Corrective Models for Speech Recognition of Inflected Languages

Izhak Shafran and **Keith Hall**
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD 21218
{zakshafran,keith_hall}@jhu.edu

Abstract

This paper presents a corrective model for speech recognition of inflected languages. The model, based on a discriminative framework, incorporates word n-grams features as well as factored morphological features, providing error reduction over the model based solely on word n-gram features. Experiments on a large vocabulary task, namely the Czech portion of the MALACH corpus, demonstrate performance gain of about 1.1–1.5% absolute in word error rate, wherein morphological features contribute about a third of the improvement. A simple feature selection mechanism based on χ^2 statistics is shown to be effective in reducing the number of features by about 70% without any loss in performance, making it feasible to explore yet larger feature spaces.

1 Introduction

N-gram models have long been the stronghold of statistical language modeling approaches. Within the n-gram paradigm, straightforward approaches for increasing accuracy include using larger training sets and augmenting the contextual information within the n-gram window. Incorporating syntactic features into the context has been at the forefront of recent research (Collins et al., 2005; Rosenfeld et al., 2001; Chelba and Jelinek, 2000; Hall and Johnson, 2004). However, much of the previous work has focused on English language syntax. This paper addresses syntax as captured by the inflectional morphology of highly inflected language.

High inflection in a language is generally correlated with some level of word-order flexibility. Morphological features either directly identify or help disambiguate the syntactic participants of a sentence. Inflectional morphology works as a proxy for structured syntax in a language. Modeling morphological features in these languages not only provides an additional source of information but can also alleviate data sparsity problems.

Czech speech recognition needs to deal with two sources of errors which are absent in English, namely, the inflectional morphology and the differences in the formal (written) and colloquial (spoken) forms. Table 1 presents an example output of our speech recognizer on an utterance from a Holocaust survivor, who is recounting General Romel's desert campaign during the Second World War. In this example, the feminine past-tense form of the Czech verb for *to be* is chosen mistakenly, which is followed by a sequence of incorrect words chosen primarily to maintain agreement with the feminine form of the verb. This is an example of what we refer to as the morphological *grouping* effect. When the acoustic model prefers a word with an incorrect inflection, the language model effectively propagates the error to later words. A language model based on word-forms prefers sequences observed in the training data, which will implicitly force an agreement with the inflections of preceding words, making it difficult to stop propagating errors. Although this analysis is anecdotal in nature, the *grouping* effect appears to be prevalent in the Czech dataset used in this work. The proposed corrective model with morphological features is expected to alleviate the *grouping* effect as well as to improve the recognition of inflected languages in general.

In the following section, we present a brief review of related work on morphological language modeling and discriminative language mod-

Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pages 390–398,
Sydney, July 2006.

REF	no	Ježíš	to	už	byl	Romel	hnedle	před	Alexandrií
gloss	well	Jesus	by that time	already	was	Romel	just	in front of	Alexandria
translation	oh Jesus, Romel was already just in front of Alexandria by that time								
HYP	no	Ježíš	to	už	byla	sama	hned	lepší	Alexandrie
gloss	well	Jesus	by that time	already	(she) was	herself	just	better	Alexandria
translation	oh Jesus, she was herself just better Alexandria by that time								

Table 1: An example of the *grouping* effect. The incorrect form of the verb *to be* begins a group of incorrect words in the hypothesis, but these words agree in their morphological inflection.

els. We begin the description of our work in section 3 with the type of morphological features modeled as well as their computation from the output word-lattices of a speech recognizer. Section 4 presents the corrective model and the training approach explored in the current work. A simple and effective feature selection mechanism is described in section 5. In section 6, the proposed framework is evaluated on a large vocabulary Czech speech recognition task. Results show that the morphological features provide a significant improvement over models lacking these features; subsequently, two different analyses are provided to understand the contribution of different morphological features.

2 Related Work

It has long been assumed that incorporating morphological features into a language models should help improve the performance of speech recognition systems. Early models for German showed little improvements over bigram language models and almost no improvement over trigram models (Geutner, 1995). More recently, morphology-based models have been shown to help reduce error rate for out-of-vocabulary words (Carki et al., 2000; Podvesky and Machek, 2005).

Much of the early work on morphological language modeling was focused on utilizing composite morphological tags, largely due to the difficulty in teasing apart the intricate interdependencies of the morphological features. Apart from a few exceptions, there has been little work done in exploring the morphological systems of highly inflected languages.

Kirchhoff and colleagues (2004) successfully incorporated morphological features for Arabic using a factored language model. In their approach, morphological inflections are modeled in a generative framework, and the space of factored morphological tags is explored using a genetic algorithm.

Adopting a different tactic, Choueiter and colleagues (2006) exploited morphological constraints to prune illegal morpheme sequences from ASR output. They noticed that the gains obtained from the application of such constraints in Arabic depends on the size of the vocabulary – an absolute gain of 2.4% in word error rate (WER) reduced to 0.2% when the size was increased from 64k to 800k.

Our approach to modeling morphology differs from that of Vergyri et al. (2004) and Choueiter et al. (2006). By choosing a discriminative framework and maximum entropy based estimation, we allow arbitrary features or constraints and their combinations without the need for explicit elaboration of the factored space and its backoff architecture. Thus, morphological features can be incorporated in the absence of knowledge about their interdependencies.

Several researchers have investigated techniques for improving automatic speech recognition (ASR) results by modeling the errors (Collins et al., 2005; Shafran and Byrne, 2004). Collins et al. (2005) present a corrective language model based on a discriminative framework. Initially, a set of hypotheses is generated by a baseline decoder with standard acoustic and language models. A corrective model is estimated such that it scores desired or oracle hypotheses higher than competing hypotheses. The parameters are learned via the perceptron algorithm which shifts weight away from features associated with poor hypotheses and towards those associated with better hypotheses. By the appropriate choice of desired hypotheses, the model parameters can be estimated to minimize WER in speech recognition. During decoding, the model can then be used to rerank a set of hypotheses, and hence, it is also known as a *reranking* framework. This paradigm allows modeling arbitrary input features, even syntactic features obtained from a parser. We adopt a variant of this framework where the corrective model is based on a conditional model estimated by the maximum entropy procedure (Charniak and John-

son, 2005) and we investigate its effectiveness in modeling morphological features for highly inflected languages, in particular, Czech.

3 Inflectional Morphology

Inflectional abundance in a language generally corresponds to some flexibility in word order. In a free word-order language, the order of sentential participants is relatively unconstrained. This does not mean a speaker of the language can arbitrarily choose an order. Word-order choice may change the semantic and/or pragmatic interpretation of an utterance. Czech is known as a free word-order language allowing for subject, object, and verbal components to come in any order. Morphological inflection in these languages must include a syntactic *case* marker to allow the determination of which participants are subjects (nominative case), objects (accusative or dative) and other such entities. Additionally, morphological inflection encodes features such as gender and number. The agreement of these features between sentential components (adjectives with nouns, subjects with verbs, etc.) may further disambiguate the target of a modifier (e.g., identifying the noun that is modified by a particular adjective).

The increased flexibility in word order aggravates the data sparsity of standard n-gram language model for two reasons: first, the number of valid configurations of a group of words increases with the free order; and second, lexical items are decorated with the inflectional morphemes, multiplying the number of word-forms that appear.

In addition to modeling sequences of word-forms, we model sequences of morphologically reduced *lemmas*, sequence of morphological *tags* and sequences of various factored representations of the morphological tags. Factoring a word into the semantics-bearing lemma and syntax-bearing morphological tag alleviates the data sparsity problem to some extent. However, the number of possible factorizations of n-grams is large. The approach adopted in this work is to provide a rich class of features and defer the modeling of their interaction to the learning procedure.

3.1 Extracting Morphological Features

The extraction of reliable morphological features critically effects further morphological modeling. Here, we first select the most likely morphological analysis for each word using a morphological tagger. In particular, we use the Czech feature-based tagger distributed with the Prague Dependency Treebank (Hajič et al., 2005). The tagger is based on a morphological analyzer which uses a lexicon and a rule-based tag guesser for words not found in the lexicon. Trained by the maximum entropy procedure, the tagger uses left and right contextual features from the input string. Currently, this is the best available Czech-language tagger. See Hajič and Vidová-Hladká (1998) for further details on the tagger.

Label	Description	# Values
lemma	Reduced lexeme	$< \|vocab\|$
POS	Coarse part-of-speech	12
D-POS	Detailed part-of-speech	65
gen	Grammatical Gender	10
num	Grammatical Number	5
case	Grammatical Case	8

Table 2: Czech morphological features used in the current work. The # Values field indicates the size of the closed set of possible values. Not all values are used in the annotated data.

A disadvantage of such an approach is that the tagger works on strings rather than the word-lattices that we expect from an ASR system. Therefore, we must extract a set of strings from the lattices prior to tagging. An alternative approach is to hypothesize all morphological analyses for each word in the lattice, thereby considering the entire set of analyses as features in the model. In the current implementation we have chosen to use a tagger to reduce the complexity of the model by limiting the number of active features while still obtaining relatively reliable features. Moreover, systematic errors in tagging can be potentially compensated by the corrective model.

The initial stage of feature extraction begins with an analysis of the data on which we train and test our models. The process follows:

1. Extract the n-best hypotheses according to a baseline model, where n varies from 50 to 1000 in the current work.

2. Tag each of the hypotheses with the morphological tagger.

3. Re-encode the original word strings along with their tagged morphological analysis in a weighted finite state transducer to allow

Word-form	to	období	bylo	poměrné	krátké
gloss	that	period	was	relatively	short
lemma	ten	období	být	poměrně	krátký
tag	PDNS1	NNNS1	VpNS-	Dg—	AAFS2

Table 3: A morphological analysis of Czech. This analyses was generated by the Hajič tagger.

form	to	období	bylo	poměrné	krátké
	to_období	období_bylo	bylo_poměrné	poměrné_krátké	
lemma	ten	období	být	poměrně	krátký
	ten_období	období_být	být_poměrně	poměrně_krátký	
tag	PDNS1	NNNS1	VpNS-	Dg—	AAFS2
	PDNS1_NNNS1	NNNS1_VpNS-	VpNS-_Dg—	Dg—_AAFS2	
POS	P	N	V	D	A
	P_N	N_V	V_D	D_A	
...			...		
case	1	1	-	-	2
	1_1	1_-	-_0	-_2	
num/case	S1	S1	S-	–	S2
	S1_S1	S1_S-	S-_–	–_S2	
...			...		

Table 4: Examples of the n-grams extracted from the Czech sentence *To období bylo poměrně krátké*. A subset of the feature classes is presented here. The morphological feature values are those assigned by the Hajič tagger.

an efficient means of projecting the hypotheses from word-form to morphology and vice versa.

4. Extract appropriately factored n-gram features for each hypothesis as described below.

Each word state in the original lattice has an associated lemma/tag from which a variety of n-gram features can be extracted.

From the morphological features assigned by the tagger, we chose to retain only a subset and discard the less reliable features which are semantic in nature. The basic morphological features used are detailed in Table 2. In the tag-based model, a string of 5 characters representing the 5 morphological fields is used as a unique identifier. The derived features include n-grams of POS, D-POS, gender (gen), number (num), and case features as well as their combinations.

POS, D-POS Captures the sub-categorization of the part-of-speech tags.

gen, num Captures complex gender-number agreement features.

num, case Captures number agreement between specific case markers.

POS, case Captures associated POS/Case features (e.g., adjectives associated with nominative elements).

The paired features allow for complex inflectional interactions and are less sparse than the composite 5-component morphological tags. Additionally, the morphologically reduced lemma and n-grams of lemmas are used as features in the models.

Table 3 presents a morphological analysis of the Czech sentence *To období bylo poměrně krátké*. The encoded tags represent the first 5 fields of the Prague Dependency Treebank morphological encoding and correspond to the last 5 rows of Table 2. Features for this sentence include the word-form, lemma, and composite tag features as well as the components of each tag and the above mentioned concatenation of tag fields. Additionally, n-grams of each of these features are included. Bigram features extracted from an example sentence are illustrated in Table 4.

The following section describes how the fea-

tures extracted above are modeled in a discriminative framework to reduce word error rate.

4 Corrective Model and Estimation

In this work, we adopt the reranking framework of Charniak and Johnson (2005) for incorporating morphological features. The model scores each test hypothesis y using a linear function, $v_\theta(y)$, of features extracted from the hypothesis $f_j(y)$ and model parameters θ_j, i.e., $v_\theta(y) = \sum_j \theta_j f_j(y)$. The hypothesis with the highest score is then chosen as the output.

The model parameters, θ, are learned from a training set by maximum entropy estimation of the following conditional model:

$$\prod_s \sum_{y_i \in Y_s : g(y_i) = max_j g(y_j)} P_\theta(y_i|Y_s)$$

Here, $Y_s = \{y_j\}$ is the set of hypotheses for each training utterance s and the function g returns an extrinsic evaluation score, which in our case is the WER of the hypothesis. $P_\theta(y_i|Y_s)$ is modeled by a maximum entropy distribution of the form, $P_\theta(y_i|Y_s) = \exp v_\theta(y_i) / \sum_j \exp v_\theta(y_j)$. This choice simplifies the numerical estimation procedure since the gradient of the log-likelihood with respect to a parameter, say θ_j, reduces to difference in expected counts of the associated feature, $E_\theta[f_j|Y_s] - E_\theta[f_j|y_i \in Y_s : g(y_i) = max_j g(y_j)]$. To allow good generalization properties, a Gaussian regularization term is also included in the cost function.

A set of hypotheses Y_s is generated for each training utterance using a baseline ASR system. Care is taken to reduce the bias in decoding the training set by following a jack-knife procedure. The training set is divided into 20 subsets and each subset is decoded after excluding the transcripts of that subset from the language model of the decoder.

The model allows the exploration of a large feature space, including n-grams of words, morphological tags, and factored tags. In a large vocabulary system, this could be an enormous space. However, in a discriminative maximum entropy framework, only the observed features are considered. Among the observed features, those associated with words that are correct in all hypotheses do not provide any additional discrimination capability. Mathematically, the gradient of the log-likelihood with respect to the parameters of these features tends to zero and they may be discarded. Additionally, the parameters associated with features that are rarely observed in the training set are difficult to learn reliably and may be discarded.

To avoid redundant features, we focus on words which are frequently incorrect; this is the *error region* we aim to model. In the training utterance, the error regions of a hypothesis are identified using the alignment corresponding to the minimum edit distance from the reference, akin to computing word error rate. To mark all the error regions in an ASR lattice, the minimum edit distance alignment is obtained using equivalent finite state machine operations (Mohri, 2002). From amongst all the error regions in the training lattices, the most frequent 12k words in error are shortlisted. Features are computed in the corrective model only if they involve words for the shortlist. The parameters, θ, are estimated by numerical optimization as in (Charniak and Johnson, 2005).

5 Feature Selection

The space of features spanned by the cross-product space of words, lemmas, tags, factored-tags and their n-gram can potentially be overwhelming. However, not all of these features are equally important and many of the features may not have a significant impact on the word error rate. The maximum entropy framework affords the luxury of discarding such irrelevant features without much bookkeeping, unlike maximum likelihood models. In the context of modeling morphological features, we investigate the efficacy of simple feature selection based on the χ^2 statistics, which has been shown to effective in certain text categorization problems. e.g. (Yang and Pedersen, 1997).

The χ^2 statistics measures the lack of independence by computing the deviation of the observed counts O_i from the expected counts E_i.

$$\chi^2 = \sum_i (O_i - E_i)^2 / E_i$$

In our case, there are two classes – oracle hypotheses c and competing hypotheses $\bar{c}$. The expected count is the count marginalized over classes.

$$\begin{aligned} \chi^2(f,c) &= \frac{(P(f,c) - P(f))^2}{P(f)} + \frac{(P(f,\bar{c}) - P(f))^2}{P(f)} \\ &+ \frac{(P(\bar{f},c) - P(\bar{f}))^2}{P(\bar{f})} + \frac{(P(\bar{f},\bar{c}) - P(\bar{f}))^2}{P(\bar{f})} \end{aligned}$$

This can be simplified using a two-way contingency table of feature and class, where A is the number of times f and c co-occur, B is the number of times f occurs without c, C is the number of times c occurs without f, and D is the number of times neither f nor c occurs, and N is the total number of examples. Then, the χ^2 is defined to be:

$$\chi^2(f,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}$$

The χ^2 statistics are computed for all the features and the features with larger value are retained. Alternatives feature selection mechanisms such as those based on mutual information and information gain are less reliable than χ^2 statistics for heavy-tailed distributions. More complex feature selection mechanism would entail computing higher order interaction between features which is computationally expensive and so is not explored in this work.

6 Empirical Evaluation

The corrective model presented in this work is evaluated on a large vocabulary task consisting of spontaneous spoken testimonies in Czech language, which is a subset of the multilingual MALACH corpus (Psutka et al., 2003).

6.1 Task

For acoustic model training, transcripts are available for about 62 hours of speech from 336 speakers, amounting to 507k spoken words from a vocabulary of 79k. A portion of this data containing speech from 44 speakers, about 21k words in all is treated as development set (dev). The test set (eval) consists of about 2 hours of speech from 10 new speakers and contains about 15k words.

6.2 Baseline ASR System

The baseline ASR system uses perceptual linear prediction (PLP) features which is computed on 44KHz input speech at the rate of 10 frames per second, and is normalized to have zero mean and unit variance per speaker. The acoustic models are made of 3-state HMM triphones, whose observation distributions are clustered into about 4500 allophonic (triphone) states. Each state is modeled by a 16 component Gaussian mixture with diagonal covariances. The parameters of the acoustic models are initially estimated by maximum likelihood and then refined by five iterations of maximum mutual information estimation (MMI).

Unlike other comparable corpora, this corpus contains a relatively high percentage of colloquial words – about 9% of the vocabulary and 7% of the tokens. For the sake of downstream application, the colloquial variants are subsumed in the lexicon. As a result, common words contain several pronunciation variants, and a few have as many as 14 variants.

For the first pass decoding, a language model was created by interpolating the in-domain model (weight=0.75), estimated from 600k words of transcripts with an out-of-domain model, estimated from 15M words of Czech National Corpus (Psutka et al., 2003). Both models are parameterized by a trigram language model with Katz back-off. The decoding graph was built by composing the language model, the lexical transducer and the context-dependent transducer (phones to triphones) into a single compact finite state machine.

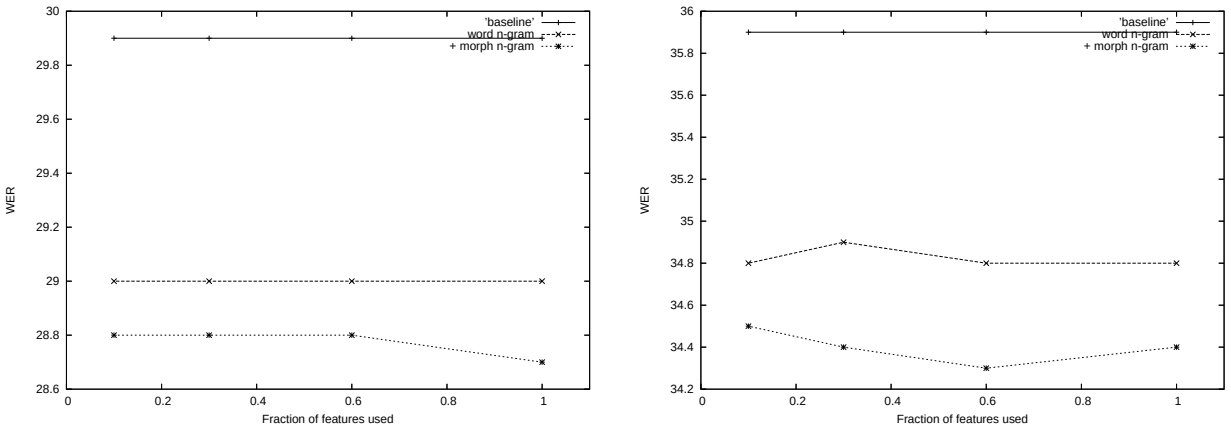
The baseline ASR system decodes test utterance in two passes. A first pass decoding is performed with MMIE acoustic models, whose output transcripts are bootstrapped to estimate two maximum likelihood linear regression transforms for each speaker using five iterations. A second pass decoding is then performed with the new speaker adapted acoustic models. The resulting performance is given in Table 5. The performance reflects the difficulty of transcribing spontaneous speech from the elderly speakers whose speech is also heavily accented and emotional in this corpus.

	1-best	1000-best
Dev	29.9	21.5
Eval	35.9	22.4

Table 5: The performance of the baseline ASR system is reported, showing the word error rate of 1-best MAP hypothesis and the oracle in 1000-best hypotheses for dev and eval sets.

6.3 Experiments With Morphology

We present a set of contrastive experiments to gauge the performance of the corrective models and the contribution of morphological features. For training the corrective models, 50 best hypotheses are generated for each utterance using the



$(a) Devel$ $(b) Eval$

Figure 1: Feature selection via χ^2 statistics helps reduce the number of parameters by 70% without any loss in performance, as observed in dev (a) and eval (b) sets.

jack-knife procedure mentioned earlier. For each hypothesis, bigram and unigram features are computed which consist of word-forms, lemmas, morphologoical tags, factored morphological tags, and the likelihood from the baseline ASR system. For testing, the baseline ASR system is used to generate 1000 best hypotheses for each utterance. These are then evaluated using the corrective models and the best scored hypothesis is chosen as the output.

Table 6 summarizes the results on two test sets – the dev and the eval set. A corrective model with word bigram features improve the word error rate by about an absolute 1% over the baseline. Morphological features provide a further gain on both the test sets consistently.

Features	Dev	Eval
Baseline	29.9	35.9
Word bigram	29.0	34.8
+ Morph bigram	28.7	34.4

Table 6: The word error rate of the corrective model is compared with that of the baseline ASR system, illustrating the improvement in performance with morphological features.

The gains on the dev set are significant at the level of $p < 0.001$ for three standard NIST tests, namely, matched pair sentence segment, signed pair comparison, and Wilcoxon signed rank tests. For the smaller eval set the significant levels were lower for morphological features. The relative gains observed are consistent over a variety of conditions that we have tested including the ones reported below.

Subsequently, we investigated the impact of reducing the number of features using χ^2 statistics, as described in section 5. The experiments with bigram features of word-forms and morphology were repeated using reduced feature sets, and the performance was measured at 10%, 30% and 60% of their original features. The results, as illustrated in Figure 1, show that the word error rate does not change significantly even after the number of features are reduced by 70%. We have also observed that most of the gain can be achieved by evaluating 200 best hypotheses from the baseline ASR system, which could further reduce the computational cost for time-sensitive applications.

6.4 Analysis of Feature Classes

The impact of feature classes can be analyzed by excluding all features from a particular class and evaluating the performance of the resulting model without re-estimation. Figure 2 illustrates the effectiveness of different features class. The y-axis shows the gain in F-score, which is monotonic with the word error rate, on the entire development dataset. In this analysis, the likelihood score from the baseline ASR system was omitted since our interest is in understanding the effectiveness of categorical features such as words, lemmas and tags.

The most independently influential feature class is the factored tag features. This corresponds with

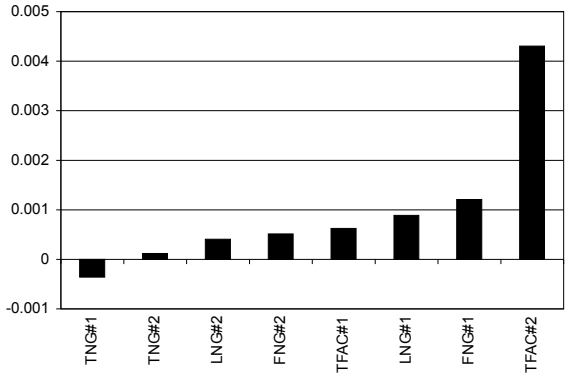


Figure 2: Analysis of features classes for a bigram form, lemma, tag, and factored tag model. Y-axis is the contribution of this feature if added to an otherwise complete model. Feature classes are labeled: TNG – tag n-gram, LNG – lemma n-gram, FNG – form n-gram and TFAC – factored tag n-grams. The number following the # represents the order of the n-gram.

our belief that modeling morphological features requires detailed models of the morphology; in this model the composite morphological tag n-gram features (TNG) offer little contribution in the presence of the factored features.

Analysis of feature reduction by the χ^2 statistics reveals a similar story. When features are ranked according to their χ^2 statistics, about 57% of the factored tag n-grams occur in the top 10% while only 7% of the word n-grams make it. The lemma and composite tag n-grams give about 6.2% and 19.2% respectively. Once again, the factored tag is the most influential feature class.

7 Conclusion

We have proposed a corrective modeling framework for incorporating inflectional morphology into a discriminative language model. Empirical results on a difficult Czech speech recognition task support our claim that morphology can help improve speech recognition results for these types of languages. Additionally, we present a feature selection method that effectively reduces the model size by about 70% while having little or no impact on recognition accuracy. Model size reduction greatly reduces training time which can often be prohibitively expensive for maximum entropy training.

Analysis of the models learned on our task show that factored morphological tags along with word-forms provide most of the discriminative power; and, in the presence of these features, composite morphological tags are of little use.

The corrective model outlined here operates on the word lattices produced by an ASR system. The morphological tags are inferred from the word sequences in the lattice. Alternatively, by employing an ASR system that models the morphological constraints in the acoustics as in (Chung and Seneff, 1999), the corrective model could be applied directly to a lattice with morphological tags.

When dealing with ASR word lattices, the efficacy of the proposed feature selection mechanism can be exploited to eliminate the intermediate tagger, a potential source of errors. Instead of considering the best morphological analysis, the model could consider all possible analyses of the words. Further, the feature space could be enriched with syntactic features which are known to be useful (Collins et al., 2005). The task of modeling is then tackled by feature selection and the maximum entropy training procedure.

8 Acknowledgements

The authors would like to thank William Byrne for discussions on modeling aspects, and Jan Hajič, Petr Němec, and Vaclav Novák for discussions regarding Czech morphology and tagging. This work was supported by the NSF (U.S.A) under the Information Technology Research (ITR) program, NSF IIS Award No. 0122466.

References

Kenan Carki, Petra Geutner, and Tanja Schultz. 2000. Turkish LVCSR: towards better speech recognition for agglutinative languages. In *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3688–3691.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*.

Ciprian Chelba and Frederick Jelinek. 2000. Structured language modeling. *Computer Speech and Language*, 14(4):283–332.

Ghinwa Choueiter, Daniel Povey, Stanley Chen, and Geoffrey Zweig. 2006. Morpheme-based language modeling for Arabic LVCSR. In *Proceedings of the 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France.

Grace Chung and Stephanie Seneff. 1999. A hierarchical duration model for speech recognition based

on the ANGIE framework. *Speech Communication*, 27:113–134.

Michael Collins, Brian Roark, and Murat Saraclar. 2005. Discriminative syntactic language modeling for speech recognition. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 507–514, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Petra Geutner. 1995. Using morphology towards better large-vocabulary speech recognition systems. In *Proceedings of the 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 445–448, Detroit, MI.

Jan Hajič and Barbora Vidová-Hladká. 1998. Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset. In *Proceedings of the COLING-ACL Conference*, pages 483–490, Montreal, Canada.

Jan Hajič, Eva Hajičová, Petr Pajas, Jarmila Panevová, Petr Sgall, and Barbora Vidová Hladká. 2005. The prague dependency treebank 2.0. http://ufal.mff.cuni.cz/pdt2.0.

Keith Hall and Mark Johnson. 2004. Attention shifting for parsing speech. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 41–47, Barcelona.

Mehryar Mohri. 2002. Edit-distance of weighted automata. In *Proceedings of the 7th International Conference on Implementation and Application of Automata, Jean-Marc Champarnaud and Denis Maurel, Eds.*

Petr Podvesky and Pavel Machek. 2005. Speech recognition of Czech—inclusion of rare words helps. In *Proceedings of the ACL Student Research Workshop*, pages 121–126, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Josef Psutka, Pavel Ircing, Josef V. Psutka, Vlasta Radovic, William Byrne, Jan Hajič, Jiri Mirovsky, and Samuel Gustman. 2003. Large vocabulary ASR for spontaneous Czech in the MALACH project. In *Proceedings of the 8th European Conference on Speech Communication and Technology*, Geneva, Switzerland.

Roni Rosenfeld, Stanley F. Chen, and Xiaojin Zhu. 2001. Whole-sentence exponential language models: a vehicle for linguistic-statistical integration. *Computers Speech and Language*, 15(1).

Izhak Shafran and William Byrne. 2004. Task-specific minimum Bayes-risk decoding using learned edit distance. In *Proceedings of the 7th International Conference on Spoken Language Processing*, volume 3, pages 1945–48, Jeju Islands, Korea.

Dimitra Vergyri, Katrin Kirchhoff, Kevin Duh, and Andreas Stolcke. 2004. Morphology-based language modeling for arabic speech recognition. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP/Interspeech 2004)*.

Yiming Yang and Jan 0. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning*, pages 412 – 420, San Francisco, CA, USA.