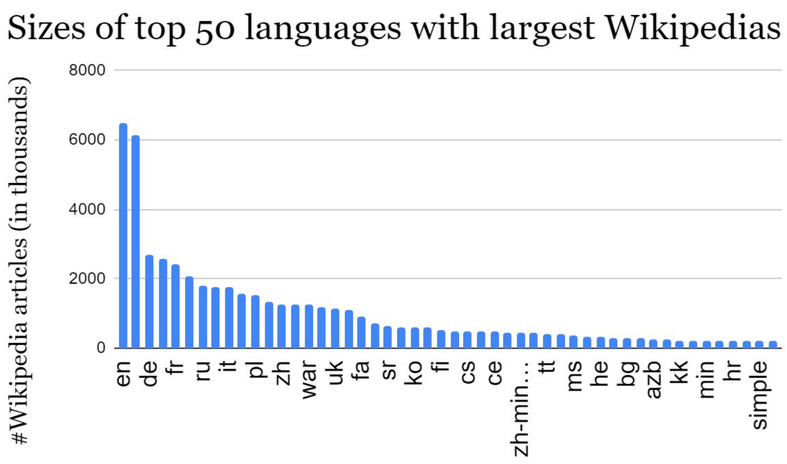
Overlap-based Vocabulary Generation Improves Cross-lingual Transfer Among Related Languages

Vaidehi Patil[1], Partha Talukdar[2], Sunita Sarawagi[1]

[1]Indian Institute of Technology Bombay, India [2]Google Research, India

vaidehipatil16@gmail.com, partha@google.com, sunita@iitb.ac.in

Challenges: Multilingual Models



Multilingual models

- Form the core of many NLP tasks
- Effective for cross-lingual transfer when there is sufficient LRL unlabeled corpus

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

Main Takeaways

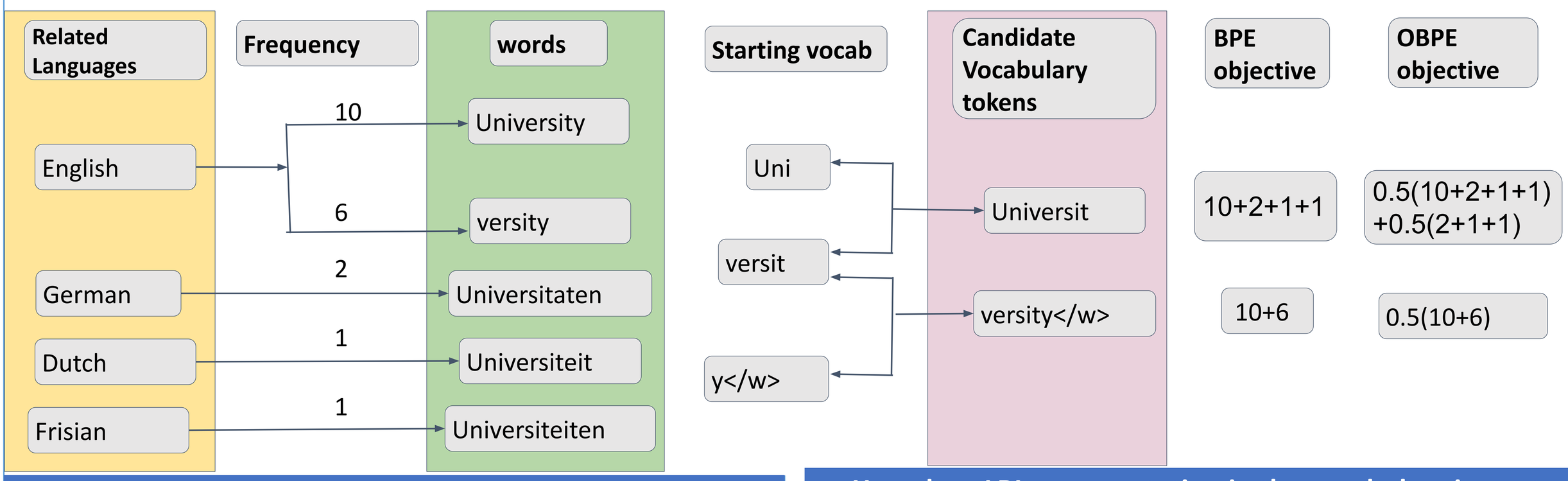
- Oversampling is less effective than exploiting token-overlap for zero-shot transfer in related languages setting
- Token overlap matters (unlike K et al., 2020) under two settings:
  - Languages are sufficiently related
  - LRL is resource-poor even in the amount of unlabeled data

Is OBPE more effective than BPE for zeroshot transfer?

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

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

How does OBPE work?

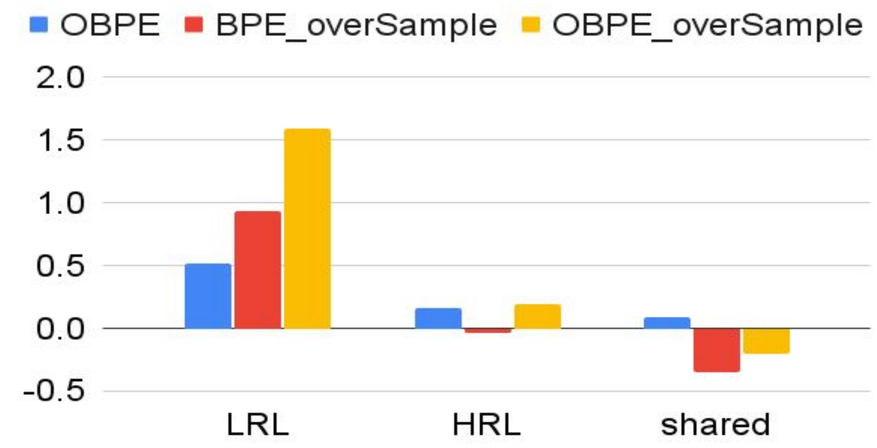


Motivation

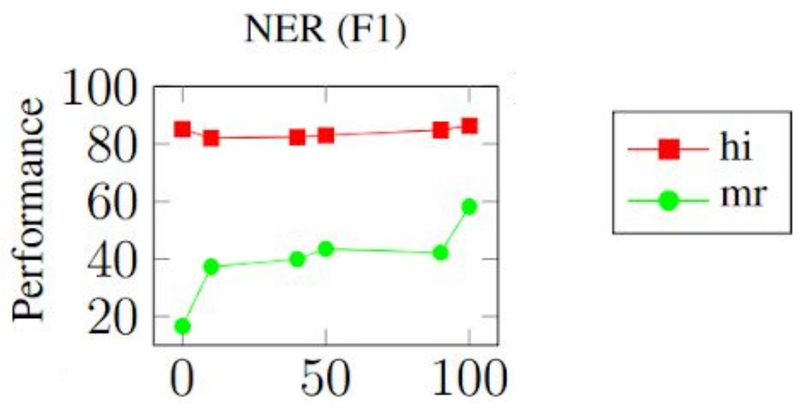
Lexically overlapping words with similar meanings across four languages in three different language families

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

How does LRL representation in the vocabulary impact accuracy?



What is the effect of token overlap on overall accuracy?



Languages	High	Low
hi-mr	-12.2	-41.6
en-es	-1.4	-11.7

Token overlap plays an important role in low resource, related languages settings

Acknowledgements : Google's TFRC program for TPUs