Factored Language Modeling for Russian LVCSR

Daria Vazhenina, Konstantin Markov

Human Interface Laboratory, The University of Aizu, Japan
(d8132102,markov)@u-aizu.ac.jp

ABSTRACT

The Russian language is characterized by very flexible word order, which limits the ability of the standard n-grams to capture important regularities in the data. Moreover, Russian is highly inflectional language with rich morphology, which leads to high out-of-vocabulary word rates. Recently factored language model (FLM) was proposed with the aim of addressing the problems of morphologically rich languages. In this paper, we describe our implementation of the FLM for the Russian language automatic speech recognition (ASR). We investigated the effect of different factors, and propose a strategy to find the best factor set and back-off path. Evaluation experiments showed that FLM can decrease the perplexity as much as 20%. This allows to achieve 4.0% word error rate (WER) relative reduction, which further increases to 6.9% when FLM is interpolated with the conventional 3-gram LM.
Index Terms: language modeling, Russian language, factored language models, inflectional languages.

1. INTRODUCTION

Russian belongs to the Slavic branch of the Indo-European group of languages, which are characterized by complex mechanism of word-formation. Several kinds of morphemes are used to produce such grammatical categories as gender, number, case, person, tense, etc. Zaliznjak's grammatical dictionary of Russian [1] contains about 150 thousand words and considering all possible word inflections an ASR vocabulary can grow up to more than 2 million words. Thus, even for systems with very large vocabulary, this leads to large number of out-of-vocabulary (OOV) words.

Another important characteristic of the Russian language is its flexible word order, which is not restricted by hard grammatical structure as in the English, German or Arabic languages. Word relations within a sentence are marked by inflections and grammatical categories such as gender, number, person, case, etc. [2]. These two problems greatly reduce the predictive power of the conventional statistical language models (LMs) [3].

To reduce vocabulary size and OOV rate, it was proposed to split words into grammatical morphemes and implement speech recognition on morphemic level adding final back-to-words step. In [4], this approach led to a significant reduction of the pronunciation lexicon, however, results of ASR experiments haven't been reported. Another study [5] found that the real-time factor can be improved using morpheme-based LM, but the recognition performance did not change. This approach was also applied to the Slovenian language, which is also the Slavic language [6]. In this case, the OOV rate was improved from 8.2% to 1.2% and WER decreased from 42.3% to 42.0% for a system with 60K lexicon, but the real-time factor (RTF) of the system has increased. Returning from morphemes back to words is not a trivial task, and is not always implemented.

In Russian large vocabulary continuous speech recognition (LVCSR) systems conventional n-grams are usually used [7, 8, 9]. An improved bi-gram was proposed in [10] whereas the counts of some of the existing n-grams are increased after syntactic analysis of the training data. Long-distance syntactical dependencies between words are identified and added as new bi-grams. This allowed to reduce the word error rate of a speech recognition system with a dictionary of 208K words from 58.4% to 56.1%.

Recently, for Arabic, which is also highly inflectional language, it was proposed to incorporate word features, called factors, into the language model [11]. This factored language model (FLM) implements a back-off procedure by excluding factors one by one or even several factors at a time without taking into account factor's distance from the predicted word. This improves the robustness of the probability estimates for rarely observed word n-grams. Using this model, relative WER reduction of 3.4% was achieved for Arabic LVCSR system with 70K vocabulary size [12]. For the Turkish language, it was reported that the FLM reduced the WER by 1.7% relative for a 200K LVCSR system [13].

This paper describes our implementation of the FLM for Russian LVCSR system of 100K words. We investigated the influence of different factors on the language model performance and propose the strategy for optimal factor set selection. The FLM is implemented using the n-best re-scoring method. The best performance, in terms of WER, was achieved using interpolation of the conventional 3-gram LM with the FLM built by the proposed strategy.

2. FACTORED LANGUAGE MODELS

Standard n-gram language models compute the probability of a word sequence $W = \{w_1, w_2 ... w_t\}$ as a product of the conditional probabilities of individual words given their histories. However, since computing word probability given the entire history is difficult, it can be approximated by the probability of word given the last few words. Thus, conditional probability of a word given one previous word is called bigram, conditional probability of word given two previous words is called trigram and so on. The probability of word sequence using trigram probabilities is expressed as:

$$p(W) \approx \prod_{i=n}^{t} p(w_i \mid w_{i-1}, w_{i-2}) \quad (1)$$

The predictive power of the language model can be evaluated without doing speech recognition experiment using perplexity measure. Improvement in perplexity doesn't guarantee better speech recognition performance, however some correlation between improvement in perplexity and recognition performance usually exists. Perplexity is defined as:

$$PP(W) = \left(\prod_{i=1}^{t} \frac{1}{P(w_i | w_1 ... w_{i-1})} \right)^{\frac{1}{t}} \quad (2)$$

In the factored language model (FLM), first introduced in [14], it is proposed to include word features, called factors, in the standard n-gram language model. Factors of a given word can be any grammatical information about the word, such as its lemma, stem, root, ending, part-of-speech, etc. In the FLM, word sequence $W = \{w_1, w_2 ... w_t\}$ is represented by a sequence of K factors for each word w_i, $f_i^{1:K} = \{f_i^1, f_i^2 ... f_i^K\}$. A probabilistic language model is estimated over the factor vectors. Using a n-gram-like formula, the general model takes the following form $P(f_t^{1:K} | f_{t-1}^{1:K}, f_{t-2}^{1:K} ... f_{t-n+1}^{1:K})$. This can be simplified to $P(f_t | f_1, f_2 ... f_m)$, where $f_t = w_t$ and $\{f_i\}$ is any combination of factors, where $i = 1...m$, $m \leq K * (n-1)$. If an n-gram of word or factor is not sufficiently observed, generalized back-off procedure is used.

In the conventional LM, zero probabilities for unseen n-gram are prevented by backing off to lower order n-gram, e.g. proceeding from a trigram to a bigram to a unigram. This is visualized on Fig.1(a) as back-off *path*. In case of FLM, during the backing off any factor can be dropped at each step in any order resulting in back-off *tree* shown on Fig.1(b). Flexible generalized parallel back-off procedure is the main advantage of the FLM.

In order to obtain a good FLM performance, we need to tune its parameters: the combination of conditioning factors (factor set) and the back-off tree. In [15], two ways were proposed to optimize these parameters:

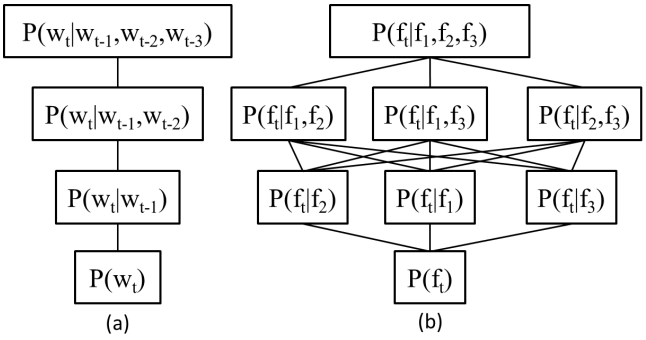

Fig. 1. N-gram and FLM back-off trees. (a) Standard n-gram back-off path has strict order of dropping words, (b) In the FLM case, any factor may be dropped at each step resulting in many possible paths.

- Manual: Choose the factor set and fix the back-off tree based on linguistic knowledge, for example, always drop syntactic before morphological factors [13].
- Automatic: Use genetic algorithm (GA) based FLM optimization method presented in [11].

In [11], it was demonstrated that the factor set and back-off tree optimized using GA-FLM can perform better in terms of perplexity than hand-selected ones. In addition, relative WER reductions presented in [11] and [12], obtained using GA, were higher than those presented in [13], obtained using manually set back-off path.

3. FACTOR SET SELECTION

The genetic algorithm for FLM optimization seeks the optimal factor set and appropriate back-off tree based on minimizing the model perplexity over some test set. However, with many factor types and longer word history the task quickly becomes computationally intractable. Given initial factor set of k factors there are 2^k possible subsets. Furthermore, for a set of m conditioning factors, there are up to $m!$ possible back-off paths. In practice, because of these reasons, a few factor types and time context of 2 (corresponding to trigram) are usually used [12, 13, 16].

To be able to use longer time context and optimal factor set without scalability problems, we propose, first, to identify those factors, which provide most useful information for the language model and, then, to extend the time history as much as possible. In order to refer to the time context more explicitly we will use $w1$ or $w2$ and f_k1 or f_k2 to denote w_{t-1} or w_{t-2} and f_{t-1}^k or f_{t-2}^k.

Our strategy to identify most optimal factor set includes the following steps:

Step 1 Define the set of possible word feature types to be used as factor types, e.g. POS category, stem, root,

inflection, etc.

Step 2 Build several small FLMs using the word[1] and one of the other factor types for time context 1 and 2. In other words, train the following FLMs $p_k(w_t|w1, f_k1, w2, f_k2)$ for $k = 1...K$. In this case, back-off path can be set manually similar to the conventional 3-gram back-off path, which has two possible variations shown on Figure 2:

- Back-off path 1: First drop the words in time distance order: $w2, w1$, then drop factors in the same order: f_k2, f_k1.
- Back-off path 2: First drop the most distant word and factor $f_k2, w2$, then less distant ones $f_k1, w1$.

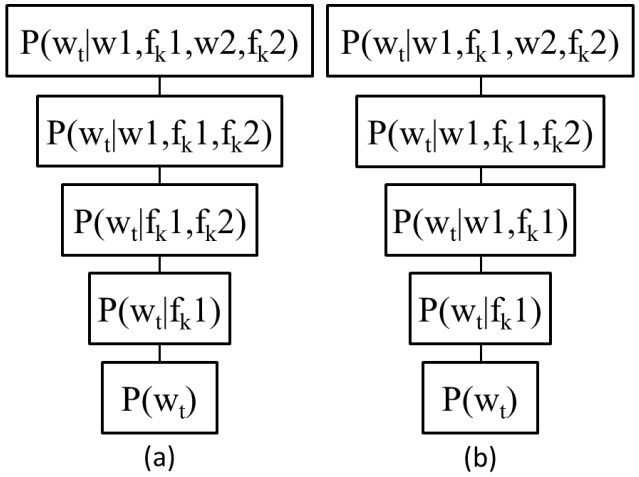

Fig. 2. Manually selected direct back-off paths: (a) - back-off path 1, (b) - back-off path 2.

Step 3 Select the best several factor types for further experiments based on their resulting model perplexities.

Step 4 For all possible 3-combinations of the selected factor types plus the word itself find the best factor set and back-off tree using the GA for the time context of 2. For instance, if factors f_x, f_y, f_z were selected, then following factor sets will be used:

- $(w1, f_x1, f_y1, w2, f_x2, f_y2)$,
- $(w1, f_x1, f_z1, w2, f_x2, f_z2)$,
- $(f_x1, f_y1, f_z1, f_x2, f_y2, f_z2)$,
- $(w1, f_y1, f_z1, w2, f_y2, f_z2)$.

[1] Actually, the word itself is a factor, but on this step we are interested in the influence of additional factor types on the FLM performance.

Step 5 From the obtained factor sets, identify factors, which were most often dropped by the GA algorithm, taking into account the time context and exclude them from further experiments. For example, if in the previous step following factor sets were selected as optimal:

- $(w1, f_x1, w2, f_x2)$,
- $(w1, f_x1, w2, f_x2)$,
- $(f_x1, f_z1, f_x2, f_y2, f_z2)$,
- $(w1, f_y1, f_z1, f_z2)$,

then factors f_y1 and f_y2 should be excluded as the most often dropped.

Step 6 Finally, using the remaining factors extend the time context and run the GA again. If most often used factors were $w1, f_x1, f_z1, f_x2, f_z2$, then factors f_x3, f_z3 may be added for time context extension.

This procedure produces many FLMs and we keep those with the lowest perplexity for further evaluation on speech test data.

4. TEXT DATA PROCESSING

Our LM training text corpus contains 11M words with vocabulary size of about ∼100K words. This corpus was assembled from recent news articles published by freely available Internet sites of several on-line Russian newspapers for the years 2006-2011. We split our corpus into 10M words train set and a test set consisting of 1M words.

Word features were obtained using the TreeTagger tool [17]. It was successfully used for tagging text data in many languages including Russian. The TreeTagger is a probabilistic tool which uses decision trees for annotating words with morphological tag and lemma information, where lemma is the canonical baseform of the word. The morphological tag contains detailed grammatical information about the word, such as POS category, gender, number, case, etc. This tagset is described and evaluated in [18]. The overall accuracy of the TreeTagger on the full tagset is 93,5%, while accuracy of tagging unknown words is 62.44%.

In addition to the word, lemma and morphological tag factor types, we use two extra types: part-of-speech category and gender-number-person factor, because it contains important grammatical information for the word relations in a sentence. The list of all factor types, we experimented with, is shown in Table 1.

The LM training corpus is preprocessed so that every word is replaced with a vector of all factor types. For instance, word 'брэнды' (brands) is replaced with the vector {W-брэнды:P-N:T-Ncmpnn:L-брэнд:G-MP}, where N

Table 1. Factor types for Russian language used in the experiments.

Factor type	Description	Size
W	word	99 958
L	lemma	23 742
T	morphological tag	819
G	gender, number and person	30
P	part-of-speech category	10

means noun POS category; Ncmpnn means noun POS category, common syntactic type, masculine gender, plural number, nominative case, not animate; MP means masculine gender, plural number.

5. EXPERIMENTS

5.1. Speech database and feature extraction

In our experiments, we used the SPIIRAS [19] and GlobalPhone [8] Russian speech databases. Speech data are collected in clean acoustic conditions. In total, there are 28671 utterances pronounced by 165 speakers (86 male and 79 female) with duration of about 38 hours. Speech test data consist of 10% of GlobalPhone corpus with duration of 1 hour 40 minutes. Test set recordings were pronounced by 5 male and 5 female speakers, and aren't used for acoustic model (AM) training.

We used the HTK toolkit to train AM [20]. The speech signal was coded with energy and 12 MFCCs and their first and second order derivatives, resulting in 39-dimension feature vector. Acoustic phoneme models were represented by three state HMMs with left-to-right topology except the silence model, which also has transition from third to the first state. The AM consists of 5342 tied states with 16 component mixture GMMs as output models. Our speech decoder (Julius ver. 4.2 [21]) produces 500-best hypothesis list, which we use for re-scoring by the selected FLMs.

As a baseline LM, we use conventional 3-gram trained using Kneser-Ney discount by the SRI-LM toolkit [22]. Its perplexity is 537 and OOV rate is 1.7%. The word error rate of the baseline speech recognition system is 35.4%.

FLMs were built and evaluated using extension of SRI-LM toolkit described in [23].

5.2. Experimental results

According to the Step 2 of the strategy described in Sect.3 and using each of the factor types from Table 1, we built several FLMs. The performance of these FLMs presented in Table 2 shows that the back-off path has big influence on the perplexity. Since the factor G showed the worst perplexity in both cases, we choose to continue with L, T and P factor types only.

Table 2. Perplexities of FLMs built at Step 2 with different back-off paths. The baseline perplexity is 537.

Factor types	Back-off path 1	Back-off path 2
WL	611	525
WT	685	488
WG	988	714
WP	652	549

Then, according to the Step 4, we compose all possible 3- and 4-combination of these factor types and W, which are shown in the first column of Table 3. Using GA, we find the optimal factor set and back-off trees for all the combinations with time context 2. For example, combination WLP results in initial factor set $W1L1P1W2L2P2$. Their corresponding FLM perplexities are shown in the last column of Table 3. As it can be seen, the usage of factor types WLT allowed to achieve the best perplexity, which is a significant 19.9% reduction relative to the baseline. Note that only in this case the GA did not reduce the initial factor set.

Table 3. Perplexities of FLMs built using all possible 3- and 4-combinations of factors with time context 2. The baseline perplexity is 537.

Factor types	Best factors after GA (# nodes of the back-off tree)	Perplexity
WLT	W1,L1,T1,W2,L2,T2 (55)	**430**
WLP	W1,L2 (4)	560
WTP	W1,P1,W2,P2,T2 (21)	756
LTP	L1,T1,L2,T2 (11)	447
WLTP	W1,L1,T1,P1,L2,T2,P2 (50)	487

There are eight different factors in this experiment (four factor types $\times$ time context 2) and from the second column of Table 3 we can see that the most often dropped factors are $P1, W2, P2$. So, we exclude them from further experiments as proposed in Step 5. Using the remaining five factors $W1, L1, T1, L2, T2$ as initial factor set we applied again a GA optimization. Finally, we extended this factor set to time context 3 (corresponding to 4-gram), as described in Step 6. Since factor types L and T were most often selected by GA in both time contexts, we added $L3$ and $T3$ factors to the initial factor set and optimized it using GA. The perplexities of the resulting two FLMs were close to the perplexity of the FLMs using factor types WLT and LTP from Table 3. Thus, for the speech recognition experiments we chose these four models. Word error rates (WERs) for those FLMs named M1, M2, M3, M4 are shown in Table 4. The lowest WER, resulting in 3.9% relative reduction, was achieved using the model M2 with extended time context.

In Table 5, performances of FLMs linearly interpolated with the baseline 3-gram models are presented. The inter-

Table 4. Performance of FLMs built using most effective factors. The baseline WER is 35.4%.

FLM	Factors (# back-off tree nodes)	ppl	WER, %
M1	W1,L1,T1,L2,T2 (20)	440	34.4
M2	W1,L1,T1,L2,T2,T3 (43)	460	**34.0**
M3	W1,L1,T1,W2,L2,T2 (55)	430	34.5
M4	L1,T1,L2,T2 (11)	447	35.6

polation coefficient is manually tuned for each model. The biggest relative WER reduction of 6.9% is obtained by the interpolation with model M1, which is built with quite small factor set and, as it is shown on Figure 3, not highly branched back-off tree. Obtained WER reductions indicate improvements in LM probability estimates due to the use of correctly preselected factor set.

Table 5. Performance of FLMs interpolated with the baseline 3-gram model.

Model	Interpolation coefficient	ppl	WER, %
3-gram (baseline)	1.0	537	35.4
M1 + 3-gram	0.45	445	**33.0**
M2 + 3-gram	0.43	463	33.4
M3 + 3-gram	0.51	437	33.3
M4 + 3-gram	0.42	453	33.6

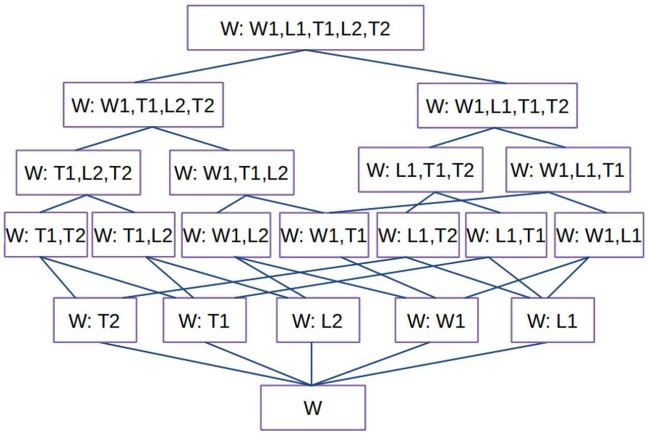

Fig. 3. Back-off tree of the model M1.

In case of interpolation with 3-gram, comparing with other FLM models, WER of the model M4 was noticeably decreased from 35.6% (M4) to 33.6% (M4 + 3-gram). In this model, in contrast with other models, word factor type W hasn't been included in the factor set. Obviously, the combination of statistical information of words *and* their grammatical features allows to implement more accurate LVCSR system.

In all previous experiments, GA was implemented using population size of 8 and 16 generations. Relaxing the GA parameters to population size 40 and number of generations 30 increased the computation time several times, but without any improvement in the performance.

6. CONCLUSIONS

This paper presents a strategy to select optimal factor set to built effective FLM. We compare performance of 3-gram language model with Kneser-Ney smoothing against factored language models and interpolation of both models. Obtained WER relative improvement of 6.9% is quite high, in comparison to improvements achieved for both the Arabic language using FLMs and Russian language using advanced language modeling techniques [12, 13, 10].

Factored language models seems to be able to capture additional information and improve LM probability estimates. Proposed strategy allows to remove not useful features at the beginning of the experiments and use longer context of more effective factors. It is important to mention, that different factor sets cannot be compared by perplexity only, because of the factor size difference. Final decision on choosing the factor set and back-off path should be made based on speech recognition results.

7. REFERENCES

[1] A. Zaliznjak, *Grammatical dictionary of the Russian language*. Moscow, Nauka, 2003.

[2] P. Cubberley, *Russian: a linguistic introduction*. Cambridge University Press, 2002.

[3] D. Vazhenina, I. Kipyatkova, K. Markov, and A. Karpov, "State-of-the-art speech recognition technologies for russian language," in *Proc. of International Conference on Human-Centered Computer Environments (HCCE)*. ACM, March 2012, pp. 59–63.

[4] I. Oparin and A. Talanov, "Stem-based approach to pronunciation vocabulary construction and language modeling for russian," in *Proc. SPECOM*, Patras, Greece, Oct 2005, pp. 575–578.

[5] A. Karpov and A. Ronzhin, "Russian speech recognition model with morphemic analysis and synthesis," in *Proc. of International Congress on Acoustics*, Madrid, Spain, Sep 2007.

[6] T. Rotovnik, M. S. Maučec, and Z. Kačič, "Large vocabulary continuous speech recognition of an inflected language using stems and endings," *Speech Communication*, vol. 49, no. 6, pp. 437–452, June 2007.

[7] E. W. Whittaker and P. C. Woodland, "Comparison of language modelling techniques for Russian and English," in *Proc. International Conference on Spoken Language Processing*, Sydney, Australia, Nov 1998.

[8] S. Stuker and T. Schultz, "A grapheme based speech recognition system for Russian," in *Proc. SPECOM*, St.Peterburg, Russia, Sep 2004, pp. 297–303.

[9] D. Vazhenina and K. Markov, "Phoneme set selection for Russian speech recognition," in *Proc. of IEEE NLP-KE*, Tokushima, Japan, Nov 2011, pp. 475–478.

[10] A. Karpov, I. Kipyatkova, and A. Ronzhin, "Very large vocabulary ASR for spoken Russian with syntactic and morphemic analysis," in *Proc. InterSpeech*, Florence, Italy, Aug 2011, pp. 3161–3164.

[11] K. Kirchhoff, D. Vergyri, J. Bilmes, K. Duh, and A. Stolcke, "Morphology-based language modelling for conversational Arabic speech recognition," *Computer Speech and Language*, vol. 20, no. 4, pp. 589–608, Oct 2006.

[12] A. El-Desoky Mousa, R. Schluter, and H. Ney, "Investigations on the use of morpheme level features in language models for Arabic LVCSR," in *Proc. IEEE ICASSP*, Kyoto, Japan, March 2012, pp. 5021–5024.

[13] H. Sak, M. Saraclar, and T. Gungor, "Morphology-based and sub-word language modelling for Turkish speech recognition," in *Proc. IEEE ICASSP*, Dallas, USA, March 2010, pp. 5402–5405.

[14] J. A. Bilmes and K. Kirchhoff, "Factored language models and generalized parallel back-off," in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 2*, Stroudsburg, PA, USA, May 2003, pp. 4–6.

[15] K. Kirchhoff, J. Bilmes, and K. Duh, "Factored language model tutorial," Department of Electrical Engineering. University of Washington., Seatle, Washington, USA, Technical report, Feb 2008.

[16] A. El-Desoky Mousa, M. A. Basha Shaik, R. Schlüter, and H. Ney, "Morpheme based factored language models for German LVCSR," in *Proc. Interspeech*, Florence, Italy, Aug 2011, pp. 1445–1448.

[17] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proc. the International Conference on New Methods of Language Processing*, Manchester, UK, 1994, pp. 44–49.

[18] S. Sharoff, M. Kopotev, T. Erjavec, A. Feldman, and D. Divjak, "Designing and evaluating russian tagsets," in *Proc. LREC*, Marrakech, Marocco, May 2008, pp. 279–285.

[19] O. Jokisch, A. Wagner, R. Sabo, R. Jaeckel, N. Cylwik, M. Rusko, A. Ronzhin, and R. Hoffmann, "Multilingual speech data collection for the assessment of pronunciation and prosody in a language learning system," in *Proc. SPECOM*, St. Petersburg, Russia, June 2009, pp. 515–520.

[20] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge University Press, 2006.

[21] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine Julius," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Sapporo, Japan, Oct 2009, pp. 131–137.

[22] A. Stolcke, J. Zheng, W. Wang, and V. Abrash, "SRILM at sixteen: Update and outlook," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop ASRU*, Hawaii, USA, Dec 2011.

[23] K. Kirchhoff, J. Bilmes, J. Henderson, R. Schwartz, M. Noamany, P. Schone, G. Ji, S. Das, M. Egan, F. He, D. Vergyri, D. Liu, and N. Duta, "Novel speech recognition models for Arabic," Johns Hopkins University, Technical report, June 2002.

Top

Program Guide

Author Index

iCAST 2013
Technical Program

UMEDIA 2013
Technical Program