vaapariyo, Marathi: Vaapartat , Punjabi: Vaaparan, Gujarati: Vaaparvana
West-Germanic	English: Category, German:Kategorie, Dutch:Categorie, Western Frisian:Kategory
Romance	French: Association, Spanish: Associacion, Portuguese: Associacao, Italian: Associazione

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

Lexical Overlap

Indo-Aryan	Hindi: Vaapariyo, Marathi: Vaapartat , Punjabi: Vaaparan, Gujarati: Vaaparvana
West-Germanic	English: Category, German:Kategorie, Dutch:Categorie, Western Frisian:Kategory
Romance	French: Association, Spanish: Associacion, Portuguese: Associacao, Italian: Associazione

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

Relatedness for vocabulary generation

PROBLEM: Resource Scarcity

Relatedness for vocabulary generation

PROBLEM: Resource Scarcity

OPPORTUNITY: Relatedness between Languages

Relatedness for vocabulary generation

PROBLEM: Resource Scarcity

OPPORTUNITY: Relatedness between Languages

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

Relatedness for vocabulary generation

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

Relatedness for vocabulary generation

PROBLEM: Resource Scarcity

Can we take advantage of this *relatedness* to overcome the barriers of resource scarcity?

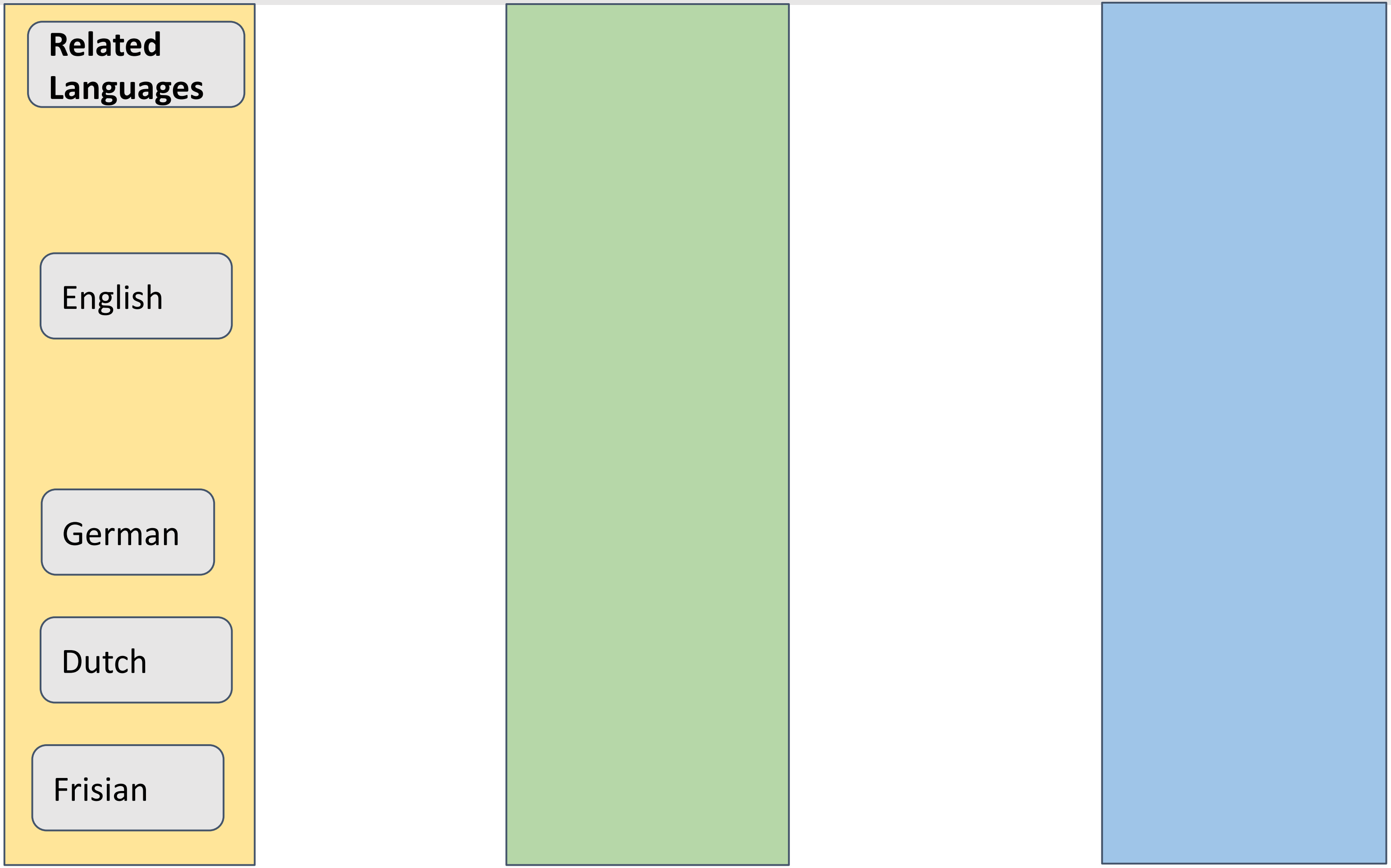
OPPORTUNITY: Relatedness between Languages

- 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

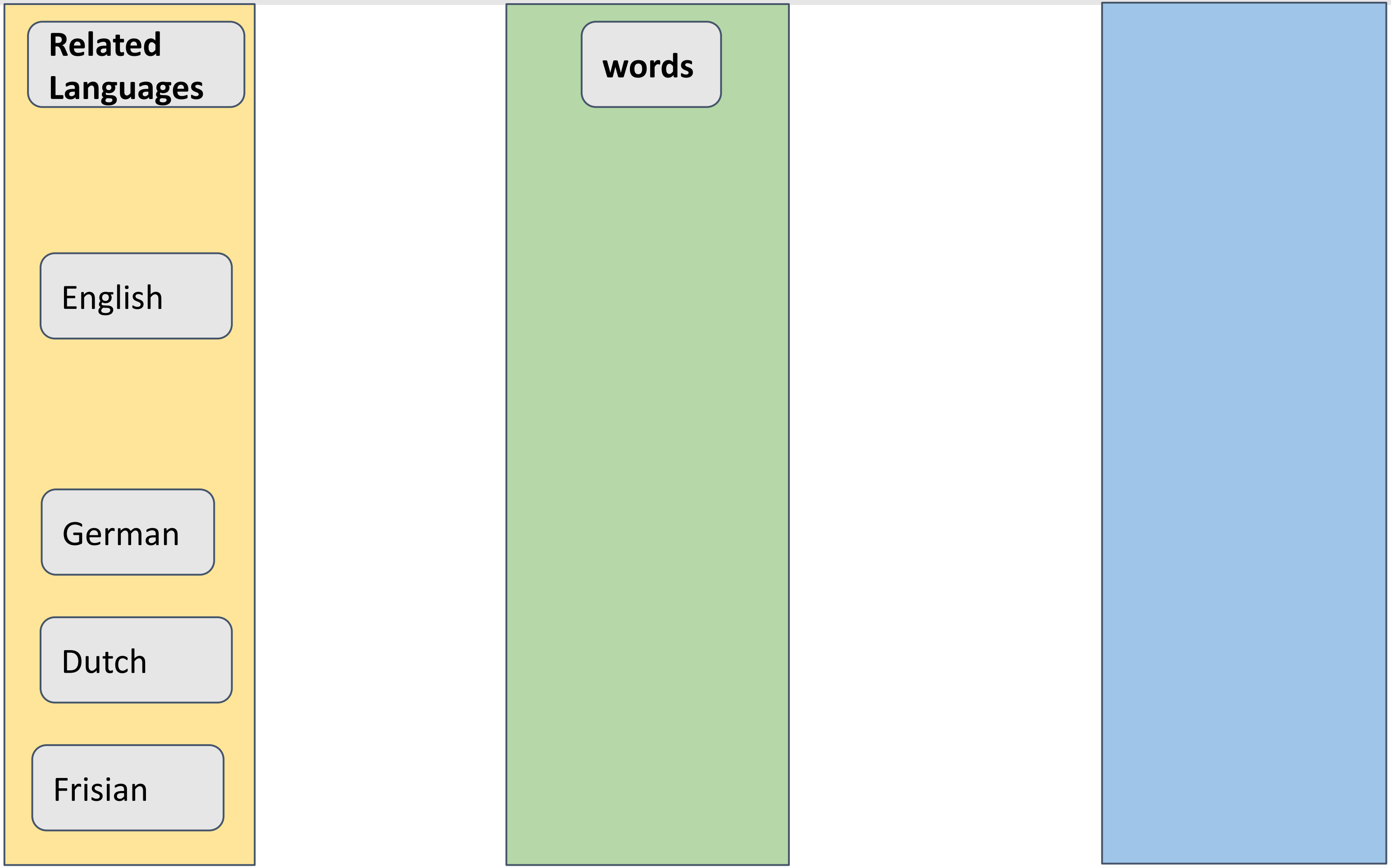
Example of BPE in action

Related Languages

Example of BPE in action



Example of BPE in action



Example of BPE in action

Related Languages

English

German

Dutch

Frisian

words

University

versity

Universitaten

Universiteit

Universiteiten

Example of BPE in action

Related Languages

Frequency

words

University

English

versity

German

Universitaten

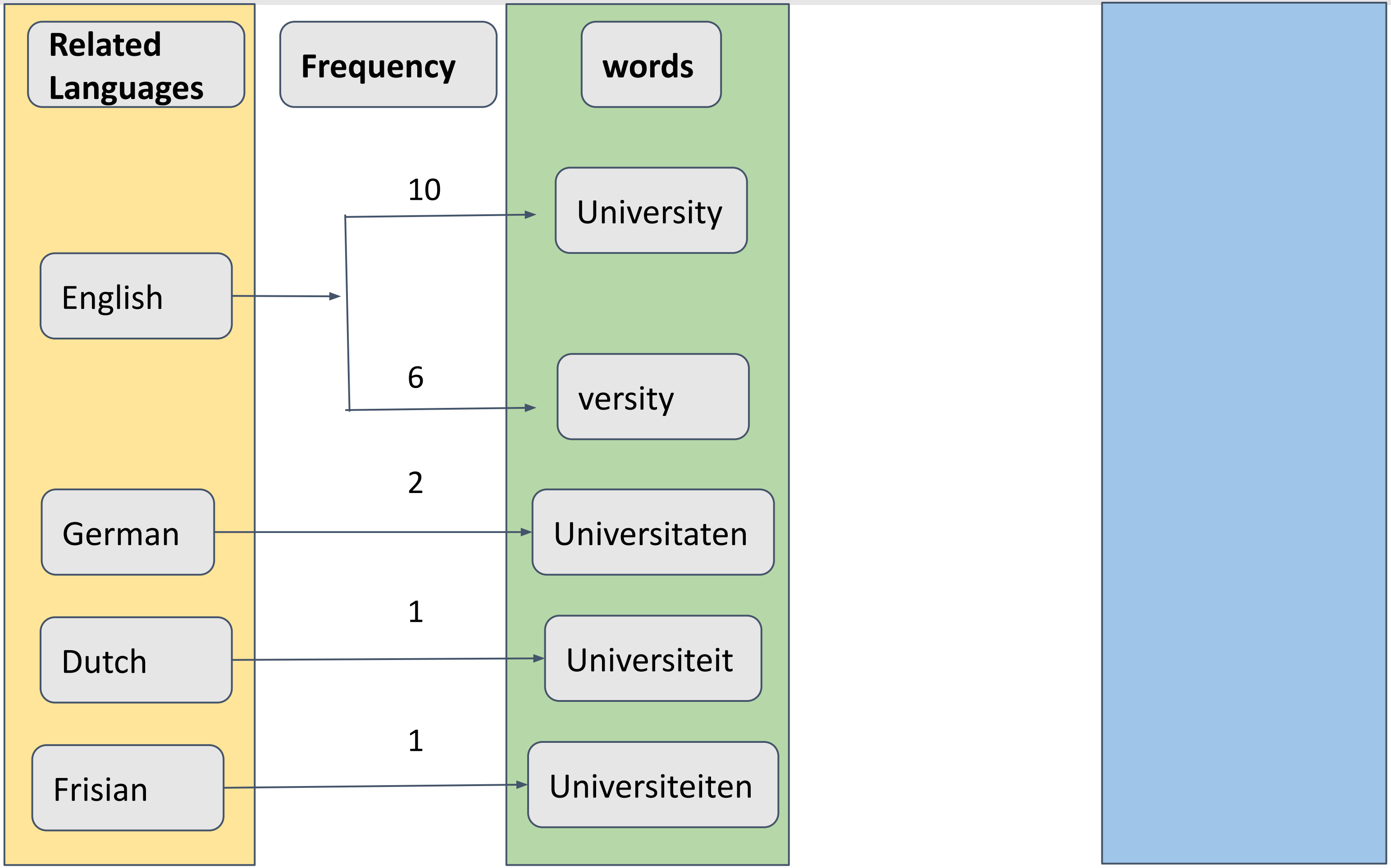
Dutch

Universiteit

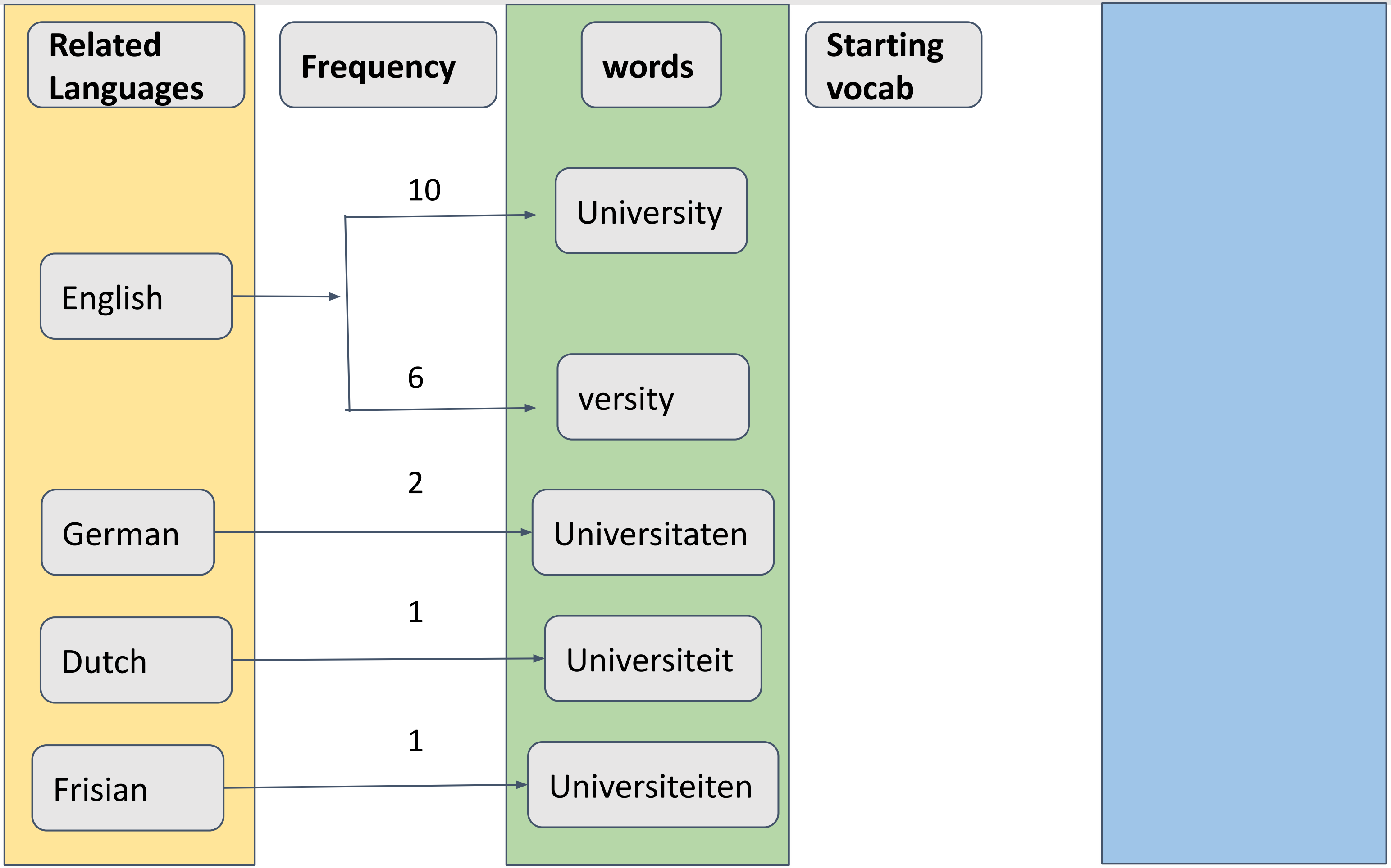
Frisian

Universiteiten

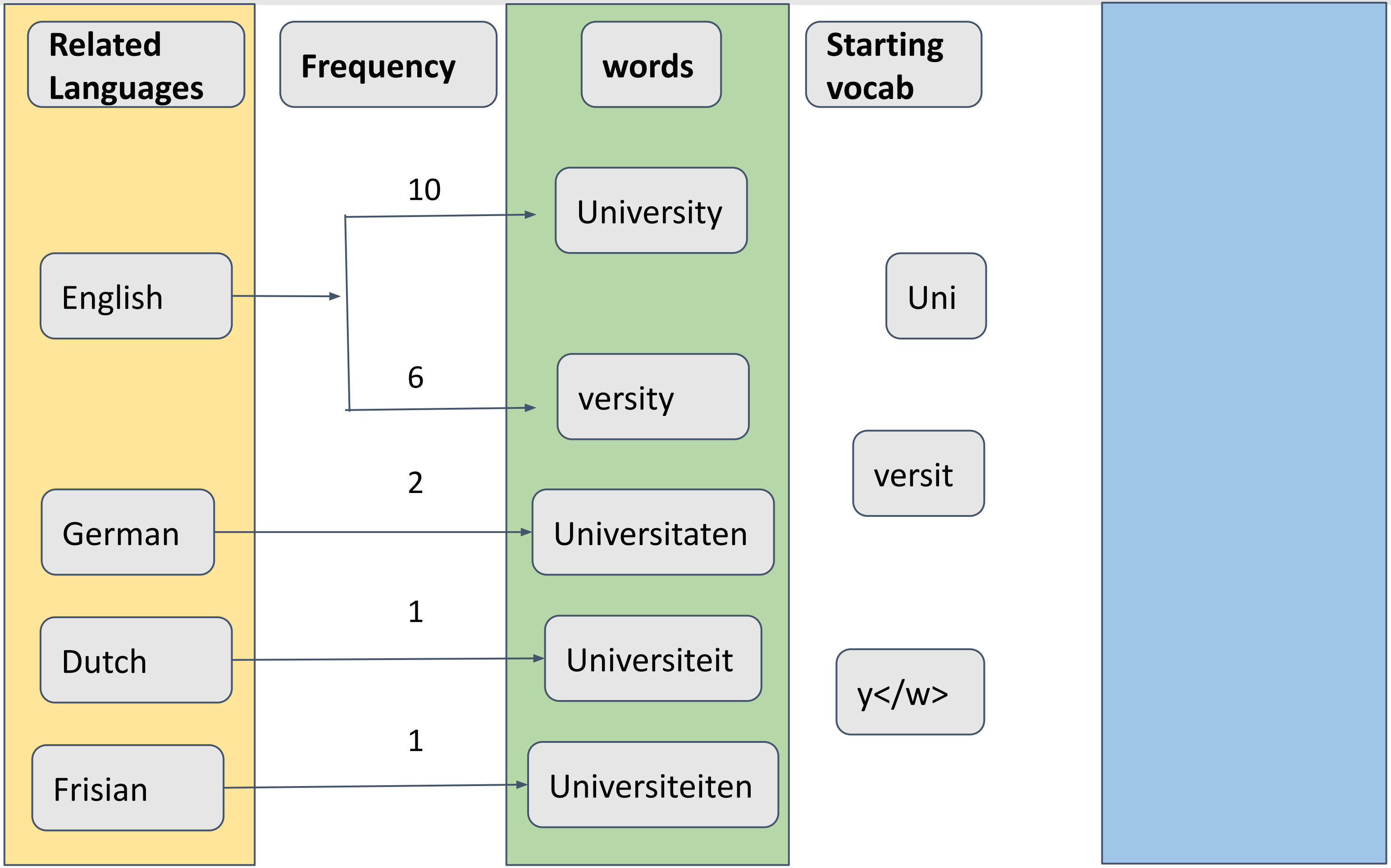
Example of BPE in action



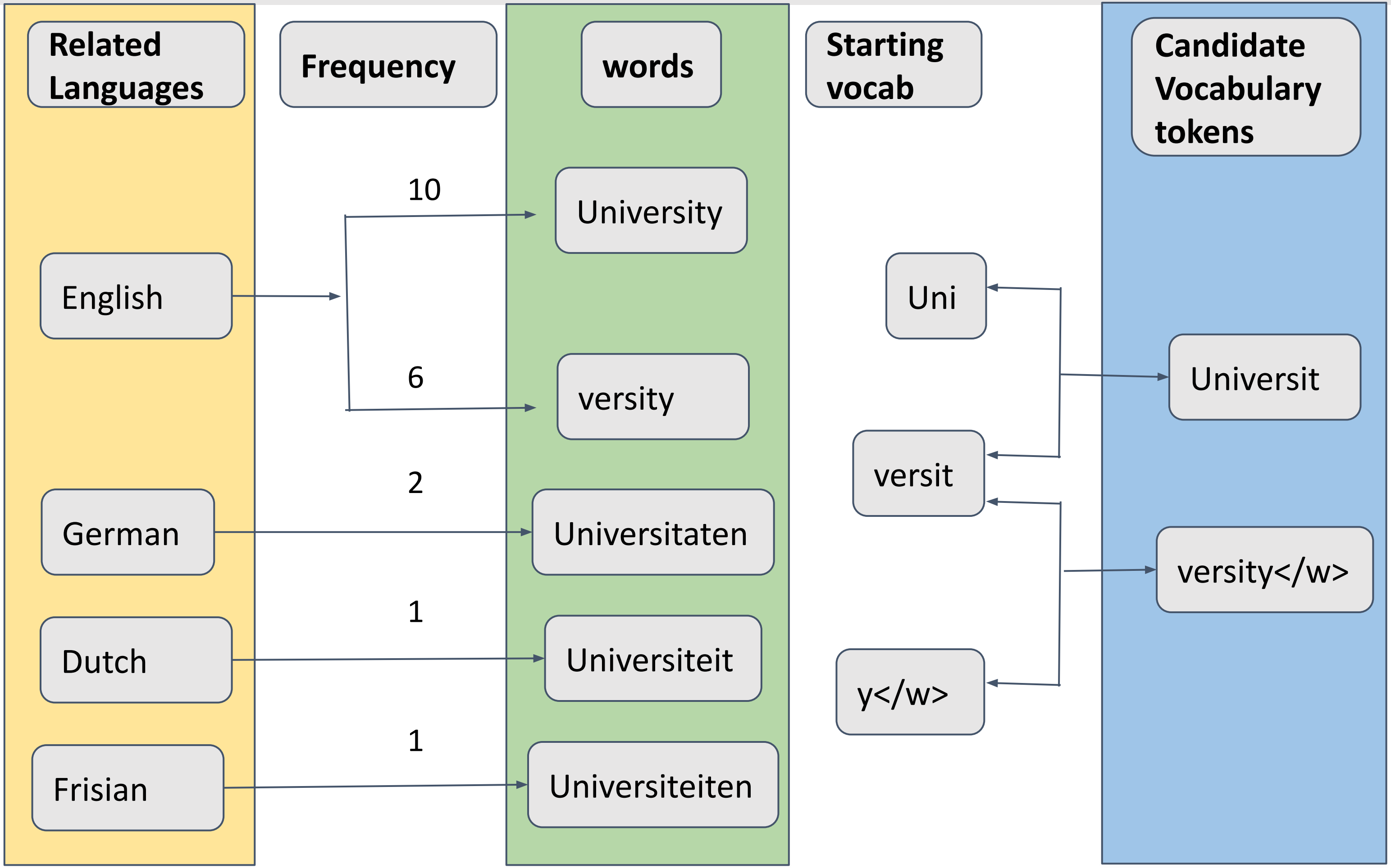
Example of BPE in action



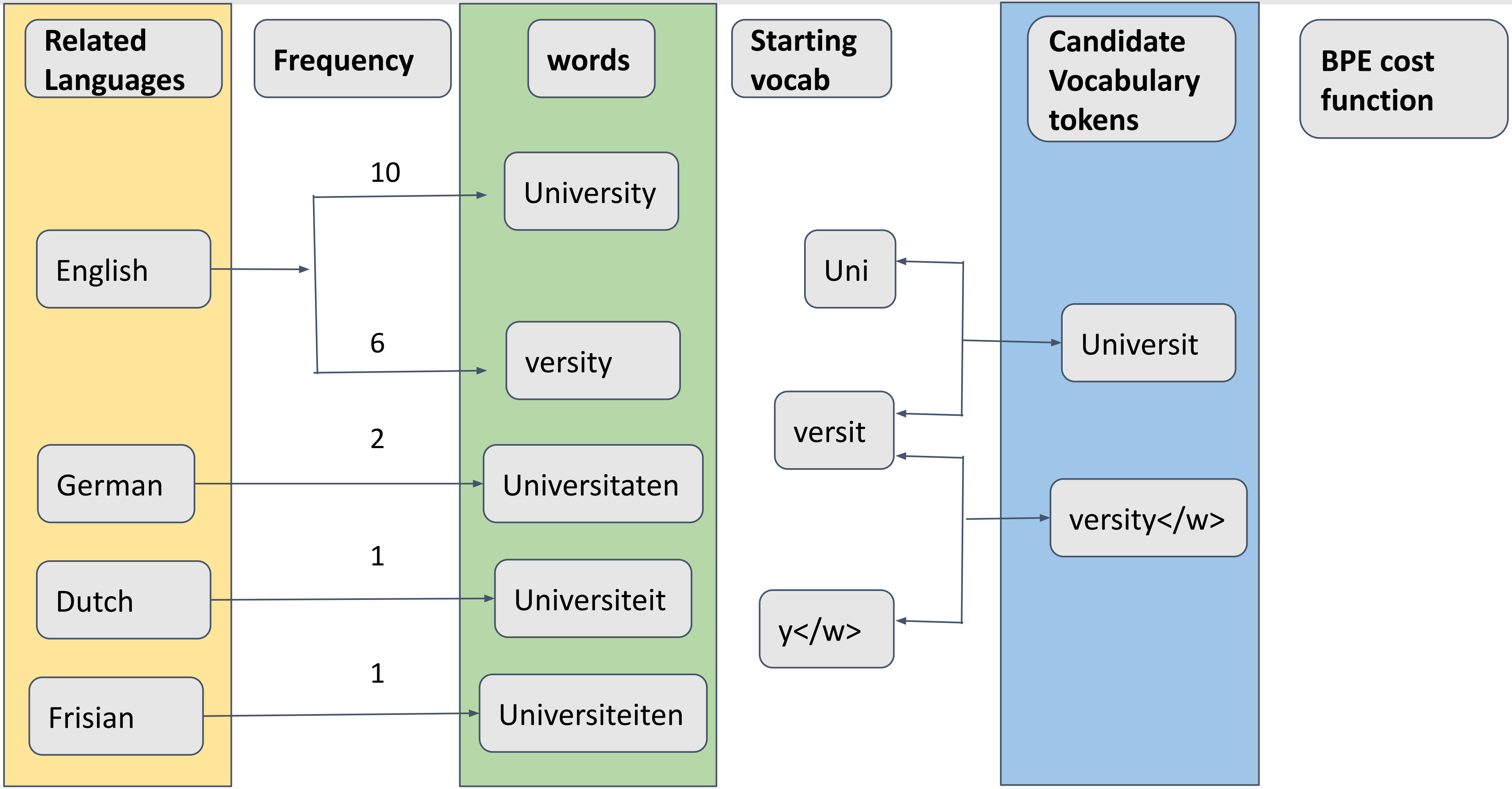
Example of BPE in action



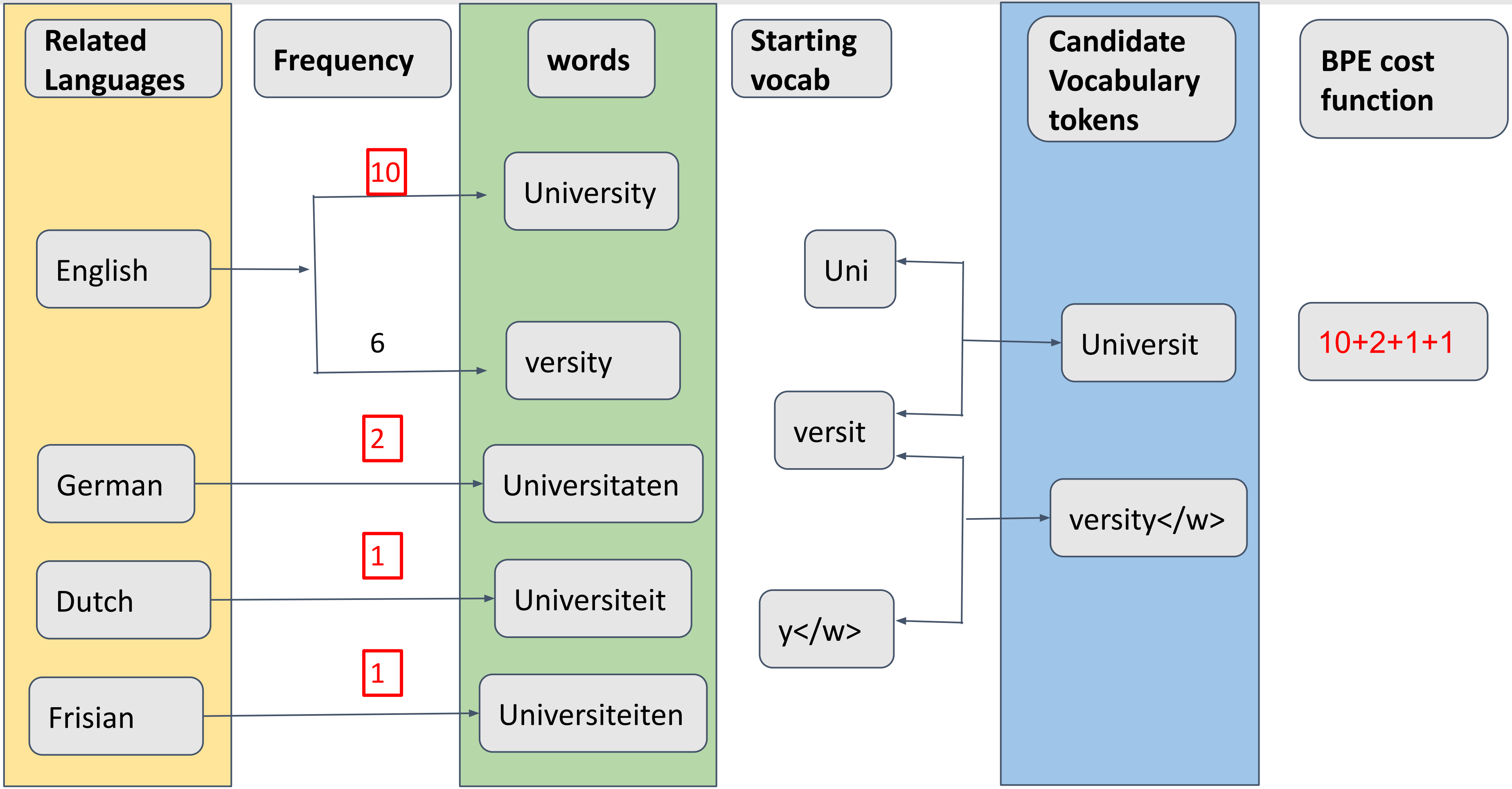
Example of BPE in action



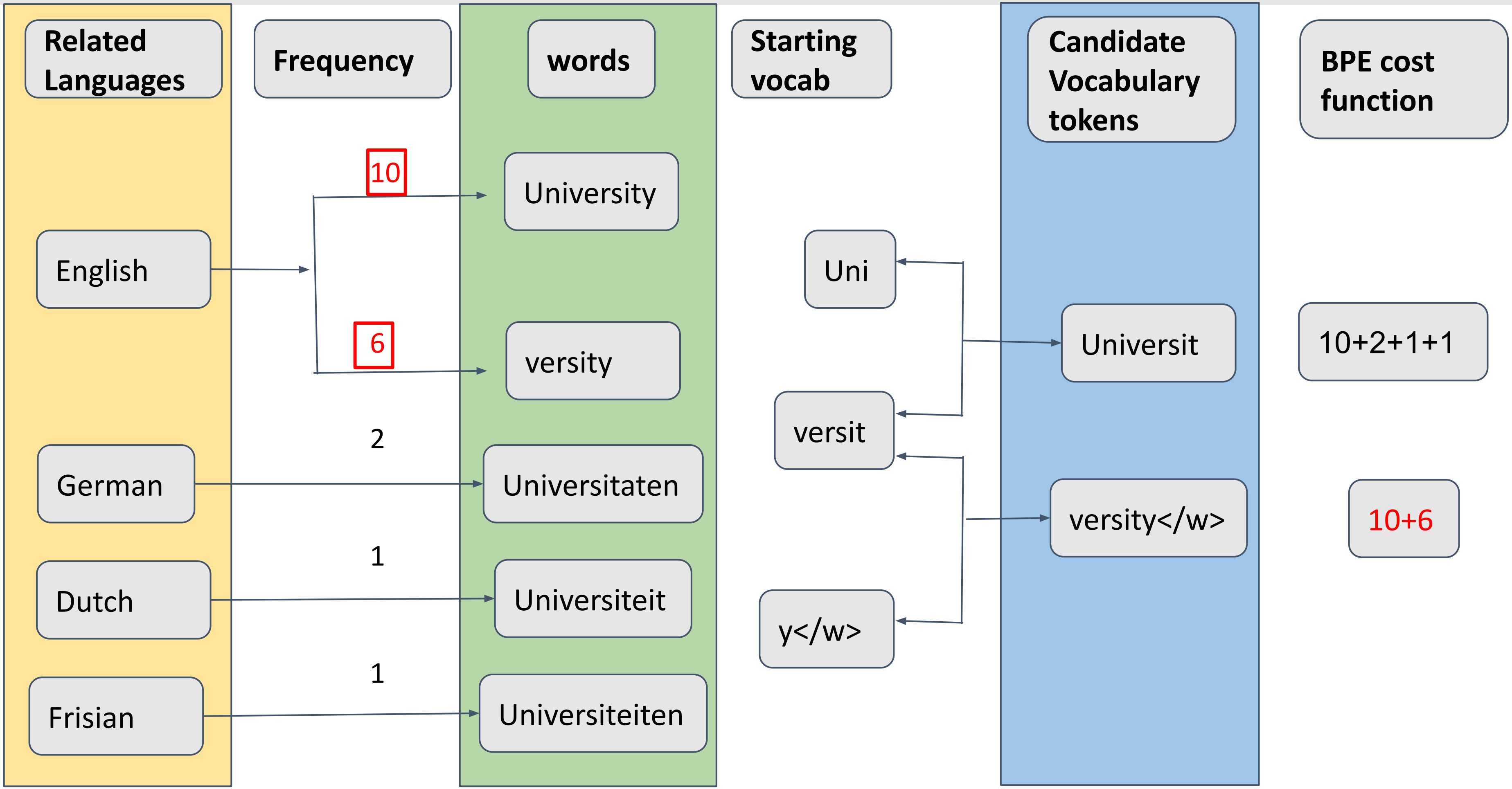
Example of BPE in action



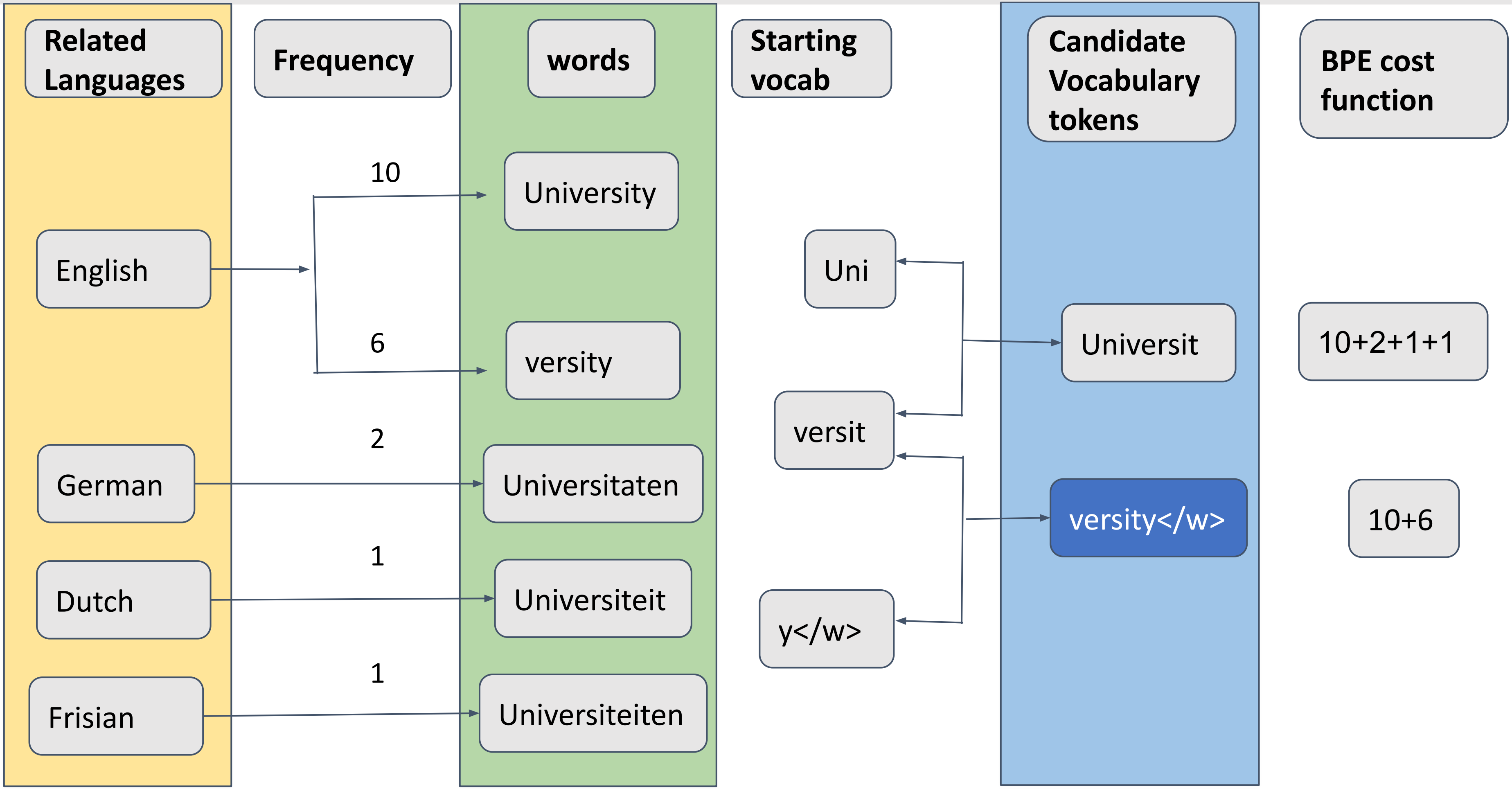
Example of BPE in action



Example of BPE in action



Example of BPE in action



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

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

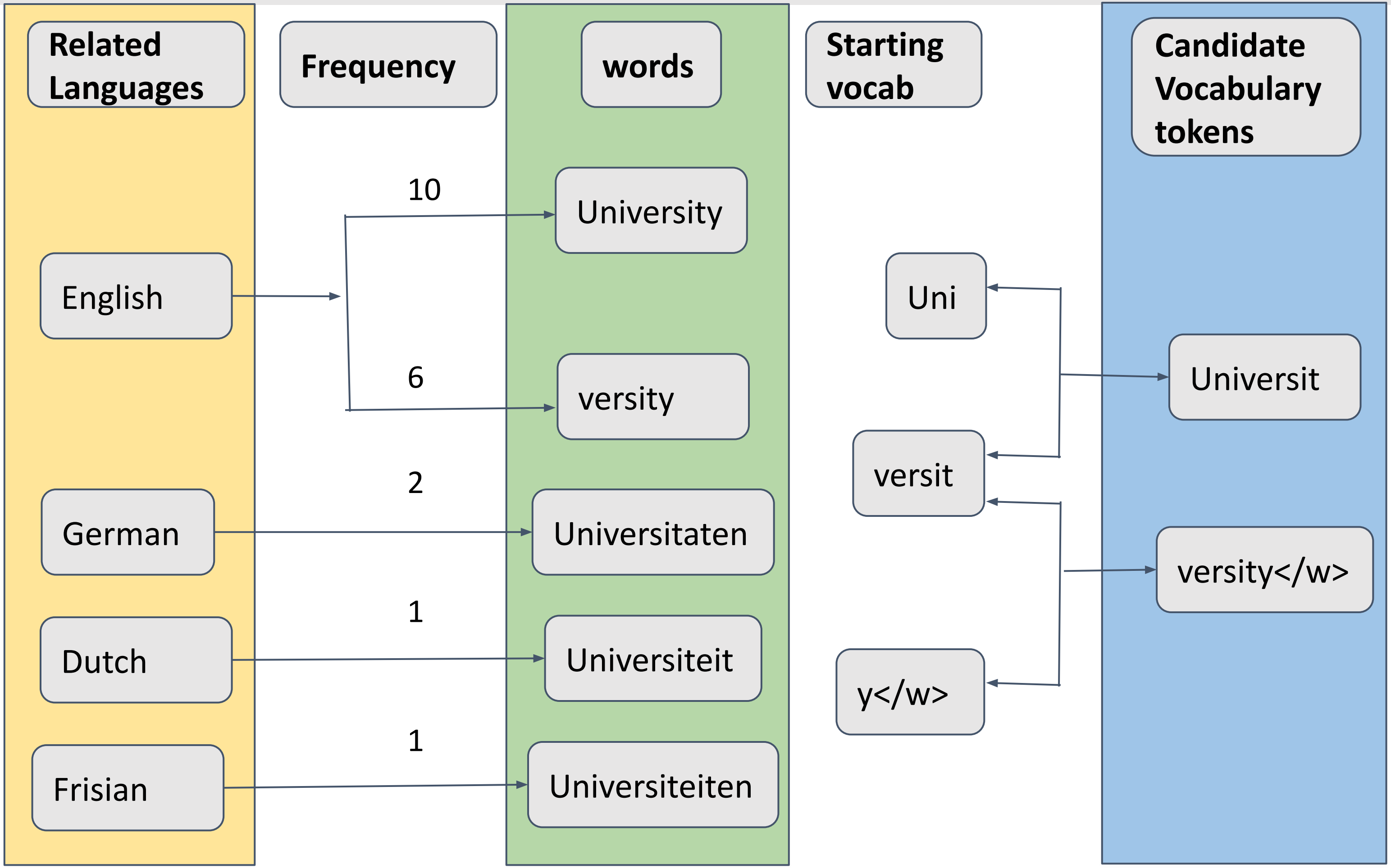
$$\text{overlap}(L_i, L_h, S) = \sum_{k \in S} \left(\frac{f_{ki}^p + f_{kh}^p}{2} \right)^{\frac{1}{p}}, \quad p \leq 1$$

.)

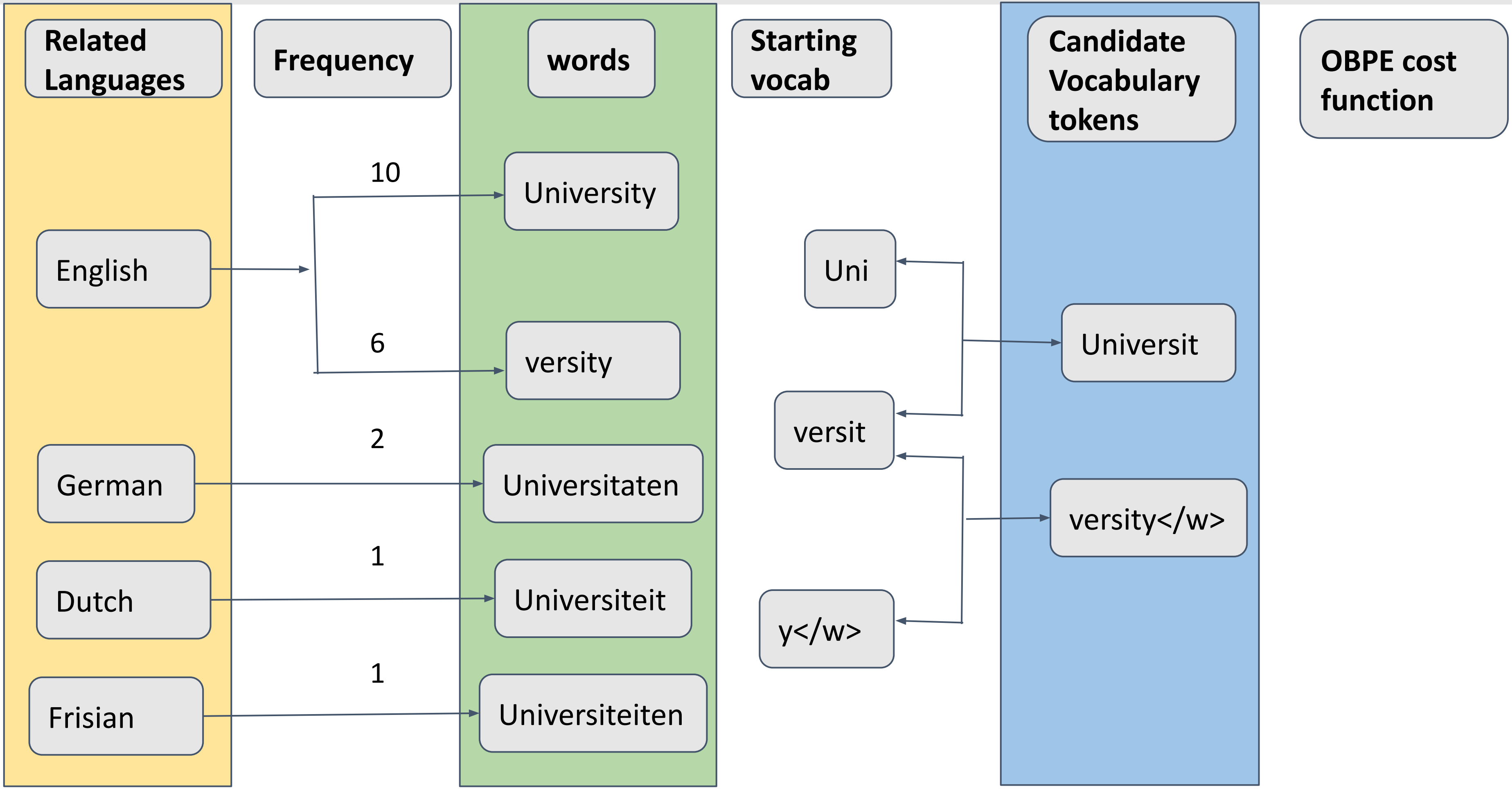
- The greedy version of the objective that controls the candidate vocabulary item to be inducted in each iteration of OBPE

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

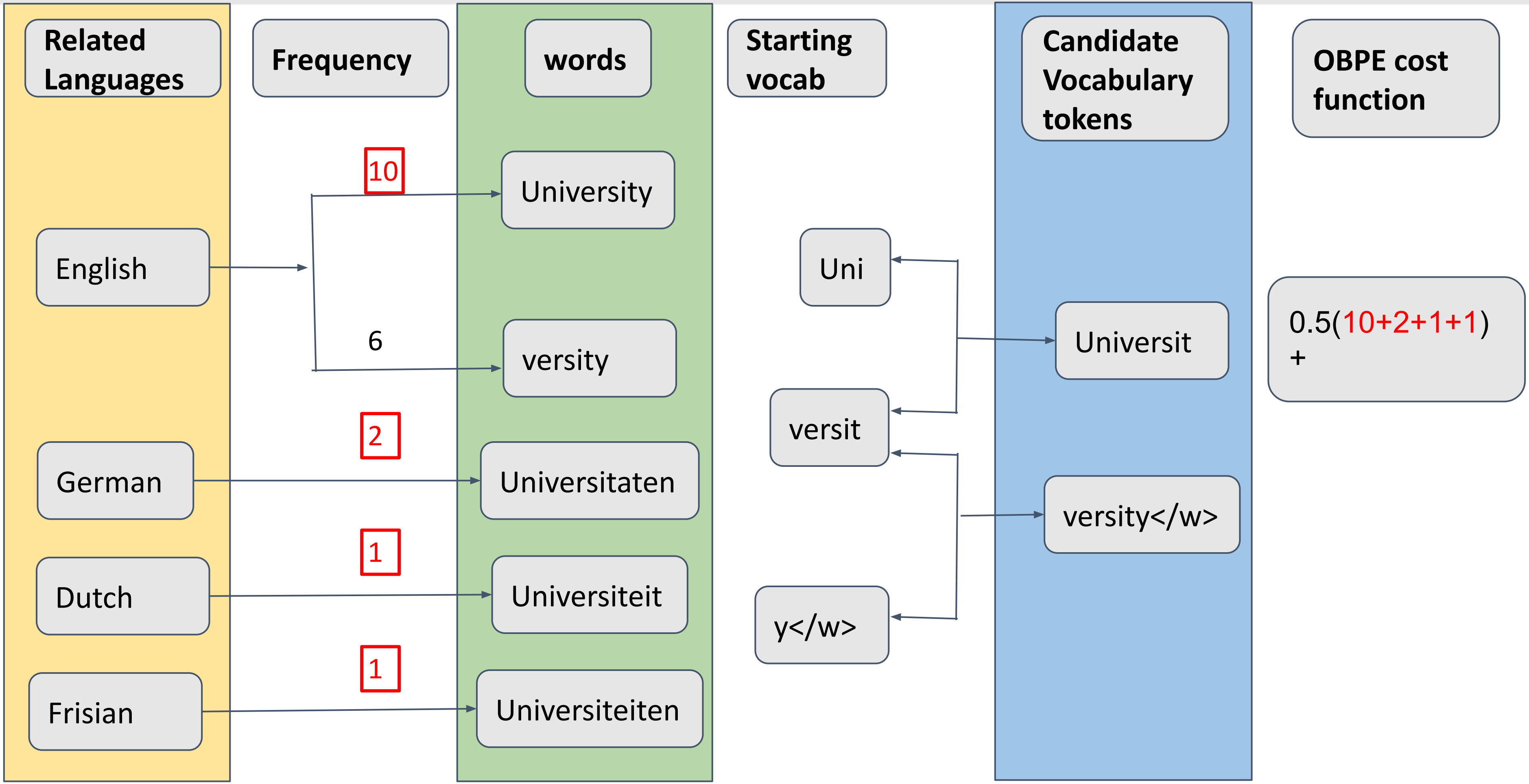
Example of OBPE in action



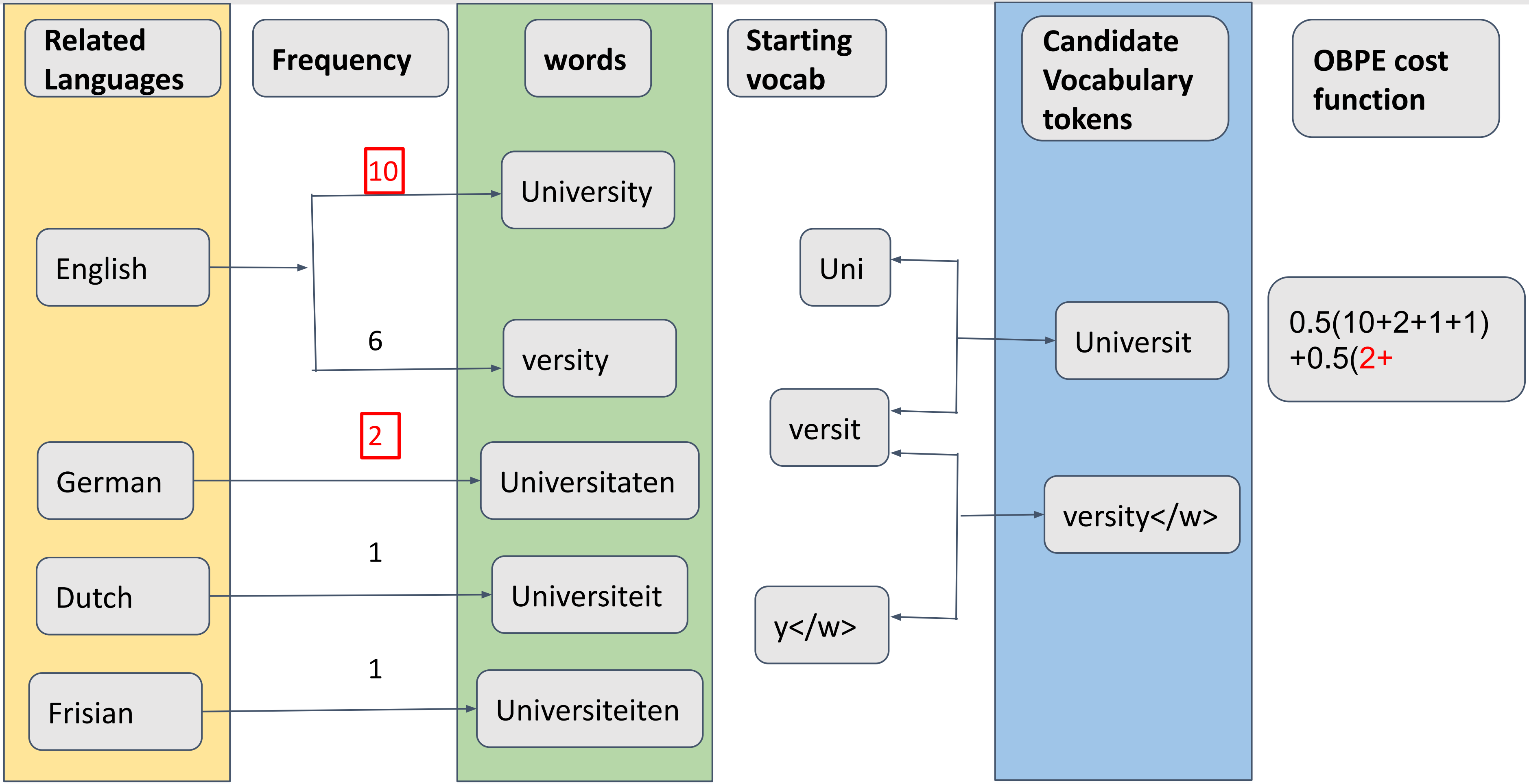
Example of OBPE in action



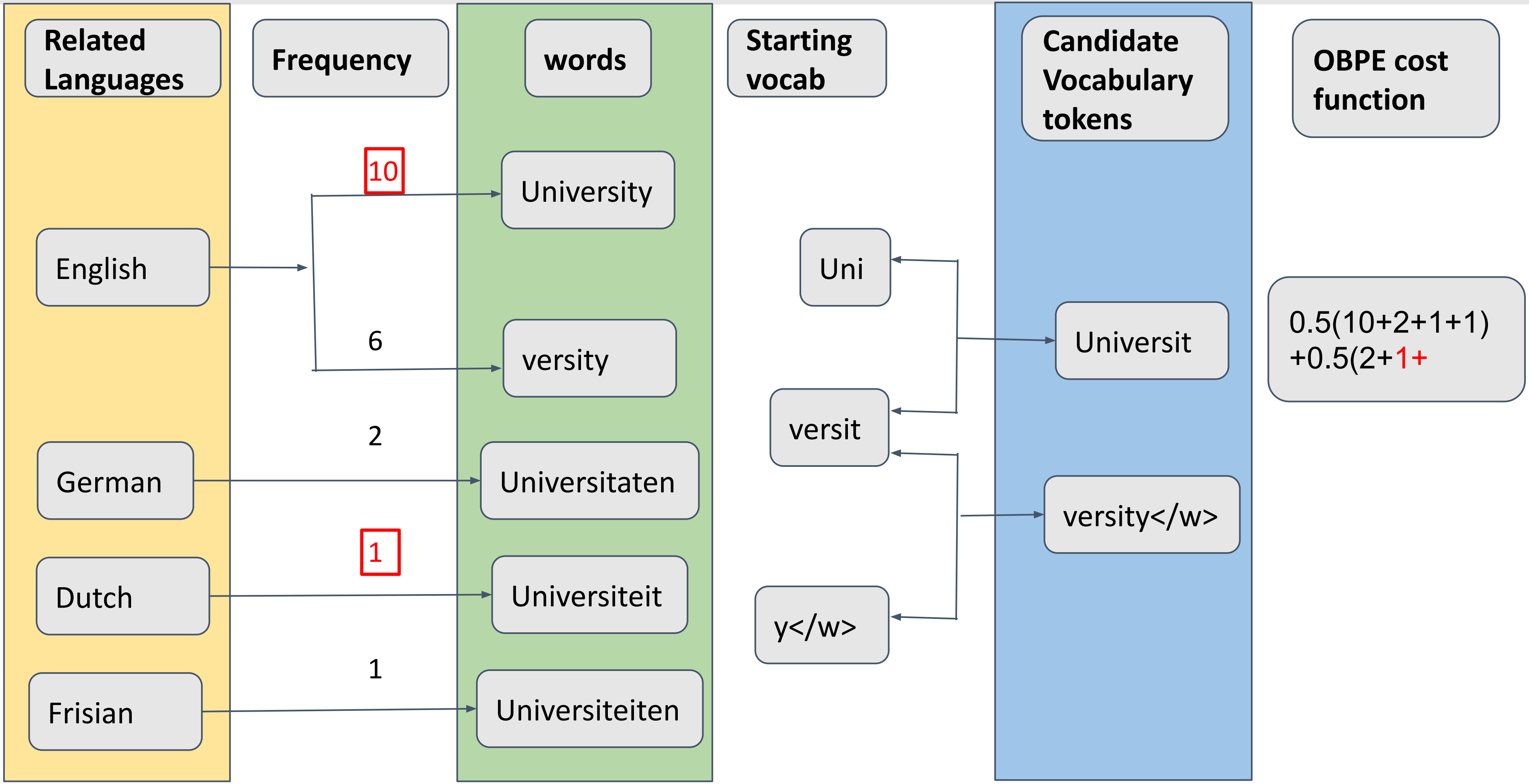
Example of OBPE in action



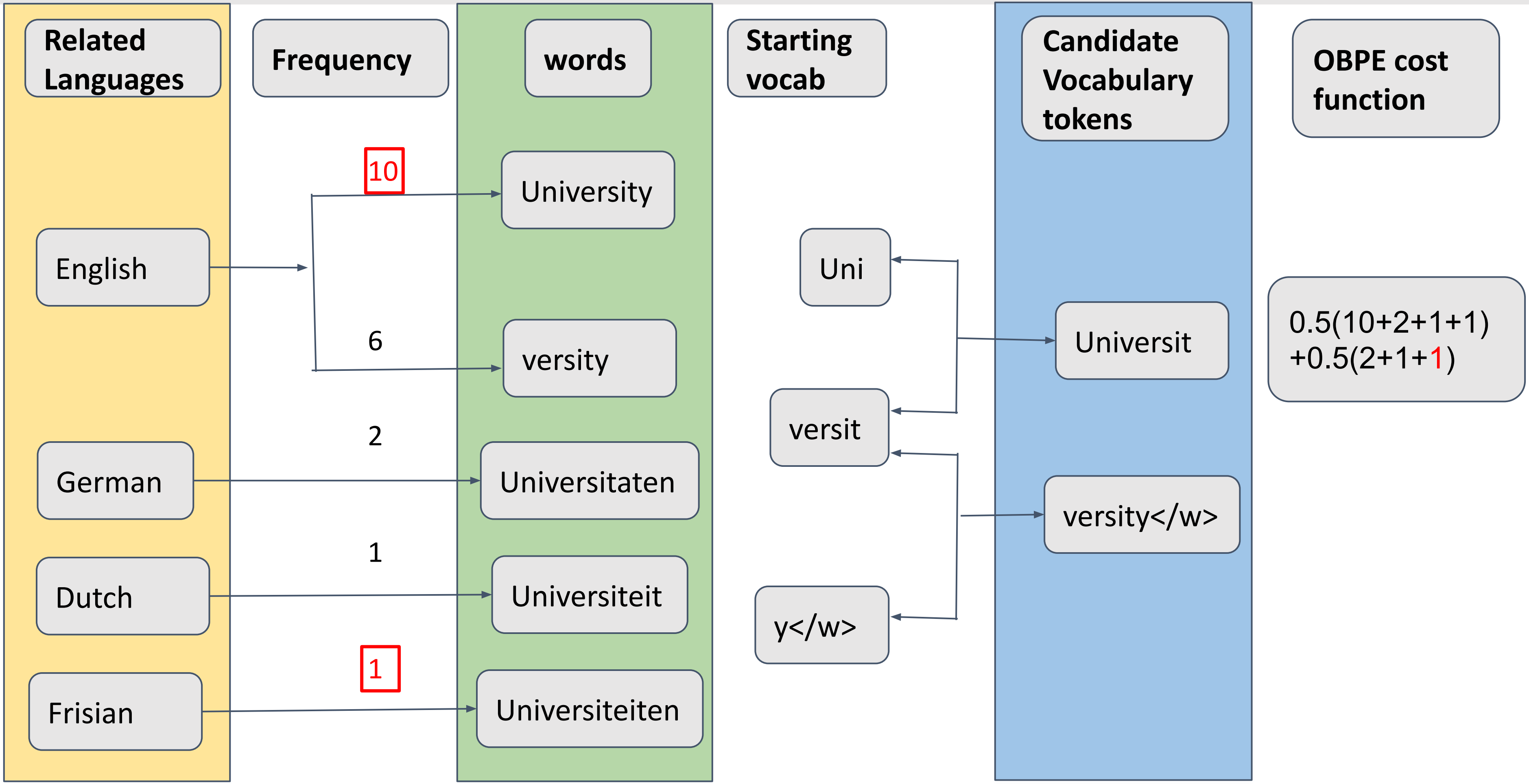
Example of OBPE in action



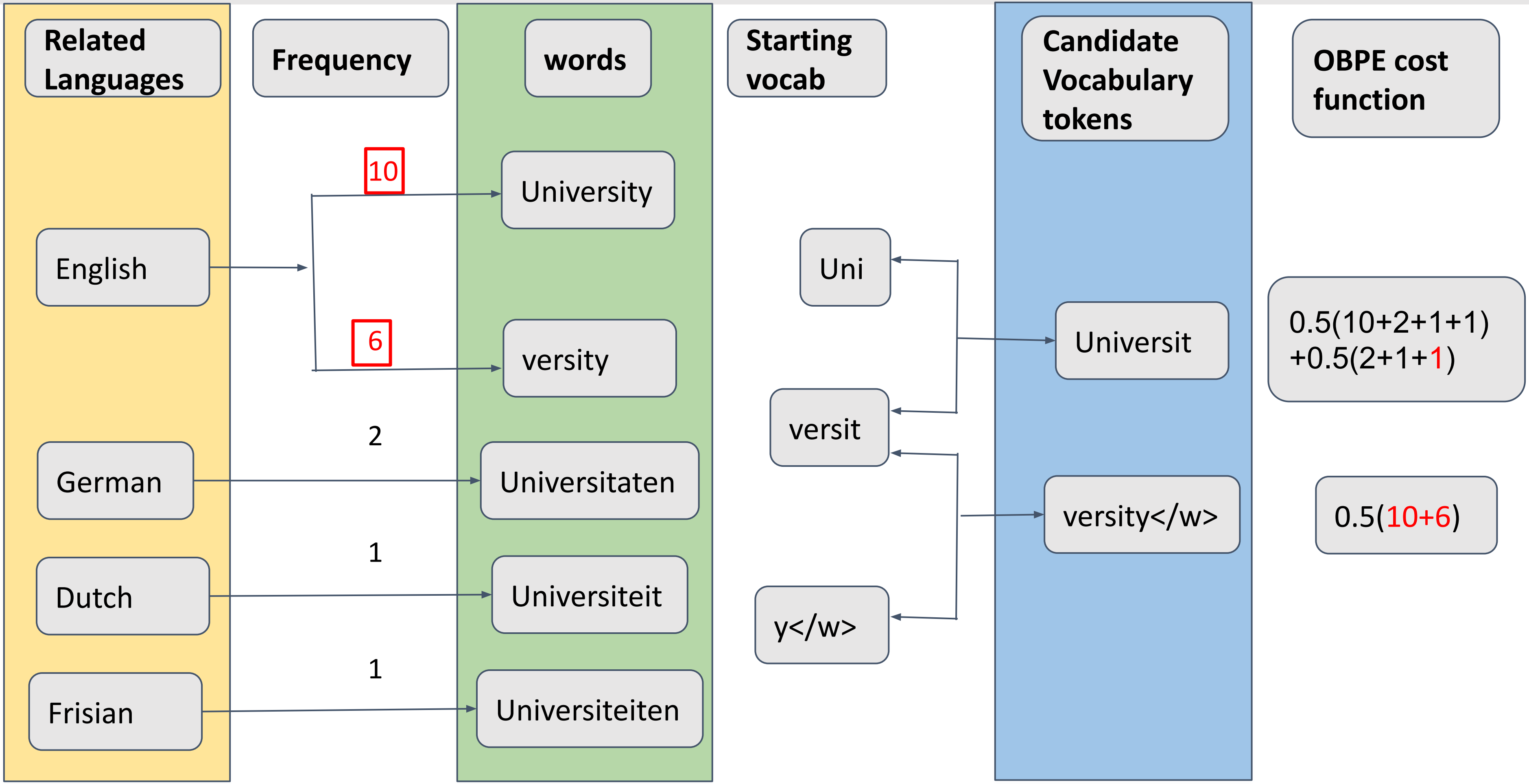
Example of OBPE in action



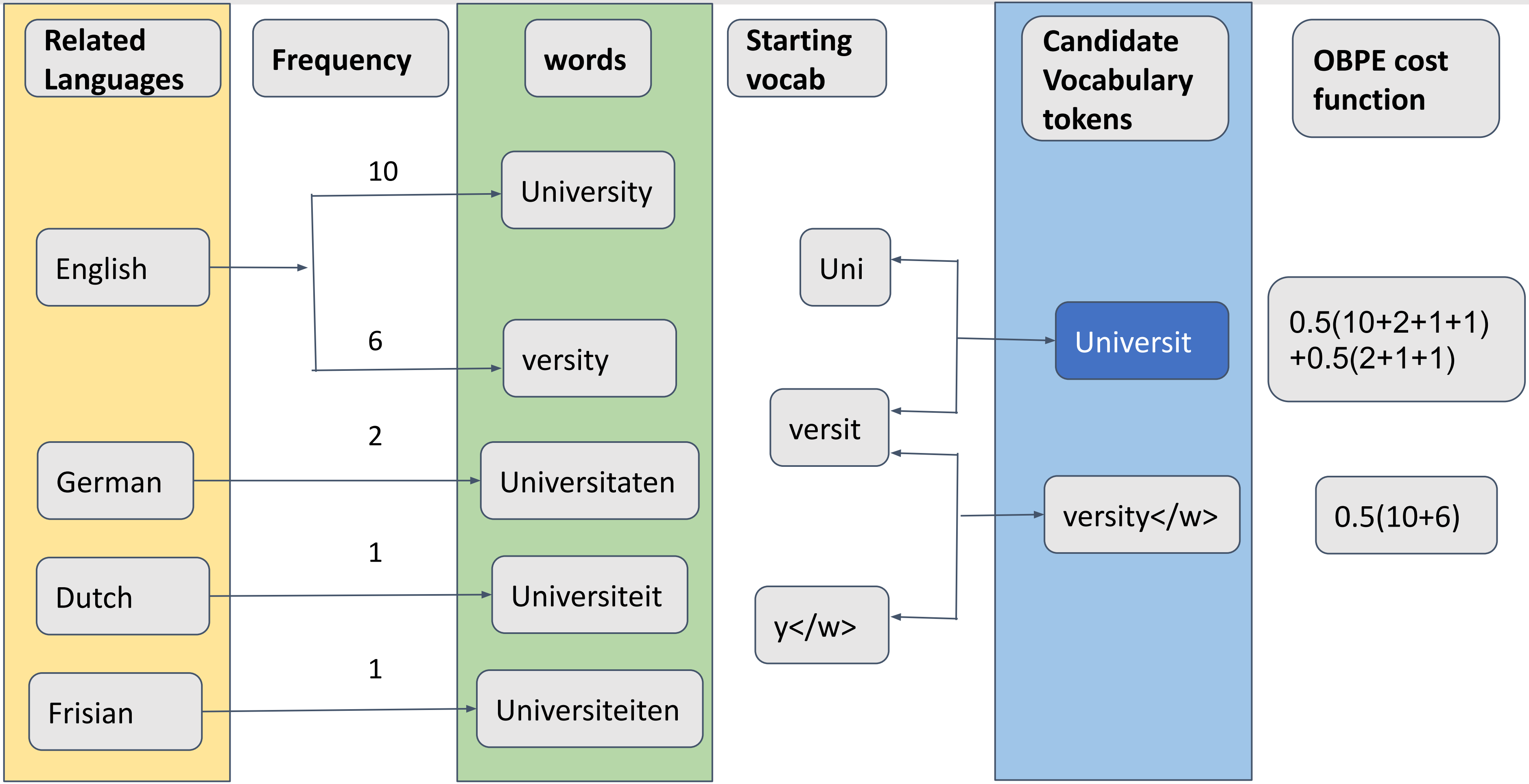
Example of OBPE in action



Example of OBPE in action



Example of OBPE in action



Experimental Setup

Family	HRL	LRLs	Number of HRL Docs	
			Balanced	Skewed
West Germanic	English (en)	German (de), Dutch (nl), Western Frisian (fy)	0.16M	1.00M
Romance	French (fr)	Spanish (es), Portuguese (pt), Italian (it)	0.16M	0.50M
Indo-Aryan	Hindi (hi)	Marathi (mr), Punjabi (pa), Gujarati (gu)	0.16M	0.16M

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

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

Balanced setting

Method	LRL Performance(↑)				HRL Performance(↑)			
	NER	TC	XNLI	POS	NER	TC	XNLI	POS
BPE	64.48	65.52	52.07	84.64	83.26	**82.07**	62.71	**95.20**
BPE-dp	63.92	64.15	52.66	84.75	81.73	81.07	63.74	94.61
CV	59.58	61.91	49.30	81.68	81.15	80.93	64.51	94.47
TokComp	63.79	65.77	53.94	**85.49**	82.43	80.93	66.10	94.86
OBPE	**65.72**	**68.02**	**54.03**	85.26	**83.98**	81.91	**66.27**	95.09

Zero-shot LRL accuracy improves compared to the baselines across all four tasks

Is OBPE more effective than BPE for zeroshot transfer?

Skewed setting

Method	LRL Performance(↑)				HRL Performance(↑)			
	NER	TC	XNLI	POS	NER	TC	XNLI	POS
BPE	52.91	51.68	48.57	74.79	81.78	80.04	64.96	95.03
CV	52.73	54.40	44.28	**76.70**	79.84	77.74	57.18	94.60
OBPE	**55.09**	**55.37**	**50.01**	75.05	**82.94**	**80.31**	**65.57**	**95.09**

Zero-shot LRL accuracy improves compared to the baselines across all four tasks

Is OBPE more effective than BPE for zeroshot transfer?

Balanced setting

Method	LRL Performance(↑)				HRL Performance(↑)			
	NER	TC	XNLI	POS	NER	TC	XNLI	POS
BPE	64.5	65.5	52.1	84.6	83.3	82.1	62.7	95.2
+overSample	64.4	67.6	52.1	84.6	82.4	82.0	62.0	95.2
OBPE	65.7	68.0	54.0	85.3	84.0	81.9	66.3	95.1
+overSample	64.6	67.9	53.5	85.1	82.7	81.7	65.7	94.8

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

Is OBPE more effective than BPE for zeroshot transfer?

Balanced setting

Method	LRL Performance(↑)				HRL Performance(↑)			
	NER	TC	XNLI	POS	NER	TC	XNLI	POS
BPE	64.5	65.5	52.1	84.6	83.3	**82.1**	62.7	**95.2**
+overSample	64.4	67.6	52.1	84.6	82.4	82.0	62.0	95.2
OBPE	**65.7**	**68.0**	**54.0**	**85.3**	**84.0**	81.9	**66.3**	95.1
+overSample	64.6	67.9	53.5	85.1	82.7	81.7	65.7	94.8

OBPE with default sampling is best for both LRLs and HRLs

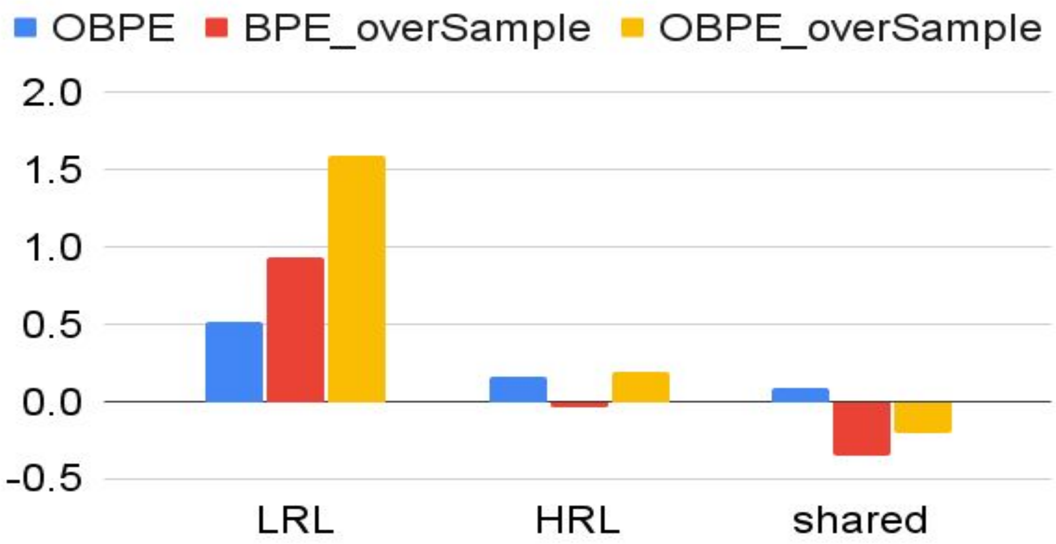
Is OBPE more effective than BPE for zeroshot transfer?

Balanced setting

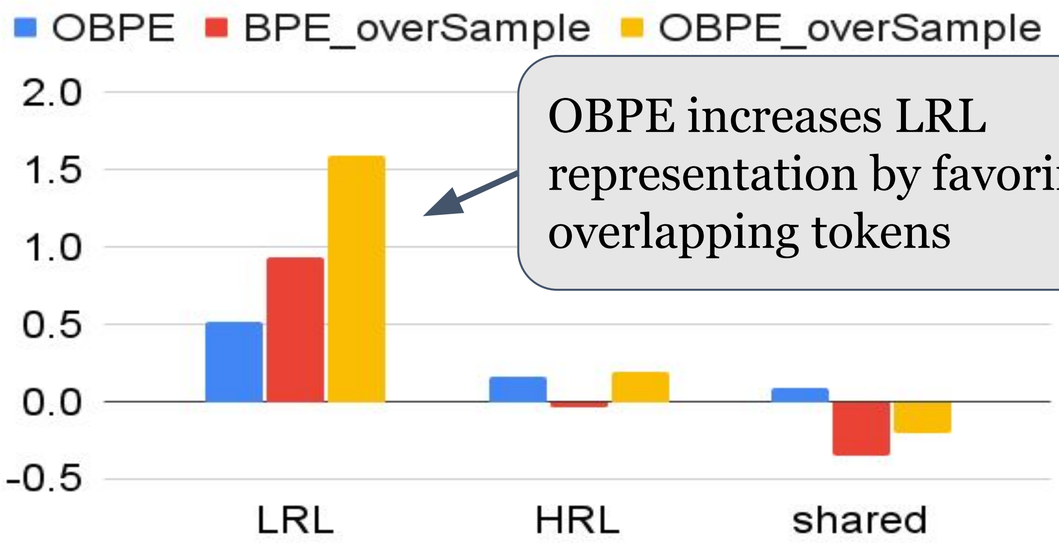
Method	LRL Performance(↑)				HRL Performance(↑)			
	NER	TC	XNLI	POS	NER	TC	XNLI	POS
BPE	64.5	65.5	52.1	84.6	83.3	82.1	62.7	95.2
+overSample	64.4	67.6	52.1	84.6	82.4	82.0	62.0	95.2
OBPE	65.7	68.0	54.0	85.3	84.0	81.9	66.3	95.1
+overSample	64.6	67.9	53.5	85.1	82.7	81.7	65.7	94.8

OBPE_overSampled is better than BPE_overSampled

How does increased LRL representation in the vocabulary impact accuracy?

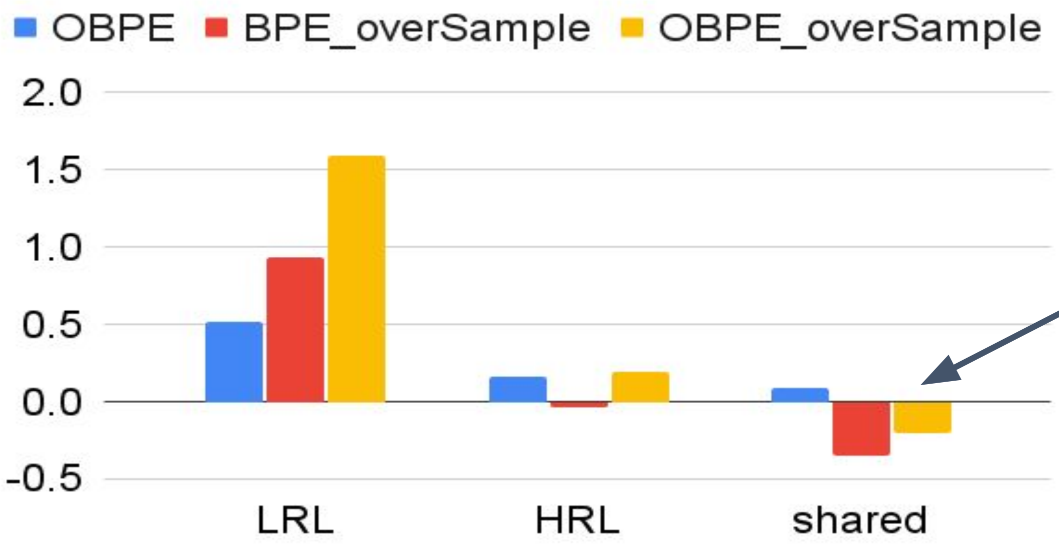


How does increased LRL representation in the vocabulary impact accuracy?



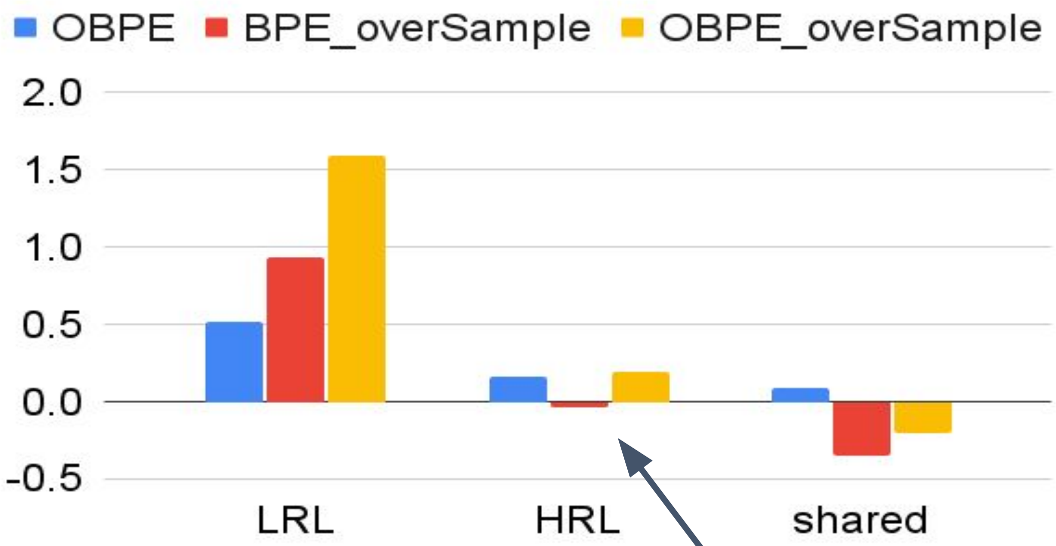
OBPE increases LRL representation by favoring overlapping tokens

How does increased LRL representation in the vocabulary impact accuracy?



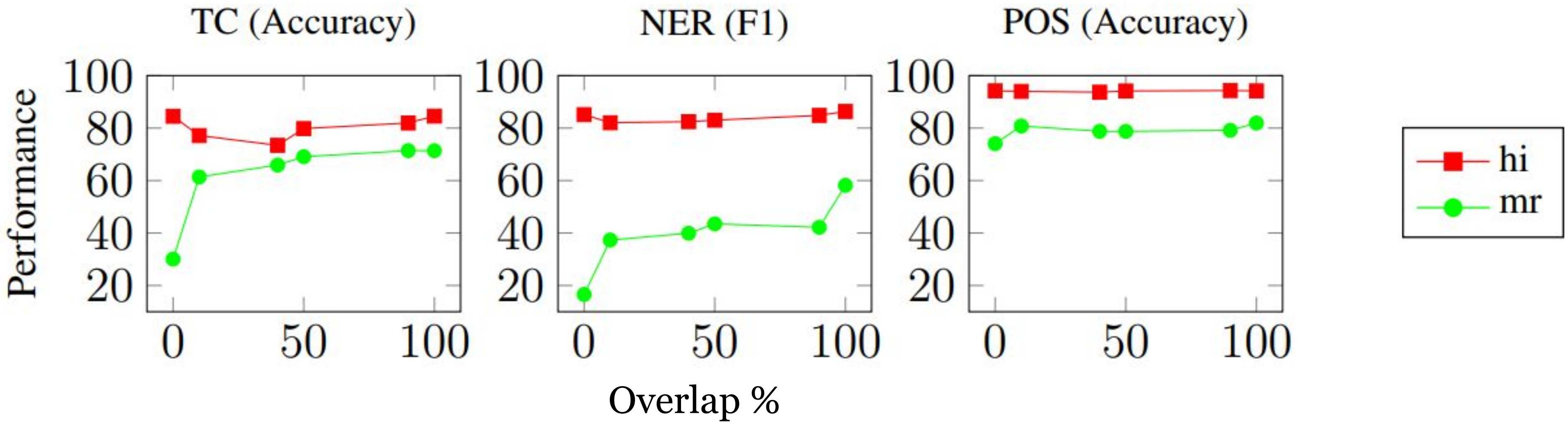
OBPE increases fraction of tokens shared across related languages, Oversampling decreases fraction of shared tokens

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?

More related | Less related

Task	High (hi: 110K)	Low (mr: 20K)	Task	High (en: 1GB)	Low (es: 20K)
NER	-12.2	-41.6	NER	-1.4	-11.7
TC	-2.7	-41.3	XNLI	-1.3	-1.3
POS	-6.6	-7.8			

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?

More related

Less related

Task	High (hi: 110K)	Low (mr: 20K)	Task	High (en: 1GB)	Low (es: 20K)
NER	-12.2	-41.6	NER	-1.4	-11.7
TC	-2.7	-41.3	XNLI	-1.3	-1.3
POS	-6.6	-7.8			

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

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

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

Paper : https://arxiv.org/abs/2203.01976
Github : https://github.com/Vaidehi99/OBPE
Contact : vaidehipatil16@gmail.com, partha@google.com, sunita@iitb.ac.in