A Comparative Investigation of Morphological Language Modeling for the Languages of the European Union

Thomas Müller, Hinrich Schütze and Helmut Schmid
Institute for Natural Language Processing
University of Stuttgart, Germany
{muellets,schmid}@ims.uni-stuttgart.de

Abstract

We investigate a language model that combines morphological and shape features with a Kneser-Ney model and test it in a large crosslingual study of European languages. Even though the model is generic and we use the same architecture and features for all languages, the model achieves reductions in perplexity for all 21 languages represented in the Europarl corpus, ranging from 3% to 11%. We show that almost all of this perplexity reduction can be achieved by identifying suffixes by frequency.

1 Introduction

Language models are fundamental to many natural language processing applications. In the most common approach, language models estimate the probability of the next word based on one or more equivalence classes that the history of preceding words is a member of. The inherent productivity of natural language poses a problem in this regard because the history may be rare or unseen or have unusual properties that make assignment to a predictive equivalence class difficult.

In many languages, morphology is a key source of productivity that gives rise to rare and unseen histories. For example, even if a model can learn that words like “large”, “dangerous” and “serious” are likely to occur after the relatively frequent history “potentially”, this knowledge cannot be transferred to the rare history “hypothetically” without some generalization mechanism like morphological analysis.

Our primary goal in this paper is not to develop optimized language models for individual languages. Instead, we investigate whether a simple generic language model that uses shape and morphological features can be made to work well across a large number of languages. We find that this is the case: we achieve considerable perplexity reductions for all 21 languages in the Europarl corpus. We see this as evidence that morphological language modeling should be considered as a standard part of any language model, even for languages like English that are often not viewed as a good application of morphological modeling due to their morphological simplicity.

To understand which factors are important for good performance of the morphological component of a language model, we perform an extensive crosslingual analysis of our experimental results. We look at three parameters of the morphological model we propose: the frequency threshold θ that divides words subject to morphological clustering from those that are not; the number of suffixes used ϕ; and three different morphological segmentation algorithms. We also investigate the differential effect of morphological language modeling on different word shapes: alphabetical words, punctuation, numbers and other shapes.

Some prior work has used morphological models that require careful linguistic analysis and language-dependent adaptation. In this paper we show that simple frequency analysis performs only slightly worse than more sophisticated morphological analysis. This potentially removes a hurdle to using morphological models in cases where sufficient resources to do the extra work required for sophisticated morphological analysis are not available.

The motivation for using morphology in language modeling is similar to distributional clustering

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 386–395,
Montréal, Canada, June 3-8, 2012.

(Brown et al., 1992). In both cases, we form equivalence classes of words with similar distributional behavior. In a preliminary experiment, we find that morphological equivalence classes reduce perplexity as much as traditional distributional classes – a surprising result we intend to investigate in future work.

The main contributions of this paper are as follows. We present a language model design and a set of morphological and shape features that achieve reductions in perplexity for all 21 languages represented in the Europarl corpus, ranging from 3% to 11%, compared to a Kneser-Ney model. We show that identifying suffixes by frequency is sufficient for getting almost all of this perplexity reduction. More sophisticated morphological segmentation methods do not further increase perplexity or just slightly. Finally, we show that there is one parameter that must be tuned for good performance for most languages: the frequency threshold θ above which a word is not subject to morphological generalization because it occurs frequently enough for standard word n-gram language models to use it effectively for prediction.

The paper is organized as follows. In Section 2 we discuss related work. In Section 3 we describe the morphological and shape features we use. Section 4 introduces language model and experimental setup. Section 5 discusses our results. Section 6 summarizes the contributions of this paper.

2 Related Work

Whittaker and Woodland (2000) apply language modeling to morpheme sequences and investigate data-driven segmentation methods. Creutz et al. (2007) propose a similar method that improves speech recognition for highly inflecting languages. They use Morfessor (Creutz and Lagus, 2007) to split words into morphemes. Both approaches are essentially a simple form of a factored language model (FLM) (Bilmes and Kirchhoff, 2003). In a general FLM a number of different back-off paths are combined by a back-off function to improve the prediction after rare or unseen histories. Vergyri et al. (2004) apply FLMs and morphological features to Arabic speech recognition.

These papers and other prior work on using morphology in language modeling have been language-specific and have paid less attention to the question as to how morphology can be useful across languages and what generic methods are appropriate for this goal. Previous work also has concentrated on traditional linguistic morphology whereas we compare linguistically motivated morphological segmentation with frequency-based segmentation and include shape features in our study.

Our initial plan for this paper was to use complex language modeling frameworks that allow experimenters to include arbitrary features (including morphological and shape features) in the model. In particular, we looked at publicly available implementations of maximum entropy models (Rosenfeld, 1996; Berger et al., 1996) and random forests (Xu and Jelinek, 2004). However, we found that these methods do not currently scale to running a large set of experiments on a multi-gigabyte parallel corpus of 21 languages. Similar considerations apply to other sophisticated language modeling techniques like Pitman-Yor processes (Teh, 2006), recurrent neural networks (Mikolov et al., 2010) and FLMs in their general, more powerful form. In addition, perplexity reductions of these complex models compared to simpler state-of-the-art models are generally not large.

We therefore decided to conduct our study in the framework of smoothed n-gram models, which currently are an order of magnitude faster and more scalable. More specifically, we adopt a class-based approach, where words are clustered based on morphological and shape features. This approach has the nice property that the number of features used to estimate the classes does not influence the time needed to train the class language model, once the classes have been found. This is an important consideration in the context of the questions asked in this paper as it allows us to use large numbers of features in our experiments.

3 Modeling of morphology and shape

Our basic approach is to define a number of morphological and shape features and then assign all words with identical feature values to one class. For the morphological features, we investigate three different automatic suffix identification algorithms: Re-s, e, d, ed, n, g, ng, ing, y, t, es, r, a, l, on, er, ion, ted, ly, tion, rs, al, o, ts, ns, le, i, ation, an, ers, m, nt, ting, h, c, te, sed, ated, en, ty, ic, k, ent, st, ss, ons, se, ity, ble, ne, ce, ess, ions, us, ry, re, ies, ve, p, ate, in, tions, ia, red, able, is, ive, ness, lly, ring, ment, led, ned, tes, as, ls, ding, ling, sing, ds, ded, ian, nce, ar, ating, sm, ally, nts, de, nd, ism, or, ge, ist, ses, ning, u, king, na, el

Figure 1: The 100 most frequent English suffixes in Europarl, ordered by frequency

ports (Keshava and Pitler, 2006), Morfessor (Creutz and Lagus, 2007) and Frequency, where Frequency simply selects the most frequent word-final letter sequences as suffixes. The 100 most frequent suffixes found by Frequency for English are given in Figure 1.

We use the ϕ most frequent suffixes for all three algorithms, where ϕ is a parameter. The focus of our work is to evaluate the utility of these algorithms for language modeling; we do not directly evaluate the quality of the suffixes.

A word is segmented by identifying the longest of the ϕ suffixes that it ends with. Thus, each word has one suffix feature if it ends with one of the ϕ suffixes and none otherwise.

In addition to suffix features, we define features that capture shape properties: capitalization, special characters and word length. If a word in the test set has a combination of feature values that does not occur in the training set, then it is assigned to the class whose features are most similar. We described the similarity measure and details of the shape features in prior work (Müller and Schütze, 2011). The shape features are listed in Table 1.

4 Experimental Setup

Experiments are performed using srilm (Stolcke, 2002), in particular the Kneser-Ney (KN) and generic class model implementations. Estimation of optimal interpolation parameters is based on (Bahl et al., 1991).

4.1 Baseline

Our baseline is a modified KN model (Chen and Goodman, 1999).

4.2 Morphological class language model

We use a variation of the model proposed by Brown et al. (1992) that we developed in prior work on English (Müller and Schütze, 2011). This model is a class-based language model that groups words into classes and replaces the word transition probability by a class transition probability and a word emission probability:

$$P_C(w_i|w_{i-N+1}^{i-1}) = \\ P(g(w_i)|g(w_{i-N+1}^{i-1})) \cdot P(w_i|g(w_i))$$

where $g(w)$ is the class of word w and we write $g(w_i \dots w_j)$ for $g(w_i) \dots g(w_j)$.

Our approach targets rare and unseen histories. We therefore exclude all frequent words from clustering on the assumption that enough training data is available for them. Thus, clustering of words is restricted to those below a certain token frequency threshold θ. As described above, we simply group all words with identical feature values into one class. Words with a training set frequency above θ are added as singletons. The class transition probability $P(g(w_i)|g(w_{i-N+1}^{i-1}))$ is estimated using Witten-Bell smoothing.[1]

The word emission probability is defined as follows:

$$P(w|c) = \begin{cases} 1 & , \ N(w) > \theta \\ \frac{N(w)}{\sum_{w \in c} N(w)} - \frac{\epsilon(c)}{|c|-1}, & \theta \geq N(w) > 0 \\ \epsilon(c) & , \ N(w) = 0 \end{cases}$$

where $c = g(w)$ is w's class and $N(w)$ is the frequency of w in the training set. The class-dependent out-of-vocabulary (OOV) rate $\epsilon(c)$ is estimated on held-out data. Our final model P_M interpolates P_C with a modified KN model:

$$P_M(w_i|w_{i-1}^{i-N+1}) = \\ \lambda(g(w_{i-1})) \cdot P_{\mathrm{C}}(w_i|w_{i-1}^{i-N+1}) \\ +(1 - \lambda(g(w_{i-1}))) \cdot P_{\mathrm{KN}}(w_i|w_{i-1}^{i-N+1}) \qquad (1)$$

This model can be viewed as a generalization of the simple interpolation $\alpha P_{\mathrm{C}} + (1 - \alpha) P_{\mathrm{W}}$ used by Brown et al. (1992) (where P_{W} is a word n-gram

[1] Witten-Bell smoothing outperformed modified Kneser-Ney (KN) and Good-Turing (GT).

$is\_capital(w)$	first character of w is an uppercase letter
$is\_all\_capital(w)$	$\forall\, c \in w : c$ is an uppercase letter
$capital\_character(w)$	$\exists\, c \in w : c$ is an uppercase letter
$appears\_in\_lowercase(w)$	$\neg capital\_character(w) \vee w' \in \Sigma_T$
$special\_character(w)$	$\exists\, c \in w : c$ is not a letter or digit
$digit(w)$	$\exists\, c \in w : c$ is a digit
$is\_number(w)$	$w \in L([+-\epsilon][0-9]\,(([.,][0-9])\vert[0-9])*)$

Table 1: Shape features as defined by Müller and Schütze (2011). Σ_T is the vocabulary of the training corpus T, w' is obtained from w by changing all uppercase letters to lowercase and $L(expr)$ is the language generated by the regular expression $expr$.

model and P_{C} a class n-gram model). For the setting $\theta = \infty$ (clustering of all words), our model is essentially a simple interpolation of a word n-gram and a class n-gram model except that the interpolation parameters are optimized for each class instead of using the same interpolation parameter α for all classes. We have found that $\theta = \infty$ is never optimal; it is always beneficial to assign the most frequent words to their own singleton classes.

Following Yuret and Biçici (2009), we evaluate models on the task of predicting the next word from a vocabulary that consists of all words that occur more than once in the training corpus and the unknown word UNK. Performing this evaluation for KN is straightforward: we map all words with frequency one in the training set to UNK and then compute $P_{\mathrm{KN}}(\mathrm{UNK}\,|h)$ in testing.

In contrast, computing probability estimates for P_{C} is more complicated. We define the vocabulary of the morphological model as the set of all words found in the training corpus, including frequency-1 words, and one unknown word for each class. We do this because – as we argued above – morphological generalization is only expected to be useful for rare words, so we are likely to get optimal performance for P_{C} if we include all words in clustering and probability estimation, including hapax legomena. Since our testing setup only evaluates on words that occur more than once in the training set, we ideally would want to compute the following estimate when predicting the unknown word:

$$P_C(\mathrm{UNK}_{\mathrm{KN}}\,|h) = \sum_{\{w:N(w)=1\}} P_C(w|h) + \sum_c P_C(\mathrm{UNK}_c\,|h) \qquad (2)$$

where we distinguish the unknown words of the morphological classes from the unknown word used in evaluation and by the KN model by giving the latter the subscript KN.

However, Eq. 2 cannot be computed efficiently and we would not be able to compute it in practical applications that require fast language models. For this reason, we use the modified class model P'_C in Eq. 1 that is defined as follows:

$$P'_C(w|h) = \begin{cases} P_C(w|h) & ,\ N(w) \geq 1 \\ P_C(\mathrm{UNK}_{g(w)}\,|h), & N(w) = 0 \end{cases}$$

P'_C and – by extension – P_M are deficient. This means that the evaluation of P_M we present below is pessimistic in the sense that the perplexity reductions would probably be higher if we were willing to spend additional computational resources and compute Eq. 2 in its full form.

4.3 Distributional class language model

The most frequently used type of class-based language model is the distributional model introduced by Brown et al. (1992). To understand the differences between distributional and morphological class language models, we compare our morphological model P_{M} with a distributional model P_{D} that has exactly the same form as P_{M}; in particular, it is defined by Equations (1) and (2). The only difference is that the classes are morphological for P_{M} and distributional for P_{D}.

The exchange algorithm that was used by Brown et al. (1992) has very long running times for large corpora in standard implementations like srilm. It is difficult to conduct the large number of clusterings necessary for an extensive study like ours using standard implementations.

Language	T/T	ϵ	#Sentences
S bg Bulgarian	.0183	.0094	**181,415**
S cs Czech	.0185	.0097	369,881
S pl Polish	.0189	.0096	358,747
S sk Slovak	.0187	.0088	368,624
S sl Slovene	.0156	.0090	365,455
G da Danish	.0086	.0077	1,428,620
G de German	.0091	.0073	1,391,324
G en English	**.0028**	**.0023**	**1,460,062**
G nl Dutch	.0061	.0048	1,457,629
G sv Swedish	.0090	.0095	1,342,667
E el Greek	.0081	.0079	851,636
R es Spanish	.0040	.0031	1,429,276
R fr French	**.0029**	**.0024**	**1,460,062**
R it Italian	.0040	.0030	1,389,665
R pt Portuguese	.0042	.0032	1,426,750
R ro Romanian	.0142	.0079	**178,284**
U et Estonian	**.0329**	**.0198**	375,698
U fi Finnish	.0231	**.0183**	1,394,043
U hu Hungarian	**.0312**	.0163	364,216
B lt Lithuanian	.0265	.0147	365,437
B lv Latvian	.0182	.0086	363,104

Table 2: Statistics for the 21 languages. S = Slavic, G = Germanic, E = Greek, R = Romance, U = Uralic, B = Baltic. Type/token ratio (T/T) and # sentences for the training set and OOV rate ϵ for the validation set. The two smallest and largest values in each column are bold.

We therefore induce the distributional classes as clusters in a whole-context distributional vector space model (Schütze and Walsh, 2011), a model similar to the ones described by Schütze (1992) and Turney and Pantel (2010) except that dimension words are immediate left and right neighbors (as opposed to neighbors within a window or specific types of governors or dependents). Schütze and Walsh (2011) present experimental evidence that suggests that the resulting classes are competitive with Brown classes.

4.4 Corpus

Our experiments are performed on the Europarl corpus (Koehn, 2005), a parallel corpus of proceedings of the European Parliament in 21 languages. The languages are members of the following families: Baltic languages (Latvian, Lithuanian), Germanic languages (Danish, Dutch, English, German, Swedish), Romance languages (French, Italian, Portuguese, Romanian, Spanish), Slavic languages (Bulgarian, Czech, Polish, Slovak, Slovene), Uralic languages (Estonian, Finnish, Hungarian) and Greek. We only use the part of the corpus that can be aligned to English sentences. All 21 corpora are divided into training set (80%), validation set (10%) and test set (10%). The training set is used for morphological and distributional clustering and estimation of class and KN models. The validation set is used to estimate the OOV rates ϵ and the optimal parameters λ, θ and ϕ. Table 2 gives basic statistics about the corpus. The sizes of the corpora of languages whose countries have joined the European community more recently are smaller than for countries who have been members for several decades.

We see that English and French have the lowest type/token ratios and OOV rates; and the Uralic languages (Estonian, Finnish, Hungarian) and Lithuanian the highest. The Slavic languages have higher values than the Germanic languages, which in turn have higher values than the Romance languages except for Romanian. Type/token ratio and OOV rate are one indicator of how much improvement we would expect from a language model with a morphological component compared to a non-morphological language model.[2]

5 Results and Discussion

We performed all our experiments with an n-gram order of 4; this was the order for which the KN model performs best for all languages on the validation set.

5.1 Morphological model

Using grid search, we first determined on the validation set the optimal combination of three parameters: (i) $\theta \in \{100, 200, 500, 1000, 2000, 5000\}$, (ii) $\phi \in \{50, 100, 200, 500\}$ and (iii) segmentation method. Recall that we only cluster words whose frequency is below θ and only consider the ϕ most

[2] The tokenization of the Europarl corpus has a preference for splitting tokens in unclear cases. OOV rates would be higher for more conservative tokenization strategies.

[4] A two-tailed paired t-test on the improvements by language shows that the morphological model significantly outperforms the distributional model with p=0.0027. A test on the Germanic, Romance and Greek languages yields p=0.19.

		$\mathrm{PP}_{\mathrm{KN}}$	θ^*_{M}	ϕ^*	M^*	PP_{C}	PP_{M}	Δ_{M}	θ^*_{D}	$\mathrm{PP}_{\mathrm{WC}}$	PP_{D}	Δ_{D}
S	bg	74	200	50	f	103	69	0.07	500	141	71	0.04
S	cs	141	500	100	f	217	129	0.08	1000	298	134	0.04
S	pl	148	500	100	m	241	134	0.09	1000	349	141	0.05
S	sk	123	500	200	f	186	111	0.10	1000	261	116	0.06
S	sl	118	500	100	m	177	107	0.09	1000	232	111	0.06
G	da	69	1000	100	r	89	65	0.05	2000	103	65	0.05
G	de	100	2000	50	m	146	94	0.06	2000	150	94	0.06
G	en	55	2000	50	f	73	53	**0.03**	5000	87	53	0.04
G	nl	70	2000	50	r	100	67	0.04	5000	114	67	0.05
G	sv	98	1000	50	m	132	92	0.06	2000	154	92	0.06
E	el	80	1000	100	f	108	73	0.08	2000	134	74	**0.07**
R	es	57	2000	100	m	77	54	0.05	5000	93	54	0.05
R	fr	**45**	1000	50	f	**56**	**43**	0.04	5000	**71**	**42**	0.05
R	it	69	2000	100	m	101	66	0.04	2000	100	66	0.05
R	pt	62	2000	50	m	88	59	0.05	2000	87	59	0.05
R	ro	76	500	100	m	121	70	0.07	1000	147	71	**0.07**
U	et	256	500	100	m	**422**	230	0.10	1000	668	248	0.03
U	fi	**271**	1000	500	f	410	**240**	**0.11**	2000	**706**	**261**	0.04
U	hu	151	200	200	m	222	136	0.09	1000	360	145	**0.03**
B	lt	175	500	200	m	278	161	0.08	1000	426	169	0.03
B	lv	154	500	200	f	237	142	0.08	1000	322	147	0.05

Table 3: Perplexities on the test set for $N = 4$. S = Slavic, G = Germanic, E = Greek, R = Romance, U = Uralic, B = Baltic. θ^*_x, ϕ^* and M^* denote frequency threshold, suffix count and segmentation method optimal on the validation set. The letters f, m and r stand for the frequency-based method, Morfessor and Reports. $\mathrm{PP}_{\mathrm{KN}}$, PP_{C}, PP_{M}, $\mathrm{PP}_{\mathrm{WC}}$, PP_{D} are the perplexities of KN, morphological class model, interpolated morphological class model, distributional class model and interpolated distributional class model, respectively. Δ_x denotes relative improvement: $(\mathrm{PP}_{\mathrm{KN}} - \mathrm{PP}_{\mathrm{x}}) / \mathrm{PP}_{\mathrm{KN}}$. Bold numbers denote maxima and minima in the respective column.[4]

frequent suffixes. An experiment with the optimal configuration was then run on the test set. The results are shown in Table 3. The KN perplexities vary between 45 for French and 271 for Finnish.

The main result is that the morphological model P_{M} consistently achieves better performance than KN (columns PP_{M} and Δ_{M}), in particular for Slavic, Uralic and Baltic languages and Greek. Improvements range from 0.03 for English to 0.11 for Finnish.

Column θ^*_{M} gives the threshold that is optimal for the validation set. Values range from 200 to 2000. Column ϕ^* gives the optimal number of suffixes. It ranges from 50 to 500. The morphologically complex language Finnish seems to benefit from more suffixes than morphologically simple languages like Dutch, English and German, but there are a few languages that do not fit this generalization, e.g., Estonian for which 100 suffixes are optimal.

The optimal morphological segmenter is given in column M^*: f = Frequency, r = Reports, m = Morfessor. The most sophisticated segmenter, Morfessor is optimal for about half of the 21 languages, but Frequency does surprisingly well. Reports is optimal for two languages, Danish and Dutch. In general, Morfessor seems to have an advantage for complex morphologies, but is beaten by Frequency for Finnish and Latvian.

5.2 Distributional model

Columns PP_{D} and Δ_{D} show the performance of the distributional class language model. As one would perhaps expect, the morphological model is superior to the distributional model for morphologically complex languages like Estonian, Finnish and Hungarian. These languages have many suffixes that have

		$\Delta_{\theta^+} - \Delta_{\theta^-}$	θ^+	θ^-	$\Delta_{\phi^+} - \Delta_{\phi^-}$	ϕ^+	ϕ^-	$\Delta_{\mathrm{M}^+} - \Delta_{\mathrm{M}^-}$	M^+	M^-
S	bg	0.03	200	5000	0.01	50	500		f	m
S	cs	0.03	500	5000		100	500		f	r
S	pl	0.03	500	5000	0.01	100	500		m	r
S	sk	0.02	500	5000		200	500	0.01	f	r
S	sl	0.03	500	5000	0.01	100	500		m	r
G	da	0.02	1000	100		100	50		r	f
G	de	0.02	2000	100		50	500		m	f
G	en	0.01	2000	100		50	500		f	r
G	nl	0.01	2000	100		50	500		r	f
G	sv	0.02	1000	100		50	500		m	f
E	el	0.02	1000	100		100	500	0.01	f	r
R	es	0.02	2000	100		100	500		m	r
R	fr	0.01	1000	100		50	500		f	r
R	it	0.01	2000	100		100	500		m	r
R	pt	0.02	2000	100		50	500		m	r
R	ro	0.03	500	5000		100	500		m	r
U	et	0.02	500	5000	0.01	100	50	0.01	m	r
U	fi	0.03	1000	100	0.03	500	50	0.02	f	r
U	hu	0.03	200	5000	0.01	200	50		m	r
B	lt	0.02	500	5000		200	50		m	r
B	lv	0.02	500	5000		200	500		f	r

Table 4: Sensitivity of perplexity values to the parameters (on the validation set). S = Slavic, G = Germanic, E = Greek, R = Romance, U = Uralic, B = Baltic. Δ_{x^+} and Δ_{x^-} denote the relative improvement of P_{M} over the KN model when parameter x is set to the best (x^+) and worst value (x^-), respectively. The remaining parameters are set to the optimal values of Table 3. Cells with differences of relative improvements that are smaller than 0.01 are left empty.

high predictive power for the distributional contexts in which a word can occur. A morphological model can exploit this information even if a word with an informative suffix did not occur in one of the linguistically licensed contexts in the training set. For a distributional model it is harder to learn this type of generalization.

What is surprising about the comparative performance of morphological and distributional models is that there is no language for which the distributional model outperforms the morphological model by a wide margin. Perplexity reductions are lower than or the same as those of the morphological model in most cases, with only four exceptions – English, French, Italian, and Dutch – where the distributional model is better by one percentage point than the morphological model (0.05 vs. 0.04 and 0.04 vs. 0.03).

Column θ^*_{D} gives the frequency threshold for the distributional model. The optimal threshold ranges from 500 to 5000. This means that the distributional model benefits from restricting clustering to less frequent words – and behaves similarly to the morphological class model in that respect. We know of no previous work that has conducted experiments on frequency thresholds for distributional class models and shown that they increase perplexity reductions.

5.3 Sensitivity analysis of parameters

Table 3 shows results for parameters that were optimized on the validation set. We now want to analyze how sensitive performance is to the three parameters θ, ϕ and segmentation method. To this end, we present in Table 4 the best and worst values of each parameter and the difference in perplexity improvement between the two.

Differences of perplexity improvement between best and worst values of θ_{M} range between 0.01

and 0.03. The four languages with the smallest difference 0.01 are morphologically simple (Dutch, English, French, Italian). The languages with the largest difference (0.03) are morphologically more complex languages. In summary, the frequency threshold θ_{M} has a comparatively strong influence on perplexity reduction. The strength of the effect is correlated with the morphological complexity of the language.

In contrast to θ, the number of suffixes ϕ and the segmentation method have negligible effect on most languages. The perplexity reductions for different values of ϕ are 0.03 for Finnish, 0.01 for Bulgarian, Estonian, Hungarian, Polish and Slovenian, and smaller than 0.01 for the other languages. This means that, with the exception of Finnish, we can use a value of $\phi = 100$ for all languages and be very close to the optimal perplexity reduction – either because 100 is optimal or because perplexity reduction is not sensitive to choice of ϕ. Finnish is the only language that clearly benefits from a large number of suffixes.

Surprisingly, the performance of the morphological segmentation methods is very close for 17 of the 21 languages. For three of the four where there is a difference in improvement of ≥ 0.01, Frequency (f) performs best. This means that Frequency is a good segmentation method for all languages, except perhaps for Estonian.

5.4 Impact of shape

The basic question we are asking in this paper is to what extent the sequence of characters a word is composed of can be exploited for better prediction in language modeling. In the final analysis in Table 5 we look at four different types of character sequences and their contributions to perplexity reduction. The four groups are alphabetic character sequences (W), numbers (N), single special characters (P = punctuation), and other (O). Examples for O would be “751st” and words containing special characters like “O’Neill”. The parameters used are the optimal ones of Table 3. Table 5 shows that the impact of special characters on perplexity is similar across languages: $0.04 \leq \Delta_P \leq 0.06$. The same is true for numbers: $0.23 \leq \Delta_N \leq 0.33$, with two outliers that show a stronger effect of this class: Finnish $\Delta_N = 0.38$ and German $\Delta_N = 0.40$.

		Δ_W	Δ_P	Δ_N	Δ_O
S	bg	0.07	0.04	0.28	0.16
S	cs	0.09	0.04	0.26	0.33
S	pl	0.10	0.05	0.23	0.22
S	sk	0.10	0.05	0.25	0.28
S	sl	0.10	0.04	0.28	0.28
G	da	0.05	0.05	0.31	0.18
G	de	0.06	0.05	0.40	0.18
G	en	0.03	0.04	0.33	0.14
G	nl	0.04	0.05	0.31	0.26
G	sv	0.06	0.05	0.31	0.35
E	el	0.08	0.05	0.33	0.14
R	es	0.05	0.04	0.26	0.14
R	fr	0.04	0.04	0.29	0.01
R	it	0.04	0.05	0.33	0.02
R	pt	0.05	0.05	0.28	0.39
R	ro	0.08	0.04	0.25	0.17
U	et	0.11	0.05	0.26	0.26
U	fi	0.12	0.06	0.38	0.36
U	hu	0.10	0.04	0.32	0.23
B	lt	0.08	0.06	0.27	0.05
B	lv	0.08	0.05	0.26	0.19

Table 5: Relative improvements of P_{M} on the validation set compared to KN for histories w_{i-N+1}^{i-1} grouped by the type of w_{i-1}. The possible types are alphabetic word (W), punctuation (P), number (N) and other (O).

The fact that special characters and numbers behave similarly across languages is encouraging as one would expect less crosslinguistic variation for these two classes of words.

In contrast, “true” words (those exclusively composed of alphabetic characters) show more variation from language to language: $0.03 \leq \Delta_W \leq 0.12$. The range of variation is not necessarily larger than for numbers, but since most words are alphabetical words, class W is responsible for most of the difference in perplexity reduction between different languages. As before we observe a negative correlation between morphological complexity and perplexity reduction; e.g., Dutch and English have small Δ_W and Estonian and Finnish large values.

We provide the values of Δ_O for completeness. The composition of this catch-all group varies considerably from language to language. For example, many words in this class are numbers with alphabetic suffixes like “2012-ben” in Hungarian and words with apostrophes in French.

6 Summary

We have investigated an interpolation of a KN model with a class language model whose classes are defined by morphology and shape features. We tested this model in a large crosslingual study of European languages.

Even though the model is generic and we use the same architecture and features for all languages, the model achieves reductions in perplexity for all 21 languages represented in the Europarl corpus, ranging from 3% to 11%, when compared to a KN model. We found perplexity reductions across all 21 languages for histories ending with four different types of word shapes: alphabetical words, special characters, and numbers.

We looked at the sensitivity of perplexity reductions to three parameters of the model: θ, a threshold that determines for which frequencies words are given their own class; ϕ, the number of suffixes used to determine class membership; and morphological segmentation. We found that θ has a considerable influence on the performance of the model and that optimal values vary from language to language. This parameter should be tuned when the model is used in practice.

In contrast, the number of suffixes and the morphological segmentation method only had a small effect on perplexity reductions. This is a surprising result since it means that simple identification of suffixes by frequency and choosing a fixed number of suffixes ϕ across languages is sufficient for getting most of the perplexity reduction that is possible.

7 Future Work

A surprising result of our experiments was that the perplexity reductions due to morphological classes were generally better than those due to distributional classes even though distributional classes are formed directly based on the type of information that a language model is evaluated on – the distribution of words or which words are likely to occur in sequence. An intriguing question is to what extent the effect of morphological and distributional classes is additive. We ran an exploratory experiment with a model that interpolates KN, morphological class model and distributional class model. This model only slightly outperformed the interpolation of KN and morphological class model (column PP_{M} in Table 3). We would like to investigate in future work if the information provided by the two types of classes is indeed largely redundant or if a more sophisticated combination would perform better than the simple linear interpolation we have used here.

Acknowledgments. This research was funded by DFG (grant SFB 732). We would like to thank the anonymous reviewers for their valuable comments.

References

Lalit R. Bahl, Peter F. Brown, Peter V. de Souza, Robert L. Mercer, and David Nahamoo. 1991. A fast algorithm for deleted interpolation. In *Eurospeech*.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Comput. Linguist.*

Jeff A. Bilmes and Katrin Kirchhoff. 2003. Factored language models and generalized parallel backoff. In *NAACL-HLT*.

Peter F. Brown, Peter V. de Souza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Comput. Linguist.*

Stanley F. Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM TSLP*.

Mathias Creutz, Teemu Hirsimäki, Mikko Kurimo, Antti Puurula, Janne Pylkkönen, Vesa Siivola, Matti Varjokallio, Ebru Arisoy, Murat Saraçlar, and Andreas Stolcke. 2007. Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *ACM TSLP*.

Samarth Keshava and Emily Pitler. 2006. A simpler, intuitive approach to morpheme induction. In *PASCAL Morpho Challenge*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *ICSLP*.

Thomas Müller and Hinrich Schütze. 2011. Improved modeling of out-of-vocabulary words using morphological classes. In *ACL*.

Ronald Rosenfeld. 1996. A maximum entropy approach to adaptive statistical language modelling. *Computer Speech & Language*.

Hinrich Schütze and Michael Walsh. 2011. Half-context language models. *Comput. Linguist.*

Hinrich Schütze. 1992. Dimensions of meaning. In *ACM/IEEE Conference on Supercomputing*, pages 787–796.

Andreas Stolcke. 2002. SRILM - An extensible language modeling toolkit. In *Interspeech*.

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

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *JAIR*.

Dimitra Vergyri, Katrin Kirchhoff, Kevin Duh, and Andreas Stolcke. 2004. Morphology-based language modeling for Arabic speech recognition. In *ICSLP*.

E.W.D. Whittaker and P.C. Woodland. 2000. Particle-based language modelling. In *ICSLP*.

Peng Xu and Frederick Jelinek. 2004. Random forests in language modeling. In *EMNLP*.

Deniz Yuret and Ergun Biçici. 2009. Modeling morphologically rich languages using split words and unstructured dependencies. In *ACL-IJCNLP*.