Web-derived Pronunciations

Arnab Ghoshal

Spoken Langauge Systems,
Saarland University

Research conducted during JHU Summer Workshop, 2008, together with:
Michael Riley, Martin Jansche, Sanjeev Khudanpur, Morgan Ulinski

October 28, 2009

Pronunciation Generation - Approaches

- **Previous Approaches:**
  - Use trained persons to manually generate pronunciations — *expensive*
  - Use rules that are hand-crafted or machine-learned from a manually-transcribed corpus — *variable quality*

Pronunciation Generation - Approaches

- **Previous Approaches:**
  - Use trained persons to manually generate pronunciations — *expensive*
  - Use rules that are hand-crafted or machine-learned from a manually-transcribed corpus — *variable quality*
- **Our Approach:** Find pronunciations derived from the web
  - IPA Pronunciations: Uses International Phonetic Alphabet: `Lorraine Albright` /ˈɔːl braɪt/
  - Ad-hoc Pronunciations: Uses informal pronunciation: `bruschetta (pronounced broo-SKET-uh)`

Web-Derived Pronunciations - Processing Steps

The following steps are needed for both web IPA and Ad-hoc pronunciations:

1. **Extraction:** Find the pronunciation and its corresponding orthographic pair on a web page.

Web-Derived Pronunciations - Processing Steps

The following steps are needed for both web IPA and Ad-hoc pronunciations:

1. **Extraction:** Find the pronunciation and its corresponding orthographic pair on a web page.
2. **Extraction Validation:** Determine if orthographic- pronunciation pair is correctly extracted - was the web page author offering a pronunciation and were the right words extracted?

   `Bazell (pronounced BRA-zell by the lisping Brokaw)`

Web-Derived Pronunciations - Processing Steps

The following steps are needed for both web IPA and Ad-hoc pronunciations:

1. **Extraction:** Find the pronunciation and its corresponding orthographic pair on a web page.
2. **Extraction Validation:** Determine if orthographic- pronunciation pair is correctly extracted - was the web page author offering a pronunciation and were the right words extracted?
   `Bazell (pronounced BRA-zell by the lisping Brokaw)`
3. **Pronunciation Validation/Normalization:** Determine if the pronunciation the web page author provided is plausible and correctly transcribed. Normalize if possible.
   `it's lunchtime, and I'm craving a nice Italian sausage (pronounced sauseege)`
   `"Hayn" is pronounced "Hawaiian"`

Letter-To-Phone Models

- **Approach:** Build *n*-gram transduction models over aligned pairs of orthographic and phone symbols. Deligne & Bimbot, 1997 Bisani & Ney, 2002
- ***N*-grams from aligned pairs:**

n	a	t	i	o	n
n	ey	sh	-	ax	n

- Same approach used for other letter-to-phone and phone-to-phone models to follow.

Extraction - Web IPA Pronunciations

- Identify terms within '[...]', '/.../' or '\...\' that contain one or more IPA Unicode symbols on English web pages.
- Use a letter-to-phone (L2P) finite-state transducer that models $\Pr(\texttt{orth} \mid \pi)$ to find the best nearby orthographic term ($\texttt{orth}$) that matches the IPA-containing phone terms (π).
- Good precision at the expense of recall.
- 3M English extractions, 370K unique ortho-pron pairs
- 165K unique words, 124K (75%) of those not in Pronlex

Extraction - Ad-hoc Pronunciations

- Identify terms that match regular expressions such as:

Pattern	Count
`\(pronounced (as |like )?([^)]+) \)`	3415K
`pronounced (as |like )?"([^"]+)"`	835K
`, pronounced (as |like )?([^,]+),`	267K

- Use a letter-to-phone finite-state transducer that models $\Pr(\texttt{orth}_2 \mid \texttt{orth}_1)$ to find the best nearby orthographic term ($\texttt{orth}_2$) that matches the ad-hoc pronunciation term ($\texttt{orth}_1$).
- $\Pr(\texttt{orth}_2 \mid \texttt{orth}_1) = \sum_\pi \Pr(\texttt{orth}_2 \mid \pi) \Pr(\pi \mid \texttt{orth}_1)$ [under a suitable independence assumption], which we create from our previous finite state models by weighted FST composition.
- 4.5M extractions, 740K unique ortho-pron pairs
- 392K unique words, 372K (95%) of those not in Pronlex

Validation of IPA Extraction

Goal: After extraction has taken place, filter out incorrect extractions.

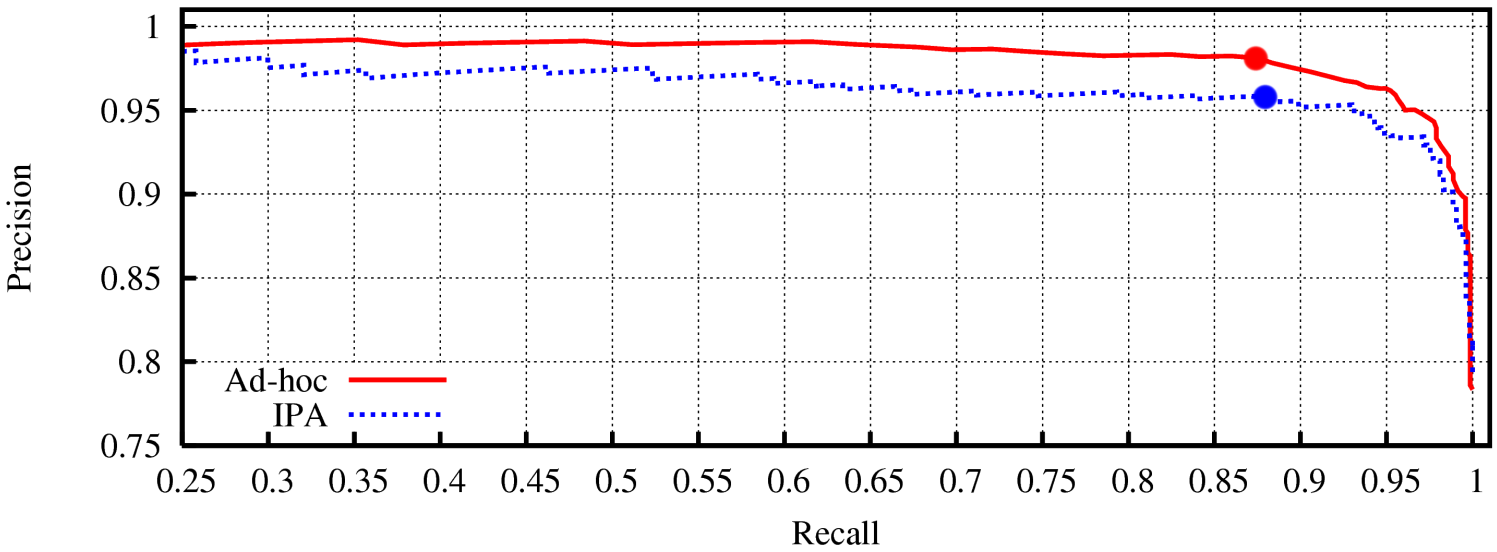
1. Hand-annotate 667 examples
2. Train SVM classifier with 16 Features
   - Language model score
   - Distance between orthography and IPA pronunciation
   - Length of orthography and IPA pronunciation
   - Presence of space in raw orthography
   - Alignment-based features
     - Use LTS model to predict pronunciation from extracted orthography
     - Align predicted pronunciation with extracted IPA
     - Divide phones into two classes, consonants and vowels
     - Use normalized consonant-vowel features
3. Results (5-fold cross-validation)
   - 85.8% accuracy, 99.6% recall, 85.0% precision

Validation of Ad hoc Extraction

Goal: After extraction has taken place, filter out incorrect extractions.

1. Hand-annotate 1000 examples
2. Train SVM classifier with 57 Features
   - Language model scores
     - Pr(ortho|adhoc) based on unigram, bigram, and trigram models
     - Per-phone alignment scores
     - Num. insertions and deletions in best orthography-adhoc alignment
   - Counts
     - Orthography, Ad hoc, Domain
   - Presence of function words and non-alphabetic characters
   - Distance between orthography and ad hoc pronunciation
   - Capitalization style of orthography and ad hoc pronunciation
   - ...
3. Results (5-fold cross-validation)
   - 93.7% accuracy, 95.9% recall, 95.3% precision

Validation of Ad hoc Extraction: Precision/Recall



- In extracting pronunciations from the web, there are always going to be errors.
- After extraction has taken place, we can successfully filter out nearly all of these errors using SVM models.

Validating Web-IPA pronunciations

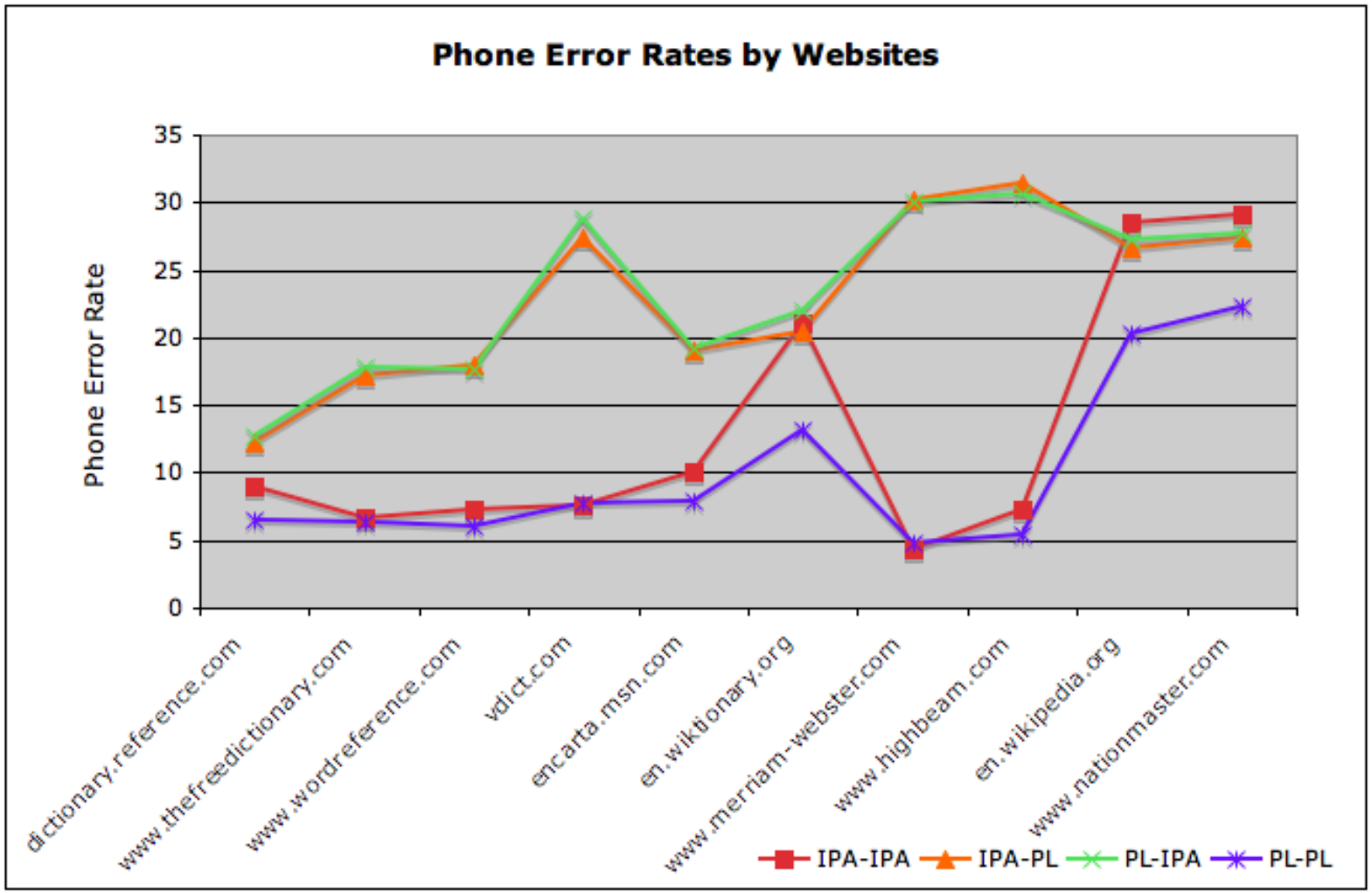
- **Experiment:** Compare L2P models built from Pronlex vs Web-IPA, on their orthographic intersection.
- **Pronlex:** 89K words, 97K pronunciations
- **Web-IPA:** 97K words, 133K pronunciations (subset)
- **Intersection between Pronlex & Web-IPA:** 30K words, 32K Pronlex pronunciations, 56K Web-IPA pronunciations
  - 5-fold cross-validation experiments done on the intersection
  - Polygram-based L2P models

Validating Web-IPA pronunciations

- **Experiment:** Compare L2P models built from Pronlex vs Web-IPA, on their orthographic intersection.
- **Pronlex:** 89K words, 97K pronunciations
- **Web-IPA:** 97K words, 133K pronunciations (subset)
- **Intersection between Pronlex & Web-IPA:** 30K words, 32K Pronlex pronunciations, 56K Web-IPA pronunciations
  - 5-fold cross-validation experiments done on the intersection
  - Polygram-based L2P models

	PL-TRN	IPA-TRN
PL-TST	6.35	17.10
IPA-TST	14.33	12.98

Pronlex vs. Web-IPA: Per site results



How to pronounce graduate?

Sources	Pronunciations
dictionary.reference.com	g r aa d y uw ey t
www.wordreference.com	g r ae d y uh ih t
en.wiktionary.org	g r ae d y uw ax t
www.thefreedictionary.com	g r ae d y uw ih t
encarta.msn.com	g r ae jh uh ax t
en.wikipedia.org	g r ae jh uw ey t
www.pearson.ch	r d uw ax t
Pronlex	g r ae jh uw ey t
Pronlex	g r ae jh uw ih t

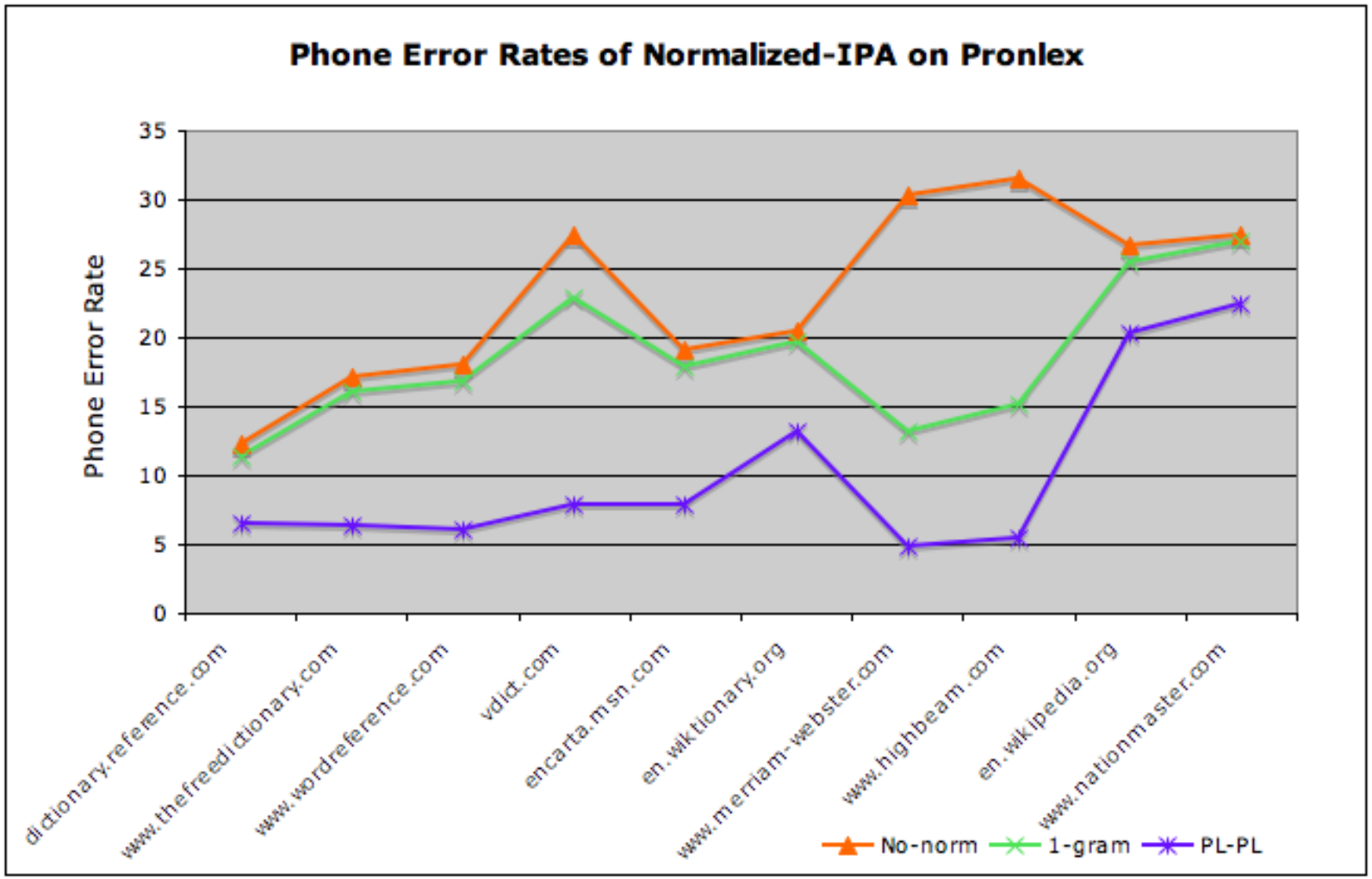
- Pronunciation variability across sources may cause systematic "errors".

How to pronounce graduate?

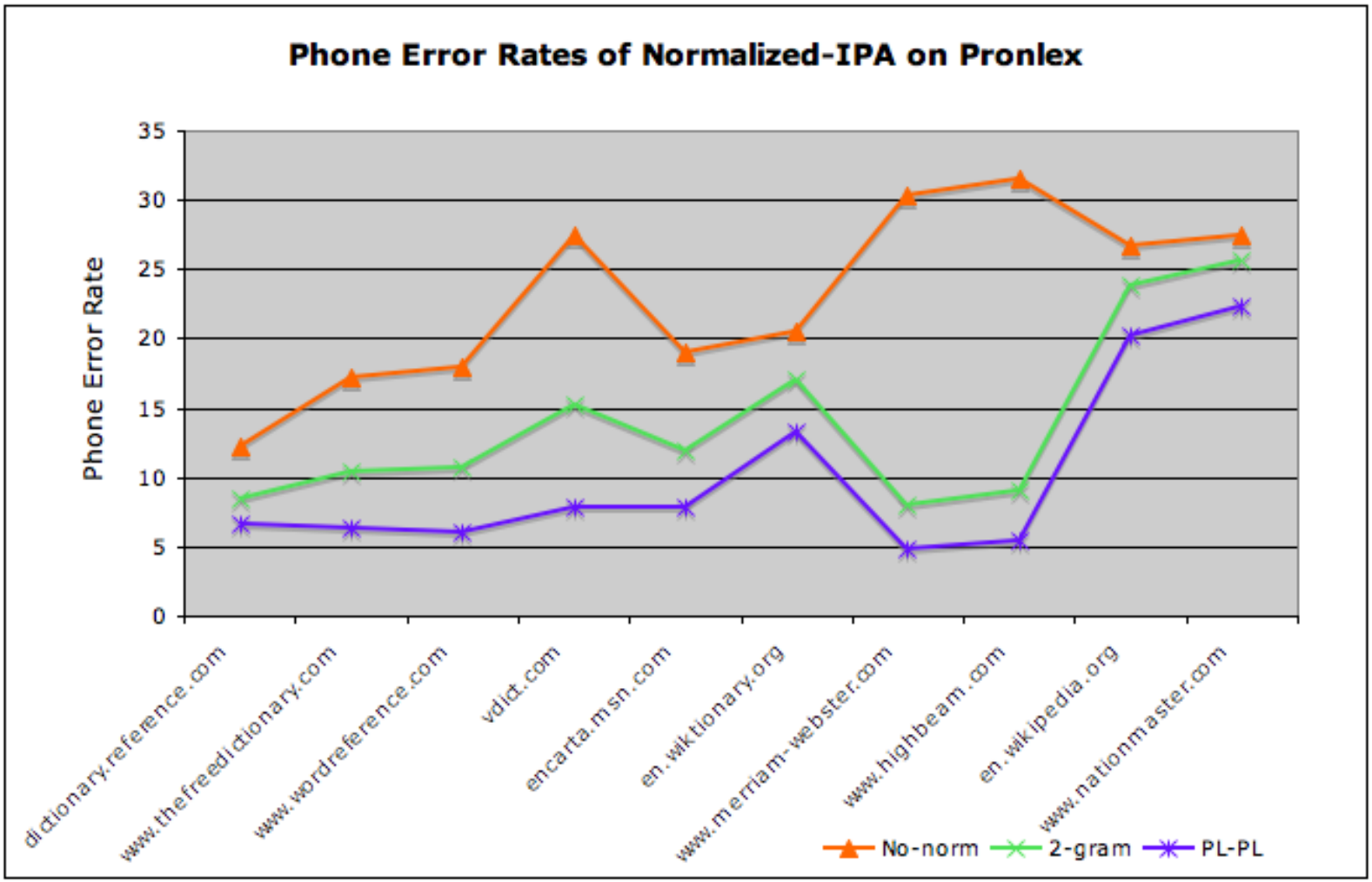
Sources	Pronunciations
dictionary.reference.com	g r ae jh uw ey t
www.wordreference.com	g r ae jh uw ey t
en.wiktionary.org	g r ae jh uw ax t
www.thefreedictionary.com	g r ae jh uw ih t
encarta.msn.com	g r ae jh uw ey t
en.wikipedia.org	g r ae jh uw ey t
www.pearson.ch	r ae jh uw ih t
Pronlex	g r ae jh uw ey t
Pronlex	g r ae jh uw ih t

- Possible to fix source-variability by normalizing Web-IPA pronunciations to Pronlex.

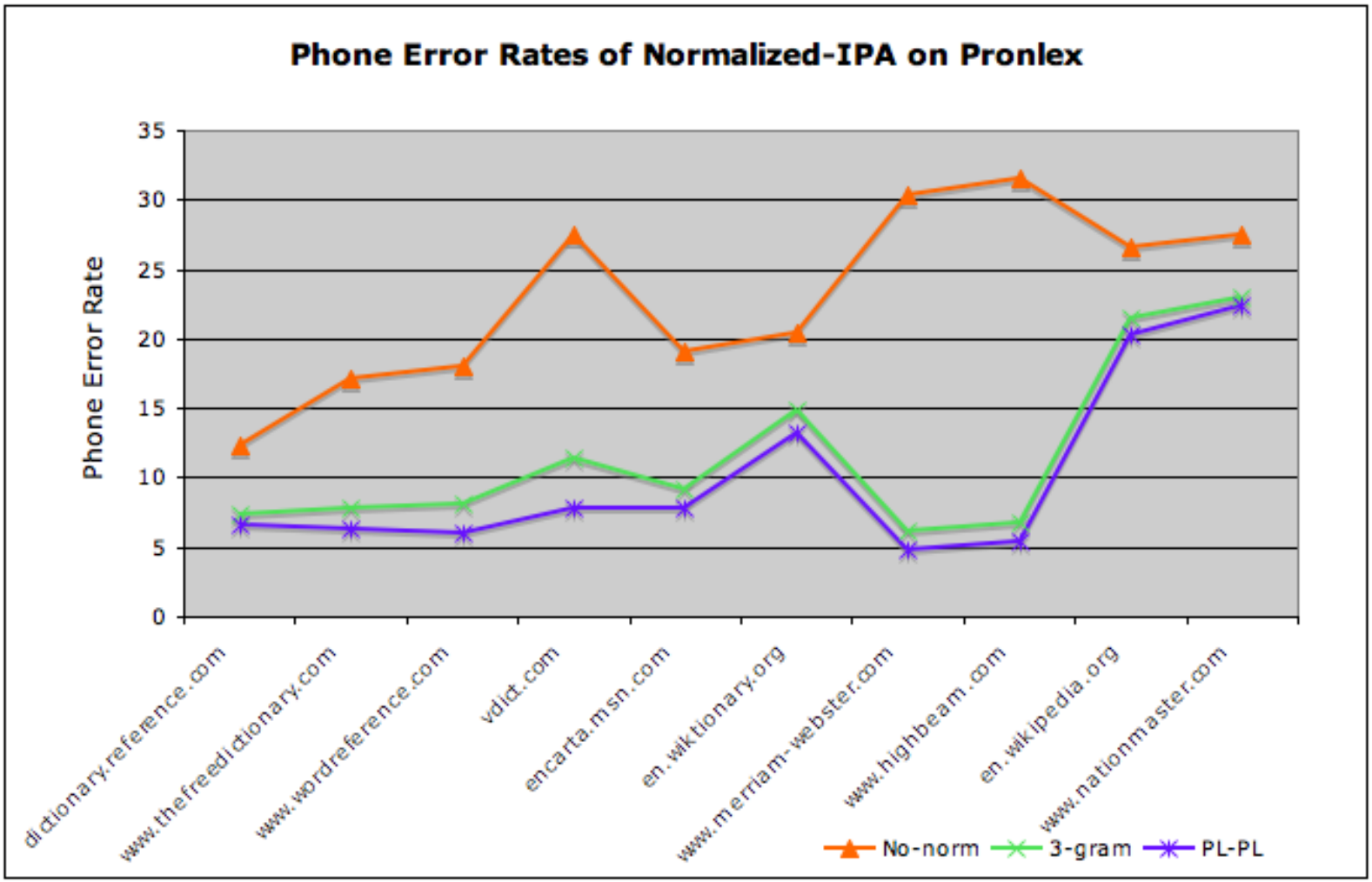
Pronlex vs. Web-IPA: Normalized phoneset 1-gram



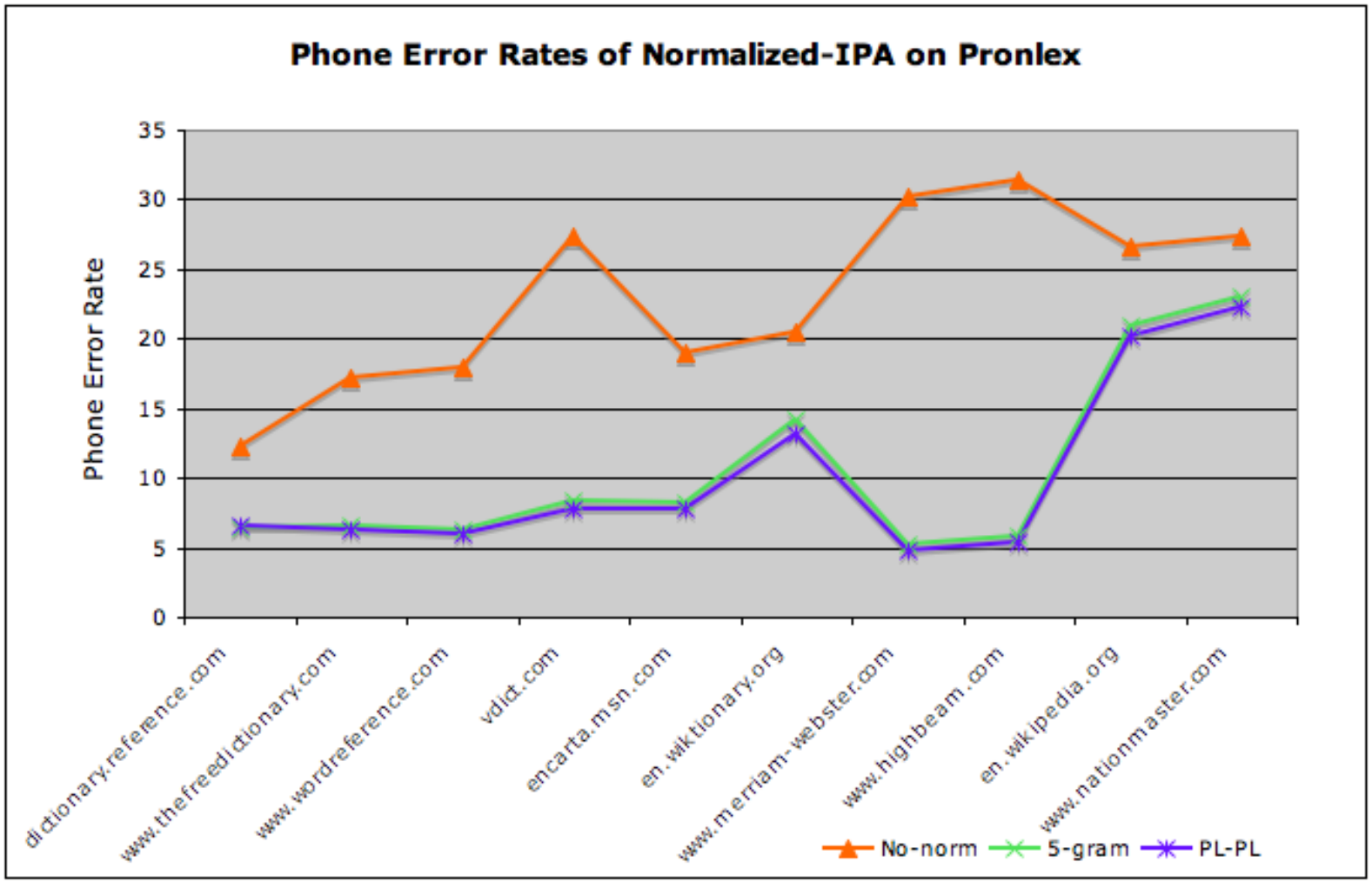
Pronlex vs. Web-IPA: Normalized phoneset 2-gram



Pronlex vs. Web-IPA: Normalized phoneset 3-gram



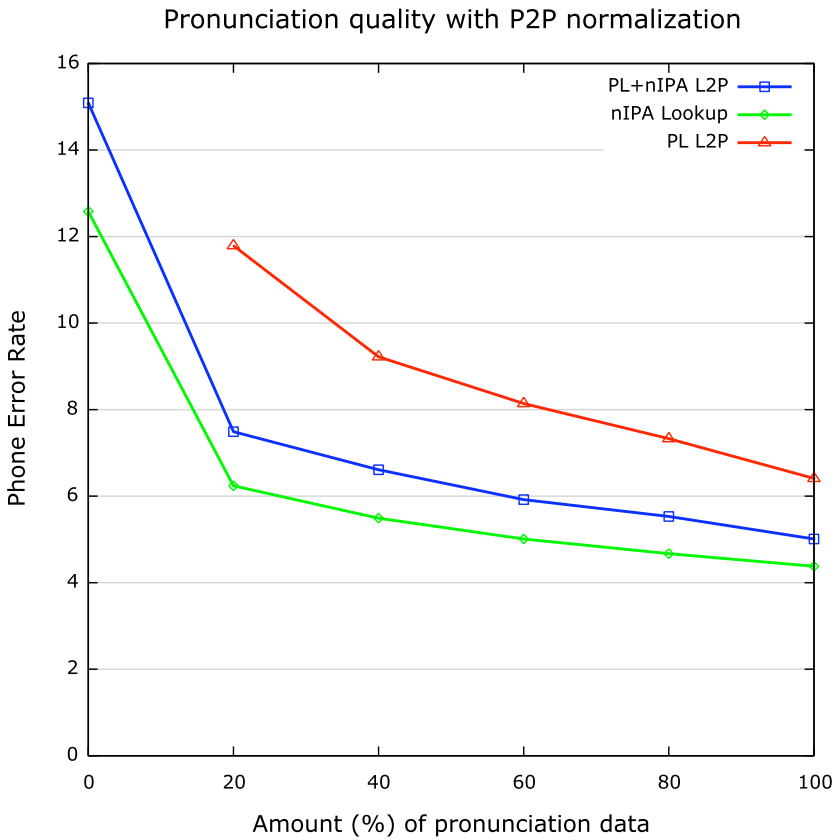
Pronlex vs. Web-IPA: Normalized phoneset 5-gram



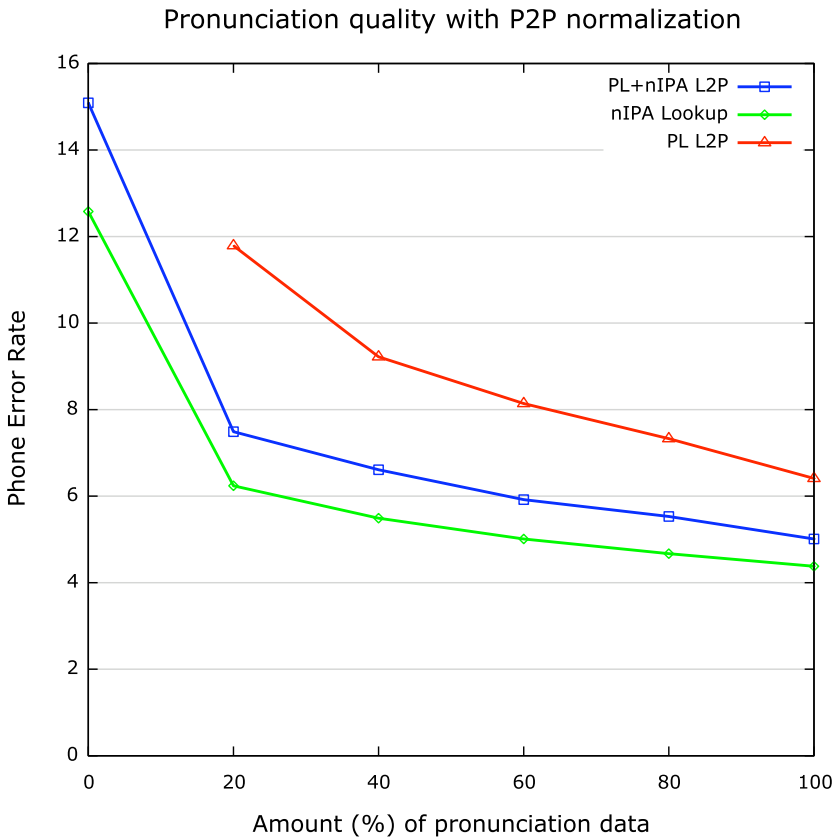
Pronlex vs. Web-IPA: On rare words

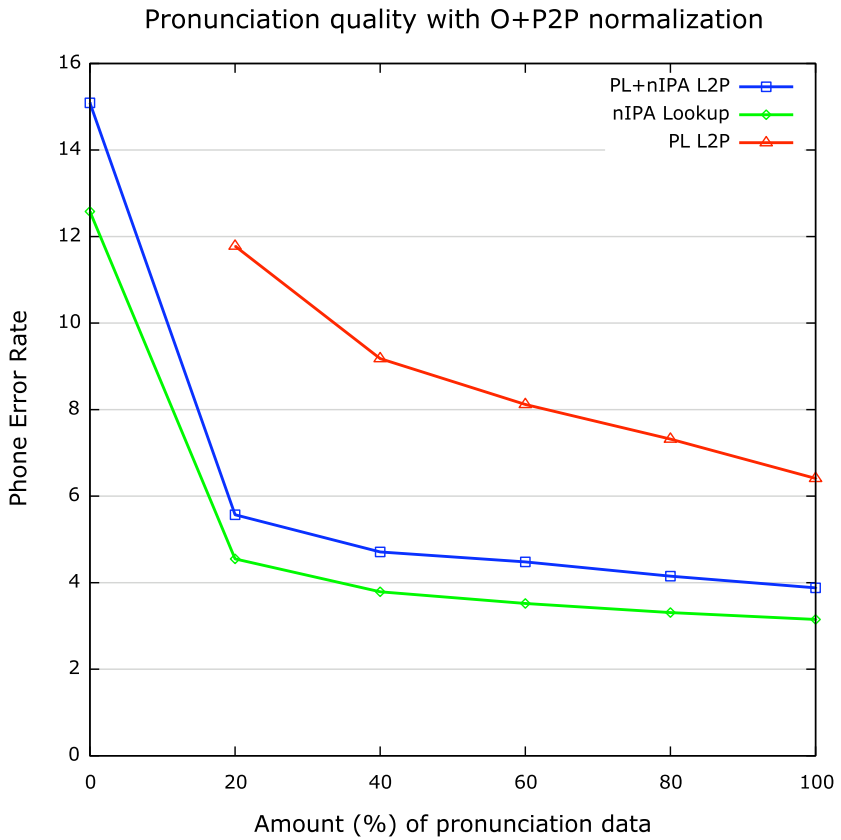
- `Pronlex Training:` Pronlex words with BN count ≥ 2, divided into subsets of size 20%, 40%, 60%, 80%, & 100%.
- `Web-IPA:` 126K words, 179K pronunciations
- `Test:` Pronlex words with BN count ≤ 1.
- **Experiments:** Find pronunciations of words in `Test` using:
  1. L2P model trained on `Pronlex Training`.
  2. Normalize `Web-IPA` using intersection with `Pronlex Training`, and look up the pronunciations of words in `Test`.
  3. L2P model trained on combination of `Pronlex Training` and normalized `Web-IPA`.

On rare words: Learning rate

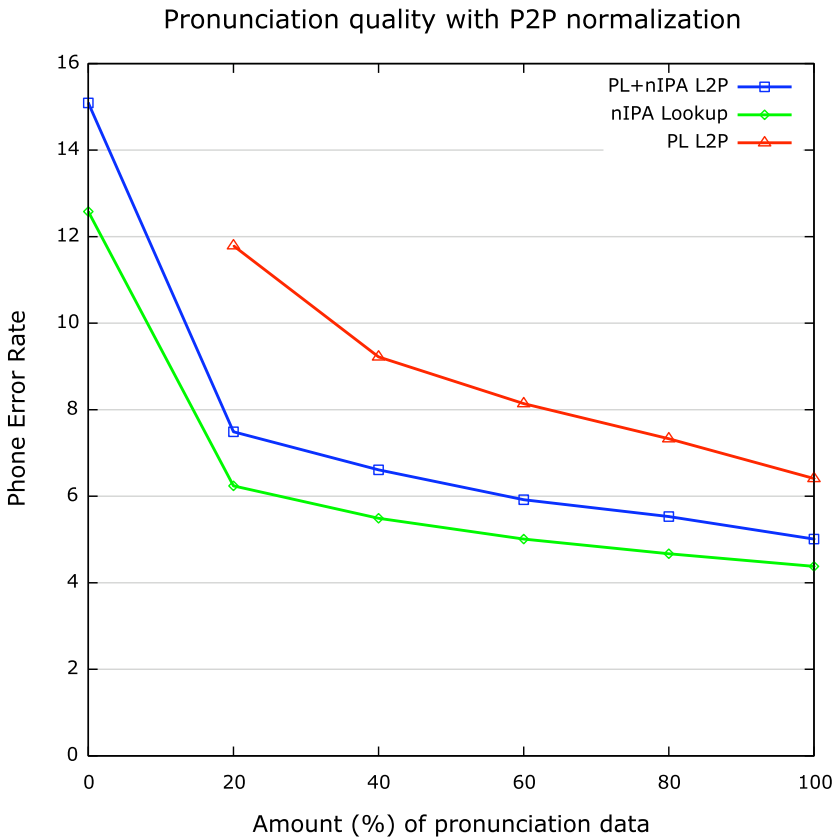


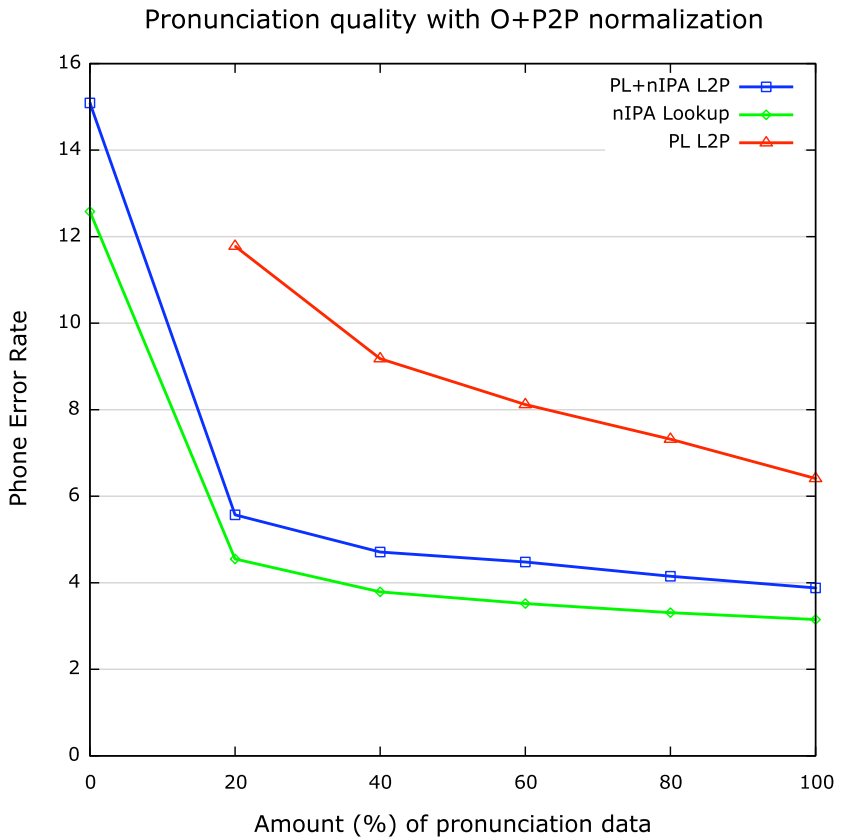
On rare words: Learning rate





On rare words: Learning rate





- Good pronunciation candidates can be found on the web (after normalization).

Ad-Hoc Pronunciation Validation: Phone Prediction

Given an ortho/ad-hoc pair, can predict phones in different ways:

1. **Models on pairs:**
   - Predict from ortho. (Train an L2P model on a standard pronunciation dictionary.)

Ad-Hoc Pronunciation Validation: Phone Prediction

Given an ortho/ad-hoc pair, can predict phones in different ways:

1. **Models on pairs:**
   - Predict from ortho. (Train an L2P model on a standard pronunciation dictionary.)
   - Predict form ad-hoc. (Pair ad-hoc with phones, based e.g. on lookup of parts, train an L2P model.)

Ad-Hoc Pronunciation Validation: Phone Prediction

Given an ortho/ad-hoc pair, can predict phones in different ways:

1. **Models on pairs:**
   - Predict from ortho. (Train an L2P model on a standard pronunciation dictionary.)
   - Predict form ad-hoc. (Pair ad-hoc with phones, based e.g. on lookup of parts, train an L2P model.)
2. **Models on triples:**
   - Use a noisy-stereo-channel model: assume a source of phones that get transmitted over two conditionally independent channels, one turning them into ortho, one turning them into adhoc. Restore phones from observed ortho/ad-hoc pairs.
     $\Pr(\texttt{ortho}, \texttt{adhoc}, \pi) = \Pr(\texttt{ortho} \mid \pi) \Pr(\texttt{adhoc} \mid \pi) \Pr(\pi)$

Ad-Hoc Pronunciation Validation: Phone Prediction

Given an ortho/ad-hoc pair, can predict phones in different ways:

1. **Models on pairs:**
   - Predict from ortho. (Train an L2P model on a standard pronunciation dictionary.)
   - Predict form ad-hoc. (Pair ad-hoc with phones, based e.g. on lookup of parts, train an L2P model.)
2. **Models on triples:**
   - Use a noisy-stereo-channel model: assume a source of phones that get transmitted over two conditionally independent channels, one turning them into ortho, one turning them into adhoc. Restore phones from observed ortho/ad-hoc pairs.
     $\Pr(\texttt{ortho}, \texttt{adhoc}, \pi) = \Pr(\texttt{ortho} \mid \pi) \Pr(\texttt{adhoc} \mid \pi) \Pr(\pi)$
   - Train a language model over aligned (ortho, ad-hoc, phone) triples. Compose with ortho, then with ad-hoc, and decode.

Ad-Hoc Pronunciation Validation: Evaluation

Evaluate on those OOV words that have ad-hoc transcriptions, with reference pronunciations generated by a human:

- Small reference dictionary (256 entries, 1181 phones).
- Difficult words: rare (e.g. *phenylpropanolamine*), unusual pronunciations (e.g. *racicot*/*roscoe*).
- Expect to see high phone error rate.

Ad-Hoc Pronunciation Validation: Evaluation

Evaluate on those OOV words that have ad-hoc transcriptions, with reference pronunciations generated by a human:

- Small reference dictionary (256 entries, 1181 phones).
- Difficult words: rare (e.g. *phenylpropanolamine*), unusual pronunciations (e.g. *racicot*/*roscoe*).
- Expect to see high phone error rate.

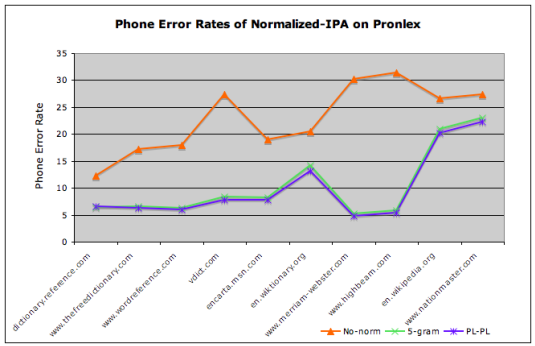
Method	Phone error rate (%)
ortho-to-phone	29.5
ad-hoc-to-phone	20.5
noisy stereo channel	19.4
language model on triples	18.8

Conclusion

1. Lots of human-supplied pronunciations are available via the web.

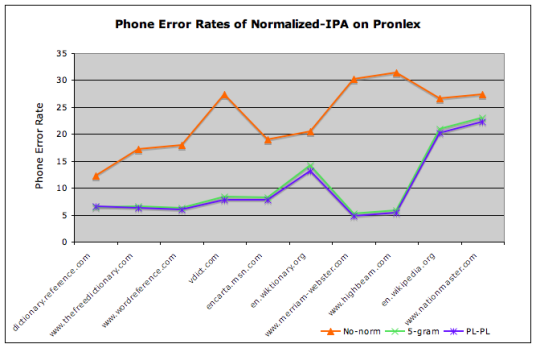
Conclusion

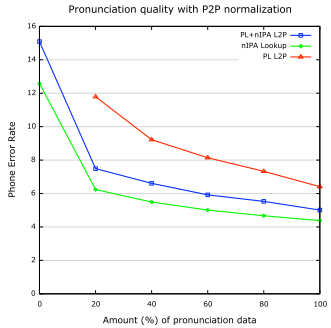
1. Lots of human-supplied pronunciations are available via the web.
2. IPA-usage often varies in site-specific manner.

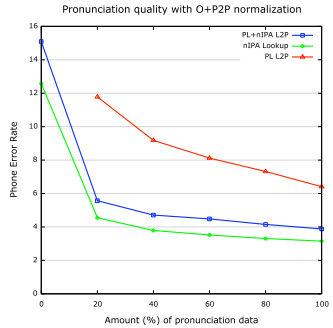


Conclusion

1. Lots of human-supplied pronunciations are available via the web.
2. IPA-usage often varies in site-specific manner.
   - Normalized web-IPA pronunciations are of high quality.

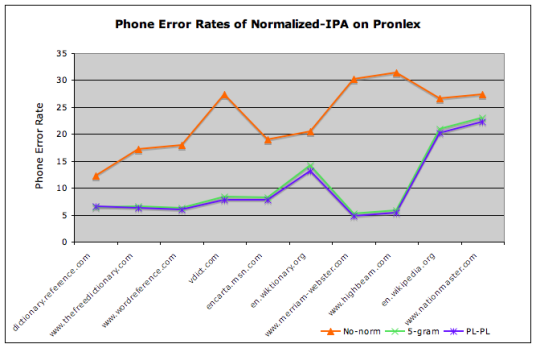


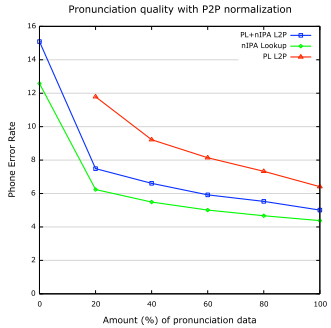


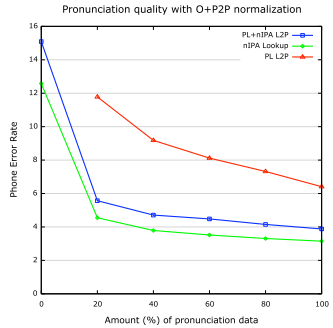


Conclusion

1. Lots of human-supplied pronunciations are available via the web.
2. IPA-usage often varies in site-specific manner.
   - Normalized web-IPA pronunciations are of high quality.
3. Ad hoc pronunciations have rich coverage of low-frequency words

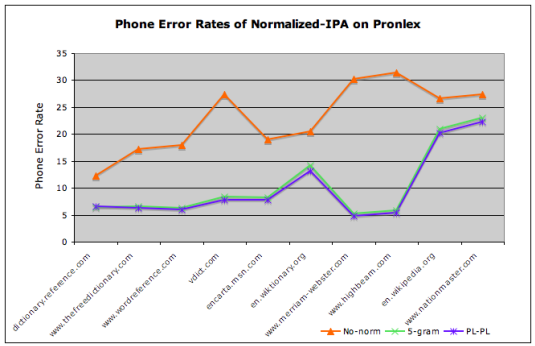


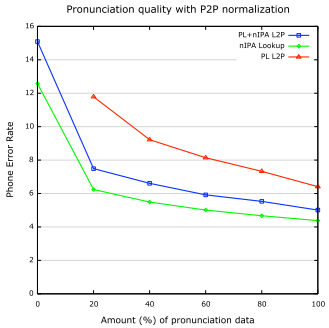


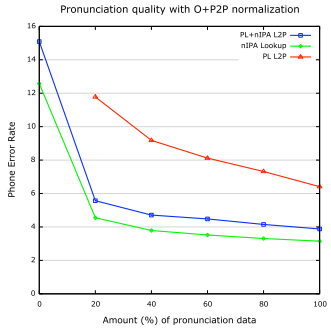


Conclusion

1. Lots of human-supplied pronunciations are available via the web.
2. IPA-usage often varies in site-specific manner.
   - Normalized web-IPA pronunciations are of high quality.
3. Ad hoc pronunciations have rich coverage of low-frequency words
   - Cover unusual pronunciations (e.g. begin vs. Begin, nice vs. Nice)
   - Problematic for intrinsic evaluation (incorrect vs. unusual pronunciation, e.g. `China (pronounced as chee-nah)`








pronunciations
Pronunciation
ad-hoc
model
Pronlex
extraction
words
transcription
extracted
candidate
models
IPA
trained
orthography
orthographic
normalization
phoneme
phonemes
data
English
training
using
Web
Web-IPA
Lexicon
pairs
normalized
validation
used
websites
transcriptions
based
experiments
PhER
described
form
rare
one
standard
conventions
intersection
available
Pru
multiple
generate
predicted
aligned
predict
terms
Cross-validation
common
e.g
features
Pr
subset
two
example
cases
use
L2P
phonemic
test
go
find
Sec
first
word
set
recall
include
see
next
5-gram
lexica
Fig
four
pages
Precision
quality
frequent
pair
train
string
given
n-gram
unigram
results
P2P
l2
ll