On the Relation between Linguistic Typology and (Limitations of) Multilingual Language Modeling

Daniela Gerz[1]*, **Ivan Vulić**[1]*, **Edoardo Maria Ponti**[1], **Roi Reichart**[2], **Anna Korhonen**[1]

[1]Language Technology Lab, DTAL, University of Cambridge
[2]Faculty of Industrial Engineering and Management, Technion, IIT
[1]{dsg40, iv250, ep490, alk23}@cam.ac.uk
[2]roiri@ie.technion.ac.il

Abstract

A key challenge in cross-lingual NLP is developing general language-independent architectures that are equally applicable to any language. However, this ambition is largely hampered by the variation in structural and semantic properties, i.e. the typological profiles of the world's languages. In this work, we analyse the implications of this variation on the language modeling (LM) task. We present a large-scale study of state-of-the art *n-gram based* and *neural* language models on 50 typologically diverse languages covering a wide variety of morphological systems. Operating in the *full vocabulary LM setup* focused on *word-level prediction*, we demonstrate that a coarse typology of morphological systems is predictive of absolute LM performance. Moreover, *fine-grained typological features* such as exponence, flexivity, fusion, and inflectional synthesis are borne out to be responsible for the proliferation of low-frequency phenomena which are organically difficult to model by statistical architectures, or for the meaning ambiguity of character n-grams. Our study strongly suggests that these features have to be taken into consideration during the construction of next-level language-agnostic LM architectures, capable of handling morphologically complex languages such as Tamil or Korean.

1 Introduction

Deep learning has allowed NLP algorithms to dispose of manually-crafted features, and to virtually achieve language independence. However, their performance still varies noticeably across languages due to different underlying data distributions (Bender, 2013; O'Horan et al., 2016). Linguistic typology, the systematic comparison of the world's languages, holds promise to explain these idiosyncrasies and interpret statistical models in terms of variation in language structures (Ponti et al., 2017).

In order to evaluate how cross-lingual structural variation hinders the design of effective general-purpose algorithms, we propose the task of language modeling (LM) as a testbed. In particular, we opt for a full-vocabulary setup where no word encountered at training time is treated as an unknown symbol, in order to **a)** ensure a fair comparison across languages with different word frequency rates and **b)** avoid setting an arbitrary threshold on vocabulary size (Cotterell et al., 2018).

Although there has recently been a tendency towards expanding test language samples, the datasets considered in previous works (Botha and Blunsom, 2014; Vania and Lopez, 2017; Kawakami et al., 2017; Cotterell et al., 2018) are not entirely adequate yet to represent the typological variation and to ground cross-lingual generalisations empirically. Hence, we test several LM architectures (including n-gram, neural, and character-aware models) on a novel and wider set of 50 languages sampled according to stratification principles.

Through this large-scale multilingual analysis, we shed new light on the current limitations of standard LM models and offer support to further developments in multilingual NLP. In particular, we demonstrate that the previous fixed-vocabulary assumption in fact ignores the limitations of language modeling for morphologically rich languages. Moreover, we find a strong correlation across the board between LM model performances and the type of morphological system adopted in each language.

To motivate this correlation we show how fine-grained typological properties interact with the frequency distribution (Zipf, 1949) by regulating word boundaries and the proliferation of word forms; and 2) with the mapping between morphemes (here intended as character n-grams) and meaning, by possibly blurring it.

The paper is organised as follows. After provid-

*Both authors equally contributed to this work.

ing a short overview of multilingual LM and its possible setups (§2), we describe the cross-lingual variation in morphological systems and propose a novel typologically diverse dataset for LM in §3. We outline the data in §4 and benchmarked language models in §5. Finally, we discuss the results in light of linguistic typology in §6.

2 Multilingual Language Modeling

A language model computes a probability distribution over sequences of word tokens, and is typically trained to maximise the likelihood of word input sequences. The LM objective is expressed as:

$$P(w_1, ...w_n) = \prod_i P(w_i|w_1, ...w_{i-1}) \quad (1)$$

w_i is a word token with the index i in the sequence. LM is considered a central task in NLP and language understanding, with applications in speech recognition (Mikolov et al., 2010), text summarisation (Filippova et al., 2015; Rush et al., 2015), and information retrieval (Ponte and Croft, 1998; Zamani and Croft, 2016). The importance of language modeling has been accentuated even more in representation learning recently, where it is used as a novel form of unsupervised pre-training (and an alternative to static word embeddings) for the benefit of a variety of NLP applications (Peters et al., 2018; Howard and Ruder, 2018).

Related Work: Datasets and Evaluation. Language modeling is predominantly tested on English and other Western European languages. Standard English LM benchmarks are the Penn Treebank (PTB) (Marcus et al., 1993) and the 1 Billion Word Benchmark (BWB) (Chelba et al., 2013). Datasets extracted from BBC News (Greene and Cunningham, 2006) and IMDB Movie Reviews (Maas et al., 2011) are also used for LM evaluation in English (Wang and Cho, 2016; Miyamoto and Cho, 2016; Press and Wolf, 2017).

For multilingual LM evaluation, Botha and Blunsom (2014) extract datasets for Czech, French, Spanish, German, and Russian from the 2013 Workshop on Statistical Machine Translation (WMT) data (Bojar et al., 2013). Kim et al. (2016) reuse these datasets and add Arabic. Ling et al. (2015) evaluate on English, Portuguese, Catalan, German and Turkish datasets extracted from Wikipedia. Kawakami et al. (2017) evaluate on 7 European languages using Wikipedia data, including Finnish.

To the best of our knowledge, the largest datasets used in previous work are from (Müller et al., 2012; Cotterell et al., 2018) and amount to 21 languages from the Europarl data (Koehn, 2005). Despite the large coverage of languages, these sets are still restricted only to the languages of the European Union. On the other hand, the most typologically diverse dataset thus far was released by Vania and Lopez (2017). It includes 10 languages representing some morphological systems.

This short survey of related work demonstrates a clear tendency towards extending LM evaluation to other languages, abandoning English-centric assumptions, and focusing on language-agnostic LM architectures. However, a comprehensive evaluation set that systematically covers a wide and balanced spectrum of typologically diverse languages is still missing. The novel dataset we discuss in this paper aims at bridging this gap (see §4).

Fixed vs. Full Vocabulary Setup. A majority of language models rely on the fixed-vocabulary assumption: they use a special symbol <UNK> that represents all words not present in the fixed vocabulary V, which are termed out-of-vocabulary (OOV). Selecting the set V typically slips under the radar, and can be seen as "something of a black art" despite its enormous impact on final LM performance (Cotterell et al., 2018).[1] Standard LM setups either fix the vocabulary V to the top n most frequent words, typically with $n = 10,000$ or $n = 5,000$ (Mikolov et al., 2010; Ling et al., 2015; Vania and Lopez, 2017; Lee et al., 2017, *inter alia*), or include in V only words with a frequency below a certain threshold (typically 2 or 5) (Heafield et al., 2013).

The rationale behind fixing the set V is **a)** to make the language model more robust to handling OOVs and to effectively bypass the problem of unreliable word estimates for low-frequency and unseen words (by ignoring them), and **b)** to enable direct comparisons of absolute perplexity scores across different models. However, this posits a critical challenge as cross-linguistic evaluation becomes uneven. In fact, we witness a larger proportion of vocabulary words replaced by <UNK> in morphologically rich languages because of their higher OOV rates (see Table 3). What is more, while the fixed-vocabulary assumption artificially

[1] For instance, Vania and Lopez (2017) report perplexity scores of ≈20 for Finnish when V is fixed to the $5k$ most frequent words. The same model in the full-vocabulary setup obtains perplexity scores of ≈2,000.

FI	*Kreikkalaiset sijoittivat geometrian synnyn muinaiseen Egyptiin , jossa sitä tarvittiin maanmittaukseen .*
FI (MIN-5)	*<UNK> <UNK> <UNK> synnyn <UNK> Egyptiin , jossa sitä tarvittiin <UNK> .*
FI (10K)	*<UNK> <UNK> <UNK> <UNK> <UNK> <UNK> , jossa sitä <UNK> <UNK> .*
KO	그 뒤 한시 백일장에서 장원하여 신동으로 알려졌다. 그러나 그의 집은 지독하게 가난했다 .
KO (MIN-5)	그 뒤 <UNK> <UNK> <UNK> <UNK> 알려졌다 . 그러나 그의 집은 <UNK> <UNK> .
KO (10K)	그 뒤 <UNK> <UNK> <UNK> <UNK> 알려졌다 . 그러나 그의 <UNK> <UNK> <UNK> .

Table 1: Examples from Finnish and Korean LM datasets after applying the standard fixed-vocabulary assumption. MIN=5: only words with corpus frequency above 5 are retained in the final fixed vocabulary V; 10K: V comprises the $10k$ most frequent words.

improves the perplexity measure, it actually makes the models less useful, especially in morphologically rich languages, as exemplified in Table 1.

Our goal is to get a clear picture on how different typological features and the corresponding corpus frequency distributions affect LM performance, without the influence of the unrealistic fixed-vocabulary assumption. Therefore, we work in the *full-vocabulary LM setup* (Adams et al., 2017; Grave et al., 2017). This means that we explicitly decide to retain also infrequent words in the modeled data: V contains all words occurring at least once in the training set, only unseen words from test data are treated as OOVs. We believe that this setup leads to an evaluation that pinpoints the crucial limitations of standard LM architectures.[2]

Why Not Open Vocabulary Setup? Recent neural LM architectures have also focused on handling large vocabularies and unseen words using character-aware modeling (Luong and Manning, 2016; Jozefowicz et al., 2016; Kawakami et al., 2017, *inter alia*). This setup is commonly referred to as the *open-vocabulary* setup. However, two distinct approaches with crucial modeling differences are referred to by the same term in the literature. **a)** *Word-level generation* constructs word vectors for arbitrary words from constituent subword-level components, but word-level prediction is still evaluated based on the fixed-vocabulary assumption. **b)** *Character-level generation* predicts characters instead of words.

Given that character-level prediction and word-level prediction operate on entirely different sets of symbols, their performance is hardly comparable. Still, Jozefowicz et al. (2016) report that, in a hybrid setup which evaluates character-level prediction based on word-level perplexity with the fixed-vocabulary assumption, current state-of-the-art word-level prediction models (i.e., the ones we discuss in §5) still significantly outperform such hybrid character-level prediction approaches. Therefore, we operate in the full-vocabulary setup.

[2]For instance, as discussed later in §3 and validated empirically in §6, the vocabularies of morphologically rich languages are inherently larger: it is simply more difficult to learn and make LM predictions in such languages.

Type	Fusion	Exponence	Flexivity	Synthesis
Isolating	low	1:1	1:1	low
Fusional	mid	many:1	1:many	mid
Introflexive	high	many:1	1:many	mid
Agglutinative	mid	1:1	1:1	high

Table 2: Traditional morphological types described in terms of selected features from WALS.

3 Typology of Morphological Systems

Aiming for a comprehensive multilingual LM evaluation in this study, we survey all possible types of morphological systems (Haspelmath and Sims, 2013), which possibly lead to different performances. Traditionally, languages have been grouped into the four main categories: *isolating*, *fusional*, *introflexive* and *agglutinative*, based on their position along a spectrum measuring the preference on breaking up concepts in many words (on one extreme) or rather compose them into single words (on the other extreme).

The mono-dimensionality of this spectrum has recently been challenged as languages exhibit a multitude of morphological features that do not covary across languages (Plank, 2017; Ponti et al., 2018). The typological database WALS (Dryer and Haspelmath, 2013) documents several of them that are relevant for LM: *inflectional synthesis*, *fusion*, *exponence*, and *flexivity*. Note that the prototypes of traditional categories can be approximated in terms of these features, as shown in Table 2, although more combinations are possible.

Languages specify different subsets of grammatical categories (such as tense for verbs, or num-

ber for nouns), and for each category different values are available in each language: for instance, Finnish has less tense values (it lacks a future), whereas Slovene has more number values (including a dual) compared to English. The feature **inflectional synthesis** for verbs (Bickel and Nichols, 2013) measures how many categories appear on the maximally inflected verb per language. More available categories enlarge the vocabulary (and consequently the OOV rate) with forms instantiating all possible combinations of their values.

Another crucial aspect is how the available grammatical categories are expressed, which can be described by fusion, exponence, and flexivity. **Fusion** measures the degree of connectedness between a grammatical marker to another word. The marker can be (from lower to higher fusion) a separate word, a clitic, an affix, or can affect the form of the root itself (e.g. an umlaut or a tone).

Exponence measures the number of categories (e.g., tense, number) a single morpheme tends to convey. Exponence is separative if one grammatical category is conveyed by one morpheme (1:1), and cumulative if multiple categories are grouped into one morpheme (many:1).

Flexivity indicates the possibility that the value of a grammatical category be mapped into different morphological forms (1:many). In other terms, lemmas belonging to the same part-of-speech are divided into inflectional classes (such as declension classes for nouns or conjugation classes for verbs), each characterised by a different paradigm, that is, a different set of value-to-form mappings.

The three last features are illustrated by the examples Ex. (2)-Ex. (5), all uttering the sentence "*I will guard the doors and I will not open (them)*".[3]

(2) *tôi sẽ bảo vệ cửa và tôi sẽ*
I FUT guard door and I FUT
không mở
NEG open (Vietnamese)

(3) *kapı-lar-ı koruy-acağ-ım ve*
door-PL-ACC guard-FUT-1SG and
aç-may-acağ-ım
open-NEG-FUT-1SG (Turkish)

(4) *sorvegli-erò le port-e e non*
guard-FUT.1SG DEF door-PL and NEG
apr-irò
open-FUT.1SG (Italian)

(5) *ʻe-šmor ʻal ha-dʻlat-ót vʻ-lo*
1SG-guard.FUT on DEF-door-PL and-NEG
ʻe-ftach otán
1SG-wait.FUT them (Hebrew)

In particular, consider how tense and person are expressed on verbs. Vietnamese in Ex. (2) puts two particles *tôi* and *sẽ* before the verb, which are distinct (separate exponence), autonomous from the root (no fusion), and fixed (absence of flexivity). Turkish in Ex. (3) attaches suffixes: *-acak-* for tense and *-ım* for person. These are distinct (separate exponence), joined to the roots (concatenative fusion), and (phonologically determined variants of) the same morpheme (1:1 flexivity). Italian in Ex. (4) uses affixes *-erò* and *-irò*: they are concatenated to the root with respect to fusion, convey both tense and person (cumulative exponence), and are dissimilar (presence of flexivity). Finally, in Ex. (5) for Hebrew the consonant pattern of the verb *š-m-r* is interdigitated by the vowel *-o-* for tense, and preceded by a prefix *ʻe-* for person. The first phenomenon alters the root itself (introflexive fusion), is distinct from the second (separate exponence), and changes its realisation based on the verb's lemma (presence of flexivity).

The above evidence strongly motivates us, as well as recent previous work (Vania and Lopez, 2017; Kawakami et al., 2017; Cotterell et al., 2018), to approach LM with models that are aware of the inner structure of their input words, and to benchmark these modeling choices on a typologically diverse range of languages, as shown in §4.

4 Data

Selection of Languages. Our selection of test languages is guided by the following goals: **a)** we have to ensure the coverage of typological properties from §3, and **b)** we want to analyse a large set of languages which extends and surpasses other work in the LM literature (see §2).

Since cross-lingual NLP aims at modeling *extant* languages rather than *possible* languages (including, e.g., extinct ones), creating a balanced sample is challenging. In fact, attested languages, intended as a random variable, are extremely sparse and not independent-and-identically-distributed (Cotterell and Eisner, 2017). First, available and reliable data exist only for a fraction of the world's languages. Second, these data are biased because their features may not stem from the underlying distribution, i.e., from what is naturally possible/frequent, but rather

[3] All morphological glosses follow the Leipzig glossing rules, listed at https://www.eva.mpg.de/lingua/resources/glossing-rules.php

can be inherited by genealogical relatedness or borrowed by areal proximity (Bakker, 2010). To mitigate these biases, theoretical works resorted to stratification approaches, where each subgroup of related languages is sampled independently. maximizing their diversity (Dryer, 1989, *inter alia*). We perform our selection in the same spirit.

We start from the Polyglot Wikipedia (PW) project (Al-Rfou et al., 2013) which provides cleaned and tokenised Wikipedia data in 40 languages. However, the majority of the PW languages are similar from the perspective of genealogy (26/40 are Indo-European), geography (28/40 are Western European), and typology (26/40 are fusional). Consequently, the PW set is not a representative sample of the world's languages.

To amend this limitation, we source additional languages with the data coming from the same domain, Wikipedia, considering candidates in descending order of corpus size cleaned and preprocessed by the Polyglot tokeniser (Al-Rfou et al., 2013). Since fusional languages are already represented in the PW, we add new languages from other morphological types: isolating (*Min Nan, Burmese, Khmer*), agglutinative (*Basque, Georgian, Kannada, Tamil, Mongolian, Javanese*), and introflexive languages (*Amharic*).

Partition. We construct datasets for all 50 languages by extracting the first 40K sentences for each language, and split them into train (34K), validation (3K), and test (3K). This choice has been motivated by the following observations: **a)** we require similarly-sized datasets from the same domain for all languages; **b)** the size of the datasets has to be similar to the standard English PTB dataset (Marcus et al., 1993) which has been utilised to guide LM development in English for more than 20 years. The final list of 50 languages along with their language codes (ISO 639-1), morphological type (i.e., isolating, fusional, introflexive, agglutinative), and corpus statistics is provided in Table 3.

5 Models and Experimental Setup

Benchmarked Language Models. The availability of LM evaluation sets in a large number of diverse languages, as described in §4, gives an opportunity to conduct a full-fledged multilingual analysis of representative LM architectures for word-level prediction. First, we evaluate a state-of-the-art model from the *n-gram* family of models (Goodman, 2001) from the KenLM package.[4] It is based on 5-grams with extended Kneser-Ney smoothing (Kneser and Ney, 1995; Heafield et al., 2013). We refer to this model as **KN5**.

Modern LM architectures are almost exclusively based on recurrent neural networks (RNNs), and especially on Long-Short-Term Memory networks (LSTMs). (Mikolov et al., 2010; Sundermeyer et al., 2015; Chen et al., 2016, *inter alia*). They map a sequence of input words to embedding vectors using a look-up matrix and then perform word-level prediction by passing the vectors to the **LSTM**.

Finally, we also evaluate a character-aware variant of the neural LSTM LM architecture. We use the **Char-CNN-LSTM** model (Kim et al., 2016) due to its public availability and strong performance in several languages. In this model, each character is embedded and passed through a convolutional neural network with max-over-time pooling (LeCun et al., 1989), followed by a highway network transformation (Srivastava et al., 2015) to build word representations from their constituent characters. By resorting to character-level information, the model is able to provide better parameter estimates for lower-frequency words, which is particularly important for morphologically rich languages. The CNN-based word representations are then processed in a sequence by a regular LSTM network to obtain word-level predictions.

Evaluation Setup. We report *perplexity* scores (Jurafsky and Martin, 2017, chapter 4.2.1) using the *full* vocabulary for each respective LM dataset. This means that we explicitly decide to retain also infrequent words in the data and analyse the difficulty of modeling such words in morphologically rich languages (see §2 for the discussion).

In the full-vocabulary setup, the set V comprises all words occurring at least once in the training set. Unseen test words are mapped to *one* <UNK> vector, sampled from the the space of trained word vectors relying on a normal distribution and the same fixed random seed for all models. On the other hand, KN5 by design has a slightly different way of handling unseen test words: they are regarded as outliers and assigned low-probability estimates.

Training Setup and Parameters. For LSTM and Char-CNN-LSTM language models, we reproduce the standard LM setup of Zaremba et al. (2015) and parameter choices of Kim et al. (2016).

[4] https://github.com/kpu/kenlm

Batch size is 20 and a sequence length is 35, where one step corresponds to one word token. The maximum word length is chosen dynamically based on the longest word in the corpus. The corpus is processed continuously; the RNN hidden states reset at the beginning of each epoch. Parameters are optimised with SGD, and the gradient is averaged over the batch size and sequence length. We then scale the averaged gradient by the sequence length (=35) and clip to 5.0 for more stable training. The learning rate is 1.0, decayed by 0.5 after each epoch if the validation perplexity does not improve. All models are trained for 15 epochs, which is typically sufficient for model convergence. Finally, KN5 is trained relying on the suggested parameters from the KenLM package.

6 Results and Discussion

In this section, we present our main empirical findings on the connection between LM performance and corpus statistics emerging from different typological profiles (see §3). Before proceeding, we stress that the absolute perplexity scores across different languages are not directly comparable, but their values provide evidence on the difficulty and limitations of language modeling in each language, considering the fact that all language models were trained on similarly-sized datasets. The results for all three benchmarked language models on all 50 languages are summarised in Table 3.

Comparison of Language Models. A quick inspection of the results from Table 3 reveals that the Char-CNN-LSTM model is the best-performing model overall. We report the best results with that model for 48/50 languages and across all traditional morphological types. Gains over the simpler recurrent LM architecture (i.e., the LSTM model) are present for all 50/50 languages. In short, this means that character-level information on the input side of neural architectures, in addition to leading to fewer parameters, is universally beneficial for the final performance of word-level prediction, as also suggested by Kim et al. (2016) on a much smaller set of languages. By relying on character-level knowledge, Char-CNN-LSTM model provides better estimates for lower-frequency words.

Moreover, the results show that KN5 is a competitive baseline for several languages (e.g., Kannada, Thai, Amharic). This further highlights the importance of testing models on a typologically diverse set of languages: despite the clear superiority of neural LM architectures such as Char-CNN-LSTM in a large number of languages, the results and the marked outliers still suggest that there is currently no "one-size-fits-all" model.

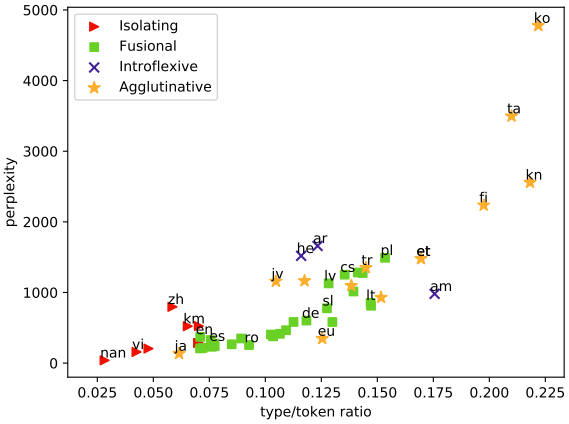

Figure 1: Perplexity scores with the Char-CNN-LSTM language model (Kim et al., 2016) on PTB-sized language modeling data in 50 languages as a function of type-to-token ratios in training data.

In general, large perplexity scores for certain languages (e.g., agglutinative languages such as Finnish, Korean, Tamil, or introflexive languages), especially when compared to performance on English on a similarly-sized dataset, clearly point at the limitations of all the "language-agnostic" LM architectures. As suggested by Jozefowicz et al. (2016), LM performance in English can be boosted by simply collecting more data and working with large vocabularies (e.g., reducing the number of relevant OOVs). However, this solution is certainly not applicable to a majority of the world's languages (Bird, 2011; Gandhe et al., 2014; Adams et al., 2017), see later in §6: *Further Discussion*.

Frequency Analysis and Traditional Morphological Types. We now analyse all languages in our collection according to word-level frequency properties also listed in Table 3 for all 50 languages. We report: 1) the vocabulary size (i.e., the total number of vocabulary words in each training dataset); 2) the total number of test words not occurring in the corresponding training data; 3) the total number of tokens in both training and test data; and finally 4) type-to-token ratios (TTR) in training data. We also plot absolute perplexity scores of Char-CNN-LSTM (Kim et al., 2016), the best-performing model overall (see §6), in relation to TTR ratios in Figure 1.

Language (code)	Data Stats					Baseline Models		
	Vocab Size (Train)	New Test Vocab	Number Tokens (Train)	Number Tokens (Test)	Type / Token (Train)	KN5	LSTM	Char-CNN-LSTM
× Amharic (am)	89749	4805	511K	39.2K	0.18	1252	1535	**981**
× Arabic (ar)	89089	5032	722K	54.7K	0.12	2156	2587	**1659**
☐ Bulgarian (bg)	71360	3896	670K	49K	0.11	610	651	**415**
☐ Catalan (ca)	61033	2562	788K	59.4K	0.08	358	318	**241**
☐ Czech (cs)	86783	4300	641K	49.6K	0.14	1658	2200	**1252**
☐ Danish (da)	72468	3618	663K	50.3K	0.11	668	710	**466**
☐ German (de)	80741	4045	682K	51.3K	0.12	930	903	**602**
☐ Greek (el)	76264	3767	744K	56.5K	0.10	607	538	**405**
☐ English (en)	55521	2480	783K	59.5K	0.07	533	494	**371**
☐ Spanish (es)	60196	2721	781K	57.2K	0.08	415	366	**275**
★ Estonian (et)	94184	3907	556K	38.6K	0.17	1609	2564	**1478**
★ Basque (eu)	81177	3365	647K	47.3K	0.13	560	533	**347**
☐ Farsi (fa)	52306	2041	738K	54.2K	0.07	355	263	**208**
★ Finnish (fi)	115579	6489	585K	44.8K	0.20	2611	4263	**2236**
☐ French (fr)	58539	2575	769K	57.1K	0.08	350	294	**231**
× Hebrew (he)	83217	3862	717K	54.6K	0.12	1797	2189	**1519**
☐ Hindi (hi)	50384	2629	666K	49.1K	0.08	473	426	**326**
☐ Croatian (hr)	86357	4371	620K	48.1K	0.14	1294	1665	**1014**
★ Hungarian (hu)	101874	5015	672K	48.7K	0.15	1151	1595	**929**
▷ Indonesian (id)	49125	2235	702K	52.2K	0.07	454	359	**286**
☐ Italian (it)	70194	2923	787K	59.3K	0.09	567	493	**349**
★ Japanese (ja)	44863	1768	729K	54.6K	0.06	169	156	**136**
★ Javanese (jv)	65141	4292	622K	52K	0.10	1387	1443	**1158**
★ Georgian (ka)	80211	3738	580K	41.1K	0.14	1370	1827	**1097**
▷ Khmer (km)	37851	1303	579K	37.4K	0.07	586	637	**522**
★ Kannada (kn)	94660	4604	434K	29.4K	0.22	**2315**	5310	2558
★ Korean (ko)	143794	8275	648K	50.6K	0.22	5146	10063	**4778**
☐ Lithuanian (lt)	81501	3791	554K	41.7K	0.15	1155	1415	**854**
☐ Latvian (lv)	75294	4564	587K	45K	0.13	1452	1967	**1129**
▷ Malay (ms)	49385	2824	702K	54.1K	0.07	776	725	**525**
★ Mongolian (mng)	73884	4171	629K	50K	0.12	1392	1716	**1165**
▷ Burmese (my)	20574	755	576K	46.1K	0.04	209	212	**182**
▷ Min-Nan (nan)	33238	1404	1.2M	65.6K	0.03	61	43	**39**
☐ Dutch (nl)	60206	2626	708K	53.8K	0.08	397	340	**267**
☐ Norwegian (no)	69761	3352	674K	47.8K	0.10	534	513	**379**
☐ Polish (pl)	97325	4526	634K	47.7K	0.15	1741	2641	**1491**
☐ Portuguese (pt)	56167	2394	780K	59.3K	0.07	342	272	**214**
☐ Romanian (ro)	68913	3079	743K	52.5K	0.09	384	359	**256**
☐ Russian (ru)	98097	3987	666K	48.4K	0.15	1128	1309	**812**
☐ Slovak (sk)	88726	4521	618K	45K	0.14	1560	2062	**1275**
☐ Slovene (sl)	83997	4343	659K	49.2K	0.13	1114	1308	**776**
☐ Serbian (sr)	81617	3641	628K	46.7K	0.13	790	961	**582**
☐ Swedish (sv)	77499	4109	688K	50.4K	0.11	843	832	**583**
★ Tamil (ta)	106403	6017	507K	39.6K	0.21	**3342**	6234	3496
▷ Thai (th)	30056	1300	628K	49K	0.05	233	241	**206**
▷ Tagalog (tl)	72416	3791	972K	66.3K	0.07	379	298	**219**
★ Turkish (tr)	90840	4608	627K	45K	0.14	1724	2267	**1350**
☐ Ukranian (uk)	89724	4983	635K	47K	0.14	1639	1893	**1283**
▷ Vietnamese (vi)	32055	1160	754K	61.9K	0.04	197	190	**158**
▷ Chinese (zh)	43672	1653	746K	56.8K	0.06	1064	826	**797**
▷ **Isolating** (avg)	40930	1825	759K	54K	0.05	440	392	**326**
☐ **Fusional** (avg)	73499	3532	689K	51.3K	0.11	842	969	**618**
× **Introflexive** (avg)	87352	4566	650K	49.5K	0.14	1735	2104	**1386**
★ **Agglutinative** (avg)	91051	4687	603K	45K	0.16	1898	3164	**1727**

Table 3: Test perplexities for 50 languages (ISO 639-1 codes sorted alphabetically) in the full-vocabulary prediction LM setup; **Left**: Basic statistics of LM evaluation data (see §4 and §5). **Right**: Results with all three language models in our comparison. Best absolute perplexity scores for each language are in bold, but note that the absolute scores in the KN5 column are not directly comparable to the scores obtained with neural models due to a different handling of OOVs at test time (see §5).

In isolating and some fusional languages (e.g., Vietnamese, Thai, English) the TTR tends to be small: we have a comparatively low number of infrequent words. Agglutinative languages such as Finnish, Estonian, and Korean are on the other side of the spectrum. Introflexive and fusional languages, typically over-represented in prior work (see the discussion in §3), are found in the middle.

This emerges clearly in Figure 1, grouping isolating languages to the left side of the x-axis, followed

by fusional languages (Germanic and Romance first to the left, and then Balto-Slavic to the right), and placing agglutinative languages towards the far right. Crucially, TTR is an excellent predictor of LM performance. To measure the correlation between this corpus statistics variable and absolute LM performance, we compute their Pearson's r correlation. We find a strong positive correlation, with a value of $r = 0.83$ and significance $p < 0.001$.

We do observe a strong link between each language's morphological type, and the corresponding perplexity score. A transition in terms of the spectrum of morphological systems (see §3) can be traced again on the y-axis of Figure 1, roughly following the reported LM performance: from isolating, over fusional and introflexive to agglutinative languages. In fact, a correlation exists also between traditional morphological types and LM performance. We assessed its strength with the one-way ANOVA statistical test, obtaining a value of $\eta^2 = 0.37$ and a significance of $p < 0.001$.

Finally, it should be noted that the choice of TTP over other corpus statistics such as vocabulary size is motivated by the fact that the corpora are comparable, and not parallel. Because of this, the variation of V may stem from the contents rather than the intrinsic linguistic properties. As a counter-check, the correlation between V and LM performance is in fact milder, with $r = 0.64$. Yet, notwithstanding the stronger correlation, TTP is unable to explain the results entirely. Only through finer-grained typological features it becomes possible to justify several outliers, as shown in the next subsection.

Fine-Grained Typological Analysis. Among the relevant typological features (see §3 and Table 2), fusion and inflectional synthesis have the largest impact on word-level predictions. In fact, the former determines the word boundaries, whereas the latter regulates the amount of possible morpheme combinations. Consider their effect on the frequency distribution of words, expressed as follows (Zipf, 1949):

$$f = \frac{\frac{1}{k^s}}{\sum_{n=1}^{V} \frac{1}{n^s}} \tag{6}$$

f is the frequency, k the rank, and $s \geq 0$ the exponent characteristic of the distribution. If high, both typological features enlarge V and s, assigning less probability mass to each word.

Low fusion means a preference for separate words (as in isolating languages such as Vietnamese and Chinese), leading to a smaller vocabulary with less (but more frequent) words. This property, additionally boosted by low inflectional synthesis, facilitates statistical language modeling in isolating languages. Vice versa, high fusion results in preference for concatenation of morphemes or introflection, and consequently sparser vocabularies. Yet, this distinction cannot justify the figures by itself, as it equates agglutinative languages and traditional fusional languages. Here, inflectional synthesis is also at play. Through the statistical test of one-way ANOVA, we found a weak effect of $\eta^2 = 0.09$ for fusion and a medium effect of $\eta^2 = 0.21$ for inflection synthesis.

On the other hand, the fine-grained typological features of exponence and flexivity play a role in the ambiguity of the mapping between morphemes and meanings or grammatical functions. This turns out to be especially relevant for character-aware models. The intuition is that if the mapping is straightforward, injecting character information is more advantageous. To validate this claim, we evaluate the ANOVA between exponence of nouns and verbs and the difference in perplexity between LSTM and Char-CNN-LSTM.[5] We report a weak, although existent, correlation with value $\eta^2 = 0.07$ and $\eta^2 = 0.04$, respectively.

Further Discussion. Importantly, our large-scale multilingual LM study strongly indicates that due to diverse typological profiles, certain languages and language groups are inherently more complex to language-model when relying on established statistical models, even when such models are constructed as widely applicable and (arguably) language-agnostic. This finding supports preliminary results from prior work (Botha and Blunsom, 2014; Adams et al., 2017; Cotterell et al., 2018), and is also backed by insights from linguistic theory on variance of language complexity in general and variance of morphological complexity in specific (McWhorter, 2001; Evans and Levinson, 2009). More broadly and along the same line, earlier research in statistical machine translation (SMT) has also shown that typological factors such as the amount of reordering, the morphological complexity, as well as genealogical relatedness of languages are crucial in predicting success in SMT (Birch et al., 2008; Paul et al., 2009; Daiber, 2018).

Our results indicate that the artificial fixed-

[5] Unfortunately no values are available in WALS for the feature of flexivity besides a limited domain.

vocabulary assumption from prior work produces overly optimistic perplexity scores, and its limitation is even more pronounced in morphologically rich languages, which inherently contain a large number of infrequent words due to their productive morphological systems. The typical solution to collect more data (Jozefowicz et al., 2016; Kawakami et al., 2017) mitigates this effect to a certain extent, but stills suffers from the Zipfian hypothesis (1949), and it cannot be guaranteed for resource-poor languages where obtaining sufficient monolingual data is also a challenge (Adams et al., 2017).

Therefore, another solution is to resort to other sources of information which are not purely contextual/distributional. For instance, a promising line of current and future research is to (learn to) exploit subword-level patterns captured in an unsupervised manner (Pinter et al., 2017; Herbelot and Baroni, 2017) or integrate existing morphological generation and inflection tools and regularities (Cotterell et al., 2015; Vulić et al., 2017; Bergmanis et al., 2017) into language models to reduce data sparsity, and improve language modeling for morphologically rich languages. For instance, a recent enhancement of the Char-CNN-LSTM language model that enforces similarity between parameters of morphologically related words leads to large perplexity gains across a large number of languages, with the most prominent gains reported for morphologically complex languages (Gerz et al., 2018).

Given the recent success and improved performance with LM-based pre-training methodology (Peters et al., 2018; Howard and Ruder, 2018) across a wide variety of syntactic and semantic NLP tasks in English, improving language models for other languages might have far-reaching consequences for multilingual NLP in general. Typological information coded in typological databases (Ponti et al., 2018) offer invaluable support to language modeling (e.g., knowledge on word ordering, morphological regularities), but such typologically-informed LM architectures are still non-existent.

7 Conclusion

In this paper, we have run a large-scale study on Language Modeling (LM) across several architectures and a collection of 50 typologically diverse languages. We have demonstrated that typological properties of languages, such as their morphological systems, have an enormous impact on the performance of allegedly "language-agnostic" models. We have found that the corpus statistics most predictive of LM performance is type-to-token ratio (TTR), as demonstrated by their strong Pearson's correlation. In turn, the value of TTR is motivated by fine-grained typological features that define the type of morphological system within a language. In fact, such features affect the word boundaries and the number of morphemes per word, affecting the word frequency distribution for each language.

We have also observed that injecting character information into word representations is always beneficial because this mitigates the above-mentioned sparsity issues. However, the extent of the gain in perplexity partly depends on some typological properties that regulate the ambiguity of the mapping between morphemes (here modeled as character n-grams) and their meaning.

We hope that NLP/LM practitioners will find the datasets for 50 languages put forth in this work along with benchmarked LMs useful for future developments in (language-agnostic as well as typologically-informed) multilingual language modeling. This study calls for next-generation solutions that will additionally leverage typological knowledge for improved language modeling. Code and data are available at: `http://people.ds.cam.ac.uk/dsg40/lmmrl.html`.

Acknowledgements

This work is supported by the ERC Consolidator Grant LEXICAL (no 648909). The authors would like to thank the anonymous reviewers for their helpful suggestions.

References

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of EACL*, pages 937–947.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of CoNLL*, pages 183–192.

Dik Bakker. 2010. Language sampling. In *The Oxford handbook of linguistic typology*, pages 100–127. Oxford University Press.

Emily M. Bender. 2013. *Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax*. Morgan & Claypool Publishers.

Toms Bergmanis, Katharina Kann, Hinrich Schütze, and Sharon Goldwater. 2017. Training data augmentation for low-resource morphological inflection. In *Proceedings of CoNLL*, pages 31–39.

Balthasar Bickel and Johanna Nichols. 2013. *Inflectional Synthesis of the Verb*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings of EMNLP*, pages 745–754.

Steven Bird. 2011. Bootstrapping the language archive: New prospects for natural language processing in preserving linguistic heritage. *Linguistic Issues in Language Technology*, 6(4).

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the 8th Workshop on Statistical Machine Translation*, pages 1–44.

Jan A. Botha and Phil Blunsom. 2014. Compositional morphology for word representations and language modelling. In *Proceedings of ICML*, pages 1899–1907.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, and Phillipp Koehn. 2013. One billion word benchmark for measuring progress in statistical language modeling. In *Proceedings of INTERPSEECH*, pages 2635–2639.

Xie Chen, Xunying Liu, Yanmin Qian, MJF Gales, and Philip C Woodland. 2016. CUED-RNNLM: An open-source toolkit for efficient training and evaluation of recurrent neural network language models. In *Proceedings of ICASSP*, pages 6000 –6004.

Ryan Cotterell and Jason Eisner. 2017. Probabilistic typology: Deep generative models of vowel inventories. In *Proceedings of ACL*, pages 1182–1192.

Ryan Cotterell, Sebastian J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of NAACL-HLT*.

Ryan Cotterell, Thomas Müller, Alexander Fraser, and Hinrich Schütze. 2015. Labeled morphological segmentation with semi-Markov models. In *Proceedings of CoNLL*, pages 164–174.

Joachim Daiber. 2018. *Typologically Robust Statistical Machine Translation*. Ph.D. thesis, University of Amsterdam.

Matthew S Dryer. 1989. Large linguistic areas and language sampling. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 13(2):257–292.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Nicholas Evans and Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5):429–448.

Katja Filippova, Enrique Alfonseca, Carlos A. Colmenares, Lukasz Kaiser, and Oriol Vinyals. 2015. Sentence compression by deletion with LSTMs. In *Proceedings of EMNLP*, pages 360–368.

Ankur Gandhe, Florian Metze, and Ian Lane. 2014. Neural network language models for low resource languages. In *Proceedings of INTERSPEECH*, pages 2615–2619.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Jason Naradowsky, Roi Reichart, and Anna Korhonen. 2018. Language modeling for morphologically rich languages: Character-aware modeling for word-level prediction. *Transactions of the ACL*, 6:451–465.

Joshua T. Goodman. 2001. A bit of progress in language modeling. *Computer Speech & Language*, 15(4):403–434.

Edouard Grave, Moustapha Cissé, and Armand Joulin. 2017. Unbounded cache model for online language modeling with open vocabulary. In *Proceedings of NIPS*, pages 6044–6054.

Derek Greene and Padraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of ICML*, pages 377–384.

Martin Haspelmath and Andrea Sims. 2013. *Understanding morphology*.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of ACL*, pages 690–696.

Aurelie Herbelot and Marco Baroni. 2017. High-risk learning: acquiring new word vectors from tiny data. In *Proceedings of EMNLP*, pages 304–309.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of ACL*, pages 328–339.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. In *Proceedings of ICML*.

Dan Jurafsky and James H. Martin. 2017. *Speech and Language Processing*, volume 3. Pearson.

Kazuya Kawakami, Chris Dyer, and Phil Blunsom. 2017. Learning to create and reuse words in open-vocabulary neural language modeling. In *Proceedings of ACL*, pages 1492–1502.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-aware neural language models. In *Proceedings of AAAI*, pages 2741–2749.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for M-gram language modeling. In *Proceedings of ICASSP*, pages 181–184.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86.

Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. 1989. Handwritten digit recognition with a back-propagation network. In *Proceedings of NIPS*, pages 396–404.

Kenton Lee, Omer Levy, and Luke Zettlemoyer. 2017. Recurrent additive networks. *CoRR*, abs/1705.07393.

Wang Ling, Tiago Luís, Luís Marujo, Ramón Fernández Astudillo, Silvio Amir, Chris Dyer, Alan W. Black, and Isabel Trancoso. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of EMNLP*, pages 1520–1530.

Minh-Thang Luong and Christopher D. Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of ACL*, pages 1054–1063.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of ACL*, pages 142–150.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

John McWhorter. 2001. The world's simplest grammars are Creole grammars. *Linguistic Typology*, 5(2):125–66.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of INTERSPEECH*, pages 1045–1048.

Yasumasa Miyamoto and Kyunghyun Cho. 2016. Gated word-character recurrent language model. In *Proceedings of EMNLP*, pages 1992–1997.

Thomas Müller, Hinrich Schütze, and Helmut Schmid. 2012. A comparative investigation of morphological language modeling for the languages of the European Union. In *Proceedings of NAACL-HLT*, pages 386–395.

Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. In *Proceedings of COLING*, pages 1297–1308.

Michael Paul, Hirofumi Yamamoto, Eiichiro Sumita, and Satoshi Nakamura. 2009. On the importance of pivot language selection for statistical machine translation. In *Proceedings of NAACL-HLT*, pages 221–224.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Yuval Pinter, Robert Guthrie, and Jacob Eisenstein. 2017. Mimicking word embeddings using subword RNNs. In *Proceedings of EMNLP*, pages 102–112.

Frans Plank. 2017. Split morphology: How agglutination and flexion mix. *Linguistic Typology*, 21(2017):1–62.

Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of SIGIR*, pages 275–281.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2018. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *CoRR*, abs/1807.00914.

Edoardo Maria Ponti, Ivan Vulić, and Anna Korhonen. 2017. Decoding sentiment from distributed representations of sentences. In *Proceedings of *SEM*, pages 22–32.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of EACL*, pages 157–163.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of EMNLP*, pages 379–389.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Highway networks. In *Proceedings of the ICML Deep Learning Workshop*.

Martin Sundermeyer, Hermann Ney, and Ralf Schluter. 2015. From feedforward to recurrent LSTM neural networks for language modeling. *IEEE Transactions on Audio, Speech and Language Processing*, 23(3):517–529.

Clara Vania and Adam Lopez. 2017. From characters to words to in between: Do we capture morphology? In *Proceedings of ACL*, pages 2016–2027.

Ivan Vulić, Nikola Mrkšić, Roi Reichart, Diarmuid Ó Séaghdha, Steve Young, and Anna Korhonen. 2017. Morph-fitting: Fine-tuning word vector spaces with simple language-specific rules. In *Proceedings of ACL*, pages 56–68.

Tian Wang and Kyunghyun Cho. 2016. Larger-context language modelling with recurrent neural network. In *Proceedings of ACL*, pages 1319–1329.

Hamed Zamani and W. Bruce Croft. 2016. Embedding-based query language models. In *Proceedings of ICTIR*, pages 147–156.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2015. Recurrent neural network regularization. In *Proceedings of ICLR (Conference Papers)*.

George Kingsley Zipf. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*.