Compositional Morphology for Word Representations and Language Modelling

Jan A. Botha JAN.BOTHA@CS.OX.AC.UK
Phil Blunsom PHIL.BLUNSOM@CS.OX.AC.UK
Department of Computer Science, University of Oxford, Oxford, OX1 3QD, UK

Abstract

This paper presents a scalable method for integrating compositional morphological representations into a vector-based probabilistic language model. Our approach is evaluated in the context of log-bilinear language models, rendered suitably efficient for implementation inside a machine translation decoder by factoring the vocabulary. We perform both intrinsic and extrinsic evaluations, presenting results on a range of languages which demonstrate that our model learns morphological representations that both perform well on word similarity tasks and lead to substantial reductions in perplexity. When used for translation into morphologically rich languages with large vocabularies, our models obtain improvements of up to 1.2 BLEU points relative to a baseline system using back-off n-gram models.

1 Introduction

The proliferation of word forms in morphologically rich languages presents challenges to the statistical language models (LMs) that play a key role in machine translation and speech recognition. Conventional back-off n-gram LMs (Chen & Goodman, 1998) and the increasingly popular vector-based LMs (Bengio et al., 2003; Schwenk et al., 2006; Mikolov et al., 2010) use parametrisations that do not explicitly encode morphological regularities among related forms, like *abstract*, *abstraction* and *abstracted*. Such models suffer from data sparsity arising from morphological processes and lack a coherent method of assigning probabilities or representations to unseen word forms.

This work focuses on continuous space language models (CSLMs), an umbrella term for the LMs that represent words with real-valued vectors. Such word representations have been found to capture some morphological regularity (Mikolov et al., 2013b), but we contend that there is a case for building *a priori* morphological awareness into the language models' inductive bias. Conversely, compositional vector-space modelling has recently been applied to morphology to good effect (Lazaridou et al., 2013; Luong et al., 2013), but lacked the probabilistic basis necessary for use with a machine translation decoder.

The method we propose strikes a balance between probabilistic language modelling and morphology-based representation learning. Word vectors are composed as a linear function of arbitrary sub-elements of the word, e.g. surface form, stem, affixes, or other latent information. The effect is to tie together the representations of morphologically related words, directly combating data sparsity. This is executed in the context of a log-bilinear (LBL) LM (Mnih & Hinton, 2007), which is sped up sufficiently by the use of word classing so that we can integrate the model into an open source machine translation decoder[1] and evaluate its impact on translation into 6 languages, including the morphologically complex Czech, German and Russian.

In word similarity rating tasks, our morpheme vectors help improve correlation with human ratings in multiple languages. Fine-grained analysis is used to determine the origin of our perplexity reductions, while scaling experiments demonstrate tractability on vocabularies of 900k types using 100m+ tokens.

2 Additive Word Representations

A generic CSLM associates with each word type v in the vocabulary $\mathcal{V}$ a d-dimensional feature vector $\mathbf{r}_v \in \mathbb{R}^d$. Regularities among words are captured in an opaque way through the interaction of these feature values and a set of transformation weights. This leverages linguistic intuitions only in an extremely rudimentary way, in contrast to hand-engineered linguistic features that target very specific phenomena, as often used in supervised-learning settings.

We seek a compromise that retains the unsupervised nature of CSLM feature vectors, but also incorporates *a priori* linguistic knowledge in a flexible and efficient manner. In particular, *morphologically related words should share statistical strength* in spite of differences in surface form.

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

[1] Our source code for language model training and integration into *cdec* is available from http://bothameister.github.io

To achieve this, we define a mapping $\mu : \mathcal{V} \mapsto \mathcal{F}^+$ of a surface word into a variable-length sequence of *factors*, i.e. $\mu(v) = (f_1, \ldots, f_K)$, where $v \in \mathcal{V}$ and $f_i \in \mathcal{F}$. Each factor f has an associated *factor feature vector* $\mathbf{r}_f \in \mathbb{R}^d$. We thereby factorise a word into its surface morphemes, although the approach could also incorporate other information, e.g. lemma, part of speech.

The vector representation $\tilde{\mathbf{r}}_v$ of a word v is computed as a function $\omega_\mu(v)$ of its factor vectors. We use addition as composition function: $\tilde{\mathbf{r}}_v = \omega_\mu(v) = \sum_{f \in \mu(v)} \mathbf{r}_f$. The vectors of morphologically related words become linked through shared factor vectors (notation: $\overrightarrow{\text{word}}$, $\overrightarrow{factor}$),

$$\begin{aligned} \overrightarrow{\text{imperfection}} &= \overrightarrow{im} + \overrightarrow{perfect} + \overrightarrow{ion} \\ \overrightarrow{\text{perfectly}} &= \overrightarrow{perfect} + \overrightarrow{ly}. \end{aligned}$$

Furthermore, representations for out-of-vocabulary (OOV) words can be constructed using their available morpheme vectors.

We include the surface form of a word as a factor itself. This accounts for noncompositional constructions ($\overrightarrow{\text{greenhouse}} = \overrightarrow{greenhouse} + \overrightarrow{green} + \overrightarrow{house}$), and makes the approach more robust to noisy morphological segmentation. This strategy also overcomes the order-invariance of additive composition ($\overrightarrow{\text{hangover}} \neq \overrightarrow{\text{overhang}}$).

The number of factors per word is free to vary over the vocabulary, making the approach applicable across the spectrum of more fusional languages (e.g. Czech, Russian) to more agglutinative languages (e.g. Turkish). This is in contrast to *factored language models* (Alexandrescu & Kirchhoff, 2006), which assume a fixed number of factors per word. Their method of concatenating factor vectors to obtain a single representation vector for a word can be seen as enforcing a partition on the feature space. Our method of addition avoids such a partitioning and better reflects the absence of a strong intuition about what an appropriate partitioning might be. A limitation of our method compared to theirs is that the deterministic mapping μ currently enforces a single factorisation per word type, which sacrifices information obtainable from context-disambiguated morphological analyses.

Our additive composition function can be regarded as an instantiation of the weighted addition strategy that performed well in a distributional compositional approach to derivational morphology (Lazaridou et al., 2013). Unlike the recursive neural-network method of Luong et al. (2013), we do not impose a single tree structure over a word, which would ignore the ambiguity inherent in words like un[[lock]able] vs. [un[lock]]able. In contrast to these two previous approaches to morphological modelling, our additive representations are readily implementable in a probabilistic language model suitable for use in a decoder.

3 Log-Bilinear Language Models

Log-bilinear (LBL) models (Mnih & Hinton, 2007) are an instance of CSLMs that make the same Markov assumption as n-gram language models. The probability of a sentence $\mathbf{w}$ is decomposed over its words, each conditioned on the n–1 preceding words: $P(\mathbf{w}) \approx \prod_i P\left(w_i | w_{i-n+1}^{i-1}\right)$. These distributions are modelled by a smooth scoring function $\nu(\cdot)$ over vector representations of words. In contrast, discrete n-gram models are estimated by smoothing and backing off over empirical distributions (Kneser & Ney, 1995; Chen & Goodman, 1998).

The LBL predicts the vector $\mathbf{p}$ for the next word as a function of the context vectors $\mathbf{q}_j \in \mathbb{R}^d$ of the preceding words,

$$\mathbf{p} = \sum_{j=1}^{n-1} \mathbf{q}_j C_j, \tag{1}$$

where $C_j \in \mathbb{R}^{d \times d}$ are position-specific transformations.

$\nu(w)$ expresses how well the observed word w fits that prediction and is defined as $\nu(w) = \mathbf{p} \cdot \mathbf{r}_w + b_w$, where b_w is a bias term encoding the prior probability of a word type. Softmax then yields the word probability as

$$P\left(w_i | w_{i-n+1}^{i-1}\right) = \frac{\exp\left(\nu(w_i)\right)}{\sum_{v \in \mathcal{V}} \exp\left(\nu(v)\right)}. \tag{2}$$

This model is subsequently denoted as **LBL** with parameters $\Theta_{\text{LBL}} = (C_j, Q, R, \mathbf{b})$, where $Q, R \in \mathbb{R}^{|\mathcal{V}| \times d}$ contain the word representation vectors as rows, and $\mathbf{b} \in \mathbb{R}^{|\mathcal{V}|}$. Q and R imply that separate representations are used for conditioning and output.

3.1 Additive Log-Bilinear Model

We introduce a variant of the LBL that makes use of additive representations (§2) by associating the *composed word vectors* $\tilde{\mathbf{r}}$ and $\tilde{\mathbf{q}}_j$ with the target and context words, respectively. The representation matrices $Q^{(f)}, R^{(f)} \in \mathbb{R}^{|\mathcal{F}| \times d}$ thus contain a vector for each factor type. This model is designated **LBL++** and has parameters $\Theta_{\text{LBL++}} = \left(C_j, Q^{(f)}, R^{(f)}, \mathbf{b}\right)$. Words sharing factors are tied together, which is expected to improve performance on rare word forms.

Representing the mapping μ with a sparse transformation matrix $M \in \mathbb{Z}_+^{\mathcal{V} \times |\mathcal{F}|}$, where a row vector $\mathbf{m}_v$ has some non-zero elements to select factor vectors, establishes the relation between word and factor representation matrices as $R = MR^{(f)}$ and $Q = MQ^{(f)}$. In practice, we exploit this for test-time efficiency—word vectors are compiled offline so that the computational cost of LBL++ probability lookups is the same as for the LBL.

We consider two obvious variations of the LBL++ to evaluate the extent to which interactions between context and

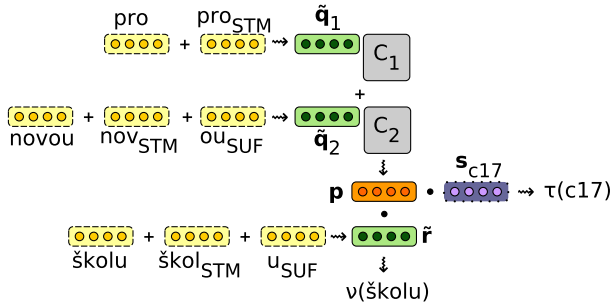

Figure 1. **Model diagram.** Illustration of how a 3-gram CLBL++ model treats the Czech phrase *pro novou školu* ('for the new school'), assuming the target word *školu* is clustered into word class 17 by the method described in §3.2.

target factors affect the model: **LBL+o** only factorises **o**utput words and retains simple word vectors for the context (i.e. $Q \equiv Q^{(f)}$), while **LBL+c** does the reverse, only factorising **c**ontext words.[2] Both reduce to the LBL when setting μ to be the identity function, such that $\mathcal{V} \equiv \mathcal{F}$.

The factorisation permits an approach to unknown context words that is less harsh than the standard method of replacing them with a global unknown symbol—instead, a vector can be constructed from the known factors of the word (e.g. the observed stem of an unobserved inflected form). A similar scheme can be used for scoring unknown target words, but requires changing the event space of the probabilistic model. We use this vocabulary stretching capability in our word similarity experiments, but leave the extensions for test-time language model predictions as future work.

3.2 Class-based Model Decomposition

The key obstacle to using CSLMs in a decoder is the expensive normalisation over the vocabulary. Our approach to reducing the computational cost of normalisation is to use a class-based decomposition of the probabilistic model (Goodman, 2001; Mikolov et al., 2011). Using Brown-clustering (Brown et al., 1992),[3] we partition the vocabulary into $|\mathcal{C}|$ classes, denoting as $\mathcal{C}_c$ the set of vocabulary items in class c, such that $\mathcal{V} = \mathcal{C}_1 \cup \cdots \cup \mathcal{C}_{|\mathcal{C}|}$.

In this model, the probability of a word conditioned on the history h of $n-1$ preceding words is decomposed as

$$P(w|h) = P(c|h)P(w|h,c). \tag{3}$$

This class-based model, **CLBL**, extends over the LBL by associating a representation vector $\mathbf{s}_c$ and bias parameter t_c to each class c, such that $\Theta_{\text{CLBL}} = (C_j, Q, R, S, \mathbf{b}, \mathbf{t})$. The same prediction vector $\mathbf{p}$ is used to compute both class score $\tau(c) = \mathbf{p} \cdot \mathbf{s}_c + t_c$ and word score $\nu(w)$, which are normalised separately:

$$P(c|h) = \frac{\exp\left(\tau(c)\right)}{\sum_{c'=1}^{|\mathcal{C}|} \exp\left(\tau(c')\right)} \tag{4}$$

$$P(w|h,c) = \frac{\exp\left(\nu(w)\right)}{\sum_{v' \in \mathcal{C}_c} \exp\left(\nu(v')\right)}. \tag{5}$$

We favour this flat vocabulary partitioning for its computational adequacy, simplicity and robustness. Computational adequacy is obtained by using $|\mathcal{C}| \approx |\mathcal{V}|^{0.5}$, thereby reducing the $\mathcal{O}(|\mathcal{V}|)$ normalisation operation of the LBL to two $\mathcal{O}(|\mathcal{V}|^{0.5})$ operations in the CLBL.

Other methods for achieving more drastic complexity reductions exist in the form of frequency-based truncation, shortlists (Schwenk, 2004), or casting the vocabulary as a full hierarchy (Mnih & Hinton, 2008) or partial hierarchy (Le et al., 2011). We expect these approaches could have adverse effects in the rich morphology setting, where much of the vocabulary is in the long tail of the word distribution.

3.3 Training & Initialisation

Model parameters Θ are estimated by optimising an L2-regularised log likelihood objective. Training the CLBL and its additive variants directly against this objective is fast because normalisation of model scores, which is required in computing gradients, is over a small number of events.

For the classless LBLs we use noise-contrastive estimation (NCE) (Gutmann & Hyvärinen, 2012; Mnih & Teh, 2012) to avoid normalisation during training. This leaves the expensive test-time normalisation of LBLs unchanged, precluding their usage during decoding.

Bias terms $\mathbf{b}$ (resp. $\mathbf{t}$) are initialised to the log unigram probabilities of words (resp. classes) in the training corpus, with Laplace smoothing, while all other parameters are initialised randomly according to sharp, zero-mean Gaussians. Representations are thus learnt from scratch and not based on publicly available embeddings, meaning our approach can easily be applied to many languages.

Optimisation is performed by stochastic gradient descent with updates after each mini-batch of L training examples. We apply AdaGrad (Duchi et al., 2011) and tune the step-size ξ on development data.[4] We halt training once the perplexity on the development data starts to increase.

4 Experiments

The overarching aim of our evaluation is to investigate the effect of using the proposed additive representations across languages with a range of morphological complexity.

[2] The +c, +o and ++ naming suffixes denote these same distinctions when used with the CLBL model introduced later.

[3] In preliminary experiments, Brown clusters gave better perplexities than frequency-binning (Mikolov et al., 2011).

[4] L=10k–40k, ξ=0.05–0.08, dependent on $|\mathcal{V}|$ and data size.

Table 1. **Corpus statistics.** The number of sentence pairs for a row X refers to the English→X parallel data (but row EN has Czech as source language).

	DATA-1M		DATA-MAIN		
	Toks.	$\|\mathcal{V}\|$	Toks.	$\|\mathcal{V}\|$	Sent. Pairs
CS	1m	46k	16.8m	206k	0.7m
DE	1m	36k	50.9m	339k	1.9m
EN	1m	17k	19.5m	60k	0.7m
ES	1m	27k	56.2m	152k	2.0m
FR	1m	25k	57.4m	137k	2.0m
RU	1m	62k	25.1m	497k	1.5m

Our intrinsic language model evaluation has two parts. We first perform a model selection experiment on small data to consider the relative merits of using additive representations for context words, target words, or both, and to validate the use of the class-based decomposition.

Then we consider class-based additive models trained on tens of millions of tokens and large vocabularies. These larger language models are applied in two extrinsic tasks: i) a word-similarity rating experiment on multiple languages, aiming to gauge the quality of the induced word and morpheme representation vectors; ii) a machine translation experiment, where we are specifically interested in testing the impact of an LBL LM feature when translating into morphologically rich languages.

4.1 Data & Methods

We make use of data from the 2013 ACL Workshop on Machine Translation.[5] We first describe data used for translation experiments, since the monolingual datasets used for language model training were derived from that. The language pairs are English→{German, French, Spanish, Russian} and English↔Czech. Our parallel data comprised the Europarl-v7 and news-commentary corpora, except for English–Russian where we used news-commentary and the Yandex parallel corpus.[6] Pre-processing involved lowercasing, tokenising and filtering to exclude sentences of more than 80 tokens or substantially different lengths.

4-gram language models were trained on the target data in two batches: DATA-1M consists of the first million tokens only, while DATA-MAIN is the full target-side data. Statistics are given in Table 1. newstest2011 was used as development data[7] for tuning language model hyperparameters, while intrinsic LM evaluation was done on newstest2012. As metric, we use model perplexity (PPL) $\exp(-\frac{1}{N}\sum_{i=1}^{N}\ln P(w_i))$, where N is the number of test tokens. In addition to contrasting the LBL variants, we also use modified Kneser-Ney n-gram models (MKNs) (Chen & Goodman, 1998) as baselines.

[5] http://www.statmt.org/wmt13/translation-task.html
[6] https://translate.yandex.ru/corpus?lang=en
[7] For Russian, some training data was held out for tuning.

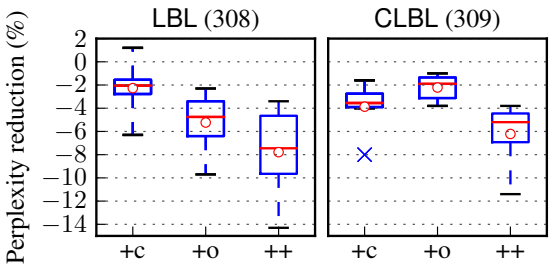

Figure 2. **Model selection results.** Box-plots show the spread, across 6 languages, of relative perplexity reductions obtained by each type of additive model against its non-additive baseline, for which median absolute perplexity is given in parentheses; for MKN, that is 348. Each box-plot summarises the behaviour of a model across languages. Circles give sample means, while crosses show outliers beyond 3× the inter-quartile range.

Language Model Vocabularies. Additive representations that link morphologically related words specifically aim to improve modelling of the long tail of the lexicon, so we do not want to prune away all rare words, as is common practice in language modelling and word embedding learning. We define a *singleton pruning rate* κ, and randomly replace that fraction of words occurring only once in the training data with a global UNK symbol. $\kappa = 1$ would imply a unigram count cut-off threshold of 1. Instead, we use low pruning rates[8] and thus model large vocabularies.[9]

Word Factorisation μ. We obtain labelled morphological segmentations from the unsupervised segmentor *Morfessor Cat-MAP* (Creutz & Lagus, 2007). The mapping μ of a word is taken as its surface form and the morphemes identified by Morfessor. Keeping the morpheme labels allows the model to learn separate vectors for, say, in^{stem} the preposition and in^{prefix} occurring as *in·appropriate*. By not post-processing segmentations in a more sophisticated way, we keep the overall method more language independent.

4.2 Intrinsic Language Model Evaluation

Results on DATA-1M. The use of morphology-based, additive representations for both context and output words (models++) yielded perplexity reductions on all 6 languages when using 1m training tokens. Furthermore, these double-additive models consistently outperform the ones that factorise only context (+c) or only output (+o) words, indicating that context and output contribute complementary information and supporting our hypothesis that is it beneficial to model morphological dependencies across words. The results are summarised in Figure 2.

For lack of space we do not present numbers for individual languages, but report that the impact of CLBL++ varies by

[8] DATA-1M: $\kappa = 0.2$; DATA-MAIN: $\kappa = 0.05$
[9] We also mapped digits to 0, and cleaned the Russian data by replacing tokens having <80% Cyrillic characters with UNK.

Table 2. **Test-set perplexities on DATA-MAIN** using two vocabulary pruning settings. Percentage reductions are relative to the preceding model, e.g. the first Czech CLBL improves over MKN by 20.8% (Rel.1); the CLBL++ improves over that CLBL by a *further* 5.9% (Rel.2).

		MKN	CLBL		CLBL++	
		PPL	PPL	Rel.1	PPL	Rel.2
κ=0.05	CS	862	683	-20.8%	643	-5.9%
	DE	463	422	-8.9%	404	-4.2%
	EN	291	281	-3.4%	273	-2.8%
	ES	219	207	-5.7%	203	-1.9%
	FR	243	232	-4.9%	227	-1.9%
	RU	390	313	-19.7%	300	-4.2%
κ=1.0	CS	634	477	-24.8%	462	-3.1%
	DE	379	331	-12.6%	329	-0.9%
	EN	254	234	-7.6%	233	-0.7%
	ES	195	180	-7.7%	180	0.02%
	FR	218	201	-7.7%	198	-1.3%
	RU	347	271	-21.8%	262	-3.4%

language, correlating with vocabulary size: Russian benefited most, followed by Czech and German. Even on English, often regarded as having simple morphology, the relative improvement is 4%.

The relative merits of the +c and +o schemes depend on which model is used as starting point. With LBL, the output-additive scheme (LBL+o) gives larger improvements than the context-additive scheme (LBL+c). The reverse is true for CLBL, indicating the class decomposition dampens the effectiveness of using morphological information in output words.

The use of classes increases perplexity slightly compared to the LBLs, but this is in exchange for much faster computation of language model probabilities, allowing the CLBLs to be used in a machine translation decoder (§4.4).

Results on DATA-MAIN. Based on the outcomes of the small-scale evaluation, we focus our main language model evaluation on the additive class-based model CLBL++ in comparison to CLBL and MKN baselines, using the larger training dataset, with vocabularies of up to 500k types.

The overall trend that morphology-based additive representations yield lower perplexity carries over to this larger data setting, again with the biggest impact being on Czech and Russian (Table 2, top). Improvements are in the 2%–6% range, slightly lower than the corresponding differences on the small data.

Our hypothesis is that the much of the improvement is due to the additive representations being especially beneficial for modelling rare words. We test this by repeating the experiment under the condition where all word types occurring only once are excluded from the vocabulary (κ=1). If the additive representations were not beneficial to rare words, the outcome should remain the same. Instead, we find the relative improvements become a lot smaller (Table 2, bottom) than when only excluding some singletons (κ=0.05), which supports that hypothesis.

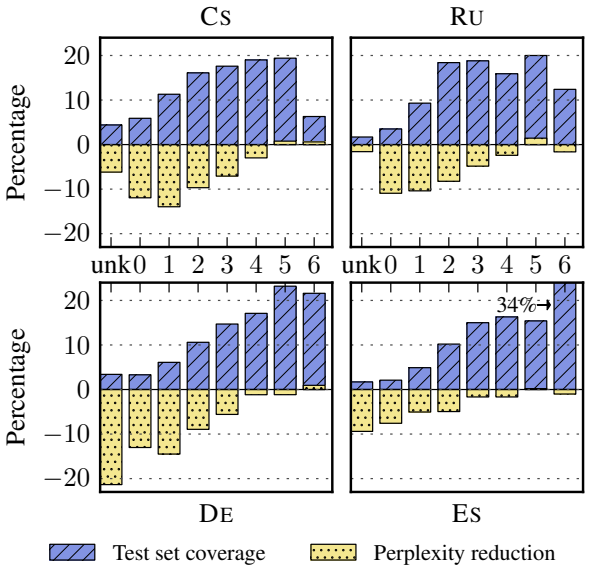

Figure 3. **Perplexity reductions by token frequency,** CLBL++ relative to CLBL. Dotted bars extending further down are better. A bin labelled with a number x contains those test tokens that occur $y \in [10^x, 10^{x+1})$ times in the training data. Striped bars show percentage of test-set covered by each bin.

Analysis. Model perplexity on a whole dataset is a convenient summary of its intrinsic performance, but such a global view does not give much insight into *how* one model outperforms another. We now partition the test data into subsets of interest and measure PPL over these subsets.

We first partition on token frequency, as computed on the training data. Figure 3 provides further evidence that the additive models have most impact on rare words generally, and not only on singletons. Czech, German and Russian see relative PPL reductions of 8%–21% for words occurring fewer than 100 times in the training data. Reductions become negligible for the high-frequency tokens. These tend to be punctuation and closed-class words, where any putative relevance of morphology is overwhelmed by the fact that the predictive uncertainty is very low to begin with (absolute PPL<10 for the highest frequency subset). For the morphologically simpler Spanish case, PPL reductions are generally smaller across frequency scales.

We also break down PPL reductions by part of speech tags, focusing on German. We used the decision tree-based tagger of Schmid & Laws (2008). Aside from unseen tokens, the biggest improvements are on nouns and adjectives (Figure 4), suggesting our segmentation-based representations help abstract over German's productive compounding.

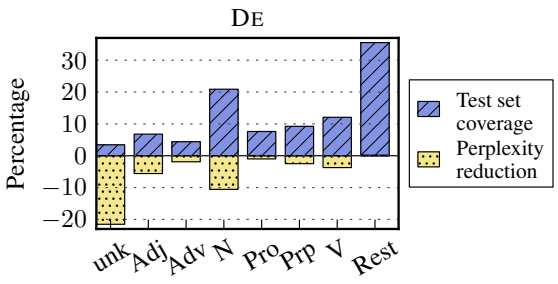

Figure 4. **Perplexity reductions by part of speech,** CLBL++ relative to CLBL on German. Dotted bars extending further down are better. Tokens tagged as foreign words or other opaque symbols resort under "Rest". Striped bars as in Figure 3

German noun phrases require agreement in gender, case and number, which are marked overtly with fusional morphemes, and we see large gains on such test n-grams: 15% improvement on adjective-noun sequences, and 21% when considering the more specific case of adjective-adjective-noun sequences. An example of the latter kind is *der ehemalig·e sozial·ist·isch·e bildung·s·minister* ('the former socialist minister of education'), where the morphological agreement surfaces in the repeated e-suffix.

We conducted a final scaling experiment on Czech by training models on increasing amounts of data from the monolingual news corpora. Improvements over the MKN baseline decrease, but remain substantial at 14% for the largest setting when allowing the vocabulary to grow with the data. Maintaining a constant advantage over MKN requires also increasing the dimensionality d of representations (Mikolov et al., 2013a), but this was outside the scope of our experiment. Although gains from the additive representations over the CLBL diminish down to 2%–3% at the scale of 128m training tokens (Figure 5), these results demonstrate the tractability of our approach on very large vocabularies of nearly 1m types.

4.3 Task 1: Word Similarity Rating

In the previous section, we established the positive role that morphological awareness played in building continuous-space language models that better predict unseen text. Here we focus on the quality of the word representations learnt in the process. We evaluate on a standard word similarity rating task, where one measures the correlation between cosine-similarity scores for pairs of word vectors and a set of human similarity ratings. An important aspect of our evaluation is to measure performance on multiple languages using a single unsupervised, model-based approach.

Morpheme vectors from the CLBL++ enable handling OOV test words in a more nuanced way than using the global unknown word vector. In general, we compose a vector $\tilde{\mathbf{u}}_v = [\tilde{\mathbf{q}}_v; \tilde{\mathbf{r}}_v]$ for a word v according to a *post hoc* word

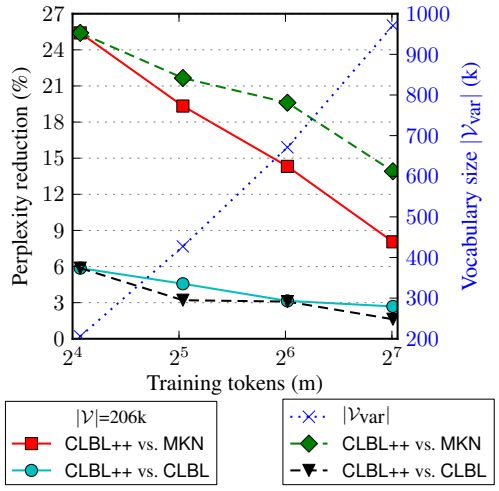

Figure 5. **Scaling experiment.** Relative perplexity reductions obtained when varying the Czech training data size (16m–128m). In the first setting, the vocabulary was held fixed as data size increased($|\mathcal{V}|$); in the second it varied freely across sizes ($|\mathcal{V}_{\mathbf{var}}|$).

map μ' by summing and concatenating the factor vectors $\mathbf{r}_f$ and $\mathbf{q}_f$, where $f \in \mu'(v) \cap \mathcal{F}$. This ignores unknown morphemes occurring in OOV words, and uses $[\mathbf{q}_{\text{UNK}}; \mathbf{r}_{\text{UNK}}]$ for $\tilde{\mathbf{u}}_{\text{UNK}}$ only if all morphemes are unknown.

To see whether the morphological representations improve the quality of vectors for known words, we also report the correlations obtained when using the CLBL++ word vectors directly, resorting to $\tilde{\mathbf{u}}_{\text{UNK}}$ for all OOV words $v \notin \mathcal{V}$ (denoted "$-$*compose*" in the results). This is also the strategy that the baseline CLBL model is forced to follow for OOVs.

We evaluate first using the English rare-word dataset (RW) created by Luong et al. (2013). Its 2034 word pairs contain more morphological complexity than other well-established word similarity datasets, e.g. crudeness—impoliteness. We compare against their context-sensitive morphological recursive neural network (csmRNN), using Spearman's rank correlation coefficient, ρ. Table 3 shows our model obtaining a ρ-value slightly below the best csmRNN result, but outperforming the csmRNN that used an alternative set of embeddings for initialisation.

This is a strong result given that our vectors come from a simple linear probabilistic model that is also suitable for integration directly into a decoder for translation (§4.4) or speech recognition, which is not the case for csmRNNs. Moreover, the csmRNNs were initialised with high-quality, publicly available word embeddings trained over weeks on much larger corpora of 630–990m words (Collobert & Weston, 2008; Huang et al., 2012), in contrast to ours that are trained from scratch on much less data. This renders our method directly applicable to languages which may not yet have those resources.

Relative to the CLBL baseline, our method performs well on

Table 3. **Word-pair similarity task.** Spearman's $\rho \times 100$ for the correlation between model scores and human ratings on the English RW dataset. The csmRNNs benefit from initialisation with high quality pre-existing word embeddings, while our models used random initialisation.

(Luong et al., 2013)		*Our models*	
HSMN	2	CLBL	18
HSMN+csmRNN	22	CLBL++	**30**
C&W	27	*−compose*	20
C&W+csmRNN	**34**		

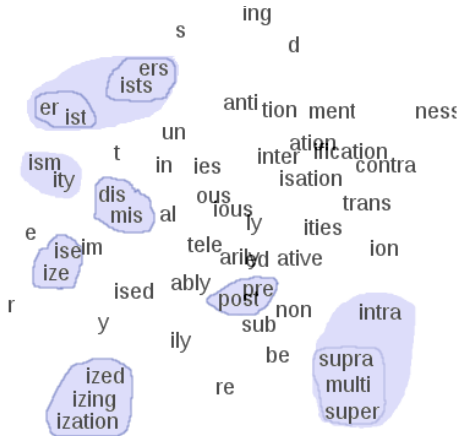

Figure 6. **English morpheme vectors learnt by CLBL++.** Dimensionality reduction was performed with t-SNE (van der Maaten & Hinton, 2008), with shading added for emphasis.

datasets across four languages. For the English RW, which was designed with morphology in mind, the gain is 64%. But also on the standard English WS353 dataset (Finkelstein et al., 2002), we get a 26% better correlation with the human ratings. On German, the CLBL++ obtains correlations up to three times stronger than the baseline, and 39% better for French (Table 4).

A visualisation of the English morpheme vectors (Figure 6) suggests the model captured non-trivial morphological regularities: noun suffixes relating to persons (writ*er*, human*ists*) lie close together, while being separated according to number; negation prefixes share a region (un-, in-, mis-, dis-); and relational prefixes are grouped (surpa-, super-, multi-, intra-), with a potential explanation for their separation from inter- being that the latter is more strongly bound up in lexicalisations (*inter*national, *inter*section).

4.4 Task 2: Machine Translation

The final aspect of our evaluation focuses on the integration of class-decomposed log-bilinear models into a machine translation system. To the best of our knowledge, this is the first study to investigate large vocabulary normalised CSLMs inside a decoder when translating into a range of morphologically rich languages. We consider 5 language pairs, translating from English into Czech, German, Russian, Spanish and French.

Table 4. **Word-pair similarity task (multi-language),** showing Spearman's $\rho \times 100$ and the number of word pairs in each dataset. As benchmarks, we include the best results from Luong et al. (2013), who relied on more training data and pre-existing embeddings not available in all languages. In the penultimate row our model's ability to compose vectors for OOV words is suppressed.

Datasets[10]	WS		Gur	RG		ZG
Model / *Language*	EN	ES	DE	EN	FR	DE
HSMN	63	–	–	63	–	–
+csmRNN	65	–	–	65	–	–
CLBL	32	26	36	**47**	33	6
CLBL++	39	**28**	**56**	41	**45**	**25**
−compose	**40**	27	44	41	41	23
# pairs	353		350	65		222

Aside from the choice of language pairs, this evaluation diverges from Vaswani et al. (2013) by using normalised probabilities, a process made tractable by the class-based decomposition and caching of context-specific normaliser terms. Vaswani et al. (2013) relied on unnormalised model scores for efficiency, but do not report on the performance impact of this assumption. In our preliminary experiments, there was high variance in the performance of unnormalised models. They are difficult to reason about as a feature function that must help the translation model discriminate between alternative hypotheses.

We use *cdec* (Dyer et al., 2010; 2013) to build symmetric word-alignments and extract rules for hierarchical phrase-based translation (Chiang, 2007). Our baseline system uses a standard set of features in a log-linear translation model. This includes a baseline 4-gram MKN language model, trained with *SRILM* (Stolcke, 2002) and queried efficiently using *KenLM* (Heafield, 2011). The CSLMs are integrated directly into the decoder as an *additional* feature function, thus exercising a stronger influence on the search than in n-best list rescoring.[11] Translation model feature weights are tuned with MERT (Och, 2003) on newstest2012.

Table 5 summarises our translation results. Inclusion of the CLBL++ language model feature outperforms the MKN-only baseline systems by 1.2 BLEU points for translation into Russian, and by 1 point into Czech and Spanish. The EN→DE system benefits least from the additional CSLM feature, despite the perplexity reductions achieved in the intrinsic evaluation. In light of German's productive compounding, it is conceivable that the bilingual coverage of

[10] ES WS353 (Hassan & Mihalcea, 2009); Gur350 (Gurevych, 2005); RG65 (Rubenstein & Goodenough, 1965) with FR (Joubarne & Inkpen, 2011); ZG222 (Zesch & Gurevych, 2006).

[11] Our source code for using CLBL/CLBL++ with *cdec* is released at http://bothameister.github.io.

Table 5. **Translation results.** Case-insensitive BLEU scores on newstest2013, with standard deviation over 3 runs given in parentheses. The two right-most columns use the listed CSLM as a feature in addition to the MKN feature, i.e. these MT systems have at most 2 LMs. Language models are from Table 2 (top).

	MKN	CLBL	CLBL++
EN→CS	12.6 (0.2)	13.2 (0.1)	**13.6** (0.0)
DE	15.7 (0.1)	**15.9** (0.2)	15.8 (0.4)
ES	24.7 (0.4)	25.5 (0.5)	**25.7** (0.3)
FR	24.1 (0.2)	24.6 (0.2)	**24.8** (0.5)
RU	15.9 (0.2)	16.9 (0.3)	**17.1** (0.1)
CS→EN	19.8 (0.4)	**20.4** (0.4)	**20.4** (0.5)

that system is more of a limitation than the performance of the language models.

On the other languages, the CLBL adds 0.5 to 1 BLEU points over the baseline, whereas additional improvement from the additive representations lies within MERT variance except for EN→CS.

The impact of our morphology-aware language model is limited by the translation system's inability to generate unseen inflections. A future task is thus to combine it with a system that can do so (Chahuneau et al., 2013).

5 Related Work

Factored language models (FLMs) have been used to integrate morphological information into both discrete n-gram LMs (Bilmes & Kirchhoff, 2003) and CSLMs (Alexandrescu & Kirchhoff, 2006) by viewing a word as a set of factors. Alexandrescu & Kirchhoff (2006) demonstrated how factorising the representations of context-words can help deal with out-of-vocabulary words, but they did not evaluate the effect of factorising output words and did not conduct an extrinsic evaluation.

A variety of strategies have been explored for bringing CSLMs to bear on machine translation. Rescoring lattices with a CSLM proved to be beneficial for ASR (Schwenk, 2004) and was subsequently applied to translation (Schwenk et al., 2006; Schwenk & Koehn, 2008), reaching training sizes of up to 500m words (Schwenk et al., 2012). For efficiency, this line of work relied heavily on small "shortlists" of common words, by-passing the CSLM and using a back-off n-gram model for the remainder of the vocabulary. Using unnormalised CSLMs during first-pass decoding has generated improvements in BLEU score for translation into English (Vaswani et al., 2013).

Recent work has moved beyond monolingual vector-space modelling, incorporating phrase similarity ratings based on bilingual word embeddings as a translation model feature (Zou et al., 2013), or formulating translation purely in terms of continuous-space models (Kalchbrenner & Blunsom, 2013). Accounting for linguistically derived information such as morphology (Luong et al., 2013; Lazaridou et al., 2013) or syntax (Hermann & Blunsom, 2013) has recently proved beneficial to learning vector representations of words. Our contribution is to create morphological awareness in a *probabilistic* language model.

6 Conclusion

We introduced a method for integrating morphology into probabilistic continuous-space language models. Our method has the flexibility to be used for morphologically rich languages (MRLs) across a range of linguistic typologies. Our empirical evaluation focused on multiple MRLs and different tasks. The primary outcomes are that (i) our morphology-guided CSLMs improve intrinsic language model performance when compared to baseline CSLMs and n-gram MKN models; (ii) word and morpheme representations learnt in the process compare favourably in terms of a word similarity task to a recent more complex model that used more data, while obtaining large gains on some languages; (iii) machine translation quality as measured by BLEU was improved consistently across six language pairs when using CSLMs during decoding, although the morphology-based representations led to further improvements beyond the level of optimiser variance only for English→Czech. By demonstrating that the class decomposition enables full integration of a normalised CSLM into a decoder, we open up many other possibilities in this active modelling space.

References

Alexandrescu, A. & Kirchhoff, K. Factored Neural Language Models. In *Proc. HLT-NAACL: short papers*. ACL, 2006.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. A Neural Probabilistic Language Model. *JMLR*, 3:1137–1155, 2003.

Bilmes, J. A. & Kirchhoff, K. Factored Language Models and Generalized Parallel Backoff. In *Proc. NAACL-HLT: short papers*. ACL, 2003.

Brown, P. F., DeSouza, P. V., Mercer, R. L., Della Pietra, V. J., & Lai, J. C. Class-Based n-gram Models of Natural Language. *Comp. Ling.*, 18(4):467–479, 1992.

Chahuneau, V., Schlinger, E., Smith, N. A., & Dyer, C. Translating into Morphologically Rich Languages with Synthetic Phrases. In *Proc. EMNLP*, pp. 1677–1687. ACL, 2013.

Chen, S. F. & Goodman, J. An Empirical Study of Smoothing Techniques for Language Modeling. Technical report, Harvard University, Cambridge, MA, 1998.

Chiang, D. Hierarchical Phrase-Based Translation. *Comp. Ling.*, 33(2):201–228, 2007.

Collobert, R. & Weston, J. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proc. ICML*. ACM, 2008.

Creutz, M. & Lagus, K. Unsupervised Models for Morpheme Segmentation and Morphology Learning. *ACM Trans. on Speech and Language Processing*, 4(1):1–34, 2007.

Duchi, J., Hazan, E., & Singer, Y. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *JMLR*, 12: 2121–2159, 2011.

Dyer, C., Lopez, A., Ganitkevitch, J., Weese, J., Ture, F., Blunsom, P., Setiawan, H., Eidelman, V., & Resnik, P. cdec: A Decoder, Alignment, and Learning Framework for Finite-State and Context-Free Translation Models. In *Proc. ACL: demonstration session*, pp. 7–12, 2010. ACL.

Dyer, C., Chahuneau, V., & Smith, N. A. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proc. NAACL*, pp. 644–648. ACL, 2013.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. Placing Search in Context: The Concept Revisited. *ACM Trans. on Information Systems*, 20 (1):116–131, 2002.

Goodman, J. Classes for Fast Maximum Entropy Training. In *Proc. ICASSP*, pp. 561–564. IEEE, 2001.

Gurevych, I. Using the Structure of a Conceptual Network in Computing Semantic Relatedness. In *Proc. IJCNLP*, pp. 767–778, 2005.

Gutmann, M. U. & Hyvärinen, A. Noise-Contrastive Estimation of Unnormalized Statistical Models , with Applications to Natural Image Statistics. *JMLR*, 13:307–361, 2012.

Hassan, S. & Mihalcea, R. Cross-lingual Semantic Relatedness Using Encyclopedic Knowledge. In *Proc. EMNLP*, pp. 1192–1201. ACL, 2009.

Heafield, K. KenLM: Faster and Smaller Language Model Queries. In *Proc. Workshop on Statistical Machine Translation*, pp. 187–197. ACL, 2011.

Hermann, K. M. & Blunsom, P. The Role of Syntax in Vector Space Models of Compositional Semantics. In *Proc. ACL*, pp. 894–904, 2013.

Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. Improving Word Representations via Global Context and Multiple Word Prototypes. In *Proc. ACL*, pp. 873–882. ACL, 2012.

Joubarne, C. & Inkpen, D. Comparison of Semantic Similarity for Different Languages Using the Google N-gram Corpus and Second- Order Co-occurrence Measures. In *Proc. Canadian Conference on Advances in AI*, pp. 216–221. Springer-Verlag, 2011.

Kalchbrenner, N. & Blunsom, P. Recurrent Continuous Translation Models. In *Proc. EMNLP*, pp. 1700–1709. ACL, 2013.

Kneser, R. & Ney, H. Improved Backing-off for m-gram Language Modelling. In *Proc. ICASSP*, pp. 181–184, 1995.

Lazaridou, A., Marelli, M., Zamparelli, R., & Baroni, M. Compositional-ly Derived Representations of Morphologically Complex Words in Distributional Semantics. In *Proc. ACL*, pp. 1517–1526, 2013. ACL.

Le, H.-S., Oparin, I., Allauzen, A., Gauvain, J.-L., & Yvon, F. Structured Output Layer Neural Network Language Model. In *Proc. ICASSP*, pp. 5524–5527, 2011. IEEE.

Luong, M.-T., Socher, R., & Manning, C. D. Better Word Representations with Recursive Neural Networks for Morphology. In *Proc. of CoNLL*, 2013.

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. Recurrent neural network based language model. In *Proc. Interspeech*, pp. 1045–1048, 2010.

Mikolov, T., Kombrink, S., Burget, L., Černocký, J., & Khudanpur, S. Extensions of Recurrent Neural Network Language Model. In *Proc. ICASSP*, 2011.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. Efficient Estimation of Word Representations in Vector Space. In *Proc. ICLR*. arXiv:1301.3781, 2013a.

Mikolov, T., Yih, W.-t., & Zweig, G. Linguistic Regularities in Continuous Space Word Representations. In *Proc. HLT-NAACL*. ACL, 2013b.

Mnih, A. & Hinton, G. Three New Graphical Models for Statistical Language Modelling. In *Proc. ICML*, pp. 641–648, 2007. ACM.

Mnih, A. & Hinton, G. A Scalable Hierarchical Distributed Language Model. In *NIPS*, pp. 1081–1088, 2008.

Mnih, A. & Teh, Y. W. A fast and simple algorithm for training neural probabilistic language models. In *Proc. ICML*, 2012.

Och, F. J. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. ACL*, pp. 160–167, 2003.

Rubenstein, H. & Goodenough, J. B. Contextual Correlates of Synonymy. *Commun. ACM*, 8(10):627–633, October 1965.

Schmid, H. & Laws, F. Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging. In *Proc. COLING*, pp. 777–784, 2008. ACL.

Schwenk, H. Efficient Training of Large Neural Networks for Language Modeling. In *Proc. IEEE Joint Conference on Neural Networks*, pp. 3059–3064. IEEE, 2004.

Schwenk, H. & Koehn, P. Large and Diverse Language Models for Statistical Machine Translation. In *Proc. IJCNLP*, 2008.

Schwenk, H., Dchelotte, D., & Gauvain, J.-L. Continuous Space Language Models for Statistical Machine Translation. In *Proc. COLING/ACL*, pp. 723–730, 2006. ACL.

Schwenk, H., Rousseau, A., & Attik, M. Large, Pruned or Continuous Space Language Models on a GPU for Statistical Machine Translation. In *In Proc. NAACL-HLT Workshop: On the Future of Language Modeling for HLT*, pp. 11–19. ACL, 2012.

Stolcke, A. SRILM – An extensible language modeling toolkit. In *Proc. ICSLP*, pp. 901–904, 2002.

van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *JMLR*, 9:2579–2605, 2008.

Vaswani, A., Zhao, Y., Fossum, V., & Chiang, D. Decoding with Large-Scale Neural Language Models Improves Translation. In *Proc. EMNLP*, 2013. ACL.

Zesch, T. & Gurevych, I. Automatically creating datasets for measures of semantic relatedness. In *Proc. Workshop on Linguistic Distances*, pp. 16–24. ACL, 2006.

Zou, W. Y., Socher, R., Cer, D., & Manning, C. D. Bilingual Word Embeddings for Phrase-Based Machine Translation. In *Proc. EMNLP*, pp. 1393–1398, 2013. ACL.