Feature-rich sub-lexical language models using a maximum entropy approach for German LVCSR

M. Ali Basha Shaik[1], *Amr El-Desoky Mousa*[1], *Ralf Schlüter*[1], *Hermann Ney*[1,2]

[1]Human Language Technology and Pattern Recognition,
Computer Science Department, RWTH Aachen University, Aachen, Germany
[2]Spoken Language Processing Group, LIMSI CNRS, Paris, France
{shaik,desoky,schlueter,ney}@cs.rwth-aachen.de

Abstract

German is a morphologically rich language having a high degree of word inflections, derivations and compounding. This leads to high out-of-vocabulary (OOV) rates and poor language model (LM) probabilities in the large vocabulary continuous speech recognition (LVCSR) systems. One of the main challenges in the German LVCSR is the recognition of the OOV words. For this purpose, data-driven morphemes are used to provide higher lexical coverage. On the other hand, the probability estimates of a sub-lexical LM could be further improved using feature-rich LMs like maximum entropy (MaxEnt) and class-based LMs. In this work, for a sub-lexical level German LVCSR task, we investigate the use of the multiple morpheme level features as classes for building class-based LMs that are estimated using the state-of-the-art MaxEnt approach. Thus, the benefits of both the MaxEnt LMs and the traditional class-based LMs are effectively combined. Furthermore, we experiment the use of Maximum a-posteriori adaptation over the MaxEnt class-based LMs. We show consistent reductions in both the OOV recognition error rate and the word error rate (WER) on a German LVCSR task from the Quaero project, compared to the traditional class-based and the N-gram morpheme based LM.
Index Terms: open-vocabulary, German LVCSR, features, maximum entropy, class-based

1. Introduction

German is one of the morphologically rich languages having high degree of word inflections, derivations and compounding leading to a wide lexical variety. This causes high OOV rates, data sparsity, and high LM perplexities in the language modeling. Conventional LVCSR systems use a fixed vocabulary consisting of full-words. The words which are not present in the vocabulary are called OOV words and cannot be recognized. In addition, during the recognition, an OOV word potentially substituted by one or more in-vocabulary words leading to neighboring word errors which increases the overall WER [1].

To overcome the OOV problem, sub-words are used in the language modeling. In general, a LM comprising sub-words with or without a fraction of full-words is called a sub-lexical LM. There are different types of sub-word units, for instance morphemes or syllables [2, 3]. A morpheme is defined as the smallest linguistic unit having a semantic meaning. In general, morphemes could be extracted using linguistic or data-driven morphological decomposition. When sub-lexical LMs are used, the data sparsity problem is relatively reduced compared to the full-word LMs, leading to lower OOV rates and higher lexical coverage. Further, as the count based statistics are improved, the LM probability estimates are better estimated [1, 2, 4]. Alternatively, the probability estimates of a sub-lexical LM could be further improved using language dependent features. One of the main objective of this work is to decrease the WER and also the OOV WER over N-gram backoff sub-lexical system using feature-rich LMs. In this work, we investigate the state-of-the-art LMs like maximum entropy LMs and class-based LMs, which provide modular structure to incorporate various knowledge sources as features.

Class-based LMs are initially introduced in [5]. A class-based LM combines the N-gram model over classes with the probability distribution of words in classes leading to the estimation of better smoothed probabilities of word sequences. Class-based LMs have a reduced parameter space due to clustering. An N-gram model is a special case of a class-based LM, where each word is a class itself. On the other hand, the fundamental principle of MaxEnt is initially introduced in [6]. The MaxEnt LMs are investigated in [7, 8, 9]. MaxEnt LM uses the information obtained from various knowledge sources as feature constraints. In general, the knowledge sources could be different types of features having different constraints (i.e., probability distribution functions). MaxEnt LM estimates a unified model in a feature space by selecting the distribution function of the highest entropy satisfying all the constraints from an intersection of all the imposed feature constraints. A token based LM is investigated, which uses word level linguistic features to generate MaxEnt LM on Wall Street Journal (WSJ) task [10]. Linguistic features are also used to estimate whole sentence MaxEnt LMs on switchboard conversational telephone speech task [11]. Recently, an improved class-based LM, namely Model-M, is introduced which combines the strengths of both the MaxEnt principle and the class-based LM [12]. Significant improvements in the results are reported in terms of the WER on the WSJ task using the Model-M. Similarly, class-based LMs for many European languages have been investigated using linguistic and data-driven features [13]. MaxEnt LMs are experimented for the Arabic task using morphological features [8]. Also, for the Greek task, class-based LMs and the MaxEnt LMs are compared by using the linguistic features [14].

2. Feature-rich MaxEnt LMs

Although, MaxEnt LMs provide the flexibility to incorporate various features, they are computationally expensive (CPU and memory resources, normalization factor, $Z(h)$: Eq. 2 in Section 3.4) depending on the vocabulary size and the number of applied feature constraints. On the other hand, for a morphologically rich language like German, using MaxEnt LMs could

help decreasing the perplexity and WER. MaxEnt LMs have been already successful on other inflectional languages for relatively small vocabularies (< 60k) [8, 14]. As we focus on a German large vocabulary task using sub-lexical LMs, we experiment the use of rich morphological features generated on a sub-lexical level using both linguistic and data-driven approaches. This paper presents an approach that attempts to gain the benefits of the feature-rich LMs using the MaxEnt and the class-based LM, while at the same time retain the advantages of sub-lexical LMs. As we experiment on a large vocabulary, certain assumptions are made due to the high resource requirements for generating a MaxEnt models. In this work, all the investigated classes (or morphemic features) are treated as independent classes. Separate MaxEnt models are trained for different features. Then, these MaxEnt models are used to construct class-based LMs. Thus, the benefits of both MaxEnt and class-based LMs are combined using interpolation to estimate better generalized LMs, as uniform as possible.

To verify our proposed approach, we select the same morphemic features for the German LVCSR system as described in our previous work [15]. Three different types of morphological features generated using both the linguistic and data-driven approaches are used. We have shown that feature-rich morpheme based LMs are better than full-word LMs in terms of perplexity and WER. Thus, we focus our investigation only on morpheme based LMs. We compare our proposed approach with the traditional class-based and the N-gram backoff LMs. The use of the *maximum a-posteriori* (MAP) adaptation is also investigated on the generated MaxEnt class-based LMs.

3. Detailed Methodology

In this section, the details of the used morphemes and their corresponding morphological features, and the definitions of the class-based LMs, MaxEnt LMs and the proposed combined approach, followed by supervised and unsupervised MAP adaptation and interpolation are explained.

3.1. Morphemes

To obtain morphemes, the words are decomposed using an open-source tool called Morfessor which is based on the Minimum Description Length principle [16]. The decomposition model is trained using a vocabulary of unique words which occur more than 5 times. The morphemes are *post-processed* in the generated model to avoid very short and noisy morphemes, which could be harmful during recognition. An example of a post-processed decomposition is: *förderung+ s+ ge+ setz* $\rightarrow$ *förderungs+ ge+ setz*. The decomposed word entries are added in the lexicon. To obtain pronunciations for missing lexicon entries, grapheme-to-phoneme (G2P) conversion is used. The pronunciations are aligned for the sub-words from its corresponding full-word pronunciation using the expectation-maximization (EM) algorithm as described in [4].

3.2. Morphological features

In this work, three types of morphological features are used as different knowledge sources. The Part-of-Speech (POS) tags are generated using the probabilistic *TreeTagger* developed by the University of Stuttgart [17]. We use the same tool to extract the lemma of the word. A lemma is the canonical *baseform* of the word. We observe that most of the post-processed morphemes are linguistically meaningful, as shown in Section 3.1. Therefore, the *TreeTagger* could be successfully used to annotate the morphemes with *POS* and *lemma* tags. We call these features as *morpheme level features*. Apart from the above mentioned linguistic features, the data-driven feature is also used, called *index*, which represents a class identity derived by applying word classification using singular value decomposition (SVD) principle. To compute the *index* of the morpheme we convert all the vocabulary entries into a real valued vectors using word-pair co-occurrence matrix and apply SVD. We classify the obtained vectors into N clusters (N=250) using k-means algorithm [18].

3.3. Class-based LMs

A standard class-based LM combines the N-gram model over classes with the probability distribution of words in classes. In principle, a word/sub-word can belong to a single class or can be shared across different classes. The former case is referred to as *hard class membership*, while the latter case is referred to as *ambiguous class membership* or *soft clustering*. In our experiments, we consider ambiguous class membership for the POS features only. If a word is represented as w and c is a class respectively, then an example bigram class-based LM is estimated as :

$$p(w_i|w_{i-1}) = \sum_{c_i,c_{i-1}} p(w_i|c_i)p(c_i|c_{i-1})p(c_{i-1}|w_{i-1}) \quad (1)$$

We create modified Kneser-Ney smoothed class-based N-gram LMs using open-source SRILM toolkit [19].

3.4. MaxEnt Class-based LMs

The formulation to generate MaxEnt LMs to use them effectively in the class-based LM methodology is described. If w is a word/morpheme taken from a vocabulary W, $f(.)$ is the feature function, λ is an optimal weight, h is the context, $Z(h)$ is the normalization factor for all the seen contexts, MaxEnt model can be computed using Eq. 2.

$$p_{me}(w|h) = \frac{e^{\sum_i \lambda_i f_i(w,h)}}{Z(h)} \quad (2)$$

$$\text{Where, } Z(h) = \sum_{w_i \epsilon W} e^{\sum_j \lambda_j f_j(w_i,h)}$$

The separate MaxEnt models for all the morphological features as discussed in Section 3.2 using N-grams as context are generated using Eq. 2. Once the optimal weights λ_i are estimated, the MaxEnt model is smoothed using Gaussian priors. To generate a MaxEnt class-based LM for a given class (feature), we combine the MaxEnt model over this class stream with a probability distribution of words in classes, using in Eq. 1.

3.5. Adaptation

In general, adapted LMs are known to perform better than non-adapted LMs in cases of domain mis-match or if the LM corpus is diverse. In general, the LM data is obtained from multiple domains for LVCSR. It is often unrealistic to significantly reduce the WER without adapting the LM to in-domain data [20]. For this purpose, we apply LM adaptation over MaxEnt class-based LMs. We perform MAP adaptation using Gaussian priors over the MaxEnt models (from Section 3.4). The MaxEnt model is trained on background data including the N-gram features of the in-domain data. The prior parameters computed from the background data are used to learn the parameters from the in-domain data. During MaxEnt training, the prior has zero mean

during Gaussian prior smoothing. But during adaptation, the prior distribution is centered at the background data parameters. The regularized log-likelihood of the adaptation training data is maximized during adaptation. As an in-domain data, we investigate two different types of adaptation, namely supervised and unsupervised [21]. In supervised adaptation, the development data is used as an in-domain data. Whereas, for an unsupervised adaptation, the automatic transcriptions are used from the first pass recognition. Here, the adaptation is performed over both morpheme and feature based MaxEnt models. The MaxEnt and adapted models are created using SRILM-extension [22].

3.6. Interpolation

N-gram backoff LMs are known to perform better in capturing the short range context dependencies. The reason is, when the data is sufficiently available, the likelihood estimates of the frequently occurring N-grams are generally better estimated and reliable. In our experiments, we interpolate morpheme LMs (N-gram and MaxEnt) with class-based LMs (N-gram and MaxEnt) as [23]:

$$p(W) = \sum_{i=1}^{I} \lambda_i p_w^{(i)}(W) + \sum_{j=1}^{J} \lambda_j p_c^{(j)}(W) \qquad (3)$$

Where, for the word or morphemic sequence W, $p_w^{(i)}(W)$ is the i^{th} word-based probability, $p_c^{(j)}(W)$ is the j^{th} class-based probability, λ_i and λ_j are the optimized interpolation weights using development data, satisfying the condition $\sum_{i=1}^{I} \lambda_i + \sum_{j=1}^{J} \lambda_j = 1$. For our experiments, we interpolate 8 LMs : 2 morphemic LMs (1 MaxEnt and 1 N-gram), 6 class-based LMs (3 MaxEnt and 3 N-gram). Alternatively in Eq. 3, MaxEnt model could also be referred to as an adapted MaxEnt model for adaptation experiments.

4. Experimental setup

For the German LVCSR experiment, across word, triphone context, Maximum Likelihood trained acoustic models are used. We use 343 hours of audio data mainly from broadcast news (BN), European parliament plenary sessions (EPPS), spoken dialogs and pod-cast data. A part of the LM training corpus from the multi-domain corpora (BN, EPPS, pod-casts, web data and blogs) is selected. This data is assumed to be close to in-domain data.

The LM training corpus consists of around 188 Million running full-words including the official Quaero project data. The top N most frequent words (N=100k) are selected as a vocabulary from the full-word text. From our previous sub-lexical LVCSR experiments, it is found that keeping 5k most frequent full-words without decomposition (out of 100k items) is quite helpful for the recognition process in terms of the WER [15]. For this reason, 100k top most frequent words (5k full-words + 95k morphemes) are selected. Alternatively, 75k baseforms, 50 POS and 250 index features are used.

Our speech recognizer works in 2 passes. A 3-gram backoff LM is created to construct the search space. The first pass uses speaker independent acoustic model. Then, speaker adaptation is applied using constrained maximum likelihood linear regression (CMLLR) and maximum likelihood linear regression (MLLR). After the second pass, the generated lattices are rescored using a 4-gram LM. In this work, all the proposed LMs are generated using 4-gram as context length. The N-best lists are rescored using the interpolated LMs as described in Section 3.6. To evaluate the performance of our proposed models, Quaero[1] Development-2009 corpus (7.5 hours) and Evaluation-2009 corpus (3.8 hours) is used.

4.1. Experimental considerations

In our sub-lexical experiments, we need to reconstruct the full-words from the morphemes. An identifier '+' is marked at the end of each non-boundary morpheme. After recognition, the recognized morphemes are combined using the pre-defined marker to regenerate the full-words. For example: *förderungs+ ge+ setz* $\rightarrow$ *förderungsgesetz*. Alternatively, the OOV rate of any corpus is computed in such a way that a word is considered an OOV if and only if it is not found in the vocabulary and it is not possible to compose it using in-vocabulary sub-words. We call this as an effective OOV rate.

5. Results

In this Section, the detailed results are described in terms of the word error rate, OOV rate, and the perplexity of the N-gram backoff full-word system (FW) and the sub-lexical system (MW) are shown in Table 1. The results of the proposed approach in comparison to the class-based LM experiments as shown in Table 2.

Table 1: *Baseline recognition results using 100k vocabulary (***FW***: full-word system,***MW***: morpheme system,***OOV***: effective Out-of-vocabulary Rate [%],* **PPL***: Perplexity,* **CER***: Character Error Rate)*

sys	dev09		eval09	
	OOV/PPL	WER/CER [%]	OOV/PPL	WER/CER [%]
FW	4.6/326	32.8/14.5	4.5/344	28.4/12.9
MW	4.1/379	32.3/14.4	3.9/406	28.0/12.7

As shown in Table 1, the development corpus and the evaluation corpus has comparable OOV rates for the full-word and the morpheme based system for the experimented vocabulary. For comparison of all the experimental results, the morpheme system is considered as our reference baseline. The perplexities of the full-word and the morpheme based system are not comparable due to the different vocabularies. For the LM adaptation, the supervised and unsupervised adaptation experiments are conducted using the development corpus and the recognition pass-1 transcriptions respectively (from Section 3.5).

6. Error Analysis

In this Section, the experimental results are validated using error analysis. Recognition related errors arising from the proposed methodology are compared to the full-word and morpheme based line systems. Different types of the errors are analyzed, namely word error rate, in-vocabulary error rate and the OOV recognition error rate followed by the significance tests for the sub-lexical experiments.

6.1. Word Error Rate (WER)

As shown in Table 2 the results are tabulated in terms of the WER. It is observed that for all the investigated features (*B,P,I*), MaxEnt models (class-based and N-gram), interpolated with the conventional class-based models and the backoff N-gram models give consistent improvements interms of reduction both in the perplexity and WER. Similar improvements interms of

[1] http://www.quaero.org

Table 2: *WERs[%] & PPLs for 100k morpheme-based system after 2^{nd} pass rescoring using* **BO***: backoff N-gram + BO class-based ,* **BO+MaxEnt***: BO N-gram + BO class-based + MaxEnt + MaxEnt class-based ;* **B***: baseform;* **P***: POS;* **I***: index ;* **sp/unsp adap***: supervised/unsupervised adaptation,* **CER***:character error rates,* **met.***:error metric*

class	met.	BO		BO+MaxEnt	
		dev09	**eval09**	**dev09**	**eval09**
B	*WER*	32.2	27.7	32.0	27.5
	CER	14.4	12.5	14.2	12.4
	PPL	*357*	*379*	*338*	*357*
P	*WER*	31.9	27.5	31.8	**27.4**
	CER	14.3	12.5	14.3	12.4
	PPL	*355*	*378*	*338*	*357*
I	*WER*	31.9	27.5	**31.7**	**27.4**
	CER	14.4	12.4	14.3	12.3
	PPL	*354*	*375*	*339*	*358*
B,P,I*	*WER*	31.8	27.5	**31.7**	**27.4**
	CER	14.2	12.5	14.3	12.4
	PPL	*343*	*362*	*330*	*346*
with sp adap	*WER*	–	–	29.9†	**27.4**
	CER	–	–	13.5	12.3
	PPL	–	–	174	*332*
with unsp adap	*WER*	–	–	–	**27.4**
	CER	–	–	–	12.4
	PPL	–	–	–	*274*

the WER are achieved using both the supervised and unsupervised adapted models. On the other hand, we notice that, although the obtained improvements are marginal (abs.: 0.1 to 0.2) they are consistent for all the investigated features under high WER conditions. Unsupervised adaptation did not help in further reducing the WER except perplexity. We emphasize that the interpolated non-adapted model (1 N-gram LM + 1 MaxEnt LM + 3 class-based N-gram LMs + 3 MaxEnt class-based LMs) provides a similar WER as the adapted models. This system is referred to as the best system in the remaining Sections of this paper (in Table 3 as s3). For this system, we report significant improvements interms of reductions in WER of around [1.9% (rel.) and 0.6% (abs.)] for the dev-09 corpus, and [2.1% (rel.) and 0.6% (abs.)] for the eval-09 corpus respectively compared to the morpheme based baseline system (MW: Table 1).

6.2. In-vocabulary Error Rate (IER)

As shown in Table 3, vocabulary related error analysis is performed. The impact of the proposed LMs is evaluated by computing the different types of the word errors, namely in-vocabulary error rate (IER) and OOV error rate (OER), apart from the WER. The IER is computed as the ratio of misrecognized in-vocabulary words to the total number of in-vocabulary words. Using the best system (s3), we report reductions in IER of around [1.3% (rel.) and 0.3% (abs.)] for the dev-09 corpus, and [3.1% (rel.) and 0.7% (abs.)] for the eval-09 corpus respectively compared to the morpheme based baseline system (s2). In general, when the sub-words are mixed with the full-words in the morpheme based LMs, an increase in IERs is observed [2, 4]. Although our best system (s3) caused more IERs compared to the full-word system (s1), they are relatively less IERs compared to the morpheme system (s2).

Table 3: *Error Analysis (***sys***: system,* **IER***: in-vocabulary error rate,* **OER***: OOV error rate,* **Tot***: Total WER,* **s1***: full-word system,* **s2***: N-gram morpheme system,* **s3***: Interpolated non-adapted BO+MaxEnt system)*

sys.	error rate [%]					
	dev09			eval09		
	IER	**OER**	**Tot**	**IER**	**OER**	**Tot**
s1	21.4	100	32.8	20.7	100	28.4
s2	24.0	65.0	32.3	22.5	60.5	28.0
s3	**23.7**	**63.5**	**31.7**	**21.8**	**59.4**	**27.4**

6.3. OOV Error Rate (OER)

The OOV error rate (OER) is computed as the ratio of misrecognized OOV words which are not present in the 100k full-word vocabulary to the total number of OOV words. As shown in Table 3, using our best system (s3), we report reductions in OER of around [2.3% (rel.) and 1.5% (abs.)] for the dev-09 corpus, and [1.8% (rel.) and 1.1% (abs.)] for the eval-09 corpus respectively compared to the baseline morpheme system (s2).

6.4. Statistical Significance Test

A significance test comparing the WER of our best system (s3) to the morpheme based baseline system (s2) was performed using the method described in [24]. The experimental results were found to be statistically significant (under 10% significant level, p-value ≤ 0.1).

7. Conclusions

For a German open vocabulary LVCSR system, morphologically rich features are effectively used as classes to estimate class-based LMs in the framework of state-of-the-art MaxEnt LMs. This helps to estimate better generalized LMs to unseen word sequences. Two linguistic features and one data-driven feature are used. It is shown that the performance of our proposed combined language modeling approach is better than the conventional class-based N-gram approach in terms of both the perplexity and the word error rate. Marginal but consistent and statistically significant WER improvements are obtained. Although small improvements are achieved (interms of the word and character related errors), their consistency confirm the usefulness of using morphological features in the MaxEnt class-based LMs for the German LVCSR task. It is verified that using morphemic features as knowledge sources in a better estimated sub-lexical LM helps to recognize significant number of OOVs. In parallel, significant reductions are also obtained for the in-vocabulary error rates and the OOV recognition error rates over both the development and the evaluation corpora compared to the morpheme based baseline system. As a future work, this approach could be further improved by taking into account the mutual dependencies among the features.

8. Acknowledgments

This work was partly funded by the European Community's 7^{th} Framework Programme under the project SCALE (FP7-213850), and partly realized under the Quaero Programme, funded by OSEO, French State agency for innovation. Hermann Ney was partially supported by a senior chair award from DIGITEO, a French research cluster in Ile-de-France.

*best system interms of the WER

†WER/PPL are not meaningful on the dev corpora. Here, they represent the difficulty of the high word error rate LVCSR task.

9. References

[1] M. Bisani and H. Ney, "Open Vocabulary Speech Recognition with Flat Hybrid Models," in *Interspeech*, Lisbon, Portugal, Sep. 2005, pp. 725 – 728.

[2] M. Shaik, A. El-Desoky, R. Schlüter, and H. Ney, "Hybrid Language Models Using Mixed Types of Sub-lexical Units for Open Vocabulary German LVCSR," in *Interspeech*, Florence, Italy, Aug. 2011, pp. 1441 – 1444.

[3] M. Shaik, A. El-Desoky, R. Schlüter, and H. Ney, "Using Morpheme and Syllable Based Sub-words for Polish LVCSR," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011, pp. 4680 – 4683.

[4] A. El-Desoky, M. Shaik, R. Schlüter, and H. Ney, "Sub-lexical language models for German LVCSR," in *IEEE Workshop on Spoken Language Technology*, Berkeley, CA, USA, Dec. 2010, pp. 159 – 164.

[5] P. Brown, P. deSouza, R. Mercer, V. D. Pietra, and J. Lai, "Class-based N-gram models of natural language," *Computational linguistics*, vol. 18, pp. 467 – 479, 1992.

[6] E. Jaynes. (1957) Information theory and statistical mechanics. Physical Review 106. [Online]. Available: http://bayes.wustl.edu/etj/articles/theory.1.pdf

[7] R. Rosenfeld, "Adaptive statistical language modeling: A Maximum Entropy approach," Ph.D. dissertation, Carnegie Mellon University, 1994.

[8] R. Sarikaya, M. Afify, Y. Deng, H. Erdogan, and Y. Gao, "Joint morphological-lexical language modeling for processing morphologically rich languages with application to dialectal Arabic," vol. 16, no. 7, pp. 1330–1339, 2008.

[9] S. F. Chen, "Performance prediction for exponential language models," in *Proc. Human Language Tech. Conf. of the North American Chapter of the ACL*, Boulder, Colorado, USA, May 2009, pp. 450 – 458.

[10] J. Cui, Y. Su, K. Hall, and F. Jelinek, "Investigating linguistic knowledge in a maximum entropy token-based language model ," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Kyoto, Japan, Dec. 2007, pp. 171 – 176.

[11] X. Zhu, S. F. Chen, and R. Rosenfeld, "Linguistic features for whole sentence maximum entropy language models," in *Proc. European Conf. on Speech Communication and Technology*, Budapest, Hungary, Sep. 1999.

[12] S. F. Chen, L. Mangu, B. Ramabhadran, R. Sarikaya, and A. Sethy, "Scaling shrinkage-based language models," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Merano, Italy, Dec. 2009, pp. 299–304.

[13] G. Maltese, P. Bravetti, H. Crépy, B. J. Grainger, M. Herzog, and F. Palou, "Combining word- and class-based language models: a comparative study in several languages using automatic and manual word-clustering techniques," in *Interspeech*, Aalborg, Denmark, Sep. 2001, pp. 21 – 24.

[14] D. Oikonomidis and V. Digalakis, "Stem-based Maximum Entropy language models for inflectional languages," in *Proc. European Conf. on Speech Communication and Technology*, Geneva, Switzerland, Sep. 2003, pp. 2285 – 2288.

[15] A. El-Desoky, M. Shaik, R. Schlüter, and H. Ney, "Morpheme Level Feature-based Language Models for German LVCSR," in *Interspeech*, Portland, OR, USA, Sep. 2012.

[16] M. Creutz and K. Lagus, "Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0," Computer and Information Science Helsinki University of Technology, Finland, Tech. Rep., Mar. 2005.

[17] H. Schmid, "Improvements in Part-of-Speech Tagging with an application to German ," in *ACL-SIGDAT-Workshop*, Dublin, Ireland, Mar. 1995, pp. 47 – 50.

[18] A. El-Desoky, M. Shaik, R. Schlüter, and H. Ney, "Morpheme Based Factored Language Models for German LVCSR," in *Interspeech*, Florence, Italy, Aug. 2011, pp. 1445 – 1448.

[19] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. Int. Conf. on Spoken Language Processing*, vol. 2, Denver, Colorado, USA, Sep. 2002, pp. 901 – 904.

[20] C. Chelba and A. Acero, "Adaptation of Maximum Entropy Capitalizer: Little data can help a lot," *Computer Speech and Language*, vol. 20, no. 4, pp. 382 – 399, 2006.

[21] M. Shaik, A. El-Desoky, R. Schlüter, and H. Ney, "Investigation of Maximum Entropy Hybrid Language Models for Open Vocabulary German and Polish LVCSR," in *Interspeech*, Portland, OR, USA, Sep. 2012.

[22] T. Alumäe and M. Kurimo, "Efficient Estimation of Maximum Entropy Language Models with N-gram features: an SRILM extension," in *Interspeech*, Chiba, Japan, September 2010.

[23] C. Samuelsson and W. Reichl, "A class-based language model for large-vocabulary speech recognition extracted from part-of-speech statistics," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Phoenix, AZ, USA, Mar. 1999, pp. 537 – 540.

[24] N. Parihar and J. Picone, "DSR Front End LVCSR Evaluation - AU/384/02," Aurora Working Group, European Telecommunications Standards Institute, France, Tech. Rep., Dec. 2002.