A Tunable Language Model for Statistical Machine Translation

Junfei Guo[1,2] guojf@ims.uni-stuttgart.de
Juan Liu[1] liujuan@whu.edu.cn
Qi Han[3] qi.han@vis.uni-stuttgart.de
Andreas Maletti[2,4] maletti@ims.uni-stuttgart.de

[1] School of Computer, Wuhan University, China
[2] Institute for Natural Language Processing, University of Stuttgart, Germany
[3] Institute for Visualization and Interactive Systems, University of Stuttgart, Germany
[4] Institute of Computer Science, University of Leipzig, Germany

Abstract

A novel variation of modified KNESER-NEY model using monomial discounting is presented and integrated into the MOSES statistical machine translation toolkit. The language model is trained on a large training set as usual, but its new discount parameters are tuned to the small development set. An in-domain and cross-domain evaluation of the language model is performed based on perplexity, in which sizable improvements are obtained. Additionally, the performance of the language model is also evaluated in several major machine translation tasks including Chinese-to-English. In those tests, the test data is from a (slightly) different domain than the training data. The experimental results indicate that the new model significantly outperforms a baseline model using SRILM in those domain adaptation scenarios. The new language model is thus ideally suited for domain adaptation without sacrificing performance on in-domain experiments.

1 Introduction

Language modeling (Manning and Schütze, 2001) is a central, important, and well-studied topic in natural language processing because the obtained language models (LM) are used in many diverse language technology tasks such as machine translation (Koehn, 2010b), speech recognition, and information retrieval (Manning et al., 2008). Most applied language models are based on n-grams, which are sequences of n consecutive words. Abstractly speaking, an n-gram language model represents a probability distribution over sequences of n words. These distributions are typically obtained with the help of maximum likelihood estimation (MLE) from large monolingual corpora. However, they are smoothed to move probability mass to n-grams that are infrequent or unseen in the training data. The most popular smoothing method in statistical machine translation is the modified KNESER-NEY model by Chen and Goodman (1996), which is implemented in language model toolkits such as SRILM by Stolcke (2002) and KenLM by Heafield (2011).

To accommodate rare n-grams, the relative frequencies of n-grams in the training data are slightly discounted. Here we replace the discounting used in the modified KNESER-NEY model by a monomial discounting. This modification allows a simple adjustment (i.e., tuning)

of the obtained language models (via their discount parameters) to different domains via a standard tuning step (similar in principle to the parameter optimization used in statistical machine translation). Our new model is trained on a large monolingual corpus as usual, but its discount parameters are tuned to a small development set, which usually coincides with the tuning set used to tune the parameters of the machine translation system. We demonstrate that very little development data is sufficient to achieve good performance.

In general, an accurate estimation of cross-domain n-grams is difficult to achieve with only knowledge about in-domain n-grams because even huge in-domain training data is typically insufficient to combat cross-domain data sparseness. The standard solution interpolates the large trained LM with an additional (usually small) LM for the target domain. Our model can utilize both types of data in a single model because the tuning step of our monomial discounting model offers a natural way to adapt it to a new domain. The basic n-gram probabilities are trained using the large amount of background training data, but the new discount parameters are adjusted using data from the target domain. The tuning is driven by perplexity (Jelinek et al., 1977) as a standard measure of language model performance. We optimize the discount parameters such that the perplexity is optimal on the development data using a simple grid search (for our two discount parameters). In experiments we observe that the LM perplexity does not deteriorate compared to the baseline, which is a modified KNESER-NEY model as implemented in SRILM. This applies to both the in-domain as well as the cross-domain setups. More precisely, we observe solid improvements in the cross-domain setups and comparable (i.e., the same) performance in in-domain setups. In addition, our model can still be interpolated with a domain-specific LM to improve it even further.

Perplexity is a rather synthetic measure and does not necessarily correlate well with the performance of downstream tasks, such as statistical machine translation, that utilize language models. Consequently, we also confirm the benefits of our new language model with the help of an evaluation of statistical machine translation performance on medium-scale translation tasks (incl. Chinese-to-English and English-to-German). In these experiments we compare our new domain adaptation LM using monomial discounting to the well-known modified KNESER-NEY model that is implemented in both SRILM and KenLM. As translation models we use the popular phrase-based model (Zens et al., 2002) and the hierarchical phrase-based model (Chiang, 2005), which are both implemented in MOSES (Koehn et al., 2007). The obtained experimental results indicate that systems using our language model significantly outperform the baselines that use the SRILM language model. The improvements are particularly pronounced in domain adaptation scenarios. Only in the biological domain for English-to-German translation we observe no improvements at all, but in this case the overall translation quality (and the trained translation model) is potentially too poor to yield reasonable translations, so that the impact of the language model might be minimal.

In summary, we present a small monomial discounting LM, which can easily be tuned to new domains and is thus ideally suited for domain adaptation. This is achieved by optimizing the LM discount parameters on a small target domain corpus. In our experiments, we compared the LM performance of our model to the LM of the popular toolkits SRILM and KenLM. It shows that our language model works well on in-domain as well as cross-domain data. We implemented our model as a new LM in the MOSES statistical machine translation framework of Koehn et al. (2007) and evaluated our model on several major translation tasks. In those experiments we observed significant improvements in most cross-domain translation setups.

2 Related Work

There exists a wealth of different language models and evaluations of them, so we can only recall the basic antecedents of our work. Kneser and Ney (1995) presented a extension

of absolute discounting method and thereby established the popular KNESER-NEY models. Chen and Goodman (1996) proposed the modified KNESER-NEY model in their study, which quickly became the dominant n-gram language model. In addition, they already noted that interpolation generally works better than backoff. Brants et al. (2007) contributed *stupid backoff*, which is slightly cheaper to estimate. Finally, Schütze (2011) proposed a recursive DUPONT-ROSENFELD model with polynomial discounting by interpolating class-based distributions (Brown et al., 1992) with the lower-order distributions. These models achieved improvements in perplexity when compared to the modified KNESER-NEY models. A simple and general scheme for the adaptation of stochastic language models was already presented by Kneser and Steinbiss (1993). Their adaptation method was used to improve a bi-gram language model.

Corresponding to the wealth of language models, there is also a wealth of implementations of them. We only mention IRSTLM by Federico et al. (2008) and MSRLM by Nguyen et al. (2007), which both implement several language models. We implemented our model in SRILM by Stolcke (2002). In our experiments we compare our model against the modified KNESER-NEY models implemented in SRILM and additionally KenLM, which is the recommended language model in the MOSES framework. SRILM is a popular toolkit for building and applying statistical n-gram-based language models and is used in speech recognition, statistical tagging and segmentation, and statistical machine translation. SRILM offers methods to compute the optimal interpolation weights for the corresponding domain models. Heafield (2011) contributed a scalable variant of the modified KNESER-NEY model that does not rely on pruning. KenLM was already evaluated in a statistical machine translation setup and significant improvements in terms of BLEU (Papineni et al., 2002) were observed (Heafield et al., 2013) at the expense of much larger language models.

3 Language Models

In this section, we recall the commonly used modified KNESER-NEY model (KN model), which is also used in our contrastive systems, and introduce the monomial discounting that we add to the KN models. This type of discounting was originally proposed for the POLKN models by Schütze (2011). Naturally, we also present the newly obtained KN models with monomial discounting in detail, which we will evaluate later on.

3.1 Modified Kneser-Ney model

The modified KN model was proposed by Chen and Goodman (1996). We present the general formulation for an n-gram language model. The model parameters are estimated on the training set, from which we extract occurrence counts $c(w)$ for all sequences $w \in \Sigma^{\leq n}$ of length at most n, where Σ is our lexicon. Given $w \in \Sigma^k$ with $k \geq 1$, we let $\mathrm{tail}(w)$ be the sub-sequence excluding just the first position; i.e., if $w = \sigma_1 \cdots \sigma_k$, then $\mathrm{tail}(w) = \sigma_2 \cdots \sigma_k$. Instead of a single discount parameter D (or a constant function D), they proposed to use three discount parameters D_1, D_2, D_3. More precisely, for every $n \geq 1$, $\sigma' \in \Sigma$, and $w' \in \Sigma^{n-1}$, let

$$p_{\mathrm{KN}}^{(n)}(\sigma' \mid w') = \frac{c(w'\sigma') - D(c(w'\sigma'))}{c(w')} + \gamma^{(n)}(w') \cdot p_{\mathrm{KN}}^{(n-1)}(\sigma' \mid \mathrm{tail}(w'))$$

$$D(k) = \begin{cases} 0 & \text{if } k = 0 \\ D_1 & \text{if } k = 1 \\ D_2 & \text{if } k = 2 \\ D_3 & \text{otherwise} \end{cases}$$

where, to make the distribution sum to 1, they set

$$\gamma^{(n)}(w') = \frac{\sum_{k\geq 0} D(k) \cdot |\{\sigma \in \Sigma \mid c(w'\sigma) = k\}|}{c(w')} \ .$$

Kneser and Ney (1995) developed an estimate for the optimal value of their discount parameter D, and Chen and Goodman (1996) derived the analogous values for the modified KN-model:

$$D_i^* = i - (i+1)\frac{n_1 n_{i+1}}{n_1 n_i + 2 n_2 n_i}$$

with $i \in \{1,2,3\}$, where $n_i = |\{w \in \Sigma^n \mid c(w) = i\}|$ is the number of n-grams that appear exactly i times in the training data.

3.2 Our Model

In our monomial-discount domain-adaptation n-gram-based language model, we use exactly the same general approach as in the modified KN-models, but we replace the discount function by the monomial discount. Schütze (2011) proposed a polynomial discounting mechanism originally for his POLKN models with the intuition that the ideal discount $D(k)$ in the model should grow monotonically with k. More precisely, he replaced the KN-discount D by the discounting function E given for two discount parameters ρ and γ by $E(k) = \rho \cdot k^\gamma$. Informally, the parameter γ controls the rate of growth of the discount as a function of k, and the parameter ρ is a classical discount factor that can be scaled for optimal performance. We assume that $0^0 = 0$, so $E(0) = 0$.

We generally compute the n^{th}-level conditional probability $p_{\text{DA}}^{(n)}(\sigma' \mid w')$ given the occurrence counts $c(w'\sigma')$ and $c(w') = \sum_{\sigma\in\Sigma} c(w'\sigma)$. In particular, we only consider lower-order levels if the n-gram was not seen in the training data (following backoff models). Note that we do not distinguish whether an n-gram occurs once or twice. The only remaining distinction is whether an n-gram occurs or not. Naturally, the number of occurrences modifies the discount E. We denote our tunable language model p_{DA} and define it for every $n \geq 1$, $\sigma' \in \Sigma$, and $w' \in \Sigma^{n-1}$ as follows.

$$p_{\text{DA}}^{(n)}(\sigma' \mid w') = \begin{cases} \frac{c(w'\sigma') - E(c(w'\sigma'))}{c(w')} & \text{if } c(w'\sigma') \neq 0 \\ \beta(w') \cdot p_{\text{DA}}^{(n-1)}(\sigma' \mid \text{tail}(w')) & \text{otherwise.} \end{cases}$$

To make the distribution sum to 1, we let

$$\beta(w') = \frac{\sum_{\sigma\in\Sigma} E(c(w'\sigma))}{c(w')} \cdot \Big(\sum_{\sigma\in\Sigma:\ c(w'\sigma)=0} p_{\text{DA}}^{(n-1)}(\sigma \mid \text{tail}(w')) \Big)^{-1} \ .$$

Overall, this LM is a simple, recursive model with monomial discount. We use a simple backoff scheme distinguishing only occurring and non-occurring n-grams. The discount parameters ρ and γ are optimized on a development set before we apply the model. For this tuning step we use heuristic grid search.

To apply our LM to a sentence, we simply multiply the conditional probabilities obtained for the various windows as usual. Let $w = \sigma_1 \cdots \sigma_k$ be the input sentence. Then

$$p_{\text{DA}}^{(n)}(w) = \prod_{i=1}^{k} p_{\text{DA}}^{(n)}(\sigma_i \mid \sigma_{i-k+1} \cdots \sigma_{i-1}) \ ,$$

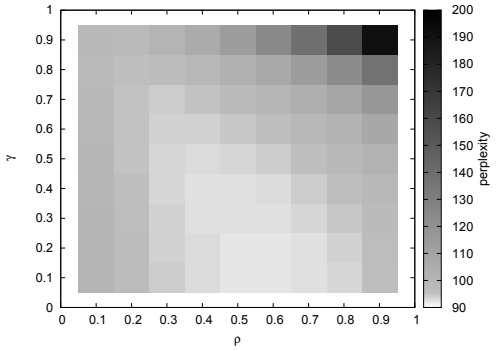

Figure 1: In-domain distribution.

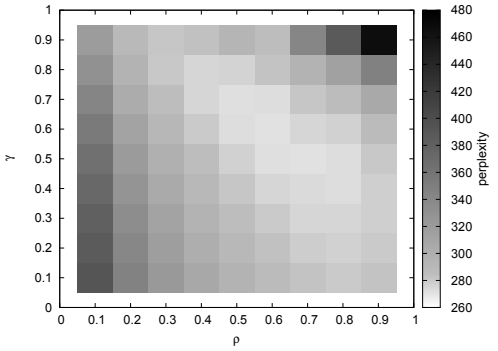

Figure 2: Cross-domain distribution.

where $\sigma_\ell = \text{NULL}$ for all $\ell \leq 0$. The same NULL-tokens are also used in the counts $c(w)$.

While the idea of the monomial discount derives from the POLKN model by Schütze (2011), which is a class-based interpolation model, we apply it without classes as a backoff model. Without the classes, our LM only relies on n-grams and is thus cheaper and easier to generate in ARPA format, which can be processed by SRILM. Since our goal was a tunable LM that can be used for domain adaptation in machine translation, our model needs to be compatible with a toolkit that is supported by the MOSES framework. We selected the SRILM toolkit, which can import our LM given in ARPA format. Naturally, our model can be used for different n-gram orders. An implementation of a class-based model or interpolation models remains future work.

3.3 Parameter Optimization

Our model has two discount parameters: ρ and γ. As already mentioned, we use a tuning step with a small amount of development data to set those parameters. We use perplexity (Jelinek et al., 1977) as the measure of language model performance. In natural language processing, perplexity is the inverse probability of the test set, normalized by the number of words (i.e., the inverse of the geometric mean of the individual word predictions). Let w be a test sentence. We assume that a language model p estimates the probability $p(w')$ of each sequence w'. The perplexity of the LM p on the test sentence w is then defined to be

$$\text{Perplexity}(p, w) = \left(\sqrt[|w|]{p(w)} \right)^{-1} .$$

Consequently, lower perplexity means that the LM is better at predicting the individual words in the test sentence.

Now that we have established our target function, we simply use heuristic grid search to optimize our parameters. More precisely, the parameters ρ and γ are selected from the range $(0, 1)$, and we explore the search space for the optimal discounting parameters with step-size 0.01 and map all the development set perplexities to a grid. To simplify this procedure, we start with step-size 0.1, which yields 81 settings in the straightforward way. Figures 1 and 2 show the perplexities of the development set in those 81 settings. In this way, we obtain the best general areas (brighter is better), for which we then lower the step-size to 0.01 to fine-tune the parameters. We found that even finer step-sizes (such as 0.001) have no effect on performance since the perplexities converge beforehand.

The obtained parameter space looks rather smooth, so we expect our obtained parameters to be close to optimal. We can naturally imagine more refined search methods for the ideal pa-

Language	Domain	Training	Development	ρ	γ
English	in-domain	WSJ	WSJ	0.61	0.03
Chinese	in-domain	MultiUN	MultiUN	0.59	0.10
German	in-domain	EuroParl	EuroParl	0.65	0.11
English	cross-domain	MultiUN	NIST	0.63	0.59
Chinese	cross-domain	MultiUN	NIST	0.64	0.57
German	cross-domain	EuroParl	khreshmoi	0.75	0.54

Table 1: Parameters of the our models. The corpora are presented in Table 2.

rameters, but we do not expect major improvements since our parameters should be almost optimal, whereas traditional methods often only yield local optima. The optimized parameters for different domains are shown in Table 1. As expected, we observe a drastic change of the parameter values for γ comparing the in-domain scenario (English: 0.03; Chinese: 0.10; German: 0.11) to the cross-domain scenario (English: 0.59; Chinese: 0.57; German: 0.54). Thus in the cross-domain scenario, our model reserves more probability mass for the unobserved n-grams in comparison to the in-domain scenario and the modified KNESER-NEY models. The parameters for the German cross-domain LM are particularly large ($\rho = 0.75$ and $\gamma = 0.54$). We can only speculate that the huge difference between the training set (European Parliament proceedings) and the development set (bio-medical data) needs huge discounts to allow for many unseen n-grams. Surprisingly, the cross-domain parameters for English ($\rho = 0.63$ and $\gamma = 0.59$) and Chinese ($\rho = 0.64$ and $\gamma = 0.57$) are very similar. Further evaluations are necessary to detect a trend here, so at present, we do not know the significance of this observation. However, we can observe that high γ-values generally indicate a domain change between the training set and the development set.

Finally, we performed a series of experiments to establish reasonable sizes for the development set. To this end, we optimized the discount parameters for different sizes of the development set. Figures 3 and 4 show the obtained perplexity on the test set for English in relation to the size of the development set. If the development set is tiny (≤ 10 sentences), then we cannot find reasonable discount parameters. However, already at sizes of 20–50 sentences, we find optimal discount parameters that yield very good perplexities also on the test set. For example, in the in-domain experiment we just need 20 sentences to find the parameters $\rho = 0.6$ and $\gamma = 0.1$. With those parameters, we achieve the perplexity 92.32 on the test set, which is already better than standard SRILM, which achieves 92.42. It might be argued that our model had access to additional training examples, but adding, for example, 100 sentences of the development set to the training set for the SRILM models does not influence their perplexity (92.42) since the training data is huge (1.6 million sentences) in comparison to those 100 sentences. In summary, the very small development data does not help as additional training data, but it is enough for our model to optimize the discount parameters, which offer an alternative way to improve the performance. The same observations are true for the cross-domain experiments (and the other languages). In all cases, approximately 100 sentences are sufficient to discover good discount parameters.

3.4 Domain Adaption

We already mentioned that even huge in-domain training data is typically insufficient to combat cross-domain data sparseness. In addition, we have seen that adding the cross-domain development set to the training set is ineffective for small development sets. The standard solution to this problem interpolates the LM for the training set with an additional LM for the target

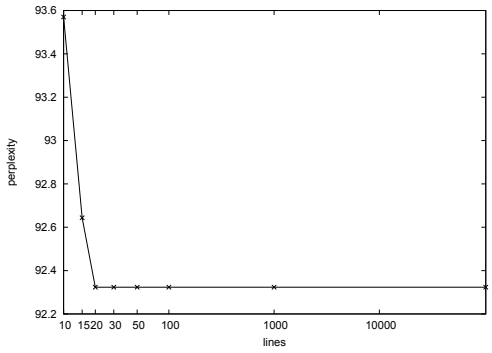

Figure 3: In-domain optimization.

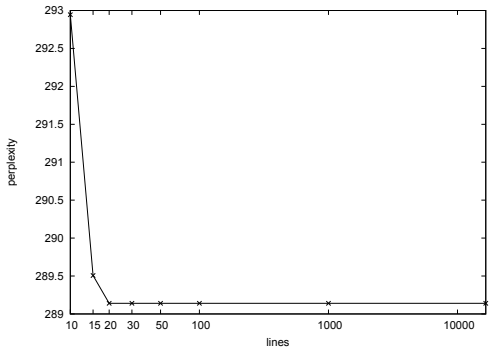

Figure 4: Cross-domain optimization.

domain. These mixture models weight the individual LM and these weights are tuned on the development set. This approach utilizes both types of data independently, which can be beneficial. However, it requires estimating an LM for the target domain, which requires substantial training data in the target domain to be effective. In our setups, in which the target domain development set is small (few thousand sentences), this approach is ineffective since the obtained target domain LM is not useful enough. In contrast, our method only needs to optimize the discount parameters on the development data. Recall that we do not update the occurrence counts of the n-grams. In addition, we actually only need a very small development set (100 sentences) to optimize our discount parameters. It is known that such very small development sets do not help the other models. To confirm these statements, we ran a preliminary experiment. We did not observe any improvements using interpolation with a target domain LM trained on less than 2,000 sentences. In addition, at 2,000 sentences our model still outperforms the interpolated models in terms of test set perplexity. Moreover, the same interpolation approach can be applied to our model, and we observe the same improvements as the development set size increases beyond 2,000 sentences.

4 Experimental Setup

4.1 Corpora

We perform two experiment types: (i) language model experiments for English, Chinese, and German evaluated by perplexity as well as (ii) machine translation experiments for the two language pairs English–Chinese and English–German evaluated by BLEU (Papineni et al., 2002). We summarize the used corpora in Table 2.

For the machine translation experiments on English–Chinese, we use the special IWSLT 2011 release of the sentence-aligned MultiUN corpus of Eisele and Chen (2010) as training data. It is a multilingual parallel corpus extracted from official documents published by the United Nations from 2000 to 2009. This corpus is available in German and all 6 official languages of the United Nations. It contains roughly 300 million words per official language. We use 2 million Chinese–English sentence pairs as training data (48,933,848 English tokens and 47,222,992 Chinese tokens) from the special release provided for IWSLT 2011. For tuning and testing we use the official NIST data provided by LDC (catalog numbers LDC-2010-T10, $\cdots$-T12, $\cdots$-T14, $\cdots$-T17, and $\cdots$-T21). Note that our test sets contain multiple reference translations. The NIST data consists of Chinese news-wire documents, human transcriptions of broadcast news as well as web newsgroup documents. Obviously, the domains of the MultiUN and the NIST data are quite different.

Corpus	Domain	Usage	Sentences	Lang.	Tokens
Wall Street Journal	news	training	1,634,529	English	39,027,486
MultiUN	official documents	training	8,820,000	English Chinese	215,096,536 206,668,165
NIST 2002, -4, -6	news	tuning	4,139	English Chinese	503,305 104,228
NIST 2005	news	test	1,082	English Chinese	139,144 30,060
NIST 2008	news	test	1,859	Chinese English	188,402 45,700
EuroParl	parliament proceedings	training	1,886,260	English German	50,406,502 47,992,387
News Commentary	news	training	200,112	German	5,020,146
Common Crawl	web	training	2,376,881	German	51,889,104
khreshmoi	medical, biology	tuning +test	500+1,000	English German	10,350+21,450 9,924+20,810

Table 2: Used corpora (parts of the tuning data serve as development data for the LM).

The corresponding experiments for English-to-German use all data present in EuroParl version 7 of Koehn (2010a) as training data. EuroParl contains the proceedings of the European parliament in 21 European languages. We use three corpora as training data for the German LM: EuroParl version 7, News Commentary and Common Crawl. The News Commentary corpus contains news text and commentaries from Project Syndicate. It is provided as training data for the shared tasks offered by the workshop on statistical machine translation (WMT). The Common Crawl corpus, which was collected from web sources, was provided as a new data resource for WMT 2013. As tuning and test data we use the bio-medical data of the *khreshmoi* project provided by the WMT 2014 shared task. Again, the domains of the training data and tuning and test data are vastly different.

4.2 Setup

As contrastive language models we use the standard modified KN language models provided by the toolkits SRILM of Stolcke (2002) and KenLM of Heafield et al. (2013). KenLM uses a no-pruning strategy, which it compensates for with its high efficiency allowing it to handle the resulting large models. Since our model works essentially as the models in SRILM, which relies on pruning to reduce the size of the models, we select the modified KN model implemented in SRILM as baseline. We currently employ the same pruning strategy as SRILM, so our models are small compared to models of KenLM and have essentially the same size as the standard SRILM models. It remains to be seen whether the reported advantages can also be obtained using a no-pruning strategy as in KenLM together with our model. For completeness' sake, we also report scores for other models.

All systems are used to generate 5-gram language models in ARPA format. We use the full monolingual data available in the training corpus (e.g., 8.8 million English sentences from

		English			Chinese		
	Model	Size	Perplexity		Size	Perplexity	
Toolkit	Smoothing Method	in GB	Dev.	Test	in GB	Dev.	Test
IRSTLM	improved KN	**208.6**	102.54	102.44	**179.2**	92.49	94.67
KenLM	interpolated mKN	621.2	**91.41**	**91.53**	483.4	82.89	86.11
SRILM	backoff mKN	217.7	94.38	94.34	187.2	84.08	85.31
SRILM	interpolated mKN	217.8	92.48	92.42	187.0	**82.52**	**83.95**
SRILM	our	217.2	92.37	92.32	187.1	83.00	84.58

Table 3: Perplexity and size of the improved Kneser-Ney, interpolated modified Kneser-Ney, backoff version of modified Kneser-Ney and our models on the in-domain data.

MultiUN). Our language model is implemented as a variant of SRILM that implements the different discounting. Our implementation is available on the homepage of the first author (JUNFEI GUO). As mentioned earlier, we use heuristic grid search with step size 0.01 during tuning to discover the optimal discount parameters (see Section 3.3) for our model. An illustration of the results of such a search is presented in Figure 1 and 2.

All the machine translation experiments use the MOSES framework of Koehn et al. (2007). It offers support for phrase-based and hierarchical phrase-based translation models and contains all tools needed to train and execute these models. In particular, it supports the ARPA format of our language models. The word segmentation of the Chinese sentences was achieved with the Stanford Word Segmenter of Chang et al. (2008). GIZA++ (Och and Ney, 2003) with the heuristic *grow-diag-final-and* (Koehn et al., 2005) was used to obtain the word alignments. All the translation models were trained on approximately 1.8 million parallel sentences after standard length-ratio filtering. They were tuned using MERT (Och, 2003) on their respective tuning sets using BLEU (Papineni et al., 2002) as score, which is also the score that we report for the test sets. Finally, the pairwise bootstrap resampling method of Koehn (2004) is used for significance testing.

5 Language Model Perplexity Experiments

First we evaluate the various language models in isolation using perplexity (Jelinek et al., 1977). Since the new feature of our model is the ability to tune the discount parameters, we perform two types of experiments: in-domain and cross-domain. In the in-domain experiments, the tuning and test data are similar to the training data, whereas in the cross-domain scenario the tuning and test data are still similar, but different to the training data. Obviously, we focus on cross-domain experiments since we expect our model to perform well there. A summary of the obtained results (together with the model sizes) is presented in Tables 3 and 4 for the in-domain and the cross-domain scenario, respectively.

5.1 In-domain

The first experiment investigates the performance of the different language models on in-domain (news) data from the *Wall Street Journal*. The training set contains more than 1.6 million sentences and both the validation and the test set have roughly 100,000 sentences. We use the same number of sentences for Chinese from the *MultiUN* corpus. Table 3 shows the performance of the language models (measured by perplexity) together with their size. The IRSTLM models, which are simplified versions of the improved KN model, are the smallest in size, but have the highest (i.e., worst) perplexity. The large unpruned KenLM models have the lowest perplexities for English, but at the expense of significantly larger sizes. The SRILM models using backoff

	ENGLISH			CHINESE			GERMAN		
		Perplexity			Perplexity			Perplexity	
Toolkit	Size	Dev.	Test	Size	Val.	Test	Size	Dev.	Test
KenLM	9.40	296.51	317.94	11.26	950.07	840.15	10.20	613.69	658.86
SRILM	2.20	289.92	312.30	**2.20**	729.40	639.42	1.29	471.39	501.28
our	**2.19**	**271.20**	**286.53**	**2.20**	**669.16**	**584.79**	**1.27**	**442.15**	**469.78**

Table 4: Perplexity and size (in GB) of the models on the cross-domain data.

(without interpolation) score consistently worse than those using interpolation, which was already observed by Chen and Goodman (1996). Our model (without interpolation) outperforms the SRILM interpolated models for English (92.48 vs. 92.37 and 92.42 vs. 92.32 on the development and test set, respectively), but is beaten by the KenLM models (91.41 vs. 92.37 and 91.53 vs. 92.32 on the development and test set, respectively). Overall, the differences between these models are rather small. In the Chinese experiments we observed similar performances. IRSTLM models are again the worst in terms of perplexity, and our model performs slightly worse than SRILM models with interpolation but always better than SRILM models with backoff only. Overall, these results suggest that in-domain our monomial discount model achieves the same performance as the modified KN models implemented in SRILM when using interpolation. Consequently, in all other experiments we use SRILM models with interpolation, which is also recommended for use in MOSES.

5.2 Cross-domain

For the cross-domain experiments we use three languages: English, Chinese, and German. The results are reported in Table 4. Before we discuss the results, let us quickly describe the experiments (see Table 2).

- For the English experiment, we train the language models on the English data of the MultiUN corpus of Eisele and Chen (2010), which contains roughly 8.8 million sentences. For the cross-domain evaluation, we use NIST data, which includes news-wire, broadcast news, and web data, so it is quite different (in style and language) from the contract documents contained in MultiUN.
- For the Chinese experiment, we use the same resources, but now the Chinese data contained in those corpora.
- For the German experiment, which we did in order to cover a morphologically rich language, the language models are trained on EuroParl version 7, News Commentary, and the Common Crawl corpus (overall 4.4 million sentences). We use the *khreshmoi* data for development and test. The data in *khreshmoi* were sampled from summaries of English medical documents.

Comparing the perplexities reported in Tables 3 and 4, we immediately observe that they increase from ≤ 100 to ≥ 200 (sometimes a lot more), which shows that the cross-domain development and test data is rather different from the training data. The results in Table 4 indicate that our model can achieve considerable perplexity improvements for cross-domain data. While our models retain the size of the models generated by SRILM, we often improve the perplexity (English: from 312.30 to 286.53; Chinese: from 639.42 to 584.79; and German: from 501.28 to 469.78). For all experiments, the perplexities computed for the development and the test set are similar because we chose similar validation and test sets. Overall, in all performed experiments, our model outperforms both the modified KN model in SRILM and KenLM (both

Language model	CHINESE → ENGLISH PBMT	CHINESE → ENGLISH HPBMT	ENGLISH → CHINESE PBMT	ENGLISH → CHINESE HPBMT	ENGLISH → GERMAN PBMT	ENGLISH → GERMAN HPBMT
KenLM	**20.59***	**21.12***	17.06	**17.97***	**13.99**	13.95
SRILM	20.05	20.64	17.04	17.68	13.83	13.92
our	20.35*	20.95*	**17.41***	17.94*	13.85	**13.96**

Table 5: BLEU-scores for the various translation experiments. Stars indicate significant improvements over the baseline SRILM (at confidence level 95%).

using interpolation). The models produced by KenLM are generally much larger (> 9 GB) than the models produced by SRILM or our variant (< 3 GB). We report the perplexity results for KenLM in this experiment since KenLM models are very popular in machine translation. Since KenLM does not prune, the KenLM models have larger vocabularies, which can be both beneficial and harmful. This might be an explanation for the poor perplexities that the KenLM models achieve. The tuning of the discount parameters in our model on the cross-domain development set seems to help our model adapt well to the new domain. Together with the results from the in-domain experiment, we can conclude that our model seems to perform as well as SRILM on in-domain data and outperforms SRILM on cross-domain data. The improvements are more pronounced the more distant the development and test data is from the training data.

6 Machine Translation Experiments

Following our LM perplexity experiments, we also want to confirm that the theoretical advantage that our model enjoys in terms of perplexity translates into an application area. Here we select statistical machine translation as an application, so we want to confirm that systems using our model achieve better BLEU-scores (Papineni et al., 2002) in a variety of translation tasks. We compare the different language models on both the phrase-based translation models [PBMT] by Zens et al. (2002) and the hierarchical phrase-based translation models [HPBMT] of Chiang (2005). Both types of translation models are implemented in MOSES toolkit of Koehn et al. (2007). The results of our evaluation are reported in Table 5.

For the Chinese-to-English experiments, we use roughly 2 million sentence pairs from the MultiUN corpus and the tuning and test data consists of the classical NIST data. Overall, the models supported by KenLM achieved the best BLEU scores and significantly beat the SRILM-based baseline, but they do not significantly outperform the models supported by our new language model. Together with our new language model, both the phrase-based and the hierarchical phrase-based models significantly outperform the SRILM-based baselines (from 20.05 to 20.35 for PBMT and from 20.64 to 20.95 for HPBMT). In this experiment, our improvement in terms of perplexity compared to the SRILM-based baseline translates well into an advantage in BLEU-score. This is not true for the perplexity advantage compared to KenLM-based models, which achieve even (insignificantly) better BLEU-scores despite worse perplexity.

For the experiments translating English to Chinese, we use the same training data, but the NIST 2008 test data, which has multiple Chinese references. The results (see Table 5) show a similar picture with one exception. The phrase-based model did not benefit from KenLM and achieves the same performance as the SRILM-based model. Otherwise, KenLM-based models and models based on our new language model achieve similar performance and both significantly outperform the SRILM-based baseline. For our models, the scores consistently improve (from 17.04 to 17.41 for PBMT and from 17.68 to 17.94 for HPBMT). Due to the particularity already mentioned, we even significantly outperform the KenLM-based phrase-based model in this task. So far, our perplexity improvements consistently yielded improvements in translation

quality when measured by BLEU.

Finally, we run experiments translating English to German. In this case, the training data is EuroParl and the tuning and test sets are from the WMT 2014 bio-medical data *khreshmoi*. The results of Table 5 show minute differences, of which none are significant. In this machine translation task, we observe no significant improvements in translation quality (measured by BLEU) even though we observed sizable LM improvements in terms of perplexity. The huge difference between the training set (European Parliament proceedings) and the test set (bio-medical data) might be the reason. The overall translation quality is very poor and potentially too poor to yield reasonable translations, which allows us to speculate that the impact of the language model might be minimal in this setup.

Overall, we demonstrated that our new language model does not harm the translation quality, but rather offers significant improvements in a number of cases. However, the improvements in terms of perplexity do not necessarily translate into BLEU-score improvements. Nevertheless, we often significantly outperformed SRILM-based models in cross-domain evaluations, which shows a nice benefit of our new language model.

7 Summary

In this paper, we introduced a tunable language model which can easily be tuned to new domains and is thus ideally suited for domain adaptation. Perplexity shows that our model outperforms the baseline model especially in domain adaptation scenarios. We implemented our model as a new language model in the MOSES statistical machine translation framework and evaluated it in machine translation task. Also there we observed significant improvements.

In future work we plan to improve the parameter optimization algorithm and implement our model with interpolation. We would also like to investigate translation from German to English and apply our model to other morphologically rich target languages.

Acknowledgments

JUNFEI GUO and ANDREAS MALETTI gratefully acknowledge the financial support by the German Research Foundation (DFG) grant MA / 4959 / 1-1. JUNFEI GUO acknowledges the support by Chinese Scholarship Council (CSC) during his PhD studies at the University of Stuttgart. All authors want to sincerely thank the colleagues at the University of Stuttgart and anonymous reviewers for their helpful comments.

References

Brants, T., Popat, A. C., Xu, P., Och, F. J., and Dean, J. (2007). Large language models in machine translation. In *Proc. EMNLP*, pages 858–867. ACL.

Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Comput. Linguist.*, 18(4):467–479.

Chang, P.-C., Galley, M., and Manning, C. D. (2008). Optimizing Chinese word segmentation for machine translation performance. In *Proc. WMT*, pages 224–232. ACL.

Chen, S. F. and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *Proc. ACL*, pages 310–318. ACL.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proc. ACL*, pages 263–270. ACL.

Eisele, A. and Chen, Y. (2010). MultiUN: a multilingual corpus from United Nation documents. In *Proc. LREC*, pages 2868–2872. ELRA.

Federico, M., Bertoldi, N., and Cettolo, M. (2008). IRSTLM: an open source toolkit for handling large scale language models. In *Proc. INTERSPEECH*, pages 1618–1621. ISCA.

Heafield, K. (2011). KenLM: faster and smaller language model queries. In *Proc. WMT*, pages 187–197. ACL.

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In *Proc. ACL*, pages 690–696. ACL.

Jelinek, F., Mercer, R. L., Bahl, L. R., and Baker, J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *J. Acoust. Soc. Am.*, 62(S1).

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Proc. ICASSP*, volume 1, pages 181–184. IEEE.

Kneser, R. and Steinbiss, V. (1993). On the dynamic adaptation of stochastic language models. In *Proc. ICASSP*, volume 2, pages 586–589. IEEE.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proc. EMNLP*, pages 388–395. ACL.

Koehn, P. (2010a). Europarl: A parallel corpus for statistical machine translation. In *Proc. MT-Summit*, pages 79–86. AAMT.

Koehn, P. (2010b). *Statistical Machine Translation*. Cambridge University Press.

Koehn, P., Axelrod, A., Callison-Burch, C., Osborne, M., and Talbot, D. (2005). Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proc. IWSLT*, pages 68–75. ISCA.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proc. ACL*, pages 177–180. ACL.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Manning, C. D. and Schütze, H. (2001). *Foundations of statistical natural language processing*. MIT Press.

Nguyen, P., Gao, J., and Mahajan, M. (2007). MSRLM: a scalable language modeling toolkit. Technical Report MSR-TR-2007-144, Microsoft Research.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proc. ACL*, pages 160–167. ACL.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318. ACL.

Schütze, H. (2011). Integrating history-length interpolation and classes in language modeling. In *Proc. ACL*, pages 1516–1525. ACL.

Stolcke, A. (2002). SRILM — an extensible language modeling toolkit. In *Proc. INTERSPEECH*, pages 901–904. ISCA.

Zens, R., Och, F. J., and Ney, H. (2002). Phrase-based statistical machine translation. In *Proc. KI*, volume 2479 of *LNCS*, pages 18–32. Springer.