Label imputation for homograph disambiguation

Theoretical and practical approaches

Jen Seale

July 19, 2021

THE GRADUATE CENTER
CITY UNIVERSITY OF NEW YORK

1. Introduction
2. Homograph typology
3. Label imputation
    - Transcribed audio
    - Parallel corpora
4. Conclusion

Introduction

Intro: Homograph disambiguation

Homograph: *lead*

1. Sha'Carri took the *lead* /ˈliːd/ in the race.
2. They considered the atomic structure of *lead* /ˈlɛd/.

Hearst (1991), Gale, Church and Yarowsky (1992), Gorman, Mazovetzkiy, and Nikolaev (2018)

Intro: Motivation

Human-labeled data for homograph disambiguation (HD):

- low resource
- imbalanced

Mihalcea and Moldovan (1999), Diab and Resnik (2002), Nishiyama et al. (2018)

Intro: Aim

Improve smaller, imbalanced
homograph disambiguation data sets
through label imputation

Intro: Wikipedia Homograph Data (WHD)

- Gorman et al. (2018); 4 annotators label ~16,000 sentences
- 162 unique homographs, ~2 pronunciation classes each
- ~100 samples per homograph

Intro: Wikipedia Homograph Data (WHD)

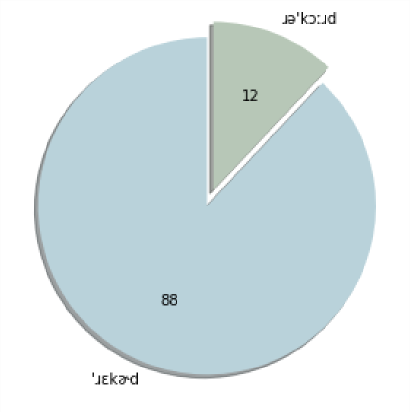


Figure 1: Homograph with median class size ratio, *record*.

Intro: Modeling with best label imputation technique

WHD
vs.
WHD + label-imputed data

absolute increase of **1.9–7.5%**
in balanced accuracy

Investigation:
use of part-of-speech (POS) in homograph disambiguation

Result:
homograph classification system

Intro: Research, Part 2

- Label imputation from transcribed audio
  - Develop a semi-automated pipeline for label imputation
  - Impute labels from Switchboard data (SWBD)
  - Evaluate the label imputation
    - Model with WHD
    - Model with WHD + label-imputed SWBD
    - Model with WHD + human-labeled SWBD
    - Compare model performance on micro and balanced accuracy

Intro: Research, Part 3

- Label imputation from parallel corpora
  - Develop hypothesis which forms the basis for label imputation
  - Impute labels from French-English European parliament proceedings (Europarl)
  - Evaluation of label imputation
    - Model with WHD
    - Model with WHD + label-imputed Europarl
    - Compare model performance on micro, balanced, and per class accuracy

Intro: Research contributions

- Novel classification system for homographs
- Formalized hypothesis of interlingual alignment between homograph pronunciations and text word forms
- Semi-automated label imputation:
    - transcribed audio
    - interlingual alignment hypothesis
- Pre-trained language models, fine-tuned as token classifier HD models
- Model performance provides evidence of the utility of the label imputation from parallel corpora
- Data sets to be made available to the research community

Typology

Typology: An important question

POS is used as a differentiating feature
for homographs, and for homonyms.

Ribeiro, Oliveira, and Trancoso (2002), Braga and Coelho (2007), Elkahky et al. (2018), Hauer and Kondrak (2020), Habibi (2020)

What happens when you rely solely on POS to disambiguate homographs?

Typology: POS & pronunciation

Homograph 1: *present*

Noun: I have a /ˈpɹɛzənt/ for you.

Verb: I have to /ˌpɹiːˈzɛnt/ information.

Noun: **I have a /ˌpɹiːˈzɛnt/ for you.

Verb: **I have to /ˈpɹɛzənt/ information.

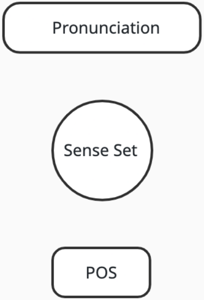
Homograph 2: *bass*

Noun: I caught a /ˈbæs/.

Noun: I play the /ˈbeɪs/.

Typology: 4 homograph types

4 homograph types : Relationships between 3 elements



Typology: Sense sets

present noun (1)

Save Word

pres·ent | \ ˈpre-zᵊnt \

Definition of *present* (Entry 1 of 4)

: something presented : GIFT

present verb

pre·sent | \ pri-ˈzent \

presented; **presenting**; **presents**

Definition of *present* (Entry 2 of 4)

transitive verb

1 : to make a gift to

2 : to give or bestow formally

3 **a** : to bring (something, such as a play) before the public

b **(1)** : to bring or introduce into the presence of someone especially of superior rank or status

(2) : to introduce socially

Typology: Type I homographs

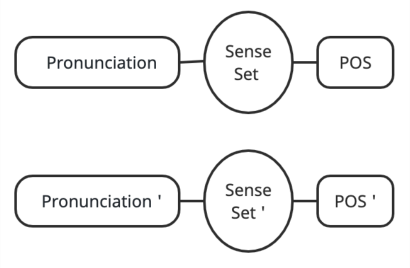


Figure 2: The defining relationships of a Type I homograph.

Typology: Type I homographs

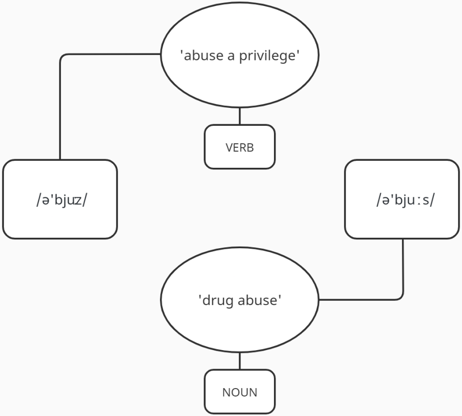


Figure 3: Type I homograph, *abuse*.

Typology: Type II homographs

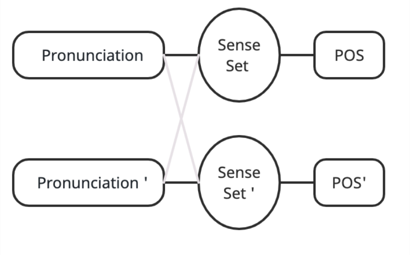


Figure 4: The defining relationships of a Type II homograph.

Typology: Type II homographs

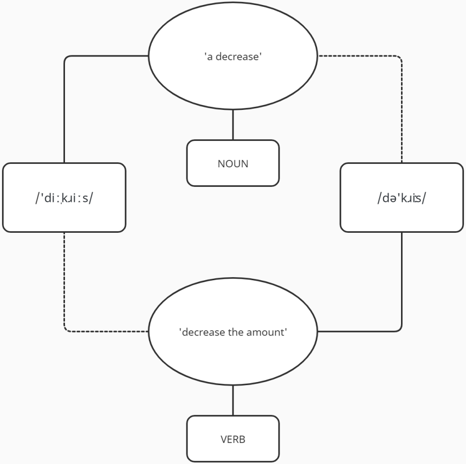


Figure 5: Type II homograph, *decrease*.

Typology: Type III homographs

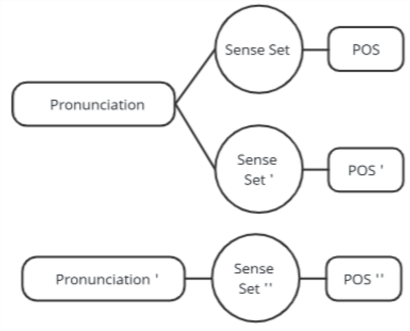


Figure 6: The defining relationships of a Type III homograph.

Typology: Type III homographs

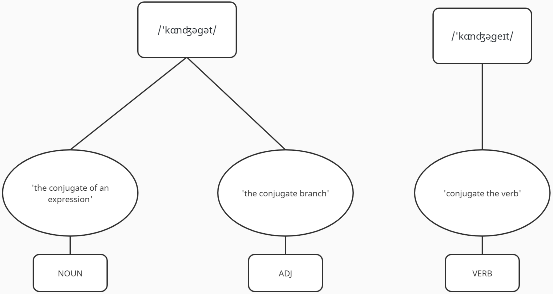


Figure 7: Type III homograph, *conjugate*.

Typology: Type IV homographs

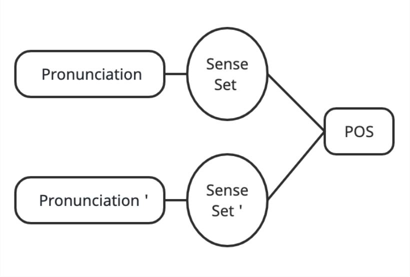


Figure 8: The defining relationships of a Type IV homograph.

Typology: Type IV homographs

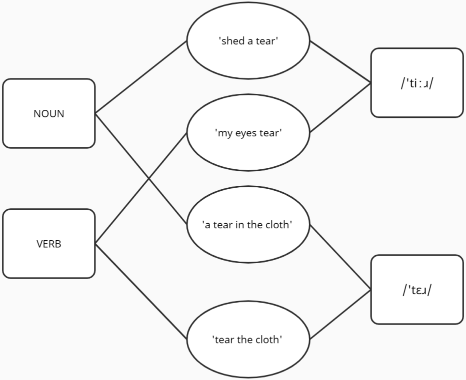


Figure 9: Type IV homograph, *tear*.

Typology: POS-based hypothetical upper bound

Type IV: 20.9% of WHD homographs

Data	Micro Accuracy	Balanced Accuracy
WHD eval set	78.69	79.01

Table 1: POS-based, upper bound micro and balanced accuracies on WHD evaluation set as released by Gorman et al. (2018)

Typology: BERT performance on 6 Type IV homographs

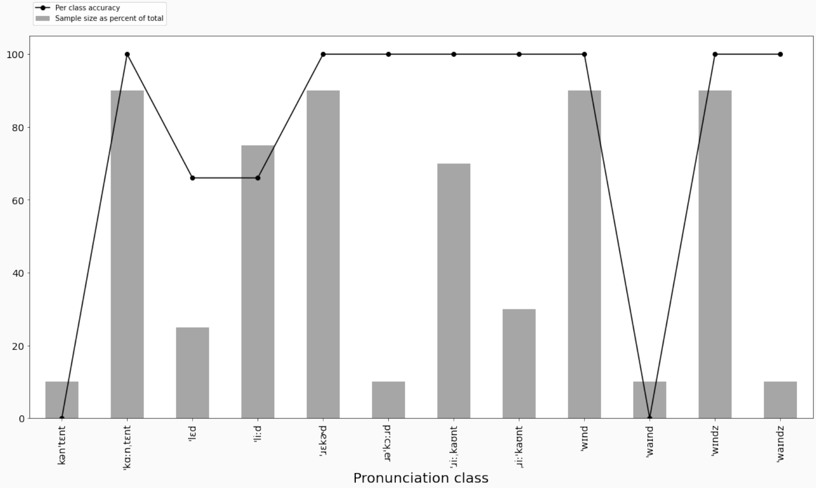


Figure 10: BERT token classifier test set performance on Type IV homographs plotted against pronunciation class sample size.

Typology: Conclusion

- Type II: Pronunciation overlap naturally occurs in audio data; can interfere with label imputation
- POS-only disambiguation:
    - easiest for Type I
    - feasible for Types II & III
    - impossible for Type IV
- Token classifier models which do not explicitly use POS feature perform relatively well on Type IV homographs

Label imputation

Label imputation: Two types

1. Transcribed audio
2. Parallel corpora

Label imputation from transcribed audio

Audio: Experiment

1. Development of semi-automated label imputation
2. BERT HD models:
    - 34-homograph WHD
    - 34-homograph WHD + semi-automatedly label-imputed SWBD data
    - 34-homograph WHD + hand-labeled SWBD data
3. Model evaluation
    - Balanced accuracy
    - Micro accuracy

Audio: Switchboard Data

- American English telephone conversations with ~3 million words
- Subset of 1 million time-stamped, transcribed word forms
- 2,935 homograph samples across 95 WHD homograph types
  - automatedly labeled
  - manually reviewed & labeled
  - semi-automatedly labeled subset

Audio: Automated labeling

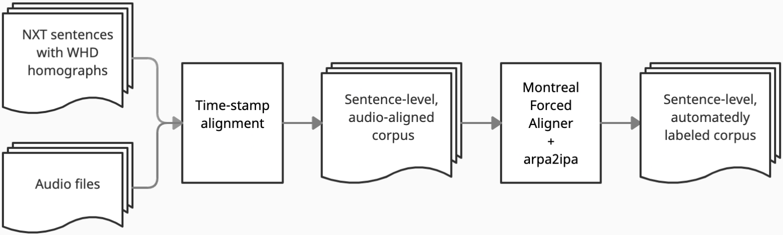


Figure 11: Automated label imputation

Audio: Automated imputation results

Do imputed labels match
WHD pronunciation labels?

No

But sometimes we come close.

Can we map imputed labels
to WHD labels?

Yes, *and no*.

Audio: Mappable distinctions

- Non-phonemic differences due to speaker idiolect and dialect
- Distinctions in notation

Audio: Mapping

Homograph	WHD IPA	Mapped IPA	Type
lead	/ˈliːd/	[lid]	Notation
uses	/ˈjuːzəz/	[ˈjuzɪz], [ˈjuzʌz]	Dialect

Table 2: Examples of WHD IPA to imputed IPA label mapping

Audio: Unmappable distinctions

Overlap in
homograph-disambiguating
sub-word elements

Audio: Unmappable distinctions

SWBD sample: “an *excuse* to do”
WHD IPA: /əkˈskjuːs/
Alternate WHD IPA: /əkˈskjuːz/
Imputed IPA: [ɪˈkskjuz]

SWBD sample: “murder would *decrease*”
WHD IPA: /dəˈkɹiːs/
Alternate WHD IPA: /ˈdiːˌkɹiːs/
Imputed IPA: [ˈdiˌkris]

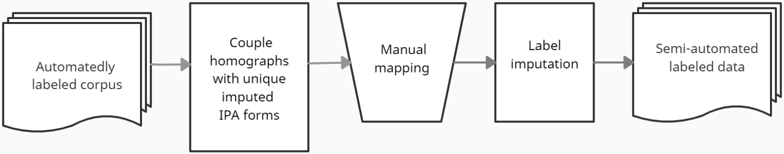


Figure 12: Semi-automated label imputation

{ juzɪz : ˈjuːzəz, juzʌzː ˈjuːzəz}

Example of mapping entries.

Audio: Subsetting

A subset of the homographs labeled in the SWBD data
is isolated for semi-automated labeling, modeling.

These homographs is constrained to those also in:

- 34 homograph subset of WHD
    - 1 low prevalence pronunciation class per homograph (less than 40% of total data; median: 11.5% of total data)

Audio: Models

BERT HD models:

- 34-homograph WHD
- 34-homograph WHD + semi-automatedly label-imputed SWBD data
- 34-homograph WHD + human-labeled SWBD data

Audio: Results

Model	Micro Acc	Balanced Acc	Bal Acc Change
BERT_WHD	93.84	84.08	-
BERT_Imputed_SWBD	95.21	84.6	.52
BERT_Human_SWBD	**95.72**	**86.47**	**2.39**

Table 3: 34 homograph-restricted BERT models' micro and balanced accuracy scores on test set. Change in balanced accuracy between baseline and semi-automatedly imputed, and hand-labeled SWBD-augmented models. Metrics averaged over four random seeds.

Audio: WHD augmentation

Data	Homograph	Low Pron	Prev	Samples
Imputed SWBD	47%	26%		+4%
Human SWBD	82%	58%		+9%

Table 4: SWBD coverage of 34 homograph-restricted WHD data set

Audio: Finding data, Nishiyama et al. (2018)

excuse /əksˈuːs/ → *justification*

1. Search: *justification*
2. Return: “They always brought up work as their justification not to spend more time with family”
3. Replace: “They always brought up work as their excuse /əksˈuːs/ not to spend more time with family”

Label imputation from parallel corpora

Background

- Different senses of a word correspond to distinct words in another language (Brown et al., 1991; Resnik and Yarowsky, 1999)
    - the *bear* : l'*ours*
    - *bear* it : le *supporter*
- Sense labeling using parallel corpora (Diab and Resnik, 2002)
    - the *bear* [animal] : l'*ours*
    - *bear* [endure] it : le *supporter*

Parallel corpora: One homonym per translation (OHPT)

Homonyms have disjoint translation sets.

'One Homonym Per Translation'
(Hauer and Kondrak, 2020)

Parallel corpora: OHPT, cont.

A homonym is a lexeme that shares a word with additional, semantically unrelated homonyms.

BANK_1	BANK_2
financial institution	sloping land
building	a heap or mass

Table 5: Homonyms with polysemous sense examples of the word ‘bank’.

If a word form has multiple semantically unrelated lexemes, but the POS for those lexemes is distinct, the lexemes are not homonyms. (Ex: bank, n and bank, v).

Parallel corpora: OHPT, cont.

OHPT excludes all homographs
that have pronunciation-sense pairings with:

- distinct POS
- related meanings

Parallel corpora: OHPAS Hypothesis

One Homograph Pronunciation Per Alignment Set (OHPAS) Hypothesis

There are disjoint sets of interlingual, aligned text word forms for each pronunciation of any homograph.

Parallel corpora: OHPAS Hypothesis

Homograph pronunciation labeling using parallel corpora:

- the *lead* /ˈlɛd/ pipe : le tuyau de *plomb*
- take the *lead* /ˈliːd/: prendre l'*initiative*

Parallel corpora: OHPAS formalization

$$pt: \mathscr{P} \mapsto \mathscr{T} \tag{1}$$

$$pt^{-1}: \mathscr{T} \mapsto \{\mathscr{P}\} \tag{2}$$

$$ps: \mathscr{P} \mapsto \{\mathscr{S}\} \tag{3}$$

$$\{S\} \in \mathscr{S} \mid \exists \{S\}, \{S'\} \in \mathscr{S}: \{S\} \cap \{S'\} = 0 \tag{4}$$

OHPAS hypothesis

$$
\begin{aligned}
\mathscr{H} \overset{\text{def}}{=} \{\mathsf{T} \in \mathscr{T} \mid \exists P, P' \in \mathscr{P} \colon (P \neq P') \wedge \\
((\mathcal{ps}(P) = \{S\}) \wedge (\mathcal{ps}(P') = \{S'\})) \wedge \\
(\mathcal{pt}(P) = \mathcal{pt}(P') = \mathsf{T})\}
\end{aligned} \tag{5}
$$

$$
\mathsf{L} \in \mathscr{L} \mid \exists\, \mathsf{L}, \mathsf{L}' \in \mathscr{L} \colon (\mathsf{L} \neq \mathsf{L}') \tag{6}
$$

Parallel corpora: OHPAS hypothesis

$$ipt\colon \mathscr{P} \in \mathsf{L} \mapsto (\forall t \in \{T\}) \in \mathsf{L}' \tag{7}$$

$$ipt^{-1}\colon (\forall t \in \{T\}) \in \mathsf{L}' \mapsto \mathscr{P} \in \mathsf{L} \tag{8}$$

Parallel corpora: OHPAS hypothesis

$$\begin{aligned} \forall\, \mathrm{H} \in \mathcal{H} \colon \forall P, P' \in pt^{-1}(\mathrm{H}) \colon P \neq P' \Rightarrow \\ (pt^{-1}(\mathrm{H}) \mapsto P \mid ipt(P)) \cap (pt^{-1}(\mathrm{H}) \mapsto P' \mid ipt(P')) = 0 \end{aligned} \tag{9}$$

Parallel corpora: OHPAS corollary

$$\begin{aligned}
\forall\, \mathsf{H} \in \mathcal{H}\colon \forall P, P' \in \mathcal{pt}^{-1}(\mathsf{H})\colon P \neq P' \Rightarrow \\
(pt^{-1}(\mathsf{H}) \mapsto P \,|\, ipt^{-1}(ipt(P)) = P) \wedge \\
(pt^{-1}(\mathsf{H}) \mapsto P' \,|\, ipt^{-1}(ipt(P')) = P')
\end{aligned} \tag{10}$$

Parallel corpora: OHPAS corollary

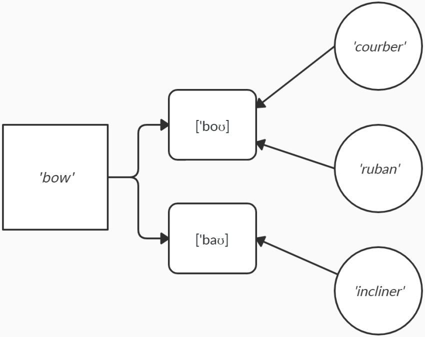


Figure 13: The English homograph ‘bow’, with unidirectional relationships from French alignments to pronunciations

Parallel corpora: AP semi-automated labeling technique

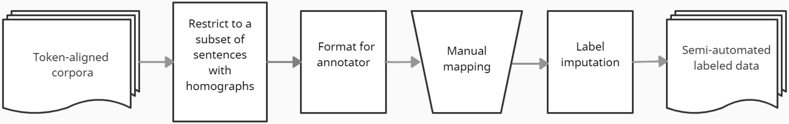


Figure 14: Alignment-to-Pronunciation (AP) labeling technique.

{ `lives-vies` : laɪvz}

Example of AP mapping entry

Parallel corpora: Augmenting WHD train data set with imputed data

Train data set size increases:
20%, *from 2,719 to 3,437 samples*

Augmentation	Homograph	Ratio	Diff
No	associate	88:12	76
Yes	abuses	83:18	65
Yes	associate	88:19	69

Table 6: Median class sample size ratios and difference

Parallel corpora: Modeling

- Regularized multinomial logistic regression (LR)
- Token classifiers

Parallel corpora: Regularized multinomial LR

Features:

- Lowercase:
  - tokens indexed 1 and 2 slots before and behind the homograph token
  - bigrams before and after the homograph token
  - skipgram around homograph token
- POS features for each of the above, and for the homograph
- A case feature for the homograph (uppercase, lowercase, titlecase, or other)

Parallel corpora: Token classifier models

- *ALBERT*, *BERT*, and `XLNet` fine-tuned for token classification
- N+1 labels, masked to reduce selection to 2 at inference

Parallel corpora: Token classification data

Token	Label
Yanowitz	O
was	O
the	O
bass	/ˈbæs/
player	O

Table 7: Example of token-level labeling for token classification task

Results with parallel corpus-based label imputation

Model Type	WHD Bal Acc	Aug WHD Bal Acc	Bal Acc Change
LR	81.23	86.46	5.23
ALBERT	85.84	93.37	7.53
BERT	92.08	94.02	1.94
XLNet	91.17	95.89	4.72

Table 8: 34-homograph-restricted models' balanced accuracy scores on test set, with change in balanced accuracy between models trained only on the WHD and models trained on the augmented WHD. Metrics taken from median of five random seeds.

Error analysis

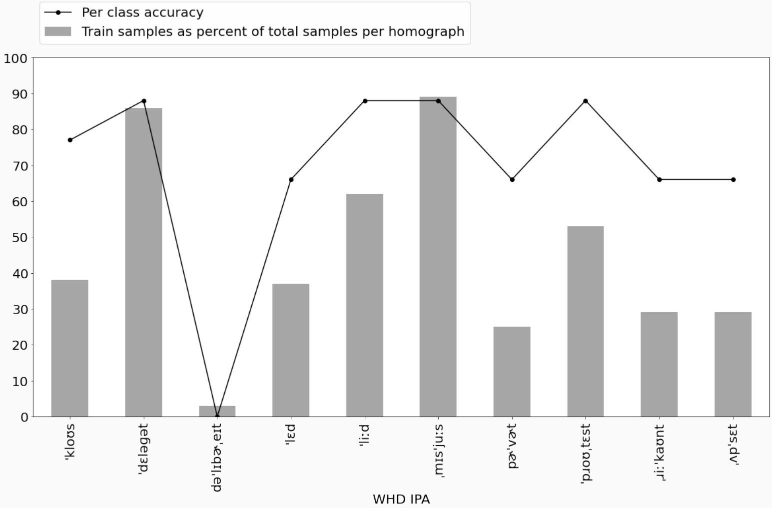


Figure 15: XLNet_Aug_Median train and test set sample sizes for pronunciation classes with under 100% accuracy, with per class accuracy.

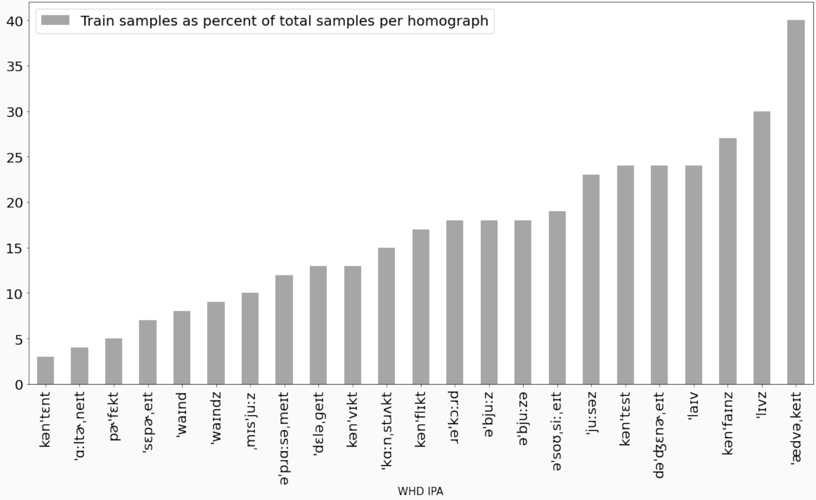


Figure 16: XLNet_Aug_Median train sample sizes for classes with 100% accuracy and training sample sizes that constitute up to 40% of the homograph's training data.

Conclusions

Conclusions

Typology: There are four distinct kinds of homographs.

- Type II homographs: difficult for audio-based label imputation.
- Type II homographs: require pronunciation selection for labeling
- Type IV homographs: Impossible to use POS only for disambiguation.

OHPAS hypothesis: One-to-one alignments exist between interlingual, aligned text word forms and homograph pronunciations

AP label imputation: Target low prevalence classes only for augmentation with imputed data.

Future research

- Non-English homographs
- Homographs with more than two pronunciations
- ‘Non-standard word’ homographs (fractions/dates, years/quantifiers; Sproat et al. 2001, Yarowsky, 1997)
- Algorithm development: Masking during fine-tuning; BiLSTMs (Gorman, p.c.)
- Implement active learning to retrieve samples with a higher likelihood of reducing model uncertainty

Thank you!