Can You Repeat That?
Using Word Repetition to Improve Spoken Term Detection

Jonathan Wintrode and **Sanjeev Khudanpur**
Center for Language and Speech Processing
Johns Hopkins University
jcwintr@cs.jhu.edu , khudanpur@jhu.edu

Abstract

We aim to improve *spoken term detection* performance by incorporating contextual information beyond traditional N-gram language models. Instead of taking a broad view of topic context in spoken documents, variability of word co-occurrence statistics across corpora leads us to focus instead the on phenomenon of word repetition within single documents. We show that given the detection of one instance of a term we are more likely to find additional instances of that term in the same document. We leverage this *burstiness* of keywords by taking the most confident keyword hypothesis in each document and interpolating with lower scoring hits. We then develop a principled approach to select interpolation weights using only the ASR training data. Using this re-weighting approach we demonstrate consistent improvement in the term detection performance across all five languages in the BABEL program.

1 Introduction

The *spoken term detection* task arises as a key subtask in applying NLP applications to spoken content. Tasks like topic identification and named-entity detection require transforming a continuous acoustic signal into a stream of discrete tokens which can then be handled by NLP and other statistical machine learning techniques. Given a small vocabulary of interest (1000-2000 words or multi-word terms) the aim of the term detection task is to enumerate occurrences of the keywords within a target corpus. Spoken term detection converts the raw acoustics into time-marked keyword occurrences, which may subsequently be fed (e.g. as a bag-of-terms) to standard NLP algorithms.

Although spoken term detection does not require the use of word-based automatic speech recognition (ASR), it is closely related. If we had perfectly accurate ASR in the language of the corpus, term detection is reduced to an exact string matching task. The word error rate (WER) and term detection performance are clearly correlated. Given resource constraints, domain, channel, and vocabulary limitations, particularly for languages other than English, the errorful token stream makes term detection a non-trivial task.

In order to improve detection performance, and restricting ourselves to an existing ASR system or systems at our disposal, we focus on leveraging *broad document context* around detection hypotheses. ASR systems traditionally use N-gram language models to incorporate prior knowledge of word occurrence patterns into prediction of the next word in the token stream. N-gram models cannot, however, capture complex linguistic or topical phenomena that occur outside the typical 3-5 word scope of the model. Yet, though many language models more sophisticated than N-grams have been proposed, N-grams are empirically hard to beat in terms of WER.

We consider term detection rather than the transcription task in considering how to exploit topic context, because in evaluating the retrieval of certain key terms we need not focus on improving the entire word sequence. Confidence scores from an ASR system (which incorporate N-gram probabilities) are optimized in order to produce the most likely sequence of words rather than the accuracy of individual word detections. Looking at broader document context within a more limited task might allow us to escape the limits of N-gram performance. We will show that by focusing on contextual information in the form of word repetition within documents, we obtain consistent improvement *across five languages* in the so called Base Phase of the IARPA BABEL program.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 1316–1325,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

1.1 Task Overview

We evaluate term detection and word repetition-based re-scoring on the IARPA BABEL training and development corpora[1] for five languages Cantonese, Pashto, Turkish, Tagalog and Vietnamese (Harper, 2011). The BABEL task is modeled on the 2006 NIST Spoken Term Detection evaluation (NIST, 2006) but focuses on limited resource conditions. We focus specifically on the so called *no target audio reuse* (NTAR) condition to make our method broadly applicable.

In order to arrive at our eventual solution, we take the BABEL Tagalog corpus and analyze word co-occurrence and repetition statistics in detail. Our observation of the variability in co-occurrence statistics between Tagalog training and development partitions leads us to narrow the scope of document context to same word co-occurrences, i.e. *word repetitions*.

We then analyze the tendency towards within-document repetition. The strength of this phenomenon suggests it may be more viable for improving term-detection than, say, topic-sensitive language models. We validate this by developing an interpolation formula to boost putative word repetitions in the search results, and then investigate a method for setting interpolation weights without manually tuning on a development set.

We then demonstrate that the method generalizes well, by applying it to the 2006 English data and the remaining four 2013 BABEL languages. We demonstrate consistent improvements in all languages in both the Full LP (80 hours of ASR training data) and Limited LP (10 hours) settings.

2 Motivation

We seek a workable definition of ***broad document context*** beyond N-gram models that will improve term detection performance on an arbitrary set of queries. Given the rise of unsupervised latent topic modeling with Latent Dirchlet Allocation (Blei et al., 2003) and similar latent variable approaches for discovering meaningful word co-occurrence patterns in large text corpora, we ought to be able to leverage these topic contexts instead of merely N-grams. Indeed there is work in the literature that shows that various topic models, latent or otherwise, can be useful for improving language model perplexity and word error rate (Khudanpur and Wu, 1999; Chen, 2009; Naptali et al., 2012). However, given the preponderance of highly frequent non-content words in the computation of a corpus' WER, it's not clear that a 1-2% improvement in WER would translate into an improvement in term detection.

Still, intuition suggests that knowing the topic context of a detected word ought to be useful in predicting whether or not a term does belong in that context. For example, if we determine the context of the detection hypothesis is about computers, containing words like 'monitor,' 'internet' and 'mouse,' then we would be more confident of a term such as 'keyboard' and less confident of a term such as 'cheese board'. The difficulty in this approach arises from the variability in word co-occurrence statistics. Using topic information will be helpful if 'monitor,' 'keyboard' and 'mouse' consistently predict that 'keyboard' is present. Unfortunately, estimates of co-occurrence from small corpora are not very consistent, and often over- or underestimate concurrence probabilities needed for term detection.

We illustrate this variability by looking at how consistent word co-occurrences are between two separate corpora in the same language: i.e., if we observe words that frequently co-occur with a keyword in the training corpus, do they also co-occur with the keywords in a second held-out corpus? Figure 1, based on the BABEL Tagalog corpus, suggests this is true only for high frequency keywords.

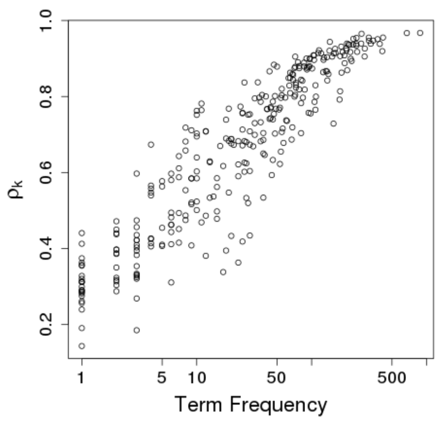
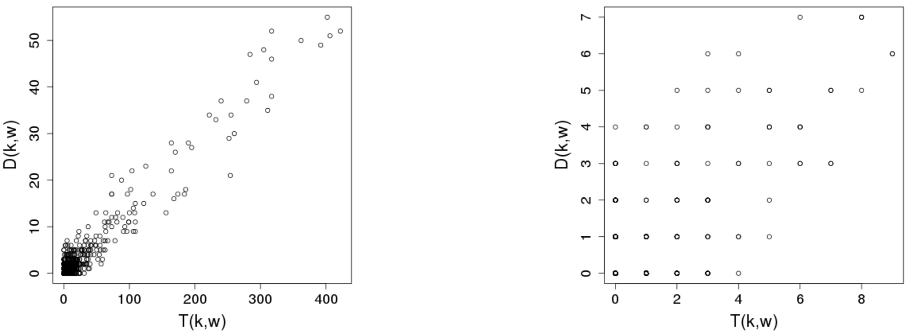


Figure 1: Correlation between the co-occurrence counts in the training and held-out sets for a fixed keyword (term) and all its "context" words.

Each point in Figure 1 represents one of 355

[1] Language collection releases IARPA-babel101-v0.4c, IARPA-babel104b-v0.4bY, IARPA-babel105b-v0.4, IARPA-babel106-v0.2g and IARPA-babel107b-v0.7 respectively.



(a) High frequency keyword 'bukas'

(b) Low frequency keyword 'Davao'

Figure 2: The number of times a fixed keyword k co-occurs with a vocabulary word w in the training speech collection — $T(k, w)$ — versus the search collection — $D(k, w)$.

Tagalog keywords used for system development by all BABEL participants. For each keyword k, we count how often it co-occurs in the same conversation as a vocabulary word w in the ASR training data and the development data, and designate the counts $T(k, w)$ and $D(k, w)$ respectively. The x-coordinate of each point in Figure 1 is the frequency of k in the training data, and the y-coordinate is the correlation coefficient ρ_k between $T(k, w)$ and $D(k, w)$. A high ρ_k implies that words w that co-occur frequently with k in the training data also do so in the search collection.

To further illustrate how Figure 1 was obtained, consider the high-frequency keyword *bukas* (count $= \mathbf{879}$) and the low-frequency keyword *Davao* (count $= \mathbf{11}$), and plot $T(k, \cdot)$ versus $D(k, \cdot)$, as done in Figure 2. The correlation coefficients ρ_{bukas} and ρ_{Davao} from the two plots end up as two points in Figure 1.

Figure 1 suggests that (k, w) co-occurrences are consistent between the two corpora ($\rho_k > 0.8$) for keywords occurring 100 or more times. However, if the goal is to help a speech retrieval system detect content-rich (and presumably infrequent) keywords, then using word co-occurrence information (i.e. topic context) does not appear to be too promising, even though intuition suggests that such information ought to be helpful.

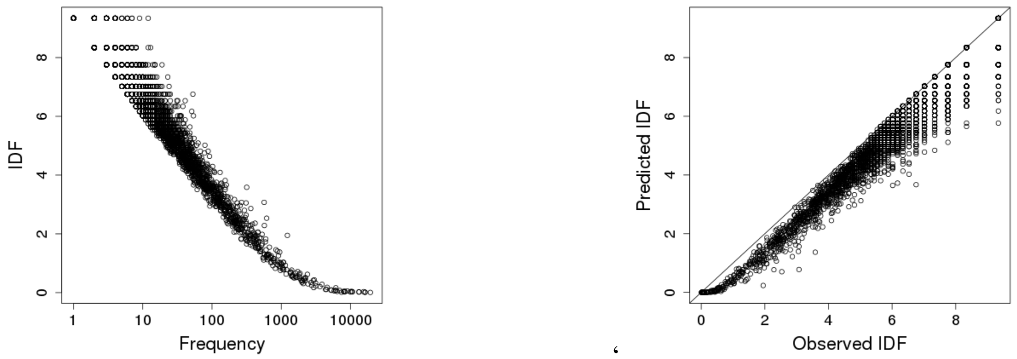
In light of this finding, we will restrict the type of ***context*** we use for term detection to the co-occurrence of the term itself elsewhere within the document. As it turns out this 'burstiness' of words within documents, as the term is defined by Church and Gale in their work on Poisson mixtures (1995), provides a more reliable framework for successfully exploiting document context.

2.1 Related Work

A number of efforts have been made to augment traditional N-gram models with latent topic information (Khudanpur and Wu, 1999; Florian and Yarowsky, 1999; Liu and Liu, 2008; Hsu and Glass, 2006; Naptali et al., 2012) including some of the early work on Probabilistic Latent Semantic Analysis by Hofmann (2001). In all of these cases WER gains in the 1-2% range were observed by interpolating latent topic information with N-gram models.

The re-scoring approach we present is closely related to adaptive or cache language models (Jelinek, 1997; Kuhn and De Mori, 1990; Kneser and Steinbiss, 1993). The primary difference between this and previous work on similar language models is the narrower focus here on the term detection task, in which we consider each search term in isolation, rather than all words in the vocabulary. Most recently, Chiu and Rudnicky (2013) looked at word bursts in the IARPA BABEL conversational corpora, and were also able to successfully improve performance by leveraging the burstiness of language. One advantage of the approach proposed here, relative to their approach, is its simplicity and its not requiring an additional tuning set to estimate parameters.

In the information retrieval community, clustering and latent topic models have yielded improvements over traditional vector space models. We will discuss in detail in the following section related works by Church and Gale (1995, 1999, and 2000). Work by Wei and Croft (2006) and Chen (2009) take a language model-based approach to



(a) f_w versus IDF$_w$

(b) Obsered versus predicted IDF$_w$

Figure 3: Tagalog corpus frequency statistics, unigrams

information retrieval, and again, interpolate latent topic models with N-grams to improve retrieval performance. However, in many text retrieval tasks, queries are often tens or hundreds of words in length rather than short spoken phrases. In these efforts, the topic model information was helpful in boosting retrieval performance above the baseline vector space or N-gram models.

Clearly topic or context information is relevant to a retrieval type task, but we need a stable, consistent framework in which to apply it.

3 Term and Document Frequency Statistics

To this point we have assumed an implicit property of low-frequency words which Church and Gale state concisely in their 1999 study of *inverse document frequency*:

> Low frequency words tend to be rich in content, and vice versa. But not all equally frequent words are equally meaningful. Church and Gale (1999).

The typical use of Document Frequency (DF) in information retrieval or text categorization is to emphasize words that occur in only a few documents and are thus more "rich in content". Close examination of DF statistics by Church and Gale in their work on Poisson Mixtures (1995) resulted in an analysis of the *burstiness* of content words.

In this section we look at DF and *burstiness* statistics applying some of the analyses of Church and Gale (1999) to the BABEL Tagalog corpus. We observe, in 648 Tagalog conversations, similar phenomena as observed by Church and Gale on 89,000 AP English newswire articles. We proceed in this fashion to make a case for why burstiness ought to help in the term detection task.

For the Tagalog conversations, as with English newswire, we observe that the document frequency, DF$_w$, of a word w is not a linear function of word frequency f_w in the log domain, as would be expected under a naive Poisson generative assumption. The implication of deviations from a Poisson model is that *words tend to be concentrated in a small number of documents* rather than occurring uniformly across the corpus. This is the *burstiness* we leverage to improve term detection.

The first illustration of word burstiness can be seen by plotting observed inverse document frequency, IDF$_w$, versus f_w in the log domain (Figure 3a). We use the same definition of IDF$_w$ as Church and Gale (1999):

$$\mathrm{IDF}_w = -\log_2 \frac{\mathrm{DF}_w}{N}, \qquad (1)$$

where N is the number of documents (i.e. conversations) in the corpus.

There is good linear correlation ($\rho = 0.73$) between $\log f_w$ and IDF$_w$. Yet, visually, the relationship in Figure 3a is clearly not linear. In contrast, the AP English data exhibits a correlation of $\rho = 0.93$ (Church and Gale, 1999). Thus the deviation in the Tagalog corpus is more pronounced, i.e. words are less uniformly distributed across documents.

A second perspective on word burstiness that follows from Church and Gale (1999) is that a Poisson assumption should lead us to predict:

$$\widehat{\mathrm{IDF}}_w = -\log_2 \left(1 - e^{-\frac{f_w}{N}}\right). \qquad (2)$$

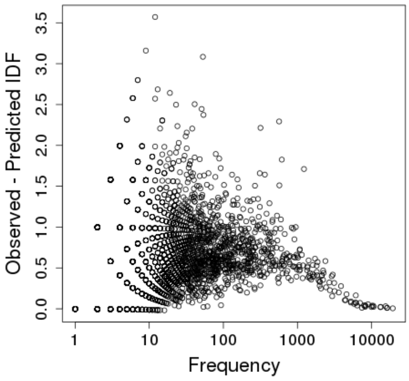


Figure 4: Difference between observed and predicted IDF_w for Tagalog unigrams.

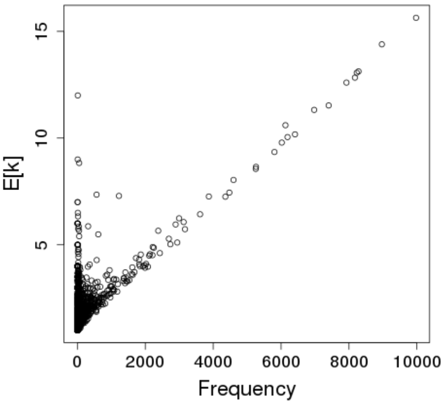


Figure 5: Tagalog *burstiness*.

For the AP newswire, Church and Gale found the largest deviation between the predicted $\widehat{\mathrm{IDF}}_w$ and observed IDF_w to occur in the middle of the frequency range. We see a somewhat different picture for Tagalog speech in Figure 3b. Observed IDF_w values again deviate significantly from their predictions (2), but all along the frequency range.

There is a noticeable quantization effect occurring in the high IDF range, given that our N is at least a factor of 100 smaller than the number of AP articles they studied: 648 vs. 89,000. Figure 4 also shows the difference between and observed IDF_w and Poisson estimate $\widehat{\mathrm{IDF}}_w$ and further illustrates the high variance in IDF_w for low frequency words.

Two questions arise: what is happening with infrequent words, and why does this matter for term detection? To look at the data from a different perspective, we consider the random variable k, which is the number of times a word occurs in a particular document. In Figure 5 we plot the following ratio, which Church and Gale (1995) define as *burstiness* :

$$E_w[k|k>0] = \frac{f_w}{\mathrm{DF}_w} \tag{3}$$

as a function of f_w. We denote this as $E[k]$ and can interpret burstiness as the expected word count given we see w at least once.

In Figure 5 we see two classes of words emerge. A similar phenomenon is observed concerning adaptive language models (Church, 2000). In general, we can think of using word repetitions to re-score term detection as applying a limited form of adaptive or cache language model (Jelinek, 1997). Likewise, Katz attempts to capture these two classes in his G model of word frequencies (1996).

For the first class, burstiness increases slowly but steadily as w occurs more frequently. Let us label these Class A words. Since our corpus size is fixed, we might expect this to occur, as more word occurrences must be pigeon-holed into the same number of documents

Looking close to the y-axis in Figure 5, we observe a second class of exclusively low frequency words whose burstiness ranges from highly concentrated to singletons. We will refer to these as Class B words. If we take the Class A concentration trend as typical, we can argue that most Class B words exhibit a larger than average concentration. In either case we see evidence that *both high and low frequency words tend towards repeating within a document.*

3.1 Unigram Probabilities

In applying the *burstiness* quantity to term detection, we recall that the task requires us to locate a particular instance of a term, not estimate a count, hence the utility of N-gram language models predicting words in sequence.

We encounter the burstiness property of words again by looking at unigram occurrence probabilities. We compare the unconditional unigram probability (the probability that a given word token is w) with the conditional unigram probability, *given the term has occurred once in the document.* We compute the conditional probability for w using frequency information.

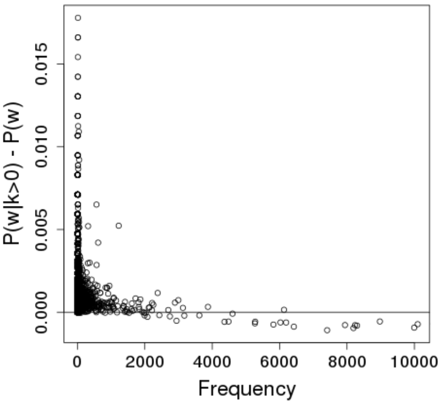


Figure 6: Difference between conditional and unconditional unigram probabilities for Tagalog

$$P(w|k>0) = \frac{f_w - \mathrm{DF}_w}{\sum_{D:w\in D}|D|} \quad (4)$$

Figure 6 shows the difference between conditional and unconditional unigram probabilities. Without any other information, Zipf's law suggests that most word types do not occur in a particular document. However, conditioning on one occurrence, most word types are more likely to occur again, due to their burstiness.

Finally we measure the *adaptation* of a word, which is defined by Church and Gale (1995) as:

$$P_{adapt}(w) = P_w(k>1|k>0) \quad (5)$$

When we plot adaptation versus f_w (Figure 7) we see that all high-frequency and a significant number of low-frequency terms have adaptation greater that 50%. To be precise, 26% of all tokens and 25% of low-frequency ($f_w < 100$) have at least 50% adaptation. Given that adaptation values are roughly an order of magnitude higher than the conditional unigram probabilities, in the next two sections we describe how we use adaptation to boost term detection scores.

4 Term Detection Re-scoring

We summarize our re-scoring of repeated words with the observation: *given a correct detection, the likelihood of additional terms in the same documents should increase*. When we observe a term detection score with high confidence, we boost the other lower-scoring terms in the same document to reflect this increased likelihood of repeated terms.

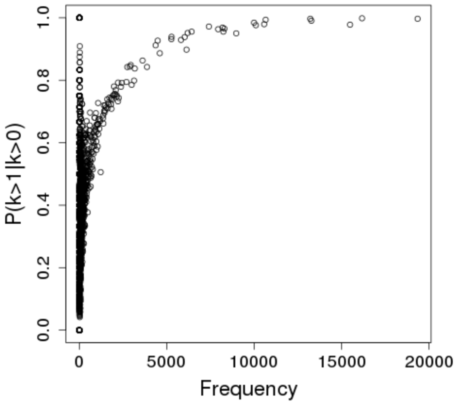


Figure 7: Tagalog word adaptation probability

For each term t and document d we propose interpolating the ASR confidence score for a particular detection t_d with the top scoring hit in d which we'll call $\widehat{t_d}$.

$$S(t_d) = (1-\alpha)P_{\mathrm{asr}}(t_d|O) + \alpha P_{\mathrm{asr}}(\widehat{t_d}|O) \quad (6)$$

We will we develop a principled approach to selecting α using the adaptation property of the corpus. However to verify that this approach is worth pursuing, we sweep a range of small α values, on the assumption that we still do want to mostly rely on the ASR confidence score for term detection. For the Tagalog data, we let α range from 0 (the baseline) to 0.4 and re-score each term detection score according to (6). Table 1 shows the results of this parameter sweep and yields us 1 to 2% absolute performance gains in a number of term detection metrics.

α	**ATWV**	P(Miss)
0.00	0.470	0.430
0.05	0.481	0.422
0.10	0.483	0.420
0.15	0.484	0.418
0.20	0.483	0.416
0.25	0.480	0.417
0.30	0.477	0.417
0.35	0.475	0.415
0.40	0.471	0.413
0.45	0.465	0.413
0.50	0.462	0.410

Table 1: Term detection scores for swept α values on Tagalog development data

The primary metric for the BABEL program, Actual Term Weighted Value (ATWV) is defined by NIST using a cost function of the false alarm probability $P(\text{FA})$ and $P(\text{Miss})$, averaged over a set of queries (NIST, 2006). The manner in which the components of ATWV are defined:

$$P(\text{Miss}) = 1 - N_{\text{true}}(\text{term})/f_{\text{term}} \quad (7)$$

$$P(\text{FA}) = N_{\text{false}}/Duration_{\text{corpus}} \quad (8)$$

implies that cost of a miss is inversely proportional to the frequency of the term in the corpus, but the cost of a false alarm is fixed. For this reason, we report both ATWV and the $P(\text{Miss})$ component. A decrease in $P(\text{Miss})$ reflects the fact that we are able to boost correct detections of the repeated terms.

4.1 Interpolation Weights

We would prefer to use prior knowledge rather than naive tuning to select an interpolation weight α. Our analysis of word burstiness suggests that *adaptation*, is a reasonable candidate. Adaptation also has the desirable property that we can estimate it for each word in the training vocabulary directly from training data and not post-hoc on a per-query basis. We consider several different estimates and we can show that the favorable result extends across languages.

Intuition suggests that we prefer per-term interpolation weights related to the term's *adaptation*. But despite the strong evidence of the adaptation phenomenon in both high and low-frequency words (Figure 7), we have less confidence in the adaptation strength of any particular word.

As with word co-occurrence, we consider if estimates of $P_{adapt}(w)$ from training data are consistent when estimated on development data. Figure 8 shows the difference between $P_{adapt}(w)$ measured on the two corpora (for words occurring in both).

We see that the adaptation estimates are only consistent between corpora for high-frequency words. Using this $P_{adapt}(w)$ estimate directly actually hurts ATWV performance by 4.7% absolute on the 355 term development query set (Table 2).

Given the variability in estimating $P_{adapt}(w)$, an alternative approach would be take $\widehat{P_w}$ as an upper bound on α, reached as the DF_w increases (cf. Equation 9). We would discount the adaptation factor when DF_w is low and we are unsure of the effect.

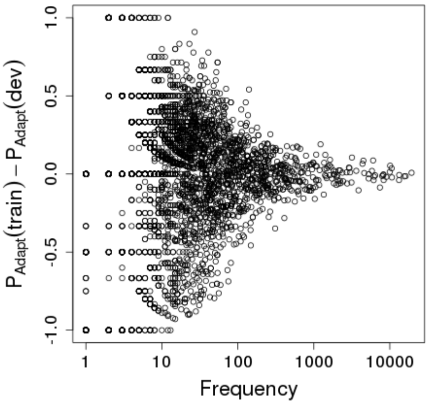


Figure 8: Difference in adaptation estimates between Tagalog training and development corpora

Interpolation Weight	**ATWV**	$P(\text{Miss})$
None	0.470	0.430
$P_{adapt}(w)$	0.423	0.474
$(1 - e^{-\text{DF}_w})P_{adapt}(w)$	0.477	0.415
$\widehat{\alpha}$ = 0.20	**0.483**	**0.416**

Table 2: Term detection performance using various interpolation weight strategies on Tagalog dev data

$$\alpha_w = (1 - e^{-\text{DF}_w}) \cdot \widehat{P}_{adapt}(w) \quad (9)$$

This approach shows a significant improvement (0.7% absolute) over the baseline. However, considering this estimate in light of the two classes of words in Figure 5, there are clearly words in Class B with high burstiness that will be ignored by trying to compensate for the high adaptation variability in the low-frequency range.

Alternatively, we take a weighted average of α_w's estimated on training transcripts to obtain a single $\widehat{\alpha}$ per language (cf. Equation 10).

$$\widehat{\alpha} = \operatorname*{Avg}_{w} \left[\left(1 - e^{-\text{DF}_w}\right) \cdot \widehat{P}_{adapt}(w) \right] \quad (10)$$

Using this average as a single interpolation weight for all terms gives near the best performance as we observed in our parameter sweep. Table 2 contrasts the results for using the three different interpolation heuristics on the Tagalog development queries. Using the mean $\widehat{\alpha}$ instead of individual α_w's provides an additional 0.5% absolute

Language	$\widehat{\alpha}$	**ATWV (%±)**	$P(\text{Miss})$ **(%±)**
Full LP setting			
Tagalog	0.20	**0.523** (+1.1)	0.396 (-1.9)
Cantonese	0.23	**0.418** (+1.3)	0.458 (-1.9)
Pashto	0.19	**0.419** (+1.1)	0.453 (-1.6)
Turkish	0.14	**0.466** (+0.8)	0.430 (-1.3)
Vietnamese	0.30	**0.420** (+0.7)	0.445 (-1.0)
English (Dev06)	0.20	**0.670 (+0.3)**	0.240 (-0.4)
Limited LP setting			
Tagalog	0.22	**0.228** (+0.9)	0.692 (-1.7)
Cantonese	0.26	**0.205** (+1.0)	0.684 (-1.3)
Pashto	0.21	**0.206** (+0.9)	0.682 (-0.9)
Turkish	0.16	**0.202** (+1.1)	0.700 (-0.8)
Vietnamese	0.34	**0.227** (+1.0)	0.646 (+0.4)

Table 3: Word-repetition re-scored results for available CTS term detection corpora

improvement, suggesting that we find additional gains boosting low-frequency words.

5 Results

Now that we have tested word repetition-based re-scoring on a small Tagalog development set we want to know if our approach, and particularly our $\widehat{\alpha}$ estimate is sufficiently robust to apply broadly. At our disposal, we have the five BABEL languages — Tagalog, Cantonese, Pashto, Turkish and Vietnamese — as well as the development data from the NIST 2006 English evaluation. The BABEL evaluation query sets contain roughly 2000 terms each and the 2006 English query set contains roughly 1000 terms.

The procedure we follow for each language condition is as follows. We first estimate adaptation probabilities from the ASR training transcripts. From these we take the weighted average as described previously to obtain a single interpolation weight $\widehat{\alpha}$ for each training condition. We train ASR acoustic and language models from the training corpus using the Kaldi speech recognition toolkit (Povey et al., 2011) following the default BABEL training and search recipe which is described in detail by Chen et al. (2013). Lastly, we re-score the search output by interpolating the top term detection score for a document with subsequent hits according to Equation 6 using the $\widehat{\alpha}$ estimated for this training condition.

For each of the BABEL languages we consider both the FullLP (80 hours) and LimitedLP (10 hours) training conditions. For the English system, we also train a Kaldi system on the 240 hours of the Switchboard conversational English corpus. Although Kaldi can produce multiple types of acoustic models, for simplicity we report results using discriminatively trained Subspace Gaussian Mixture Model (SGMM) acoustic output densities, but we do find that similar results can be obtained with other acoustic model configurations.

Using our final algorithm, we are able to boost repeated term detections and improve results in **all languages and training conditions**. Table 3 lists complete results and the associated estimates for $\widehat{\alpha}$. For the BABEL languages, we observe improvements in ATWV from 0.7% to 1.3% absolute and reductions in the miss rate of 0.8% to 1.9%. The only test for which $P(\text{Miss})$ did not improve was the Vietnamese Limited LP setting, although overall ATWV did improve, reflecting a lower $P(\text{FA})$.

In all conditions we also obtain α estimates which correspond to our expectations for particular languages. For example, adaptation is lowest for the agglutinative Turkish language where longer word tokens should be less likely to repeat. For Vietnamese, with shorter, syllable length word tokens, we observe the lowest adaptation estimates.

Lastly, the reductions in $P(\text{Miss})$ suggests that we are improving the term detection metric, which is sensitive to threshold changes, by doing what we set out to do, which is to boost lower confidence repeated words and correctly asserting them as true hits. Moreover, we are able to accomplish this in a wide variety of languages.

6 Conclusions

Leveraging the **burstiness** of content words, we have developed a simple technique to consistently boost term detection performance across languages. Using word repetitions, we effectively use a broad document context outside of the typical 2-5 N-gram window. Furthermore, we see improvements across a broad spectrum of languages: languages with syllable-based word tokens (Vietnamese, Cantonese), complex morphology (Turkish), and dialect variability (Pashto).

Secondly, our results are not only effective but also intuitive, given that the interpolation weight parameter matches our expectations for the burstiness of the word tokens in the language on which it is estimated.

We have focused primarily on re-scoring results for the term detection task. Given the effectiveness of the technique across multiple languages, we hope to extend our effort to exploit our human tendency towards redundancy to decoding or other aspects of the spoken document processing pipeline.

Acknowledgements

This work was partially supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD / ARL) contract number W911NF-12-C-0015. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

Insightful discussions with Chiu and Rudnicky (2013) are also gratefully acknowledged.

References

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Guoguo Chen, Sanjeev Khudanpur, Daniel Povey, Jan Trmal, David Yarowsky, and Oguz Yilmaz. 2013. Quantifying the value of pronunciation lexicons for keyword search in low resource languages. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Berlin Chen. 2009. Latent topic modelling of word co-occurence information for spoken document retrieval. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3961–3964. IEEE.

Justin Chiu and Alexander Rudnicky. 2013. Using conversational word bursts in spoken term detection. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association*, pages 2247–2251. ISCA.

Kenneth Church and William Gale. 1995. Poisson Mixtures. *Natural Language Engineering*, 1(2):163–190.

Kenneth Church and William Gale. 1999. Inverse Focument Frequency (IDF): A measure of deviations from Poisson. In *Natural Language Processing Using Very Large Corpora*, pages 283–295. Springer.

Kenneth Church. 2000. Empirical estimates of adaptation: the chance of two Noriegas is closer to p/2 than p 2. In *Proceedings of the 18th Conference on Computational Linguistics*, volume 1, pages 180–186. ACL.

Radu Florian and David Yarowsky. 1999. Dynamic nonlocal language modeling via hierarchical topic-based adaptation. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pages 167–174. ACL.

Mary Harper. 2011. IARPA Solicitation IARPA-BAA-11-02. `http://www.iarpa.gov/solicitations_babel.html`.

Thomas Hofmann. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196.

Bo-June Paul Hsu and James Glass. 2006. Style & topic language model adaptation using HMM-LDA. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. ACL.

Fred Jelinek. 1997. *Statistical Methods for Speech Recognition*. MIT Press.

Slava Katz. 1996. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, 2(1):15–59.

Sanjeev Khudanpur and Jun Wu. 1999. A maximum entropy language model integrating n-grams and topic dependencies for conversational speech recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 553–556. IEEE.

Reinhard Kneser and Volker Steinbiss. 1993. On the dynamic adaptation of stochastic language models. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 586–589. IEEE.

Roland Kuhn and Renato De Mori. 1990. A cache-based natural language model for speech recognition. *Transactions on Pattern Analysis and Machine Intelligence*, 12(6):570–583.

Yang Liu and Feifan Liu. 2008. Unsupervised language model adaptation via topic modeling based on named entity hypotheses. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, pages 4921–4924. IEEE.

Welly Naptali, Masatoshi Tsuchiya, and Seiichi Nakagawa. 2012. Topic-dependent-class-based n-gram language model. *Transactions on Audio, Speech, and Language Processing*, 20(5):1513–1525.

NIST. 2006. The Spoken Term Detection (STD) 2006 Evaluation Plan. `http://www.itl.nist.gov/iad/mig/tests/std/2006/docs/std06-evalplan-v10.pdf`. [Online; accessed 28-Feb-2013].

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Xing Wei and W Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 178–185. ACM.