LABEL IMPUTATION FOR HOMOGRAPH DISAMBIGUATION: THEORETICAL AND PRACTICAL APPROACHES

by

JEN M. SEALE

A dissertation submitted to the Graduate Faculty in Linguistics in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2021

This manuscript has been read and accepted by the Graduate Faculty in Linguistics in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Professor William Gregory Sakas

Date	Chair of Examining Committee

Professor Cecelia Cutler

Date	Executive Officer

Professor William Gregory Sakas

Professor Kyle Gorman

Professor Alla Rozovskaya

Supervisory Committee

Abstract

LABEL IMPUTATION FOR HOMOGRAPH DISAMBIGUATION: THEORETICAL AND PRACTICAL APPROACHES

by

JEN M. SEALE

Supervisor: Professor William Gregory Sakas

This dissertation presents the first implementation of label imputation for the task of homograph disambiguation using 1) transcribed audio, and 2) parallel, or translated, corpora. For label imputation from parallel corpora, a hypothesis of interlingual alignment between homograph pronunciations and text word forms is developed and formalized. Both audio and parallel corpora label imputation techniques are tested empirically in experiments that compare homograph disambiguation model performance using: 1) hand-labeled training data, and 2) hand-labeled training data augmented with label-imputed data. Regularized, multinomial logistic regression and pre-trained `ALBERT`, `BERT`, and `XLNet` language models fine-tuned as token classifiers are developed for homograph disambiguation. Model performance after training on parallel corpus-based, label-imputed augmented data shows improvement over training on hand-labeled data alone in classes with low prevalence samples. Four homograph disambiguation data sets generated during the work on the dissertation are made available to the research community. In addition, this dissertation offers a novel typology of homographs with practical implications for both the label imputation process and homograph disambiguation.

Acknowledgments

With my sincerest gratitude to Professor William Sakas for welcoming me as a learner, for challenging and sharpening my work, and for keeping me on the path. For Professor Kyle Gorman, without whom this dissertation would not have been written. His own work has led the way. And to Professor Alla Rozovskaya, whose keen insight helped separate the dark from the dark for me, and led to the writing of one of my favorite chapters.

Contents

List of Tables

List of Figures

List of Abbreviations

AL	**A**ctive **L**earning
AP	**A**lignment-to-**P**ronunciation
ASR	**A**utomated **S**peech **R**ecognition
GMN	**G**orman, **M**azovetskiy, and **N**ikolaev
HD	**H**omograph **D**isambiguation
IPA	**I**nternational **P**honetic **A**lphabet
ML	**M**achine **L**earning
MT	**M**achine **T**ranslation
OHPT	**O**ne **H**omonym **P**er **T**ranslation
OHPAS	**O**ne **H**omograph **P**ronunciation **P**er **A**lignment **S**et
SOTA	**S**tate-**O**f-**T**he-**A**rt
SWBD	**S**witchboard
TTS	**T**ext-**T**o-**S**peech
WHD	**W**ikipedia **H**omograph **D**ata
WSD	**W**ord **S**ense **D**isambiguation

List of Symbols

$\mathcal{H}$	homograph text word forms in a language
$\mathcal{L}$	elements of a language
$\mathcal{P}$	pronunciations in a language
$\mathcal{S}$	senses, or meanings
$\mathcal{T}$	text word forms in a language
ipt	function that maps a pronunciation to any text word form in its set of aligned, interlingual text word forms
ipt^{-1}	function that maps a text word form to its aligned, interlingual pronunciation
ps	function that maps a pronunciation to a set of senses
ps^{-1}	function that maps a sense to its pronunciation
pt	function that maps a pronunciation to its text word form
pt^{-1}	function that maps a text word form to its pronunciations

Dedicated to my family—to those who have passed away,

those who share the present, and those who are to come.

Chapter 1

Introduction

1.1 Problem and approaches

This work advances the state of empirical research on the development of label imputation in addressing low resource data with imbalanced classes for the purpose of homograph disambiguation (HD). A homograph is a text word form that has multiple pronunciations associated with distinct sense sets. The task of disambiguating homographs is that of determining which pronunciation is correct given the text surrounding the homograph. The homographs in this work are taken from the Wikipedia Homograph Data (WHD), released by Gorman et al. (2018). The WHD exhibits two properties which this research is designed to address, 1) it is a low-resource data set, with around 100 data samples per homograph, and 2) its pronunciation classes are predominately imbalanced. Over 65% of the homographs exhibit very low resource pronunciation classes, with smaller classes in the WHD containing as little as one sample per data set split. Either one of these issues present problems for supervised machine learning (ML), and both issues are prevalent in a great deal of training data in use for natural language modeling. Supervised machine learning models trained

on imbalanced classes tend to learn to predict the larger of the classes. Low sample size restricts the capacity for model generalization.

In order to address challenges due to low resource data in homograph disambiguation modeling, two distinct approaches to label imputation are taken, 1) automated and semi-automated pronunciation label generation from transcribed audio data, 2) semi-automated label imputation from parallel corpora using an imputation technique developed on a novel hypothesis of interlingual homograph alignment. While the first approach proves promising, the second approach is proven to be a useful technique in label imputation. The results of the hypothesis-based semi-automated label imputation are investigated to see if they align with the hypothesis through manual inspection of the label-imputed training data, and a comparison of baseline models trained solely on the human-labeled WHD to ML models trained on the WHD augmented with label-imputed data. Balanced accuracy is used to treat imbalance in the classes, rectifying an obfuscation of performance on lower prevalence classes that occurs when using metrics that do not take sample size into account. In addition, per class accuracy and error analysis provide an in depth look at the impact of modeling with the augmented data. The results provide evidence towards the hypothesis. The generated data conforms to expectations, and model performance is seen to improve over the established baseline.

1.2 Motivation

While machine learning in natural language processing (NLP) has made a good deal of advancement in the last few decades, there is something that has not yet changed, the need for both quality and quantity in the data from which these models generalize. If anything, the need for quantity has grown, and while human labeling has remained the most trusted source for this data, there are many reasons to pursue a more automated approach in order to obtain the amount of data useful for contemporary NLP models. This section provides a discussion on reasons for embarking on ho-

mograph disambiguation in the first place, followed by the motivating factors for the development of homograph pronunciation label imputation.

1.2.1 Reasons for homograph disambiguation research

There is a great deal of opportunity to increase the accessibility of technology through smart home devices and computers equipped for voice interaction. Spoken communication is more intuitive for users and requires less dedicated attention, freeing up one's hands from a device interface. Visually impaired populations, as well as more elderly populations with less technical acumen, stand to benefit a good deal from such advances in usability.

Performance of dialogue-based interfaces is degraded when the device generates incorrect pronunciations for words with multiple pronunciations and meanings, which is why homograph disambiguation is important in text-to-speech (TTS) applications. While Gorman et al. (GMN; 2018) achieve micro and macro accuracies of 99% on the Wikipedia homograph disambiguation task with hybrid rule-based and logistic regression (LR) models, and 95% using LR models alone, these approaches involve feature and rule development that, while providing greater model interpretability, are not easily generalizable outside the task for which they're initially developed and can be time, labor and cost intensive. Current state-of-the-art (SOTA) transformer neural net models are of interest for use in text-to-speech applications as the models learn directly from, albeit labeled, natural language data, rather than language-derived features or human-generated rules.

There has also not yet been a great deal published on the use of SOTA models for homograph disambiguation. Dai et al. (2019) and Sun et al. (2019) explore the use of BERT (Devlin et al., 2019) in various ways for pronunciation selection, but do not fine-tune for a named entity/token classification task, or investigate newer BERT variants (both are done in this work). Tangentially, Emelin et al. (2020) investigate transformer models for the prediction of word sense disambiguation errors. There is also a good bit of work with transformer models for the purpose of grapheme to phoneme conversion in the findings of the SIGMORPHON 2020 task (Gorman et al., 2020).

However, homograph disambiguation itself, specifically framed as a token classification, or named entity, task using transformer models, is, to the best of the author's knowledge, not currently attested to in peer reviewed literature or otherwise published online.

In addition to its importance in TTS applications, homograph disambiguation is relevant to automatic speech recognition (ASR). For example, the sentential context of somewhat garbled or otherwise less interpretable frequencies of transcribed spoken homographs can be used to boost the probability of pronunciation selection. Homograph disambiguation is also a subset of word sense disambiguation (WSD). The pronunciations of homographs are associated with one or multiple meanings, and so the selection of a pronunciation is a selection of a sense or group of related senses. As the capacity to disambiguate word senses has relevance for many natural language processing and understanding tasks, advances in homograph disambiguation contributes to our understanding and capabilities in those areas as well (p.c., Gorman 2021).

1.2.2 Reasons for label imputation development

As mentioned in section 1.2.1, at the current time most SOTA natural language processing models require labeled language data for parameter estimation. Label imputation is the process of automatedly generating labels, which can mitigate issues surrounding cost and time in annotation initiatives that involve developing annotation instruction materials, training annotators, and discerning inter-rater agreement along with human-applied label accuracy. Often enough, even at great cost (especially for annotation requiring specialized subject matter experts, such as medical doctors), relatively little human-labeled data is generated compared to the amount needed for models to generalize well. With low resource data, where the sample size per class is low, the models often do not have enough data to generalize from, and overfit to the specific examples available in the training data, making for poor performance at inference time. These realities make label imputation, or automated labeled data generation, an attractive goal of research today, as can be seen with work such as that coming out of Stanford with Snorkel (Ratner et al., 2020) to generate, clean,

and train with imputed data, also known as silver label or weak supervision data. While automated label generation is desirable, it doesn't have the same authoritative quality that human-selected labels can have. In the research done for this dissertation, human input is leveraged automatedly to increase the amount of work done while decreasing the over all amount of human effort required, and adding more reliability to the imputation. The semi-automated approaches leveraging this human effort are now introduced in more depth.

1.3 Semi-automated label imputation

While pronunciation label imputation from audio data is mentioned for future work by Gorman et al. (2018), the research recorded herein is, to the best of the author's knowledge, the first recorded attempt at this task. During the audio-based homograph pronunciation label imputation IPA representations are generated without human input. However, while some of the representations come close to the WHD labels, none of them match completely. Due to this mismatch, a manual mapping is made between generated IPA and WHD IPA, and used to label homographs in the transcribed audio. While enough data to significantly improve accuracy metrics via the augmented data is not generated, some increase in accuracy is seen, indicating that further research and development may prove profitable. In addition, 1) human-labeled data is generated and used for modeling and model comparison, 2) observations as to the usefulness of the kind of language in the audio are made based on modeling using the hand-labeled data, and 3) groundwork is set for further efforts into audio-based label imputation.

In contrast, semi-automatedly label-imputed data obtained through the *Alignment-to-Pronunciation* (AP) label imputation technique (described in section 5.3) increases balanced accuracy metrics by up to 7.5% when used to augment hand-labeled data. This technique is based on the *One Homograph Pronunciation Per Alignment Set* (OHPAS) hypothesis found in sections 5.1–5.2.1. The OHPAS hypothesis is formalized, the AP technique developed, and empirical evidence towards

both is provided through manual inspection of the generated data and model performance.

1.4 Research contributions

This dissertation serves as the first academic work focused on label imputation specifically for homograph disambiguation using transcribed audio, and parallel corpora. Its contributions include: a typology of homographs with implications for both label imputation and for homograph disambiguation, a formalized hypothesis of interlingual alignment between homograph pronunciations and text word forms, pronunciation label imputation from both audio and parallel corpora coupled with empirical evidence on the efficacy of the techniques, the development of multiclass token classifier homograph disambiguation models, and the generation of new data sets to be made publicly available for further experimentation by the NLP research community.

In addressing the use of POS as a vehicle for disambiguation, a typology of homographs is developed based on the relationships between homograph POS, sense sets, and pronunciations. This typology clearly delineates between those homographs for which POS may be used as the sole tool to select pronunciations and those for which it may not while simultaneously offering a rationale for these expressions. Additionally, the typology illuminates difficulties encountered in the imputation of labels from recorded conversational speech. The research then goes on to implement the first recorded experiment in academic literature in homograph label imputation from transcribed audio data. A pipeline for automated label imputation is developed, revised to incorporate human decision-making, and employed to generate labeled data which is then used in disambiguation modeling, the performance of which is compared against baselines. While the audio-based, semi-automated pronunciation labeling does not lead to a significant increase in model performance, the study breaks ground for future research.

A separate approach to label imputation is then implemented based on a novel hypothesis of interlingual homograph pronunciation and text word form alignment, the One Homograph Per

Alignment Set (OHPAS) hypothesis. The hypothesis is formalized and empirical evidence towards its accuracy is provided in an analysis of mapped pronunciations to aligned, interlingual tokens. A semi-automated labeling technique, Alignment-to-Pronunciation (AP) labeling, developed to leverage OHPAS-hypothesized alignment, is used to obtain imputed data. Increased performance in balanced accuracy for a number of different ML homograph disambiguation models then provides evidence of the utility of the AP labeling technique in providing data which improves approximation on lower prevalence classes when it is used to augment human-labeled data.

Four data sets generated during the course of this work are made publicly available on GitHub.[1] The 34-homograph, three split version of the original two split Wikipedia Homograph Data used in this work is released along with the three augmented train splits used for model comparison in the experimentation. Three additional data sets, labeled using the WHD pronunciation label inventory, are also made available. One of these data sets consists of hand-labeled, sentence-level transcribed audio from the Switchboard corpus (Godfrey et al., 1992), and two contain subsets of the French-English Europarl corpus (Koehn, 2005), one hand-labeled and the other semi-automatedly labeled.

In sum, this work provides insight into both the classification of homographs and a phenomenon of interlingual alignment in ways that have direct application for pronunciation label imputation. It pioneers audio-based label imputation and goes on to effectually address issues arising from low resource homograph disambiguation training data through a semi-automated label imputation technique developed on a novel hypothesis of interlingual homograph pronunciation alignment. The successful technique improves balanced accuracy metrics for multiple ML homograph disambiguation models trained on human-labeled data augmented with the label-imputed training data. Finally, this research fosters further innovation through making new homograph disambiguation data sets available to the public.

The remainder of the dissertation is organized as follows. Chapter 2 contextualizes this research

[1] See: `https://github.com/jseale/homograph_label_imputation_data` for the labeled homograph data.

within the fields of homograph and homonym disambiguation as well as label imputation in word sense disambiguation. Chapter 3 addresses the use of POS in disambiguation and delineates four types of homographs based on POS, pronunciation and sense set relationships. Chapter 4 then provides an account of experimentation in audio-based label imputation. The OHPAS hypothesis is formalized and the AP labeling technique outlined in chapter 5. Empirical evidence towards the efficacy of AP labeling in improving accuracy for pronunciation classes with low prevalence sample sizes is recorded in chapter 6. A discussion of the work including paths for future research is then found in chapter 7.