**Knowledge-Rich Morphological Priors for Bayesian Language Models**

**Victor Chahuneau Noah A. Smith Chris Dyer**

Language Technologies Institute Carnegie Mellon University Pittsburgh, PA 15213, USA

{vchahune,nasmith,cdyer}@cs.cmu.edu

**Abstract**

We present a morphology-aware nonparamet- ric Bayesian model of language whose prior distribution uses manually constructed finite- state transducers to capture the word forma- tion processes of particular languages. This relaxes the word independence assumption and enables sharing of statistical strength across, for example, stems or inflectional paradigms in different contexts. Our model can be used in virtually any scenario where multinomial distributions over words would be used. We obtain state-of-the-art results in language modeling, word alignment, and un- supervised morphological disambiguation for a variety of morphologically rich languages.

**1 Introduction**

Despite morphological phenomena’s salience in most human languages, many NLP systems treat fully inflected forms as the atomic units of language. By assuming independence of lexical stems’ vari- ous surface forms, this *avoidance approach* exacer- bates the problem of data sparseness. If it is em- ployed at all, morphological analysis of text tends to be treated as a preprocessing step to other NLP modules. While this latter *disambiguation approach* helps address data sparsity concerns, it has substan- tial drawbacks: it requires supervised learning from expert-annotated corpora, and determining the op- timal morphological granularity is labor-intensive (Habash and Sadat, 2006).

Neither approach fully exploits the finite-state transducer (FST) technology that has been so suc- cessful for modeling the mapping between surface

forms and their morphological analyses (Karttunen and Beesley, 2005), and the mature collections of high quality transducers that already exist for many languages (e.g., Turkish, Russian, Arabic). Much linguistic knowledge is encoded in such FSTs.

In this paper, we develop morphology-aware non- parametric Bayesian language models that bring to- gether hand-written FSTs with statistical modeling and require no token-level annotation. The sparsity issue discussed above is addressed by hierarchical priors that share statistical strength across different inflections of the same stem by backing off to word formation models that piece together morphemes us- ing FSTs. Furthermore, because of the nonparamet- ric formulation of our models, the regular morpho- logical patterns found in the long tail of word types will rely more heavily on deeper analysis, while fre- quent and idiosyncratically behaved forms are mod- eled opaquely.

Our prior can be used in virtually any generative model of language as a replacement for multino- mial distributions over words, bringing morphologi- cal awareness to numerous applications. For various morphologically rich languages, we show that:

*•* our model can provide rudimentary unsuper-

vised disambiguation for a highly ambiguous

analyzer;

*•* integrating morphology into *n*-gram language models allows better generalization to unseen words and can improve the performance of ap- plications that are truly open vocabulary; and

*•* bilingual word alignment models also bene- fit greatly from sharing translation information

1206

*Proceedings of NAACL-HLT 2013*, pages 1206–1215,

Atlanta, Georgia, 9–14 June 2013. *Q*c 2013 Association for Computational Linguistics

across stems.

We are particularly interested in low-resource sce- narios, where one has to make the most of the small quantity of available data, and overcoming data sparseness is crucial. If analyzers exist in such settings, they tend to be highly ambiguous, and an- notated data for learning to disambiguate are also likely to be scarce or non-existent. Therefore, in our experiments with Russian, we compare two analyz- ers: a rapidly-developed *guesser*, which models reg- ular inflectional paradigms but contains no lexicon or irregular forms, and a high-quality analyzer.

**2 Word Models with Morphology**

In this section, we describe a generative model of word formation based on Pitman-Yor processes that generates word types using a finite-state morpho- logical generator. At a high level, the process first produces lexicons of stems and inflectional patterns; then it generates a lexicon of inflected forms us- ing the finite-state generator. Finally, the inflected forms are used to generate observed data. Different independence assumptions can be made at each of these levels to encode beliefs about where stems, in- flections, and surface forms should share statistical strength.

**2.1 Pitman-Yor Processes**

**2.2 Unigram Morphology Model**

The most basic expression of our model is a uni- gram model of text. So far, we only assume that each word can be analyzed into a stem and a se- quence of morphemes forming an inflection pattern. Let *Gs* be a distribution over stems, *Gp* be a distribu- tion over inflectional patterns, and let GENERATE be

a deterministic mapping from *(*stem*,* pattern*)* pairs

to inflected word forms.1 An inflected word *type* is

generated with the following process, which we des- ignate MP(*Gs, Gd,* GENERATE):

stem *∼ Gs*

pattern *∼ Gp*

word = GENERATE(stem*,* pattern)

For example, in Russian, we might sample stem

= прочий,2 pattern = STEM+Adj+Pl+Dat, and obtain word = прочим.

This model could be used directly to generate ob- served tokens. However, we have said nothing about *Gs* and *Gp*, and the assumption that stems and pat- terns are independent is clearly unsatisfying. We therefore assume that both the stem and the pattern distributions are generated from PY processes, and that MP(*Gs, Gp,* GENERATE) is itself the base dis- tribution of a PYP.

Our work relies extensively on Pitman-Yor pro- cesses, which provide a flexible framework for ex- pressing backoff and interpolation relationships and extending standard models with richer word distri- butions (Pitman and Yor, 1997). They have been shown to match the performance of state-of-the-art language models and to give estimates that follow appropriate power laws (Teh, 2006).

A draw from a Pitman-Yor process (PYP), de- noted *G ∼* PY(*d, θ, G*0), is a discrete distribution

over a (possibly infinite) set of events, which we de- note abstractly *E* . The process is parameterized by a discount parameter 0 *≤ d <* 1, a strength parameter *θ > −d*, and a base distribution *G*0 over the event space *E* .

In this work, our focus is on the base distribution

*G*0. We place vague priors on the hyperparameters

*d ∼* U([0*,* 1]) and (*θ* + *d*) *∼* Gamma(1*,* 1). Infer-

ence in PYPs is discussed below.

*Gs ∼* PY(*ds, θs, G*0)

*Gp ∼* PY(*dp, θp, G*0)

*s*

*p*

*Gw ∼* PY(*d, θ,* MP(*Gs, Gp,* GENERATE))

A draw *Gw* from this PYP is a unigram distribu- tion over tokens.

**2.3 Base Stem Model** *G*0

*s*

In general there are an unbounded number of stems possible in any language, so we set *G*0 to be charac- ter trigram model, which we statically estimate, with Kneser-Ney smoothing, from a large corpus of word types in the language being modeled. While using fixed parameters estimated to maximize likelihood is

*s*

1 The assumption of determinism is only inappropriate in cases of inflectional spelling variants (e.g., *modeled* vs. *mod- elled*) or pronunciation variants (e.g., reduced forms in certain environments).

2 прочий (pronounced [pr5tCij]) = *other*

questionable from the perspective of Bayesian learn- ing, it is tremendously beneficial for computational reasons. For some applications (e.g., word align-

available to compute the marginal base word distri- bution:

ment), the set of possible stems for a corpus *S* can be precomputed, so we will also experiment with using a uniform stem distribution based on this set.

*p*(*w | G*0 ) =

GENERATE(*s,p*)=*w*

*w*

*p*(*s | Gs*) *p*(*p | Gp*)

**2.4 Base Pattern Model** *G*0

*p*

Several choices are possible for the base pattern dis- tribution:

**MP**0 We can assume a uniform *G*0 when the num- ber of patterns is small.

*p*

**MP**1 To be able to generalize to new patterns, we can draw the length of the pattern from a Poisson distribution and generate morphemes one by one from a uniform distribution.

**MP**2 A more informative prior is a Markov chain of morphemes, where each morpheme is generated conditional on the preceding morpheme.

The choice of the base pattern distribution could depend on the complexity of the inflectional patterns produced by the morphological analyzer, reflecting the type of morphological phenomena present in a given language. For example, the number of possi- ble patterns can practically be considered finite in Russian, but this assumption is not valid for lan- guages with more extensive derivational morphol- ogy like Turkish.

**2.5 Posterior Inference**

For most applications, rather than directly gener- ating from a model using the processes outlined above, we seek to infer posterior distributions over latent parameters and structures, given a sample of data.

Although there is no known analytic form of the PYP density, it is possible to marginalize the draws from it and to work directly with observa- tions. This marginalization produces the classi- cal Chinese restaurant process representation (Teh,

Since our approach encodes morphology using

FSTs, which are invertible, this poses no problem.

To illustrate, consider the Russian word прочим, which may be analyzed in several ways:

прочий +Adj +Sg +Neut +Instr прочий +Adj +Sg +Masc +Instr прочий +Adj +Pl +Dat

прочить +Verb +Pl +1P

прочее +Pro +Sg +Ins

Because the set of possible analyses is in general small, marginalization is fast and complex blocked sampling is not necessary.

Finally, to infer hyperparameter values (*d, θ, . . .*), a Metropolis-Hastings update is interleaved with Gibbs sampling steps for the rest of the hidden vari- ables.3

Having described a model for generating words, we now show its usage in several contexts.

**3 Unsupervised Morphological**

**Disambiguation**

Given a rule-based morphological analyzer encoded as an unweighted FST and a corpus on which the analyzer has been run – possibly generating multi- ple analyses for each token – we can use our un- igram model to learn a probabilistic model of dis- ambiguation in an unsupervised setting (i.e., with- out annotated examples). The corpus is assumed to be generated from the unigram distribution *Gw* , and the base stem model is set to a fixed character tri- gram model.4 After learning the parameters of the model, we can find for each word in the vocabulary its most likely analysis and use this as a crude dis- ambiguation step.

2006). When working with the morphology mod-

3 The proposal distribution for Metropolis-Hastings is a Beta

els we are proposing, we also need to marginalize

the different latent forms (stems *s* and patterns *p*) that may have given rise to a given word *w*. Thus, we require that the inverse relation of GENERATE is

distribution (*d*) or a Gamma distribution (*θ* + *d*) centered on the previous parameter values.

4 Experiments suggest that this is important to constrain the

model to realistic stems.

**3.1 Morphological Guessers**

Finite-state morphological analyzers are usually specified in three parts: a **stem lexicon**, which de- fines the words in the language and classifies them into several categories according to their grammat- ical function and their morphological properties; a set of **prefixes and suffixes** that can be applied to each category to form surface words; and possibly **alternation rules** that can encode exceptions and spelling variations. The combination of these parts provides a powerful framework for defining a gener- ative model of words. Such models can be reversed to obtain an analyzer. However, while the two latter parts can be relatively easy to specify, enumerating a comprehensive stem lexicon is a time consuming and necessarily incomplete process, as some cate- gories are truly open-class.

To allow unknown words to be analyzed, one can use a *guesser* that attempts to analyze words missing in the lexicon. Can we eliminate the stem lexicon completely and use only the guesser? This is what we try to do by designing a lexicon-free analyzer for Russian. A guesser was developed in three hours; it is prone to over-generation and produces ambiguous analyses for most words but covers a large number of morphological phe- nomena (gender, case, tense, etc.). For example, the word иврите5 can be correctly analyzed as иврит+Noun+Masc+Prep+Sg but also as the in- correct forms: иврить+Verb+Pres+2P+Pl, иврита+Noun+Fem+Dat+Sg, иври- тя+Noun+Fem+Prep+Sg, and more.

**3.2 Disambiguation Experiments**

We train the unigram model on a 1.7M-word cor- pus of TED talks transcriptions translated into Rus- sian (Cettolo et al., 2012) and evaluate our ana- lyzer against a test set consisting of 1,500 gold- standard analyses obtained from the morphology disambiguation task of the DIALOG 2010 confer- ence (Lyaševskaya et al., 2010).6

Each analysis is composed of a lemma (иврит), a part of speech (Noun), and a sequence of ad- ditional functional morphemes (Masc,Prep,Sg). We consider only open-class categories: nouns, ad-

5 иврите = *Hebrew* (masculine noun, prepositional case)

6 [http://ru-eval.ru](http://ru-eval.ru/)

jectives, adverbs and verbs, and evaluate the output of our model with three metrics: the lemma accu- racy, the part-of-speech accuracy, and the morphol- ogy *F* -measure.7

As a baseline, we consider picking a random anal- ysis from output of the analyzer or choosing the most frequent lemma and the most frequent morpho- logical pattern.8 Then, we use our model with each of the three versions of the pattern model described in §2.2. Finally, as an upper bound, we use the gold standard to select one of the analyses produced by the guesser.

Since our evaluation is not directly comparable to the standard for this task, we use for reference a high-quality analyzer from Xerox9 disambiguated with the MP0 model (all of the models have very close accuracy in this case).

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Lemma | POS | Morph. |
| Random  Frequency | 29.8  31.1 | 70.9  74.4 | 50.2  48.8 |
| Guesser MP0  Guesser MP1  Guesser MP2  Guesser oracle | 60.0  58.9  59.9  68.4 | 82.2  82.5  82.4  84.9 | 66.3  69.5  65.5  83.0 |
| Xerox MP0 | 83.6 | 96.4 | 78.1 |

Table 1: Russian morphological disambiguation.

Considering the amount of effort put in develop- ing the guesser, the baseline POS tagging accuracy is relatively good. However, the disambiguation is largely improved by using our unigram model with respect to all the evaluation categories. We are still far from the performance of a high-quality analyzer but, in absence of such a resource, our technique might be a sensible option. We also note that there is no clear winner in terms of pattern model, and con- clude that this choice is task-specific.

7 *F* -measure computed for the set of additional morphemes and averaged over the words in the corpus.

8 We estimate these frequencies by assuming each analysis of

each token is uniformly likely, then summing fractional counts.

9 <http://open.xerox.com/Services/>

fst-nlp-tools/Pages/morphology

**4 Open Vocabulary Language Models**

We now integrate our unigram model in a hierar- chical Pitman-Yor *n*-gram language model (Fig. 1). The training corpus words are assumed to be

**4.1 Language Modeling Experiments**

We train several trigram models on the Russian TED talks corpus used in the previous section. Our base- line is a hierarchical PY trigram model with a tri-

generated from a distribution *Gn*

*w*

drawn from

gram character model as the base word distribution.

PY(*dn, θn, Gn−*1), where *Gn−*1

is defined recur-

We compare it with our model using the same char-

*w w*

sively down to the base model *G*0 . Previous work

*w*

acter model for the base *stem* distribution. Both of

Teh (2006) simply used *G*0

*w*

= U(*V* ) where *V* is

the morphological analyzers described in the previ-

the word vocabulary, but in our case *G*0

*w*

defined in §2.2.

is the MP

ous section help obtaining perplexity reductions (Ta-

ble 2). We ran a similar experiment on the Turkish version of this corpus (1.6M words) with a high-

*d*3 *, ✓*3 *d*2 *, ✓*2 *d*1 *, ✓*1

*ds , ✓s*

quality analyzer (Oflazer, 1994) and obtain even larger gains (Table 3).

3 2 1 0

|  |  |
| --- | --- |
| Model | ppl |
| PY-character LM  Guesser MP2  Xerox MP2 | 563  530  522 |

*G*

*G*

*G*

*G*

*G*

*w w w s s*

*Gp G*0

*p*

*dp , ✓p*

Table 2: Evaluation of the Russian *n*-gram model.

Figure 1: The trigram version of our language model rep-

resented as a graphical model. *G*1

|  |  |
| --- | --- |
| Model | ppl |
| PY-character LM  Oflazer MP2 | 1,640  1,292 |

*w*

is the unigram model

of §2.2.

We are interested in evaluating our model in an open vocabulary scenario where the ability to ex- plain new unseen words matters. We expect our model to be able to generalize better thanks to the combination of a morphological analyzer and a stem distribution which is less sparse than the word dis- tribution (for example, for the 1.6M word Turkish

corpus, *|V | ≈* 3*.*5*|S| ≈* 140k).

To integrate out-of-vocabulary words in our eval-

Table 3: Evaluation of the Turkish *n*-gram model.

These results can partly be attributed to the high OOV rate in these conditions: 4% for the Russian corpus and 6% for the Turkish corpus.

**4.2 Predictive Text Input**

It is difficult to know whether a decrease in perplex-

uation, we use infinite base distributions: *G*0

*w*

(in the

ity, as measured in the previous section, will result in

baseline model) or *G*0 (in the MP) are character tri-

*s*

gram models. We define perplexity of a held-out test corpus in the standard way:

a performance improvement in downstream applica- tions. As a confirmation that correctly modeling new words matters, we consider a predictive task with a truly open vocabulary and that requires *only* a lan-

ppl = exp

1 *N*

*− N*

*i*=1

log *p* (*wi | wi−n*+1 *· · · wi−*1)

guage model: predictive text input.

Given some text, we encode it using a lossy de- terministic character mapping, and try to recover the

but compared to the common practice, we do not need to discount OOVs from this sum since the model vocabulary is infinite. Note that we also marginalize by summing over all the possible analy- ses for a given word when computing its base prob- ability according to the MP.

original content by computing the most likely word sequence. This task is inspired by predictive text input systems available on cellphones with a 9-key keypad. For example, the string gave me a cup is encoded as 4283 63 2 287, which could also be decoded as: hate of a bus.

Silfverberg et al. (2012) describe a system de- signed for this task in Finnish, which is composed of a weighted finite-state morphological analyzer trained on IRC logs. However, their system is re- stricted to words that are encoded in the analyzer’s lexicon and does not use context for disambiguation.

In our experiments, we use the same Turkish TED talks corpus as the previous section. As a baseline, we use a trigram character language model. We pro- duce a character lattice which encodes all the pos- sible interpretations for a word and compose it with a finite-state representation of the character LM us- ing OpenFST (Allauzen et al., 2007). Alternatively, we can use a unigram word model to decode this lat- tice, backing off to the character language model if no solution is found. Finally, to be able to make use of word context, we can extract the *k* most likely paths according to the character LM and produce a word lattice, which is in turn decoded with a lan- guage model defined over the extracted vocabulary.

regard, our model is a minor variant on traditional *n*- gram models that work with “opaque” word forms. How to best relax this assumption in a computation- ally tractable way is an important open question left for future work.

**5 Word Alignment Model**

Monolingual models of language are not the only models that can benefit from taking into account morphology. In fact, alignment models are a good candidate for using richer word distributions: they assume a target word distribution conditioned on each source word. When the target language is mor- phologically rich, classic independence assumptions produce very weak models unless some kind of pre- processing is applied to one side of the corpus. An alternative is to use our unigram model as a word translation distribution for each source word in the corpus.

Our alignment model is based on a simple variant of IBM Model 2 where the alignment distribution is only controlled by two parameters, *λ* and *p*0 (Dyer et al., 2013). *p*0 is the probability of the null alignment. For a source sentence *f* of length *n*, a target sentence *e* of length *m* and a latent alignment *a*, we define the

following alignment link probabilities (*j /*= 0):

Table 4: Evaluation of Turkish predictive text input.

*p*(*ai* = *j | n, m*) *∝* (1 *− p*0) exp

|  |  |  |
| --- | --- | --- |
| Model | WER | CER |
| Character LM  1-gram + character LM  1-gram MP2  3-gram + character LM  3-gram MP2 | 48.37  8.50  **6.46**  7.86  **5.73** | 16.72  3.28  **2.37**  3.07  **2.15** |

( \

*−λ*

*j*

*i*

We measure word and character error rate (WER, CER) on the predicted word sequence and observe large improvements in both of these metrics by mod- eling morphology, both at the unigram level and when context is used (Table 4).

Preliminary experiments with a corpus of 1.6M Turkish tweets, an arguably more appropriate do- main this task, show smaller but consistent improv- ing: the trigram word error rate is reduced from 26% to 24% when our model is used.

**4.3 Limitations**

While our model is an important step forward in practical modeling of OOVs using morphological

*m − n*

*λ* controls the flatness of this distribution: larger val- ues make the probabilities more peaked around the diagonal of the alignment matrix.

Each target word is then generated given a source word and a latent alignment link from the word

translation distribution *p*(*ei | fai , Gw* ). Note that

this is effectively a unigram distribution over tar-

get words, albeit conditioned on the source word *fj* . Here is where our model differs from classic alignment models: the unigram distribution *Gw* is assumed be generated from a PY process. There are two choices for the base word distribution:

*•* As a baseline, we use a uniform base distribu-

*w*

processes, we have made the linguistically naive as-

tion over the target vocabulary: *G*0

= U(*V* ).

sumption that morphology applies inside the lan- guage’s lexicon but has no effect on the process that put inflected lexemes together into sentences. In this

*•* We define a stem distribution *Gs*[*f* ] for each

source word *f* , a shared pattern distribution *Gp*,

and set *G*0 [*f* ] = MP(*Gs*[*f* ]*, Gp*). In this case,

*w*

we obtain the model depicted in Fig. 2. The stem and the pattern models are also given PY priors with uniform base distribution (*G*0 = U(*S*)).

*s*

Finally, we put uninformative priors on the align- ment distribution parameters: *p*0 *∼* Beta(*α, β*) is collapsed and *λ ∼* Gamma(*k, θ*) is inferred using

Metropolis-Hastings.

*↵,*

*p*0

from Bojar and Prokopová (2006). The morpholog- ical analyzer is provided by Xerox.

**Results** Results are shown in Table 5. Our lightly parameterized model performs much better than IBM Model 4 in these small-data conditions. With an identical model, we find PY priors outperform traditional multinomial distributions. Adding mor- phology further reduced the alignment error rate, for both languages.

|  |  |  |
| --- | --- | --- |
|  | AER | |
| Model | en-tr | en-cs |
| Model 4  EM  PY-U(*V* )  PY-U(*S*) | 52.1  43.8  39.2  **33.8** | 34.5  28.9  25.7  **24.8** |

*dw , ✓w*

*ds , ✓s*

*f Gw*

*Gs* 0

*Gp* 0

*G*

*G*

*s*

*p*

*dp , ✓p*

Table 5: Word alignment experiments on English-Turkish

(en-tr) and English-Czech (en-cs) data.

As an example of how our model generalizes bet- ter, consider the sentence pair in Fig. 3, taken from the evaluation data. The two words composing the Turkish sentence are not found elsewhere in the cor-

Figure 2: Our alignment model, represented as a graphi- cal model.

**Experiments**

We evaluate the alignment error rate of our models for two language pairs with rich morphology on the target side. We compare to alignments inferred us- ing IBM Model 4 trained with EM (Brown et al.,

1993),10 a version of our baseline model (described

above) without PY priors (learned using EM), and the PY-based baseline. We consider two language pairs.

**English-Turkish** We use a 2.8M word cleaned version of the South-East European Times corpus (Tyers and Alperen, 2010) and gold-standard align- ments from Çakmak et al. (2012). Our morphologi- cal analyzer is identical to the one used in the previ- ous sections.

**English-Czech** We use the 1.3M word News

Commentary corpus and gold-standard alignments

10 We use the default GIZA++ stage training scheme: Model 1 + HMM + Model 3 + Model 4.

pus, but several related inflections occur.11 It is

therefore trivial for the stem-base model to find the correct alignment (marked in black), while all the other models have no evidence for it and choose an arbitrary alignment (gray points).

I was not able to

finish my homework

ödevimi bitiremedim

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

Figure 3: A complex Turkish-English word alignment (alignment points in gray: EM/PY-U(*V* ); black: PY- U(*S*)).

**6 Related Work**

Computational morphology has received consider- able attention in NLP since the early work on *two- level morphology* (Koskenniemi, 1984; Kaplan and

11 ödevinin, ödevini, ödevleri; bitmez, bitirileceg˘ inden, bitmesiyle, ...

Kay, 1994). It is now widely accepted that finite- state transducers have sufficient power to express nearly all morphological phenomena, and the XFST toolkit (Beesley and Karttunen, 2003) has con- tributed to the practical adoption of this modeling approach. Recently, open-source tools have been re- leased: in this paper, we used Foma (Hulden, 2009) to develop the Russian guesser.

Since some inflected forms have several possible analyses, there has been a great deal of work on se- lecting the intended one in context (Hakkani-Tür et al., 2000; Hajicˇ et al., 2001; Habash and Rambow,

2005; Smith et al., 2005; Habash et al., 2009). Our disambiguation model is closely related to genera- tive models used for this purpose (Hakkani-Tür et al., 2000).

Rule-based analysis is not the only approach to modeling morphology, and many unsupervised models have been proposed.12 Heuristic segmenta- tion approaches based on the minimum description length principle (Goldsmith, 2001; Creutz and La- gus, 2007; de Marcken, 1996; Brent et al., 1995) have been shown to be effective, and Bayesian model-based versions have been proposed as well (Goldwater et al., 2011; Snyder and Barzilay, 2008; Snover and Brent, 2001). In §3, we suggested a third way between rule-based approaches and fully un- supervised learning that combines the best of both worlds.

Morphological analysis or segmentation is crucial to the performance of several applications: machine translation (Goldwater and McClosky, 2005; Al- Haj and Lavie, 2010; Oflazer and El-Kahlout, 2007; Minkov et al., 2007; Habash and Sadat, 2006, *in- ter alia*), automatic speech recognition (Creutz et al.,

2007), and syntactic parsing (Tsarfaty et al., 2010). Several methods have been proposed to integrate morphology into *n*-gram language models, includ- ing factored language models (Bilmes and Kirch- hoff, 2003), discriminative language modeling (Arı- soy et al., 2008), and more heuristic approaches (Monz, 2011).

Despite the fundamentally open nature of the lex- icon (Heaps, 1978), there has been distressingly lit-

12 Developing a high-coverage analyzer can be a time- consuming process even with the simplicity of modern toolkits, and unsupervised morphology learning is an attractive problem for computational cognitive science.

tle attention to the general problem of open vocabu- lary language modeling problem (most applications make a closed-vocabulary assumption). The classic exploration of open vocabulary language modeling is Brown et al. (1992), which proposed the strategy of interpolating between word- and character-based models. Character-based language models are re- viewed by Carpenter (2005). So-called *hybrid* mod- els that model both words and sublexical units have become popular in speech recognition (Shaik et al.,

2012; Parada et al., 2011; Bazzi, 2002). Open- vocabulary language language modeling has also re- cently been explored in the context of assistive tech- nologies (Roark, 2009).

Finally, Pitman-Yor processes (PYPs) have be- come widespread in natural language processing since they are natural power-law generators. It has been shown that the widely used modified Kneser- Ney estimator (Chen and Goodman, 1998) for *n*- gram language models is an approximation of the posterior predictive distribution of a language model with hierarchical PYP priors (Goldwater et al., 2011; Teh, 2006).

**7 Conclusion**

We described a generative model which makes use of morphological analyzers to produce richer word distributions through sharing of statistical strength between stems. We have shown how it can be in- tegrated into several models central to NLP appli- cations and have empirically validated the effective- ness of these changes. Although this paper mostly focused on languages that are well studied and for which high-quality analyzers are available, our mod- els are especially relevant in low-resource scenarios because they do not require disambiguated analyses. In future work, we plan to apply these techniques to languages such as Kinyarwanda, a resource-poor but morphologically rich language spoken in Rwanda. It is our belief that knowledge-rich models can help bridge the gap between low- and high-resource lan- guages.

**Acknowledgments**

We thank Kemal Oflazer for making his Turkish lan- guage morphological analyzer available to us and Bren- dan O’Connor for gathering the Turkish tweets used in

the predictive text experiments. This work was spon- sored by the U. S. Army Research Laboratory and the U. S. Army Research Office under contract/grant number W911NF-10-1-0533.

**References**

H. Al-Haj and A. Lavie. 2010. The impact of Arabic morphological segmentation on broad- coverage English-to-Arabic statistical machine trans- lation. *Proc. of AMTA*.

Cyril Allauzen, Michael Riley, Johan Schalkwyk, Woj-

ciech Skut, and Mehryar Mohri. 2007. OpenFst: A general and efficient weighted finite-state transducer library. In *Implementation and Application of Au- tomata*, pages 11–23.

Ebru Arısoy, Brian Roark, Izhak Shafran, and Murat

Saraçlar. 2008. Discriminative *n*-gram language mod- eling for Turkish. In *Proc. of Interspeech*.

Issam Bazzi. 2002. *Modelling out-of-vocabulary words*

*for robust speech recognition*. Ph.D. thesis, MIT.

K.R. Beesley and L. Karttunen. 2003. *Finite-state mor- phology: Xerox tools and techniques*. CSLI, Stanford. Jeff A. Bilmes and Katrin Kirchhoff. 2003. Factored language models and generalized parallel backoff. In

*Proc. of NAACL*.

Ondˇrej Bojar and Magdalena Prokopová. 2006. Czech- English word alignment. In *Proc. of LREC*.

Michael R. Brent, Sreerama K. Murthy, and Andrew

Lundberg. 1995. Discovering morphemic suffixes: A case study in MDL induction. In *Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*.

Peter F. Brown, Vincent J. Della Pietra, Stephen A.

Della Pietra, Robert L. Mercer, and Jennifer C. Lai.

1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1):31–40.

P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L.

Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Lin- guistics*, 19(2):263–311.

Bob Carpenter. 2005. Scaling high-order character lan-

guage models to gigabytes. In *Proceedings of the ACL Workshop on Software*.

Mauro Cettolo, Christian Girardi, and Marcello Federico.

2012. WIT3: Web inventory of transcribed and trans- lated talks. In *Proc. of EAMT*.

Stanley F. Chen and Joshua Goodman. 1998. An empiri-

cal study of smoothing techniques for language model- ing. Technical Report TR-10-98, Harvard University.

Mathias Creutz and Krista Lagus. 2007. Unsupervised

models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1).

M. Creutz, T. Hirsimäki, M. Kurimo, A. Puurula, J. Pylkkönen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraçlar, and A. Stolcke. 2007. Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *ACM Transactions on Speech and Language Processing*, 5(1):3.

Carl G. de Marcken. 1996. *Unsupervised Language Ac- quisition*. Ph.D. thesis, MIT.

Chris Dyer, Victor Chahuneau, and Noah A. Smith.

2013. A simple, fast, and effective reparameterization of IBM Model 2. In *Proc. of NAACL*.

J. Goldsmith. 2001. Unsupervised learning of the mor- phology of a natural language. *Computational Lin- guistics*, 27(2):153–198.

S. Goldwater and D. McClosky. 2005. Improving statis- tical MT through morphological analysis. In *Proc. of EMNLP*.

Sharon Goldwater, Thomas L. Griffiths, and Mark John- son. 2011. Producing power-law distributions and damping word frequencies with two-stage language models. *Journal of Machine Learning Research*,

12:2335–2382.

Nizar Habash and Owen Rambow. 2005. Arabic tok- enization, part-of-speech tagging, and morphological disambiguation in one fell swoop. In *Proc. of ACL*.

Nizar Habash and Fatiha Sadat. 2006. Arabic prepro- cessing schemes for statistical machine translation. In *Proc. of NAACL*.

Nizar Habash, Owen Rambow, and Ryan Roth. 2009.

MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In *Proceedings of the Second International Conference on Arabic Lan- guage Resources and Tools*.

Jan Hajicˇ, P. Krbec, P. Kveˇtonˇ , K. Oliva, and V. Petrovicˇ.

2001. Serial combination of rules and statistics. In

*Proc. of ACL*.

D. Z. Hakkani-Tür, Kemal Oflazer, and G. Tür. 2000.

Statistical morphological disambiguation for aggluti- native languages. In *Proc. of COLING*.

Harold Stanley Heaps. 1978. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press.

M. Hulden. 2009. Foma: a finite-state compiler and li- brary. In *Proc. of EACL*.

Ronald M. Kaplan and Martin Kay. 1994. Regular mod- els of phonological rule systems. *Computational Lin- guistics*, 20(3):331–378.

Lauri Karttunen and Kenneth R. Beesley. 2005. Twenty- five years of finite-state morphology. In *Inquiries into Words, Constraints and Contexts*, pages 71–83. CSLI. Kimmo Koskenniemi. 1984. A general computational model for word-form recognition and production. In

*Proc. of ACL-COLING*.

O. Lyaševskaya, I. Astaf’yeva, A. Bonch-Osmolovskaya, A. Garejšina, Y. Grišina, V. D’yacˇkov, M. Ionov, A. Koroleva, M. Kudrinskij, A. Lityagina, Y. Lucˇina, Y. Sidorova, S. Toldova, S. Savcˇuk, and S. Ko- val’. 2010. Ocenka metodov avtomaticˇeskogo analiza teksta: morfologicˇeskie parseri russkogo yazyka. *Komp’juternaya lingvistika i intellektual’nye texnologii (Computational linguistics and intellectual technologies)*.

Einat Minkov, Kristina Toutanova, and Hisami Suzuki.

2007. Generating complex morphology for machine translation. In *Proc. of ACL*.

Christof Monz. 2011. Statistical machine translation with local language models. In *Proc. of EMNLP*.

Kemal Oflazer and ˙Ilknur Durgar El-Kahlout. 2007. Ex- ploring different representational units in English-to- Turkish statistical machine translation. In *Proc. of StatMT*.

K. Oflazer. 1994. Two-level description of Turk- ish morphology. *Literary and Linguistic Computing*,

9(2):137–148.

Carolina Parada, Mark Dredze, Abhinav Sethy, and Ariya Rastrow. 2011. Learning sub-word units for open vo- cabulary speech recognition. In *Proc. of ACL*.

Jim Pitman and Marc Yor. 1997. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25(2):855–90.

Brian Roark. 2009. Open vocabulary language model- ing for binary response typing interfaces. Technical Report CSLU-09-001, Oregon Health & Science Uni- versity.

M. Ali Basha Shaik, David Rybach, Stefan Hahn, Ralf Schluüter, and Hermann Ney. 2012. Hierarchical hy- brid language models for open vocabulary continuous speech recognition using wfst. In *Proc. of SAPA*.

M. Silfverberg, K. Lindén, and M. Hyvärinen. 2012.

Predictive text entry for agglutinative languages using unsupervised morphological segmentation. In *Proc. of Computational Linguistics and Intelligent Text Pro- cessing*.

Noah A. Smith, David A. Smith, and Roy W. Tromble.

2005. Context-based morphological disambiguation with random fields. In *Proc. of EMNLP*.

Matt G. Snover and Michael R. Brent. 2001. A Bayesian model for morpheme and paradigm identification. In *Proc. of ACL*.

Benjamin Snyder and Regina Barzilay. 2008. Unsuper- vised multilingual learning for morphological segmen- tation. In *Proc. of ACL*.

Yee Whye Teh. 2006. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proc. of ACL*.

Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kübler, Marie Candito, Jennifer Foster, Yannick Vers- ley, Ines Rehbein, and Lamia Tounsi. 2010. Statistical parsing of morphologically rich languages: What, how and whither. In *Proc. of Workshop on Statistical Pars- ing of Morphologically Rich Languages*.

F. Tyers and M.S. Alperen. 2010. South-east european times: A parallel corpus of Balkan languages. In *Proceedings of the LREC workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages*.

M. Talha Çakmak, Süleyman Acar, and Güls¸en Eryig˘ it.

2012. Word alignment for English-Turkish language pair. In *Proc. of LREC*.